Universitat d'Alacant
Universidad de Alicante

INTEGRACIÓN DE FUENTES DE INFORMACIÓN FORMALES E INFORMALES PARA LA IDENTIFICACIÓN DEL FOCO GEOGRÁFICO EN EL TEXTO

Fernando Samuel Peregrino Torregrosa

Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Universitat d'Alacant
Universidad de Alicante

INTEGRACIÓN DE FUENTES DE INFORMACIÓN FORMALES E INFORMALES PARA LA IDENTIFICACIÓN DEL FOCO GEOGRÁFICO EN EL TEXTO

Fernando Samuel Peregrino Torregrosa

EDUA Escola de Doctorat
Escuela de Doctorado
edua.ua.es

Fernando Samuel Peregrino Torregrosa

Tesis doctoral
Alicante, julio 2016

Tesis doctoral
Alicante, julio 2016

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

INTEGRACIÓN DE FUENTES DE INFORMACIÓN FORMALES E INFORMALES PARA LA IDENTIFICACIÓN

Fernando Samuel Peregrino Torregrosa

PROGRAMA DE DOCTORADO EN APLICACIONES DE LA INFORMÁTICA

Tesis presentada para aspirar al grado de
DOCTOR O DOCTORA POR LA UNIVERSIDAD DE ALICANTE

Tesis dirigida por
Dr. David Tomás Díaz
Dr. Fernando Llopis Pascual

Esta tesis ha sido financiada por el Ministerio de Economía y Competitividad a través del proyecto TIN2009-13391-C04-01: *Las tecnologías del lenguaje humano ante los nuevos retos de la comunicación digital.*

Integración de fuentes de información formales e informales para la identificación del foco geográfico en el texto

12 de julio de 2016

Agradecimientos

Gracias, de corazón, a mis tutores, el doctor Fernando Llopis Pascual por sacar tiempo de donde no lo tenía para guiarme y exigirme en mi trabajo; y al doctor David Tomás Díaz por su paciencia, dedicación, motivación, criterio y aliento, pero sobre todo por poder contar con su amistad. Ha sido un privilegio el tener su guía y ayuda.

Gracias a todos mis compañeros y excompañeros de laboratorio, a los cuales no me atrevo a nombrar por temor a que mi memoria me juegue una mala pasada. Me habéis tenido que soportar durante todo este tiempo y aún así habéis logrado que fuera a trabajar con una sonrisa gracias a que disteis verdadero significado a la palabra *compañerismo*.

Gracias también al resto de miembros del *grupo de procesamiento del lenguaje natural y sistemas de información*, especialmente a quien me empujó a emprender este camino, el doctor Patricio Manuel Martínez Barco. Sin su apoyo desde el primer instante este trabajo no hubiera sido posible.

Gracias a los miembros del laboratorio de geomática de la Universidad de Alicante. Con ellos me pude iniciar en el mundo de la investigación y, lo que es más importante, contar con su amistad.

Y, ¿cómo no?, un millón de gracias a todos mis amigos y familiares por haberme apoyado y animado en este largo viaje.

Por último, me veo obligado a hacer un agradecimiento especial y pedir perdón a esas personas tan queridas para mi por haberles robado tanto tiempo y dedicación. Mi madre, Lucía, la cual me ha ayudado no sólo con su apoyo y fe ciega, sino que también regalándome tiempo para que pudiera centrarme en mi trabajo. Mi mujer, Nuria, quien ha sido mi espejo y bastón durante todos estos años. Sin su apoyo y comprensión nunca lo hubiera logrado. Mi hijo, Daniel, la persona con la que más en deuda estoy ya que es a quien más tiempo le ha quitado mi trabajo. Mi hija, Nadia, por no haberle ofrecido toda la dedicación que merecía en su primer año de vida.

Para todos vosotros esta dedicatoria de tesis. Gracias a vosotros nunca me faltaron motivos para seguir adelante.

Alicante, 12 de julio de 2016

Fernando S. Peregrino
Grupo de Procesamiento del Lenguaje
y Sistemas de Información
Departamento de Lenguajes
y Sistemas Informáticos
Universidad de Alicante
Tel. (+34) 965 90 27 37
Fax: (+34) 965 90 93 26

Índice general

Índice de figuras

Índice de figuras

Índice de tablas

1 Introducción

Hoy en día vivimos inmersos en la sociedad de la información y del conocimiento, la cual se caracteriza por disponer de una ingente cantidad de información y por la necesidad de permitir su acceso y divulgación con celeridad. Con este fin, resulta crucial que el almacenaje de dicha información sea llevado a cabo en formato digital. El acceso a esta información digital ha de ser lo más rápido y preciso posible, dando significado así a la expresión "*sociedad del conocimiento*". Con esta motivación surgen los sistemas de recuperación de información (*IR* - *Information Retrieval*), también conocidos como *motores de búsqueda*.

La recuperación de la información (*IR*) es la ciencia de la búsqueda de información en documentos electrónicos dando como resultado un conjunto de estos mismos documentos ordenados según la relevancia que tengan con la consulta formulada.

La *IR* se ocupa de la representación, almacenamiento, organización y acceso a elementos de información tales como documentos, páginas web, catálogos en línea, registros estructurados, registros semiestructurados y objetos multimedia. La representación y la organización de los elementos de información deben ser tales que proporcionen a los usuarios un fácil acceso a la información de su interés (Baeza-Yates et al., 1999).

El núcleo de la mayoría de estos sistemas, descrito en líneas generales, se divide en dos fases: indexación y búsqueda. En una primera instancia indexan todos los términos encontrados en los documentos del corpus donde se realizará la búsqueda posteriormente (páginas web en el caso de los buscadores de Internet), registrando en qué documentos aparecen y con qué frecuencia, así como la longitud en palabras o términos de cada uno de estos documentos.

En un segundo paso, se le introduce una consulta al sistema con el fin de que devuelva los documentos por orden de relevancia a la consulta. Para llevar a cabo esta tarea, los sistemas de *IR* suelen hacer un emparejado de los términos que aparecen en la consulta con los términos que tienen indexados, devolviendo así los documentos por orden de relevancia donde un mayor número de términos de la consulta tenga más peso.

Entre toda la información que los sistemas *IR* tienen que tratar podemos encontrarnos con la información geográfica. Esta información geográfica comprende datos tales como: topónimos, códigos postales, direcciones, coordenadas, etc. Los sistemas de *IR* convencionales tratan este tipo de información del mismo modo que el resto de información debido a su funcionamiento intrínseco, es decir, dichos sistemas no diferencian entre datos geográficos y cualquier otro término que aparezca en los documentos de texto a analizar.

Según un estudio realizado por Zhang et al. (2006), el 12,7 % sobre 4 millones de consultas de ejemplo contenía un topónimo[1]. Esto es corroborado por el trabajo llevado a cabo en Gan et al. (2008), donde se analizaron 36 millones de consultas ejecutadas en un sistema de *IR*, deduciendo que entre el 18 % y el 22 % de las consultas lanzadas en dicho motor de búsqueda estaban acotadas geográficamente.

De trabajos como los citados anteriormente se desprende que la información geográfica también se ve involucrada en *IR*. Consultas del tipo "*Catedrales en Europa*", o "*Dónde murió Osama bin Laden*", hacen necesaria la intervención de dicha materia para ser resueltas de la manera adecuada.

La recuperación de información geográfica (*GIR - Geographical Information Retrieval*) es una especialización de la *IR* con metadatos geográficos asociados. Los sistemas de *IR*, generalmente, ven los documentos como una colección o "bolsa de palabras", es decir, que simplemente cuentan las palabras que aparecen en cada documento, el número de veces que éstas aparecen y la longitud de dicho documento. Por el contrario, los sistemas *GIR* necesitan información semántica, es decir, necesitan de un lugar o rasgo geográfico asociado a un documento. Debido a esto, es común en los sistemas *GIR* que se separe el análisis y la indexación de texto de la indexación geográfica, es decir, la indexación de cada documento según el lugar sobre el que verse el texto contenido en él.

Para la indexación de la parte geográfica se hace necesaria la detección del foco o focos geográficos, o lo que es lo mismo, el ámbito o ámbitos geográficos en los que está centrado el texto para que, como se ha comentado en el párrafo previo, los documentos se almacenen acorde al área geográfica sobre la que versa el texto. Para la detección de estas áreas resulta fundamental la detección de topónimos existentes en dicho texto, lo cual en sí resulta ser un arduo problema.

Lo primero a lo que hay que enfrentarse para detectar los topónimos es poder identificar en los textos las entidades nombradas, pudiendo ser éstas de distinta índole, tales como geográficas, personas, organizaciones, expresiones temporales, cantidades, valores monetarios, porcentajes, etc.

[1] La toponimia es el estudio del origen, uso, significación y tipología de los nombres propios de lugar (topónimos).

Para esta detección se suele emplear el denominado reconocimiento de entidades nombradas (*NER - Named Entities Recognition*).

El *NER* es una subtarea de la extracción de la información, la cual intenta localizar y clasificar elementos en el texto dentro de unas categorías predefinidas, tales como las mencionadas en el párrafo anterior.

Una vez detectadas las entidades, si nos centramos meramente en las geográficas, pese a que el resto también pueden ser de gran utilidad tal y como se demuestra en Peregrino et al. (2012c), surge un nuevo problema, la ambigüedad, la cual, en los sistemas *GIR* consiste en los topónimos compartidos por más de una localidad.

El problema de la desambiguación de topónimos ha sido ampliamente tratado casi siempre de manera aislada en campos como la lingüística, el aprendizaje automático (*Machine Learning*) o la inteligencia artificial, así como por investigadores en los campos de la *IR* y de la gestión del conocimiento (*Knowledge Management*). Las dos principales aproximaciones que abordan dicha problemática están basadas en el aprendizaje automático, el cual reconoce nombres a partir de la estructura del texto y de su contexto, y la aproximación basada en glosarios y *gazetteers* o nomenclátores[2] que ayudan a dilucidar qué lugar está siendo referenciado. Esta última aproximación no puede encontrar nombres que no estén en sus listados, tal y como sí hace la primera, aunque normalmente se compone por algoritmos más simples y no necesita de datos de entrenamiento, los cuales suelen ser difíciles de obtener (Amitay et al., 2004).

Como aproximación más extendida entre los sistemas *GIR* más efectivos, una vez localizados y desambiguados los topónimos que se mencionan en el texto, se procede a determinar el foco geográfico. Para ello hay que discernir qué topónimos son relevantes y en qué grado, y cuáles no. Si se encuentra un texto como el siguiente: "*La multinacional* ***china*** *ha creado su segunda mayor sede, después de la de* ***San Francisco*** *en* ***EEUU****, en* ***Madrid****, la cual creará miles de puestos de trabajo no sólo para* ***la capital****, sino también para* ***las ciudades de los alrededores****, tales como* ***Leganés*** *y* ***Getafe****...*", el foco geográfico debería ser la comunidad de Madrid, desechando así topónimos encontrados en el texto, tales como *china*, *San Francisco* o *EEUU*, para el ejemplo dado.

Finalmente, habiendo determinado el foco geográfico del texto analizado, ya se puede proceder a indexarlo acorde a éste, pudiéndose crear, de este modo, el índice geográfico de un sistema *GIR* que posteriormente devolverá los documentos que mejor satisfagan los requisitos demandados, geográficos y no geográficos, por las consultas lanzadas al sistema.

Dada la gran importancia que tiene en los resultados finales de un sistema de *GIR* la correcta detección de dichos focos geográficos, en esta tesis se profundizará en esta materia con el fin de detectar las debilidades

[2]Diccionarios geoespaciales de topónimos.

de las actuales aproximaciones para la detección del ámbito geográfico, y se mostrarán aproximaciones novedosas que pueden ayudar a mejorar la detección de dichos ámbitos geográficos en los textos, ya no sólo formales, sino los textos informales, los cuales han irrumpido con gran ímpetu en la sociedad actual gracias a la omnipresencia de las redes sociales. El núcleo de las investigaciones llevadas a cabo en esta tesis está basado en la clasificación geográfica de los textos con la asistencia de textos procedentes de distintas fuentes.

También se va a realizar un estudio en el que se compararan textos de diversas fuentes con el fin de poder analizar aspectos tales como la correlación entre los textos emitidos por dichas fuentes, la evolución temporal que sufre el lenguaje, qué importancia tienen ciertos términos dependiendo del área geográfica en la que se encuentre, qué similitudes existen entre los términos empleados en zonas geográficas cercanas y qué términos son los más representativos de cada lugar.

En lo que resta de capítulo, se hablará de las motivaciones que impulsan esta tesis y de los objetivos planteados, así como de la metodología seguida y la estructura de este trabajo.

1.1. Motivaciones

La clasificación automática de textos[3] es ampliamente conocida y usada en multitud de situaciones. Dicha clasificación puede ser afrontada desde distintos puntos de vista, siendo uno de los más usados la clasificación por ámbito geográfico.

De sobra son conocidas las clasificaciones geográficas realizadas por los buscadores de Internet y por los periódicos, las cuales agrupan un conjunto de páginas web o noticias acorde al ámbito geográfico que cubren.

A la hora de mostrar o buscar estos textos acorde a unos criterios geográficos, los distintos proveedores de los textos (periódicos, blogs, motores de búsqueda web, etc.) hacen uso de sistemas *GIR*.

Los sistemas *GIR* son un campo de investigación en auge en los últimos años debido a la falta de buenos resultados cuando se realiza una búsqueda centrada en una ubicación específica. Son diversas las competiciones que se han organizado alrededor de este tipo de sistemas.

El ciclo de conferencias *CLEF*[4] (*CLEF - Conference and Labs of the Evaluation Forum*), anteriormente conocido como *Cross-Language Evaluation Forum*, es una organización que promueve la investigación en el acceso de la información multilingüe, y que actualmente se centra en las

[3] La clasificación automática de textos consiste en dividir un conjunto de textos acorde a ciertos criterios dentro de una serie de categorías tales como temática, geografía, fechas, etc.

[4] http://www.clef-campaign.org

lenguas europeas. Sus funciones específicas son mantener un marco de referencia base para probar los sistemas de recuperación de información, y la creación de repositorios de datos para que los investigadores lo utilicen en el desarrollo de normas comparables. La organización lleva a cabo una reunión del foro cada mes de septiembre en Europa. Antes de cada congreso, los participantes reciben un conjunto de tareas de competición. Las tareas están diseñadas para evaluar diversos aspectos de los sistemas de recuperación de información y fomentar su desarrollo. Grupos de investigadores proponen y organizan campañas para satisfacer esas tareas. Los resultados se utilizan como puntos de referencia para conocer el estado de la cuestión en áreas específicas.

Dentro de estas conferencias *CLEF* se agregó una rama geográfica, *GeoCLEF*[5]. El propósito de *GeoCLEF* era experimentar y evaluar la recuperación de documentos obtenidos, los cuales estaban orientados hacia lugares geográficos que eran descriptivos de los documentos. La idea principal era ver si la adición de operadores geográficos y localizaciones mejorarían la precisión y especificidad de la recuperación de los documentos pertinentes.

A raíz de esta rama geográfica, nacieron otro tipos de tareas como *GikiP*, la cual fue una tarea piloto en *GeoCLEF* 2008 pasando en 2009 a llamarse *GikiCLEF* y ser una tarea propia dentro de las conferencias *CLEF*. Esta tarea consistía en encontrar entradas o documentos en la *Wikipedia*[6] que contestaran a una serie de consultas que requería de algún tipo de razonamiento geográfico.

Además, dentro de la conferencia *NTCIR* (*NII Test Collection for IR Systems*)[7] se creó la tarea *GeoTime*[8]. Esta tarea combinaba *GIR* con búsqueda basada en el tiempo para encontrar eventos específicos en una colección de textos.

Por último, también cabe destacar la creación de *workshops* específicos en la materia, como *Geographic Information Retrieval*[9].

Por otro lado, la incipiente y enérgica aparición de las redes sociales con su lenguaje altamente informal ha hecho que las técnicas utilizadas para la clasificación geográfica automática de textos formales tengan que ser readaptadas con el propósito de obtener unos mejores resultados. Así pues, los *NERs* convencionales obtienen un pobre resultado cuando se trata de identificar las entidades, geográficas o de cualquier otra índole, que aparecen en el texto informal a tratar.

[5] http://ir.shef.ac.uk/geoclef
[6] http://www.wikipedia.org
[7] http://research.nii.ac.jp/ntcir/index-en.html
[8] http://metadata.berkeley.edu/NTCIR-GeoTime
[9] http://www.geo.uzh.ch/~rsp/gir15

Estos malos resultados con textos informales se acentúan aún más si cabe cuando dichos textos tienen una longitud muy limitada, como ocurre en el caso de los tuits[10].

Este panorama nos deja los siguientes frentes abiertos:

- Un gran número de los sistemas de *GIR* no detectan el foco geográfico de cada texto, es decir, son incapaces de identificar la principal área geográfica sobre la que versa el texto, y se limitan a recoger todas las menciones a topónimos asignándoles el mismo peso dentro de esta detección del foco geográfico.

- Hasta la fecha, no se ha creado ningún sistema de *GIR* capaz de manejarse con éxito con textos informales.

- Habitualmente, los sistemas de *GIR* que realizan indexación geográfica, para detectar el foco geográfico de cada texto utilizan exclusivamente información puramente geográfica, sin tener en cuenta la aportación de otro tipo de información como podría ser los nombres de personas, empresas, expresiones y declinaciones lingüísticas de cada región geográfica, etc.

- Para la clasificación geográfica de los textos, la inmensa mayoría de aproximaciones se basa puramente en textos de la misma fuente, los cuales raramente están etiquetados geográficamente, o en textos estructurados como son los *gazetteers*, lo cual hace que sean totalmente dependientes del idioma, sin tener en cuenta la aportación que podrían realizar otras fuentes de textos.

- La evaluación de la precisión de los sistemas de detección del ámbito geográfico en textos que hacen uso de herramientas externas tales como *NERs* es, cuando menos, difícil de comprobar, debido al amplio margen de error que se arrastra a través de estas herramientas.

Por otro lado, si se centra la atención en otras aproximaciones basadas en técnicas de aprendizaje automático (*ML*), se puede apreciar cómo una de las partes críticas a la hora de detectar el ámbito geográfico (o cualquier otro problema afrontado desde la perspectiva de *ML*) es la obtención de un buen conjunto de entrenamiento correctamente etiquetado. Dicho conjunto de entrenamiento no tiene por qué provenir de la misma fuente de datos a evaluar, es decir, si se quiere clasificar geográficamente noticias, no tiene por qué entrenarse única y exclusivamente con un corpus de noticias geográficamente etiquetadas, sino que se podría hacer con otras fuentes tales como noticias de otras fuentes, artículos de *Wikipedia*, etc., pudiendo usarse todas estas fuentes en conjunto o por separado.

[10] Mensajes de textos de la red social ***Twitter***, de hasta 140 caracteres.

Tanto si se utiliza una única fuente como conjunto de entrenamiento de un sistema *GIR*, como si se utilizan diversas fuentes, es necesaria la realización de un estudio de al menos la fuentes a clasificar con el fin de comprender la naturaleza de las misma y poder así detectar rasgos capaces de ayudar en la tarea de la detección del ámbito geográfico de los textos.

1.2. Objetivos

A la luz de las motivaciones expuestas en el apartado anterior, en esta tesis se pretende mostrar las posibles mejoras que podrían aportar distintas fuentes de texto en el entrenamiento a la hora de determinar el ámbito geográfico de un texto. Concretamente, la tesis se dividirá en tres grandes bloques:

1. **Determinar la relevancia de la información no geográfica en la detección del foco geográfico en textos.** Este bloque se dividirá a su vez en dos secciones, donde cada una se centrará en textos formales e informales respectivamente. Así pues, las cuestiones que se afrontarán en cada una de estas secciones son:
   - ¿Qué aporta la información no geográfica procedente de fuentes de texto formales a la detección del foco geográfico sobre textos formales?
   - ¿Qué aporta la información no geográfica procedente de fuentes de texto formales a la detección del foco geográfico sobre textos informales?
2. **Determinar qué pueden aportar distintas fuentes de texto, tanto formales como informales, para la detección del foco geográfico en textos.** Este bloque, al igual que el anterior se dividirá en dos secciones, donde cada una se centrará en textos formales e informales respectivamente. Así pues, las cuestiones que se afrontarán en cada una de estas secciones son:
   - ¿Qué aportan otras fuentes de información formales, distintas a las que se pretenden clasificar, a la detección del foco geográfico?
   - ¿Qué aportan otros textos informales, distintos a los que se pretenden clasificar, a la clasificación geográfica de textos?
   - ¿Qué aporta la combinación de los recursos formales e informales?
3. **Analizar corpus de texto formales e informales.** El objetivo de este bloque es el de conocer la relación existente entre la terminología empleada en estos corpus para poder mejorar así la detección del ámbito geográfico para futuros sistemas de recuperación de información geográfica.

1.3. Metodología

A continuación se describirá la metodología seguida para la consecución de los objetivos marcados en el punto anterior. Así pues, para alcanzar estos objetivos se puede dividir esta metodología en los siguientes apartados:

1. **Identificación del foco geográfico en textos formales.** Este bloque se dividirá a su vez en cuatro secciones, donde cada una de ellas intentará determinar la importancia de los distintos corpus de entrenamiento y categorías gramaticales de los términos que componen dichos corpus contestando a las siguientes cuestiones:

   a) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a la misma fuente de texto que se pretende clasificar, ¿qué aproximación es la más apropiada para clasificar dichos textos?* Para ello se utilizará el corpus de noticias como fuente de texto formal. Con esta fuente de textos formales se dará respuesta a las siguientes preguntas:

      - ¿Qué aproximación es la más apropiada para clasificar geográficamente textos formales?
      - ¿Qué aportan cada una de las principales categorías gramaticales a la detección del foco geográfico en textos formales?
      - ¿Qué mejora se puede lograr utilizando las características más relevantes del corpus para el entrenamiento del sistema de clasificación geográfica?
      - ¿Cómo se comporta el sistema dependiendo del volumen de textos de entrenamiento?

   b) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a otra fuente de textos formales distinta a la que se pretende clasificar, ¿qué precisión tendría el sistema utilizando dicho corpus?* Para resolver esta cuestión se utilizará un corpus de artículos de *Wikipedia* geográficamente etiquetados. Con dicho corpus se dará contestación a las siguientes preguntas:

      - ¿Cómo afecta al rendimiento del sistema la utilización de las diversas categorías gramaticales?
      - ¿Resulta útil esta fuente de textos formales como selector de características para entrenar al sistema?

   c) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a una fuente de textos informales, ¿qué precisión tendría el sistema utilizando dicho corpus?* Para resolver esta cuestión se utilizará un corpus de mensajes de *Twitter* geográficamente etiquetados. Con dicho corpus se dará contestación a las siguientes preguntas:

- ¿Resulta útil esta fuente de textos informales como selector de características para entrenar al sistema?
- ¿Qué ocurre si se agrupan las distintas muestras de entrenamiento en una única según su lugar de procedencia?

d) *¿Cómo funcionaría un sistema que combinara diversas fuentes de texto que a su vez tienen distintos grados de formalidad?* En esta ocasión se mostrarán los resultados obtenidos cuando se combinan los corpus utilizados previamente para clasificar los textos de noticias. Para ello se llevarán a cabo experimentos que prueben las distintas combinaciones de los corpus tratados:

- Combinación de un corpus formal con el que se pretende clasificar geográficamente.
- Combinación de un corpus informal con el que se pretende clasificar geográficamente.
- Combinación de un corpus formal y otro informal con el que se pretende clasificar geográficamente.

2. **Identificación del foco geográfico en textos informales.** Al igual que sucediera con el apartado anterior, este bloque se dividirá a su vez en cuatro grandes secciones, donde cada una de ellas intentará determinar la importancia de los distintos corpus de entrenamiento para determinar el foco geográfico de conjuntos de tuits agrupados por usuario y ciudad, teniendo en cuenta el nivel de actividad de los usuarios que han emitido estos textos. En cada una de estas secciones se dará respuesta a las siguientes cuestiones:

a) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a la misma fuente de texto que se pretende clasificar, ¿qué aproximación es la más apropiada para clasificar dichos textos?* Para ello se utilizará un corpus de textos geográficamente localizados emitidos en la red social *Twitter* como fuente de textos informales a clasificar. Con esta fuente de textos se dará respuesta a las siguientes preguntas:

- ¿Qué aproximación es la más apropiada para clasificar geográficamente textos informales?
- ¿Cómo afecta que se agrupen los textos de entrenamiento en únicas muestras de entrenamiento por cada lugar a clasificar, o utilizar dichos textos de manera individual para entrenar al sistema?

b) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a una fuente de textos formales, ¿qué precisión tendría el sistema utilizando dicho corpus?* Para resolver esta

cuestión se utilizará un corpus de artículos de *Wikipedia* geográficamente etiquetados. Con dicho corpus se dará contestación a las siguientes preguntas:

- ¿Cómo afecta al rendimiento del sistema la utilización de las diversas categorías gramaticales de *Wikipedia*?
- Puesto que se pretende clasificar textos informales utilizando textos formales como entrenamiento del sistema, ¿resulta útil esta fuente de textos formales como selector de características para entrenar al sistema?

c) *Si se dispone de un conjuntos de textos geográficamente etiquetados, pertenecientes a otra fuente de textos informales distinta a la que se pretende clasificar, ¿qué precisión tendría el sistema utilizando dicho corpus?* Para resolver esta cuestión se utilizará un corpus de textos procedentes de *Flickr* geográficamente etiquetados. Con dicho corpus se dará contestación a las siguientes preguntas:

- ¿Resulta útil esta fuente de textos informales como selector de características para entrenar al sistema?
- ¿Qué campo de texto de *Flickr* es el que más ayuda a la clasificación de tuits?
- ¿Qué ocurre si se agrupan los distintos campos de texto?

d) *¿Cómo funcionaria un sistema que combinara diversas fuentes de texto que a su vez tienen distintos grados de formalidad?* En esta ocasión se mostrarán los resultados obtenidos cuando se combinan los corpus utilizados previamente para clasificar los conjuntos de tuits. Para ello se llevarán a cabo experimentos que prueben las distintas combinaciones de los corpus tratados:

- Combinación de un corpus formal con el que se pretende clasificar geográficamente.
- Combinación de un corpus informal con el que se pretende clasificar geográficamente.
- Combinación de un corpus formal y otro informal con el que se pretende clasificar geográficamente.

3. **Análisis de los corpus de noticias y de *Twitter*.** Este bloque se dividirá a su vez en cinco partes, donde en cada una de estas partes se intentará dar respuesta a las siguientes cuestiones:

   - *¿Existe correlación en la terminología empleada en los corpus de noticias y Twitter?* Para contestar a esta pregunta se compararan los textos de ambos corpus pertenecientes tanto al mismo como a distintos periodos temporales.

- *¿Qué relevancia adquieren o pierden los términos con el transcurrir del tiempo?* Para este experimento se mostrará cómo va ganando o perdiendo relevancia una serie de términos a lo largo del periodo temporal que comprenden los distintos corpus.
- *¿Dónde adquiere más relevancia ciertos términos?* Para ello se mostrará la relevancia de una serie de términos con respecto a las distintas áreas geográficas.
- *¿Existe alguna relación entre la terminología empleada según el área geográfica?* En esta ocasión se agruparán las áreas geográficas que contengan una terminología similar.
- *¿Cuáles son los términos que mejor ayudan a clasificar geográficamente cada una de las distintas áreas geográficas?* Para afrontar esta pregunta se obtendrán los términos más discriminativos por área geográfica.

1.4. Estructura de la tesis

A continuación se esboza la organización y contenido del resto de capítulos que conforman este trabajo de tesis:

- En el capítulo 2 (*Recuperación de Información Geográfica*), se detallará qué es un sistema de *GIR* y cómo funciona, así como cuáles son las partes esenciales que hay que abordar en la construcción de un sistema de *GIR*.
- En el capítulo 3 (*Detección del foco geográfico*), se describirá con detalle las técnicas empleadas para la obtención del ámbito geográfico en textos exponiendo de este modo el estado de la cuestión actual en esta área.
- En el capítulo 4 (*Agregación de recursos*), se detallarán otras fuentes de información textual desestructurada que pueden ser útiles a la hora de clasificar textos geográficamente.
- En el capítulo 5 (*Experimentación*), se mostrarán los experimentos llevados a cabo y se detallarán los resultados y conclusiones obtenidas.
- En el capítulo 6 (*Análisis geográfico del lenguaje*), se mostrará la evolución del lenguaje según las distintas áreas geográficas así como la correlación entre el lenguaje empleado en una fuente de textos formales y otra informal.
- En el capítulo 7 (*Conclusiones y Trabajo Futuro*), se hará un resumen de lo obtenido con esta investigación y se trazarán las futuras líneas a seguir partiendo del trabajo aquí realizado.

2

Recuperación de Información Geográfica

Hoy en día disponemos de una gran cantidad de información en documentos de texto y otros tipos de objetos multimedia, la cual, normalmente, no está estructurada. Dicha información, cuando es objeto de consulta, en la inmensa mayoría de las ocasiones hace referencia a algún espacio geográfico. Según Wang et al. (2005a), el 79 % de las páginas web bajo el dominio *.gov* contienen al menos una referencia geográfica, tales como topónimos, direcciones postales o números de teléfono. Otro estudio realizado por los mismos autores (Wang et al., 2005b) muestra que un 14 % de las consultas realizadas en un motor de búsqueda de internet tienen restricciones geográficas. Un ejemplo de dichas consultas son: "*Restaurantes en el centro de Toledo*" o "*Las mejores calas de Menorca*".

La disciplina de la Recuperación de Información Geográfica (*GIR - Geographic Information Retrieval*) afronta esta demanda de información mediante la creación de métodos que sean capaces de acceder a la información geográfica en documentos multimedia, estructurados y semiestructurados, tales como documentos de texto e imágenes. El objetivo es pues, el desarrollo de un sistema de información que pueda interpretar automáticamente la terminología geográfica y los conceptos espaciales que la gente utiliza cuando guarda y consulta dicha información, así como el poder recuperar la información que sea relevante para las necesidades de información geográfica de los usuario. *GIR* se ubica en la intersección de la Recuperación de Información (*IR - Information Retrieval*) y de los Sistemas de Información Geográfica (*GIS - Geographic Information Systems*) y se centra especialmente en investigar y desarrollar sistemas para la extracción de la geo-información, indexación espacio-textual, recuperación de la información espacial y su visualización.

En los últimos años ha habido un incremento en la investigación dedicada a la recuperación de información geográfica dado su gran interés mercantil. Los grandes motores comerciales de búsqueda de internet (*Google*[1], *Yahoo!*[2]

[1] http://www.google.com
[2] http://www.yahoo.com

y *Bing*[3]) han desarrollado herramientas para poder manejar información geográfica.

Así pues, podemos ver como *Google* a creado herramientas como: *Google Earth*[4], *Google Maps*[5], *Google Map Maker*[6], *Google Fusion Tables*[7] y *Google Geo APIs*[8].

Por otro lado, *Yahoo!* ha creado herramientas que directamente tratan con los textos desestructurados y los enriquecen con información geográfica. Entre las más destacadas están: *Yahoo! Placemaker*, el cual está actualmente integrado dentro de *Yahoo! Boss Geo*[9]. *Yahoo! Placemaker* es un servicio web de *geo-parsing* de libre disposición. Es útil para desarrolladores que quieren hacer aplicaciones basadas en localización espacial mediante la identificación de topónimos existentes en textos no estructurados (p. ej. *feeds*, páginas web, noticias, etc.), de los que es capaz de devolver metadatos geográficos asociados a dichos textos. La aplicación identifica los topónimos en el texto, los desambigua, y devuelve identificadores de lugar únicos para cada uno de los lugares que tiene en su base de datos. También aporta otro tipo de información como, cuántas veces aparece el lugar en el texto, en qué lugar del texto se encontró, etc. *Yahoo! Placemaker* devuelve un documento *XML* por cada texto que se le pase.

En este capítulo se va a mostrar una descripción de los aspectos que con más frecuencia se suelen tratar en la literatura cuando se pretende crear un sistema *GIR*. Para ello se hará un recorrido por cada uno de los estos aspectos que se han extraído como factor común de las aproximaciones más exitosas.

En este análisis, también se mostrarán los recursos geográficos utilizados con más frecuencia por los principales sistemas *GIR*.

2.1. Aspectos a tratar por un sistema GIR

En el trabajo llevado a cabo en Wang et al. (2005a), se clasificaron las áreas de investigación en esta materia en tres grandes grupos:

- Identificación y desambiguación de topónimos (etiquetado geográfico). Donde se lleva a cabo el trabajo de detectar las entidades geográficas nombradas de manera inequívoca, tal y como se explicará más detalladamente en la secciones 2.1.1 y 2.1.2.

[3] http://www.bing.com
[4] http://earth.google.com
[5] http://maps.google.com
[6] http://www.google.com/mapmaker
[7] http://tables.googlelabs.com
[8] http://code.google.com/apis/maps y http://code.google.com/apis/earth
[9] http://developer.yahoo.com/boss/geo/

- Desarrollo de herramientas informáticas para el manejo de la información geográfica. Dentro de esta área se analizan las utilidades y funcionalidades que las distintas aplicaciones deberían de ofrecer al usuario con el fin de obtener un correcto y preciso manejo de la información geográfica. La parte visual de estas herramientas, las interfaces, se verá con más detalle en la sección 2.1.6

- Explotación de las diversas fuentes de recursos geográficos. Dentro de este grupo, los autores destacan principalmente los *gazetteers*[10] y recursos web, los cuales serán ampliamente tratados en la sección 2.2.

Aunque, si hay un proyecto que es referencia obligatoria para todo aquel que se quiera iniciar en la materia, y que aún hoy en día sigue marcando la pauta a seguir por el resto de investigadores en *GIR*, ese es el proyecto *SPIRIT*[11] (Purves et al., 2007). En este proyecto se crearon herramientas software y técnicas que pueden ser usadas para crear motores de búsqueda y sitios web que muestren inteligencia en el reconocimiento de terminología geográfica. Con el fin de demostrar y evaluar los resultados del proyecto, se construyó un prototipo de motor de búsqueda *GIR*, el cual se utilizó como plataforma para probar y evaluar nuevas técnicas en recuperación de información geográfica. Este proyecto aborda plenamente todos los frentes abiertos en la investigación en *GIR*.

A raíz de este proyecto, parte del equipo del mismo elaboró un estudio (Jones and Purves, 2009) donde se puede ver una disección más exhaustiva de los principales asuntos que se pueden abordar en los sistemas *GIR*:

- Detección de referencias geográficas.
- Desambiguación de topónimos.
- Terminología geográfica vaga o imprecisa.
- Indexación espacial y textual.
- Ranking por relevancia geográfica.
- Interfaces de usuario.
- Métodos de evaluación de los sistemas *GIR*.

Basándonos en este estudio, a continuación se detallará en qué consisten las distintas aproximaciones que se han llevado a cabo en cada uno de estos siete apartados.

[10] Un nomenclátor (o índice de topónimos, o *gazetteer*) es un catálogo de nombres geográficos, el cual, en conjunto con un mapa, constituye una importante referencia sobre lugares y sus nombres.

[11] http://www.geo-spirit.org/

2.1.1. Detección de referencias geográficas

El reconocimiento del lenguaje espacial en los documentos de texto (conocido en inglés como *geoparsing*, *geotagging*, *georecognition* o *toponym recognition*) en el procesamiento del lenguaje natural (*NLP - Natural Language Processing*) es una extensión de los reconocedores de entidades nombradas (*NER - Named Entity Recognition*), y versa sobre el análisis de los textos con el fin de identificar la presencia inequívoca de topónimos (también conocido en inglés por *toponym grounding* o *toponym disambiguation*) (Buscaldi and Magnini, 2010) como se verá con más detalle en la sección 2.1.2.

Para llevar a cabo con éxito dicha tarea se debe hacer frente a problemas como la metonimia[12], en situaciones tales como "*no aceptaremos órdenes de Madrid*", en cuyo caso "*Madrid*" hace referencia al gobierno central español. En Leveling and Hartrumpf (2006) se muestra un enfoque para solucionar el problema de la metonimia a través del análisis de rasgos superficiales.

Según el trabajo elaborado en Leidner and Lieberman (2011), principalmente existen 3 enfoques distintos para abordar la problemática del reconocimiento del lenguaje geográfico en textos:

1. **Búsquedas basadas en *gazetteers*.** Estas búsquedas se realizan recorriendo el texto, bien palabra por palabra, bien carácter a carácter, buscando ocurrencias de un conjunto de topónimos predefinidos. Estos topónimos están almacenados en un *gazetteer*.

   Los *gazetteers* se almacenan típicamente en *tries*[13], *hash tables*[14] y/o bases de datos *SQL* como almacenamiento secundario. Puede ser necesario un trato especial para topónimos que abarcan más de una palabra, p. ej. "*Ciudad de Nueva York*", donde una búsqueda básica con un único término no daría los resultados esperados.

   Si el conjunto de topónimos no se organiza en una lista plana, sino como una jerarquía, entonces recibe el nombre de *geoparsing* basado en ontología. Una ontología es una forma más sofisticada de representar el conocimiento geográfico. Éstas suelen organizar las localizaciones en estructuras jerárquicas, para las que proveen relaciones topológicas entre ellas (Fu et al., 2005). Los trabajos recientes con ontologías se han centrado en campos como el modelado espacio-temporal (Mostern

[12] Tropo que consiste en designar algo con el nombre de otra cosa tomando el efecto por la causa o viceversa, el autor por sus obras, el signo por la cosa significada, etc.

[13] Un *trie* es una estructura de datos de tipo árbol que permite la recuperación de información. Es muy útil para conseguir búsquedas eficientes en repositorios de datos muy voluminosos.

[14] Una *hash table*, *tabla hash*, *mapa hash*, *tabla de dispersión* o *tabla fragmentada* es una estructura de datos que asocia llaves o claves con valores. La operación principal que soporta de manera eficiente es la búsqueda.

and Johnson, 2008), interoperabilidad (Janowicz and Keßler, 2008) y la construcción automática de gazetteers (Nadeau et al., 2006).

La calidad de los datos es algo que también hay que tener en cuenta ya que la naturaleza incompleta o ruidosa de los datos de los *gazetteers* puede llevar a dar falsos positivos y falsos negativos. Además, los nombres de lugar y los límites administrativos están constantemente cambiando, y la gestión de un proceso de actualización del *gazetteer* requiere un flujo de trabajo integrado y automatizado. Los *gazetteers* que se suelen utilizar con más frecuencia son *NGA's GNS*[15], *USGS's GNIS*[16] y *GeoNames*[17].

Ejemplo: Búsqueda de "*Alicante*" en *GeoNames*: ***Name:*** *Alicante (ALC, Akra Leuke, Alacant, Alakanto, Alicante, Alikante, Alíkante, Lucentum, a li kan te, alykanth, arikante)*; ***Country:*** *Spain, Valencia, Alicante > Alicante*; ***Feature class:*** *seat of a second-order administrative division population 334,757*; ***Latitude:*** *N 38° 20'42"* ***Longitude:*** *W 0° 28'53"*

2. **Búsquedas basadas en reglas.** Un conjunto de reglas simbólicas en un lenguaje específico de dominio codifica un procedimiento de decisión que crea un intérprete para decidir si una palabra es un topónimo o no. Por lo general, se utilizan expresiones regulares (*REs - Regular Expressions*), que corresponden a estados autómatas finitos (*FSA - Finite State Automata*), o gramáticas libres de contexto (*CFGs - Context-Free Grammars*), que se corresponden a autómatas de pila (*PDA - Push-Down Automaton*). Esto permite búsquedas rápidas y pequeños almacenamientos al mismo tiempo (Beesley and Karttunen, 2003), aunque sólo permiten patrones con profundidad máxima predefinida y formas limitadas de anidación (Hopcroft, 2007). Por otro lado, las *DCGs* (*Definite Clause Grammar*) son una extensión de *CFGs* implementadas por *PROLOG*[18], las cuales permiten la formulación de gramáticas menos eficientes aunque más expresivas (Bilhaut et al., 2003)(Schilder et al., 2008).

   Ejemplo: "*la ciudad de ?* $\rightarrow$ *<TOPÓNIMO>*" en *DIAL*, el lenguaje de reglas de *OpenCalais*[19].

   Ejemplo: Un autómata que representa "*[AZ].+shire*", obteniendo topónimos terminados en "*shire*".

[15] http://earth-info.nga.mil/gns/html
[16] http://geonames.usgs.gov/domestic
[17] http://www.geonames.org
[18] Lenguaje para programar artefactos electrónicos mediante el paradigma lógico con técnicas de producción final interpretada. http://www.swi-prolog.org/
[19] http://www.opencalais.com/

3. **Búsquedas basadas en el aprendizaje automático (*ML - Machine Learning*)**. Una ventana deslizante se mueve sobre el texto, y en cada posición se calcula un conjunto de propiedades conocidas como características. Las características pueden comprender la comprobación de cadenas específicas, cálculos de longitud, capitalización y similares, y suelen ser pruebas *booleanas*. Partiendo de un corpus de entrenamiento que contiene datos etiquetados, se extraen las configuraciones de características que están más altamente correlacionadas con los topónimos. Cuando se ejecuta sobre un corpus de prueba no anotado, las mismas características del conjunto de entrenamiento calculan la clase más probable para cada palabra (es decir, topónimo o no topónimo) utilizando, por ejemplo, la inferencia estadística (Curran et al., 2007) (Finkel et al., 2005) (Manning and Schütze, 1999).

   Ejemplo: Una característica *F(W[*+1] == "Ave") == true, P(LOC—F) = 0,9918*.

   En este caso, para cada palabra en el documento, la palabra siguiente *(W[*+1])* se comprueba si coincide con la cadena "*Ave*", abreviatura en inglés de "*avenida*" ("*avenue*"), lo cual comprende una prueba *booleana* de si encuentra o no una secuencia como "Madison Ave", en cuyo caso el valor de esta característica sería cierto, y falso en caso contrario. Cada característica se correlaciona estadísticamente con el resultado obtenido durante la fase de entrenamiento, lo que permite la predicción de la categoría en tiempo de ejecución.

2.1.2. Desambiguación de topónimos

Una vez obtenido de manera inequívoca el nombre del lugar, hay que desambiguarlo, ya que muchos de los topónimos existentes son compartidos por varios lugares (p. ej. *Granada*, *Springfield*, etc.). Para dicha tarea se ha optado por diversas estrategias tales como la identificación por medio del resto de topónimos del texto, es decir, obteniendo el ámbito del que se está hablando para desambiguar cada uno de los lugares (Wang et al., 2005a).

La identificación del foco o ámbito geográfico de un texto consiste en asignar una localización geográfica y un nivel de confianza a cada una de las huellas geográficas (*footprints*)[20] del texto, obteniendo así uno o varios focos geográficos para el texto completo (Amitay et al., 2004).

Por ejemplo, si en un texto aparecen únicamente varias localidades pertenecientes a una provincia, el foco geográfico de dicho texto correspondería con el de la provincia.

[20] Las huellas geográficas o *footprints* son el conjunto de topónimos recogidos en un texto.

Otra de las estrategias seguidas en esta materia, al hilo de la anterior, es el esclarecimiento del lugar nombrado jugando con las entidades geográficas de orden superior o inferior mencionadas en el texto (Silva et al., 2006).

Normalmente, los algoritmos de desambiguación de topónimos (*toponym grounding*) hacen uso de nomenclátores (*gazetteers*) con el fin de verificar los nombres geográficos, a la vez que utilizan la información del contexto con el fin de poder así inferir el correcto significado del nombre geográfico encontrado. Dicha problemática dista mucho de ser algo trivial, dado que en numerosas ocasiones se puede dar el caso en el que los nombres de personas, organizaciones, clubes deportivos, etc., son compartidos por los topónimos.

La desambiguación de topónimos resulta de gran utilidad para un extenso número de aplicaciones, ya que la asignación de unos valores latitud/longitud a estos términos, también conocido como geocodificación (*geocoding* en inglés) (Buscaldi, 2011), permite la conexión entre campos de textos desestructurados y los datos ya estructurados de los Sistemas de Información Geográfica (*SIG*) (Leidner, 2008).

La inmensa mayoría del público desconoce la latitud y longitud de sus hogares, por no hablar de sus coordenadas *UTM*[21]. Por lo tanto, con el fin de permitir la creación de datos geográficos para el público en general, se hace indispensable la creación de un conjunto de herramientas que sean capaces de identificar las coordenadas de los términos geográficos sobre la superficie de la tierra (Goodchild, 2007).

La ambigüedad no existe únicamente entre topónimos (*geo/geo*) como puede ser el caso de "*Granada*" o "*Springfield*" mencionados anteriormente, sino que también se puede encontrar entre otro tipo de nombres o entidades (*geo/non-geo)*, como es el caso de "*Jack* ***London***"[22] o "*Madrid*", el cual puede hacer referencia no sólo a la capital de España, sino también al gobierno central del mismo país (metonimia), al club deportivo *Real Madrid*, etc., o incluso pueden ser *geo/geo/non-geo* como es el caso expuesto anteriormente de *Granada*, que además de existir ciudades, provincias o distritos con el mismo nombre, también puede tener distintos significados no geográficos, como es el caso de la fruta, el proyectil, etc.

2.1.3. Terminología geográfica difusa

Otra problemática adicional es la de las expresiones geográficas difusas, es decir, aquellas que describen lugares imprecisos que no podemos encontrar en ninguna base de datos geográfica (*gazetteer*). Estas expresiones serían del tipo "Estaciones de esquí en el norte de España", donde la entidad geográfica

[21] El sistema de coordenadas universal transversal de Mercator (*UTM - Universal Transverse Mercator*) es un sistema de coordenadas basado en la proyección cartográfica transversa de Mercator, que se construye como la proyección de Mercator normal, pero en vez de hacerla tangente al Ecuador, se la hace tangente a un meridiano.

[22] Escritor británico

en la que deberíamos buscar ("norte de España") resulta imprecisa (Peregrino et al., 2011a).

Los trabajos previos realizados en el campo de la definición automática de regiones geográficas difusas han seguido dos líneas fundamentales a la hora de obtener la información necesaria para definir dichas regiones. La primera aproximación se centra en la obtención de información a partir de la consulta directa a un conjunto de usuarios reales, que son los encargados de delimitar la región a estudio. La segunda aproximación está enfocada a la obtención de información a partir de fuentes de información no estructurada para la identificación de estas regiones.

Dentro del primer grupo se encuentra el trabajo de Montello et al. (2003). En este trabajo se propone una aproximación probabilística basada en frecuencias, donde la inclusión de una determinada localización dentro de una región difusa viene condicionada por el número de usuarios que considera su pertenencia a la misma. Para ello, los autores pidieron a un grupo de ciudadanos de Santa Bárbara (California, EEUU) escogidos al azar que trazaran los límites de lo que consideraban el centro de la ciudad, donde pudieron comprobar que la intersección creada por estos ciudadanos era cláramente el centro de la ciudad y el resto una región difusa. Estos valores sirven para generar una curva de nivel que delimita la región difusa y proporciona una probabilidad de pertenencia a las localizaciones que se hallan en su interior.

El objetivo de la segunda aproximación es recuperar suficiente información de la web para poder definir espacialmente las regiones difusas estudiadas. En esta línea, en Clough (2005) se creó un sistema donde las coordenadas de las localidades contenidas en dichas regiones se empleaban para la definición de polígonos representativos de dicha región. Una extensión de este trabajo se puede encontrar en Jones et al. (2008), donde utilizan esta aproximación para identificar tanto regiones difusas como precisas. Estas investigaciones se basan en la premisa de que la terminología geográfica difusa suele venir acompañada de terminología precisa que les permite trazar los límites de estas áreas difusas.

2.1.4. Indexación espacial y textual

Existen una serie de técnicas para la indexación textual de documentos. Dichas técnicas, tal y como podemos ver en Baeza-Yates et al. (1999), suelen basarse en la creación de índices inversos, es decir, se añaden al índice todas las palabras encontradas en el corpus, indicando en qué documentos aparece cada palabra y con qué frecuencia, para su posterior recuperación mediante la intersección con la consulta del usuario.

En cuanto a la indexación espacial, es decir, la indexación de la información espacial tal como el foco geográfico de cada documento o los topónimos en él mentados, los sistemas de información geográfica (*GIS* -

Geographical Information System) son los que han tratado con más éxito dicho asunto. La dificultad subyace en conseguir mezclar ambos índices con acierto (Cardoso and Santos, 2008).

La técnica más común para afrontar dicha mezcla ha sido la obtención de "huellas" (*footprints*) espaciales de los documentos, para posteriormente poder indexar esas huellas por documentos, es decir, análogamente al índice textual, el índice geográfico estará compuesto de una lista de nombres de lugar, los cuales estarán asociados a los documentos del corpus donde aparezcan dichos lugares. Estas huellas espaciales no son más que los topónimos y focos geográficos encontrados en cada documento. El problema es que un alto número de huellas por documento (según Vaid et al. (2005) hay unas 21 por documento web) puede hacer intratable el problema, por lo que han habido muchos trabajos orientados a obtener un mínimo número de huellas representativas por documento (Wang et al., 2005a)(Silva et al., 2006).

En Vaid et al. (2005) podemos ver tres estilos diferentes de conseguir llevar a cabo dicha tarea con celeridad y mejorando los resultados de un sistema de *IR* genérico. En dicho trabajo se comparó la indexación textual corriente (*PT - Pure Text*), contra tres tipos de indexaciones textuales y espaciales:

1. Indexación texto-espacial (*TS - Text-primary spatio-textual indexing*), donde primero se hace una recuperación de los documentos acorde al filtro textual (tal cual se hace en *PT*) y posteriormente se le aplica el filtro geográfico.

2. Indexación espacio-textual (*ST - Space-primary spatio-textual indexing*), donde primero se recuperan los documentos por orden de relevancia geográfica y posteriormente, sobre los documentos obtenidos se recuperan los más relevantes textualmente.

3. Indexación separada (*T - Separate text and spatial indexes*), donde la recuperación geográfica y textual de los documentos se hace por separado y posteriormente se unen los resultados.

Estas indexaciones dieron como resultado una gran mejora en la cobertura (*recall*), aunque las indexaciones *TS* y *ST* supusieron un considerable aumento en el espacio requerido para su indexación.

2.1.5. Ranking por relevancia geográfica

La clasificación de documentos relevantes o ranking, determina la manera en la que se debe de puntuar un documento según su idoneidad para con la consulta lanzada para su posterior devolución al usuario. La manera más usual con la que suelen afrontar el ranking de documentos los principales motores de búsqueda es mediante la intersección de los términos de la

consulta con los de los documentos, dando un mayor valor a aquellos términos que ocurran en un menor número de documentos, a los términos que con más frecuencia aparezcan en un documento, y a la proporción de apariciones de un término en un documento dada la longitud de éste último (Robertson et al., 1999).

Tal y como se refleja en Van Kreveld et al. (2005), en el caso de los sistemas *GIR* no hay que devolver una simple intersección de términos, sino que hay que saber tratar semánticamente los topónimos referenciados en la consulta de tal forma que se capturen las huellas existentes en ésta y se comparen con las de los documentos del corpus, teniendo en cuenta que dichas huellas pueden pertenecer a un ámbito geográfico:

- Inferior. Si se buscan "*estaciones de esquí en España*", probablemente obtengamos resultados de estaciones de esquí tales como *Sierra Nevada*, la cual es una entidad geográfica que está por debajo de *Granada*, *Andalucía* y *España*.
- Superior. Si se buscan "*playas de la provincia de Alicante*" podremos encontrar documentos que versen sobre playas del litoral valenciano, el cual es una entidad geográfica que se encuentra por encima de la provincia de Alicante, pero que comprende dicha provincia también.
- Intersectante. Si se buscan "*Montañas de la península Ibérica*", podríamos obtener documentos que listen montañas de *España*, los cuales cubren en parte la península Ibérica, aunque también algunas islas que no pertenecen a la península, así como las montañas que pertenecen a Portugal o Andorra.

2.1.6. Interfaces de usuario

El desarrollo de una interfaz de usuario eficaz para los sistemas *GIR* es un tema que se ha tratado con escaso éxito dada su dificultad. La mayoría de consultas geográficas, como ya se ha indicado anteriormente, son del modo <*qué se busca*> + <*relación*> + <*localización*>, por lo que parece sencillo crear una interfaz con tres campos, pero habría que saber tratar cualquier tipo de "*relación*", y también habría que tener en cuenta que no todas las consultas son del tipo descrito anteriormente (p. ej. "*Dónde murió Osama bin Ladem*"). Otra aproximación que se ha hecho en Purves et al. (2007) es permitir al usuario trazar una zona en un mapa indicando los puntos del área donde se quiere obtener los resultados e introducir la consulta en sí.

Por otro lado, nos encontramos con la problemática de cómo devolver los resultados. ¿Se debe adjuntar un mapa sobre los sitios de los que habla cada documento? ¿A qué escala? ¿Un mapa por documento o un único mapa para todos los documentos relevantes? ¿Se muestra en la máquina local o desde un servidor?

Figura 2.1: Mapa de calor que muestra la densidad de los lugares mencionados en el *Domesday Book*. A más oscuro, mayor densidad.

Hay un límite a la cantidad de información que se puede visualizar simultáneamente en un mapa. Generalmente, se han probado dos métodos para esto, tal y como podemos ver en Clough et al. (2011). La primera aproximación se basa en unir los puntos que están cercanos entre sí en uno solo. Esto reduce el número de puntos que aparecen en el mapa, sin embargo, cuando el usuario hace clic en uno de los grupos es necesario mostrar algún tipo de interfaz donde el usuario puede decidir qué punto quiere el usuario ver en realidad.

El segundo enfoque es crear una representación en el servidor cuando hay demasiada información para mostrar en forma de puntos. Para hacer esto se utilizan los llamados "mapas de calor" (*heatmaps*), que indican dónde hay más información y dónde menos (Hill, 2006). En la figura 2.1 podemos ver como los autores de Clough et al. (2011) muestran un mapa de calor de los topónimos mencionados en el *Domesday Book*[23]. El usuario puede utilizar este tipo de mapas para dirigirse al lugar donde hay más información. Cuando el número de puntos es lo suficientemente bajo como para que puedan aparecer en el mapa, los mapas de calor pasan a ser mapas de puntos (ver figura 2.2).

Los autores de esta aproximación utilizaron un mapa interactivo basado en *OpenLayers*[24]. Diferentes tipos de información se organizan en forma

[23] El *Domesday Book* (también conocido como *Domesday*, *Doomsday* o *Libro de Winchester*) fue el principal registro de Inglaterra, completado en 1086 por orden del rey Guillermo I de Inglaterra. Este registro era similar a los censos nacionales que se realizan hoy en día.

[24] *OpenLayers* es una librería de *JavaScript* de código abierto para la visualización de la información de mapas de diversas de fuentes

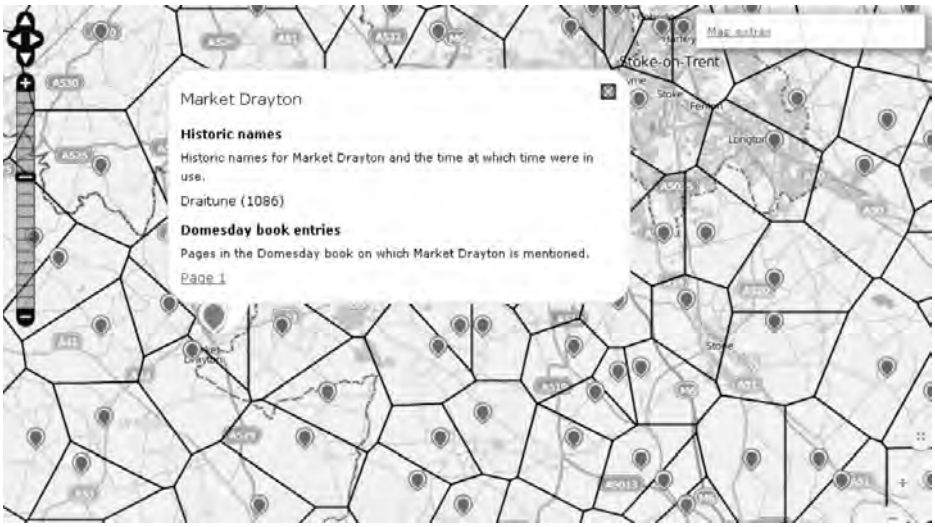


Figura 2.2: Mapa de puntos que muestra entradas individuales de los lugares mencionados en el *Domesday Book*.

de capas, por ejemplo, en el sistema expuesto en este trabajo se creó un prototipo de dos capas: la capa del mapa y la de los resultados. Eligieron *OpenStreetMap*[25] para obtener el mapa utilizado en OpenLayers. El mapa soporta interacciones básicas tales como zooms y panorámicas. Este prototipo permite dos formas principales de interacción. En primer lugar, el usuario simplemente podía navegar a través de los archivos, el prototipo mostraba el registro de archivos disponibles para el área visible en el mapa. El segundo modo de interacción combina el mapa de navegación con la capacidad de búsqueda por palabra clave. El usuario puede introducir una palabra clave ("*cricket*" en el ejemplo de la figura 2.3) y la base de datos es consultada para obtener los documentos con esa palabra clave, estando los resultados restringidos a la zona que está visible en el mapa (véase la figura 2.3). A medida que el usuario navega por el mapa, los resultados de la búsqueda se actualizan continuamente, de modo que el usuario sólo ve los resultados que se relacionan con el área visible en el mapa. Esta capacidad de navegar a lo largo tanto de la dimensión del contenido, utilizando la funcionalidad de búsqueda, como de la dimensión geográfica, mediante la interacción con el mapa, proporciona un método muy poderoso para interrogar a grandes conjuntos de datos.

[25] http://www.openstreetmap.org/

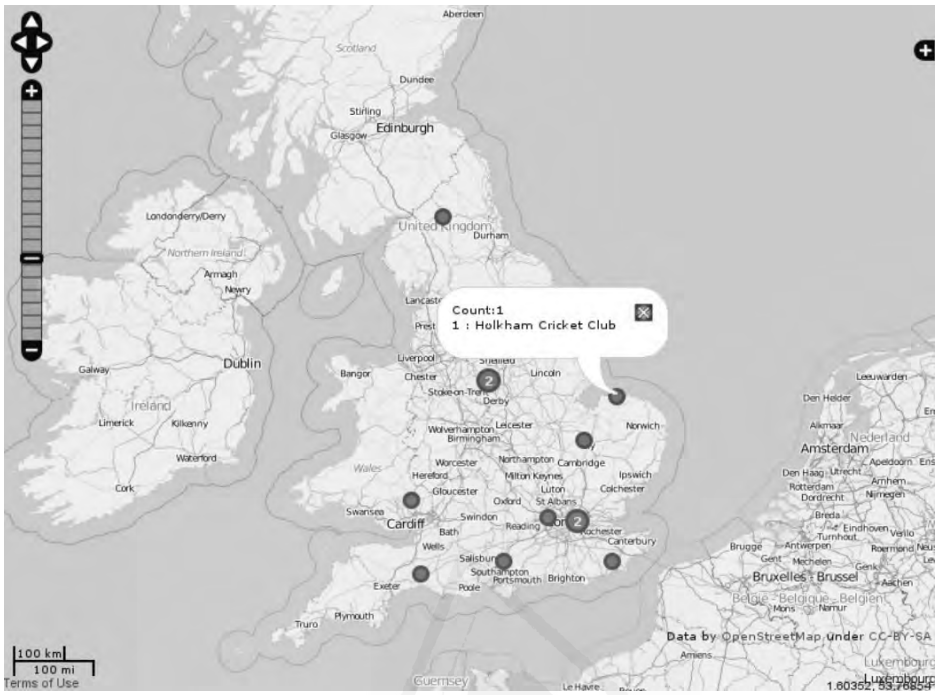


Figura 2.3: Ejemplo de resultados con una interfaz de un sistema *GIR*.

2.1.7. Métodos de evaluación de los sistemas GIR

La evaluación de los resultados devueltos por un sistema *GIR*, al igual que en los sistemas de *IR*, es una tarea costosa dado que hay que ir examinando manualmente, uno a uno, todos los documentos devueltos por un sistema para comprobar la relevancia de cada uno de los documentos devueltos por el sistema respecto a cada una de las preguntas que se le han lanzado. En tareas como la del *Geo-CLEF* o *NTCIR GeoTime*, previamente descritas en la sección 1.1, dichas evaluaciones se han resuelto mediante valores binarios, es decir, diciendo si un documento es relevante o no para la pregunta realizada. Posteriormente, con dichos valores binarios se utilizó el *MRR*[26] (ver ecuación 2.1), aunque también es cierto que en el *NTCIR GeoTime*, dicho juicio se hizo añadiendo una mayor escala de valores (completamente relevante, goegráficamente relevante, temporalmente relevante, relevante de algún modo y no relevante) según lo relevante que fuera.

[26] *MRR* (*Mean reciprocal rank*) es una medida estadística de la evaluación de cualquier proceso que genera una lista de posibles respuestas a una muestra de consultas, ordenada por probabilidad de corrección. El rango recíproco de una respuesta de la consulta es el inverso multiplicativo de la fila de la primera respuesta correcta. El rango de reciprocidad media es el promedio de los rangos recíprocas de resultados para una muestra de consultas Q.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{2.1}$$

donde $|Q|$ es el conjunto de consultas que se van a evaluar en el sistema y $rank_i$ hace referencia a la posición del primer documento relevante devuelto por el sistema para la pregunta i.

En *GIR* hay que tener en cuenta que se añade una nueva dimensión a evaluar, la geográfica, por lo que un documento puede ser relevante en cierto grado geográficamente hablando, y en otro grado en cuanto a contenido.

Por ejemplo, en la tarea *GeoTime* del *NTCIR* 2011 (Mitamura et al., 2008), se emplearon tres métricas distintas para la evaluación de los documentos:

1. *Average Precision (AP)*, es la que tradicionalmente utilizan en las conferencias *TREC*[27], *NTCIR* y *CLEF*. La precisión promedio (*Average Precision - AP*) son métricas basadas en toda la lista de documentos retornada por el sistema dada una consulta. Para sistemas que crean un ranking de los documentos retornados para una consulta es deseable considerar además el orden en que los documentos retornados son presentados. Si se computa la precisión y los documentos recuperados en cada posición de la secuencia de documentos con ranking, podemos trazar la curva de precisión-cobertura, trazando la precisión $p(r)$ como una función de los documentos obtenidos. La *AP* computa el promedio de los valores de $p(r)$ sobre la integral desde $r = 0$ hasta $r = 1$.

2. *Q-measure (Q)* (Sakai, 2007), a diferencia de la medida anterior, añade la posibilidad de que cada valor devuelto pueda tener un mayor rango de valores, es decir, los documentos devueltos no tienen por qué ser solamente relevantes o irrelevantes, sino que también pueden ser parcialmente relevantes, como es en el caso de los documentos del *NTCIR-GeoTime*. Por ejemplo, si se plantea la pregunta "*¿Cuándo y dónde tuvo lugar el accidente del transbordador espacial Columbia?*", y la respuesta devuelta por el sistema comprende documentos que indican el lugar del accidente pero no la fecha, o viceversa, la respuesta se debe considerar como parcialmente relevante.

3. *normalised Discounted Cumulative Gain (nDCG)* (Järvelin and Kekäläinen, 2002), es una medida de la calidad del ranking obtenido que está basada en *DCG*. En la *IR*, *DCG* a menudo se utiliza para medir la eficacia de los algoritmos de los motor de búsqueda web o aplicaciones relacionadas. Utilizando una escala de

[27] *Text REtrieval Conference*(*TREC*) está formado por una serie de *workshops* que se centran en una lista de diferentes tareas relacionadas con la investigación en el área recuperación de información (*IR*). http://trec.nist.gov/

relevancia graduada de los documentos devueltos por un motor de búsqueda, *DCG* mide la utilidad o ganancia de un documento sobre la base de su posición en la lista de resultados. La ganancia se acumula desde la parte superior de la lista de documentos devueltos por el sistema *IR*, hasta la parte inferior de dicha lista, con la ganancia de cada resultado previo descontada. Las listas de resultados de búsqueda varían en longitud dependiendo de la consulta. Al comparar el rendimiento de un motor de búsqueda de una consulta con la siguiente no se puede lograr un uso consistente utilizando solamente *DCG*, por lo que se hace necesaria una normalización de la ganancia acumulada en cada posición, pasándose de *DCG* a *nDCG*.

La evaluación en la tarea *GeoTime* del *NTCIR* se efectuó recogiendo los n primeros documentos que cada uno de los sistemas de los participantes entendió que eran relevantes para cada una de las consultas efectuadas. Dichos documentos se evaluaban de la siguiente manera:

- Relevante. Si contestaba a la pregunta dónde y cuándo.
- Parcialmente relevante (dónde). Si contestaba a la pregunta dónde.
- Parcialmente relevante (cuándo). Si contestaba a la pregunta cuándo.
- Parcialmente relevante (otro). Si contestaba de alguna manera a la pregunta.
- Irrelevante. Si no contestaba a la pregunta.

2.2. Recursos Geográficos

Los recursos geográficos son un conjunto de herramientas que están disponibles para su utilización y asistencia en gran parte de las tareas que debe satisfacer un sistema *GIR* (ver sección 2.1).

Dentro de los recursos geográficos, existen dos grandes divisiones, los nomemclátores (*gazetteers*) y los reconocedores de entidades nombradas (*NERs*).

2.2.1. Nomenclátores

El principal y más recurrente recurso que se utiliza en varios de los procesos realizados por un sistema *GIR*, tales como la detección de topónimos, desambiguación de topónimos, etc., son los nomenclátors (*gazetteers*). Un *gazetteer* es un diccionario geográfico o directorio usado junto a un mapa o atlas. Los *gazetters* típicamente contienen información relativa a la distribución geográfica, las estadísticas sociales y las características físicas de un país, región o continente. El contenido de estos diccionarios geográficos

pueden tener la ubicación de los términos expuestos en ellos, la dimensiones de las montañas y vías navegables, población, el PIB de un lugar, su índice de alfabetización, etc. Esta información está normalmente dividida por temas, con cada una de sus entradas ordenadas alfabéticamente.

Entre los *gazetteers* más utilizados por la comunidad científica nos encontramos con:

- "***Alexandria Digital Research Library***"[28] (Frew et al., 1998), la cual contiene alrededor de 5,9 millones de nombres de lugares, con distintos tipos de información y características para todas las entradas. El *gazetteer* está principalmente centrado en Estados Unidos (se utilizan topónimos de los Estados Unidos de la base de datos *GNIS* del Servicio Geológico de los EE.UU.), aunque también contiene alrededor de 4 millones de topónimos de fuera de los Estados Unidos provenientes de la *GEOnet Names Server*.

- ***GeoNames***[29] es una base de datos (creciente) de más de 10 millones de nombres de lugares (y códigos postales) de los países de todo el mundo. Las localizaciones se representan mediante puntos y contienen una amplia gama de tipos de entidad (por ejemplo, países, regiones, ciudades, acuíferos, etc.) La base de datos se puede descargar y utilizar de forma gratuita. Los datos se basan en varias fuentes, entre ellas el *GEOnet Names Server* y *Wikipedia*. Este es posiblemente el recurso más utilizado por la comunidad científica.

La base de datos geográfica de *GeoNames* está disponible para su descarga gratuita bajo licencia *Creative Commons*. Contiene más de 10 millones de nombres geográficos y se compone de más de 8 millones de características únicas, es decir, características georeferenciadas y por ende desambiguadas, de las cuales, 2,8 millones son nombres de poblaciones y 5,5 millones son nombres alternativos. Todas sus características están clasificadas dentro de una de las nueve clases principales, habiendo más niveles de subcategorización en sus 645 códigos de características.

Los datos son accesibles de forma gratuita a través de una serie de servicios web y una base de datos guardada diariamente. *GeoNames* atiende más de 30 millones de solicitudes de servicios web al día.

GeoNames integra datos geográficos tales como nombres de lugares en varios idiomas, datos de elevación, población y muchos otros más de diversa índole. Todas las coordenadas latitud-longitud están en formato *WGS84* (*World Geodetic System 1984*). Los usuarios pueden editar manualmente, corregir y añadir nuevos nombres utilizando una interfaz de usuario *wiki*. *GeoNames* tiene embajadores en muchos países que colaboran con su ayuda y experiencia.

[28] http://www.alexandria.ucsb.edu/
[29] http://www.geonames.org/

2.2.2. Reconocedores de entidades nombradas

Una de las principales utilidades de los *gazetteers* es la de la asistencia en la detección de entidades geográficas, para lo cual, hay diversas herramientas disponibles que realizan esta función. Estas herramientas hacen el trabajo de un *NER* aunque teniendo sólo en cuenta las entidades geográficas. También suelen desempeñar las funciones de un desambiguador, e incluso en ocasiones pueden acotar zonas geográficas difusas (ver secciones 2.1.1, 2.1.2 y 2.1.3).

Un ejemplo de de estas herramientas es *Yahoo! Placemaker*, el cual está actualmente integrado dentro de *Yahoo! Boss Geo* con el nombre de *Yahoo BOSS PlaceSpotter*. Esta herramienta consta de una API[30] de libre disposición que permite a los desarrolladores y editores crear aplicaciones y conjuntos de datos con conciencia geográfica determinando la ubicación del contenido desestructurado (publicaciones en blogs, artículos de noticias, textos de canales web (*feed*), páginas web, etc.). Para utilizarlo se le ha de pasar un texto desestructurado al sistema. A continuación, *Placemaker* identifica, desambigua y extrae las referencias geográficas que encuentra, devolviendo, finalmente, los metadatos geográficos que ubican tanto a los contenidos estructurados como a los desestructurados.

2.3. Conclusiones

Como se ha comentado en este capítulo, los sistemas de recuperación de información geográfica (*GIR*) resultan de gran utilidad a la hora de manejar consultas que comprenden un grado de inteligencia geográfica, dado los pobres resultados obtenidos cuando se realizan dichas búsquedas mediante sistemas de recuperación de información (*IR*) convencionales.

Así pues, a la hora de crear un sistema *GIR*, se ha de afrontar teniendo en cuenta los diversos problemas que se han de superar:

- **Detección de referencias geográficas**, donde hay que saber qué términos encontrados en el texto analizado representan alguna ubicación geográfica.
- **Desambiguación de topónimos**, dónde se debe identificar de forma inequívoca a qué lugar se está refiriendo la referencia geográfica encontrada y el texto del documento completo en sí (foco geográfico) para su posterior indexación geográfica.
- **Terminología geográfica vaga o imprecisa**, donde se ha de hacer frente a zonas geográficas que no se encuentran delimitadas ni en ningún *gazetteer*.

[30] Interfaz de programación de aplicaciones (*API*) (del inglés *Application Programming Interface*) es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta librería para ser utilizado por otro software como una capa de abstracción.

- **Indexación espacial y textual**, dónde se ha de combinar la indexación del texto de los sistemas de *IR* convencionales con la geográfica.
- **Ranking por relevancia geográfica**, el cual debe satisfacer las restricciones geográficas de la consulta además de las específicas de ésta.
- **Interfaces de usuario**, las cuales deben permitir tanto el poder realizar una consulta geográficamente acotada, como su correcta visualización.
- **Métodos de evaluación de los sistemas *GIR***, donde debe comprobarse qué aportan dichos sistemas frente a los sistemas de *IR* tradicionales.

Por otro lado, para la correcta creación de un sistema *GIR* se hace imprescindible el uso de recursos geográficos, entre los que cabe destacar los nomenclátores (*gazetteers*) y los reconocedores de entidades nombradas (*NERs*) orientados a entidades geográficas.

3

Detección del foco geográfico

Como ya se ha explicado en el capítulo anterior, los sistemas de recuperación de información geográfica (sistemas *GIR*) se ocupan de la obtención de documentos de un corpus dado, por orden de relevancia a la temática y restricciones geográficas pertinentes, en respuesta a una pregunta de la forma <*qué se busca*> + <*relación*> + <*localización*> (ver sección 2.1.6). Por ejemplo, "*Templos* (<*qué se busca*>) *en un radio de 5 km. de* (<*relación*>) *Tokio* (<*localización*>)" (Larson and Larson, 1996).

Los sistemas *GIR*, como las librerías geográficas digitales y los motores de búsqueda web dotados de consciencia espacial, se basan en un conjunto de recursos de información geo-referenciados y en métodos para buscar espacialmente dentro de estos recursos, siendo el foco geográfico clave para dichas tareas.

Uno de los objetivos principales de esta tesis es la detección del foco geográfico en textos. Por este motivo se ha considerado oportuna la inclusión de un capítulo completo que determine el estado de la cuestión actual en este campo, resumiendo así los mejores trabajos realizados hasta la fecha en dicho campo.

Un recurso de información se considera que está geo-referenciado si está espacialmente indexado (ubicado) en una o más regiones del planeta, donde las localizaciones específicas de estas regiones están codificadas directamente como coordenadas espaciales o indirectamente por el nombre del lugar (Hill, 2006). Sin embargo, a fin de que los nombres de lugares puedan ayudar a los sistemas *GIR* a tener un enfoque espacial, éstos deben estar asociados a un modelo geográfico, es decir, cada uno de los nombres de lugar existentes en el texto debe estar desambiguado y por ende estar establecida su ubicación concreta.

La identificación del foco geográfico de un documento consiste en determinar la principal o principales localizaciones a las que hace referencia un texto de entre todas las que se nombran en él (Amitay et al., 2004).

De acuerdo con la primera ley de la geografía, o principio de autocorrelación espacial: "Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes." (Tobler, 1970). De este principio se puede extraer directamente la gran importancia

que tiene la detección del foco geográfico en un documento de texto para saber si la materia sobre la que trata dicho documento puede estar más o menos relacionada con la que busca una consulta dada.

Las búsquedas con restricciones o inteligencia geográfica son bastante frecuentes en un amplio abanico de campos. Un gran número de documentos (webs, noticias, etc.) tienen un foco geográfico.

La idea que subyace a la hora de determinar el foco geográfico de un documento es la de que si varios topónimos de una misma región son mencionados, esto debe significar que esta región es el foco geográfico.

En ocasiones no se puede decir que una página tenga solamente un único foco geográfico, ya que podría estarse hablando de dos países separados por miles de kilómetros, por ejemplo, aunque conviene tener en cuenta la premisa de que si una región más pequeña es el foco geográfico real de la página, no se debe devolver un ámbito mayor.

El mayor problema a la hora de determinar el foco geográfico de un documento son los errores que se arrastran de las fases previas a esta detección, es decir, los errores a la hora de detectar inequívocamente todos los topónimos que aparecen en el texto.

Para extraer correctamente el foco geográfico en un texto se debe determinar el grado de relevancia que tienen para un documento dado las entidades geográficas presentes en él. Si bien la desambiguación de topónimos es una tarea ampliamente tratada dentro del campo de la *GIR*, no todos los sistemas de este tipo determinan el grado de relevancia de las entidades geográficas que en él aparecen.

Entre los problemas que se deben afrontar a la hora de determinar el foco geográfico de un texto se encuentra la omisión de la mención explícita de dicho foco, es decir, un texto puede estar asociado a uno o más lugares, a pesar de que su foco geográfico pueda comprender áreas no mencionadas directamente en el texto.

Por ejemplo, si un texto trata sobre países mediterráneos, y países como España, Francia o Italia no se mencionaran explícitamente en dicho texto, éstos estarán asociados al ámbito geográfico del documento. Por tanto, si una consulta comparte temática con dicho documento, debería ser devuelto por el sistema *GIR* siempre y cuando la restricción geográfica comprenda alguno de los países implícitamente incluidos en el ámbito geográfico del texto, sin ser necesario que la consulta abarque todo el ámbito completo. Poniendo un ejemplo más concreto, supongamos que se formula la siguiente consulta: "*Cruceros por el mar Mediterráneo*", y que nos encontramos con un documento que habla sobre cruceros entre la costa levantina y las islas baleares (ambas en España). Según lo aquí expuesto, dicho documento tendría una alta relevancia debido a que coincide la temática de la pregunta y a que el ámbito o foco geográfico del documento está implícitamente incluido en el de la consulta.

De igual modo, cada lugar dentro de cada uno de los países del área mencionada también tiene asociado un valor de relevancia, es decir, cada lago, río, ciudad, punto de interés, etc., dentro de estos países (costa mediterránea española en el ejemplo mostrado) implícitamente mencionados, tienen un valor de relevancia o peso asociado acorde a su importancia dentro de la región geográfica tratada. En el ejemplo anterior, lugares como la ciudad de Palma de Mallorca, la ciudad de las artes y las ciencias de Valencia, etc., adquirirán cierta relevancia por hallarse dentro del área geográfica comprendida por el documento y la consulta dada.

Igualmente, también podríamos encontrarnos con la situación opuesta, es decir, que una consulta haga referencia a un lugar el cual es una entidad geográfica superior a la del foco geográfico de un documento. Siguiendo con el ejemplo anterior, si existe un documento que versa sobre un tema que tiene su foco geográfico en Europa, y hay una consulta que pregunta por el mismo tema del documento pero con restricción geográfica de los países mediterráneos, el sistema debería darle un valor de relevancia a dicho documento.

Así pues, un sistema de *GIR* no tiene únicamente que buscar documentos geográficamente indexados que coincidan plenamente con la consulta lanzada, sino que tiene que buscar una intersección de dichas restricciones geográficas, amén de las propias de la temática de la consulta, tal y como haría un sistema convencional de recuperación de información.

La inmensa mayoría de aproximaciones que intentan detectar el foco geográfico en textos formales se componen de tres fases:

1. **Detección de topónimos**. Esta tarea suele ser afrontada mediante la utilización de *gazetteers* que permitan cruzar los términos encontrados en el texto de los documentos con los términos existentes en el *gazetteer*. Los *gazetteers* también son de gran utilidad para la obtención de una visión jerárquica de los topónimos encontrados. Con el fin de solventar el problema de los topónimos que están compuestos por más de un término, así como los que puede que no aparezcan en los *gazetters*, se suelen utilizar los reconocedores de entidades nombradas, ya que estos aplican técnicas de *PLN* y patrones lingüísticos para identificar los nombres de lugar.

2. **Desambiguación de los topónimos encontrados**. La investigación en el campo de la detección del foco geográfico en textos tiene como uno de sus mayores desafíos la resolución de topónimos. En cuanto a lo que ambigüedad geográfica se refiere, nos podemos encontrar con dos tipos distintos de ésta:

    a) *geo/non-geo*, donde un topónimo ambiguo puede pertenecer también a otro tipo de categoría gramatical del idioma del texto. Por ejemplo, si encontramos la palabra '*Granada*', ésta puede

hacer referencia a un topónimo, aunque también podría tener diversos significados no relacionados con la geografía, tales como la fruta, el proyectil, el adjetivo, etc.

b) *geo/geo*, donde el mismo nombre de un topónimo existe en más de un lugar. Siguiendo con el ejemplo anterior, si se sabe con certeza que el término '*Granada*' representa un topónimo, habría que saber a qué lugar hace referencia, la ciudad andaluza, su provincia homónima, cualquiera de las *Granadas* del continente americano, etc.

Para resolver la problemática *geo/geo* se suele tener en cuenta las 3 siguientes pautas:

- Priorización de localizaciones, donde en caso de ambigüedad se suele optar por la población con más habitantes.
- Búsqueda de desambiguadores, para lo cual se suele buscar en la jerarquía del topónimo a desambiguar con el fin de que haya una coincidencia de alguna entidad geográfica superior de dicho topónimo ambiguo que aparezca en el texto.
- Minimalidad espacial, donde se debe priorizar por los topónimos que hagan que el foco geográfico sea lo más estrecho posible. Para ello, se suelen basar en los topónimos más cercanos no ambiguos y/o en la combinación de los topónimos ambiguos que minimicen el ámbito geográfico.

En este sentido, también es muy común el intentar desambiguar los topónimos mediante los términos colindantes al supuesto topónimo en busca de una aclaración que lo identifique inequívocamente.

3. **Cálculo del foco geográfico**. La buena detección del ámbito geográfico depende en gran medida del acierto en los dos pasos previos.

   Esta detección puede dar como resultado más de un foco geográfico, aunque habría que tener como premisa obtener la región más pequeña posible que represente el ámbito geográfico del texto analizado. Para ello se hace necesaria la introducción de la jerarquía geográfica que permita detectar ámbitos geográficos que pueden incluso no ser mencionados explícitamente en los textos, tales como entidades geográficas superiores o las inferiores.

En las próximas secciones se podrá ver de qué manera se ha tratado dicha problemática en la literatura así como la evaluación realizada sobre los resultados obtenidos en esta tarea.

3.1. Detección de topónimos

Tal y como ya se describió en la sección 2.1.1, la detección de topónimos (*geotagging*) intenta identificar los nombres de lugares que aparecen en un texto dado. Para dicho cometido, se suelen emplear 3 aproximaciones distintas, las cuales pueden combinarse:

1. Búsquedas basadas en *gazetteers*. Éstas búsquedas emparejan los términos existentes en los *gazetteers* con los existentes en los textos, siendo también de gran utilidad la jerarquía geográfica que los propios *gazetteers* aportan.

2. Búsquedas basadas en reglas. En esta ocasión la búsqueda de topónimos se suele basar en técnicas de *PLN* que son capaces de detectar cuándo uno o más términos son nombres de lugar por el contesto que los rodea. Para unos mejores resultados se suele utilizar en combinación con las otras dos técnicas aquí expuestas, especialmente con los *gazetteers*.

3. Búsquedas basadas en aprendizaje automático. Se valen de conjuntos de entrenamiento donde están anotados los topónimos para saber en un futuro qué términos pueden ser nombres de lugar.

Si nos centramos en los trabajos más relevantes en la detección del foco geográfico, podemos ver cómo las aproximaciones más exitosas en primer lugar tratan la tarea de la detección de topónimos.

Así pues, en el trabajo realizado por Ding et al. (2000) para la detección del foco geográfico basándose meramente en el contenido textual, los autores tratan la problemática de la detección de topónimos utilizando el reconocedor de entidades nombradas (*NER*) del sistema *Alembic Workbench*[1] desarrollado por *The MITRE* (Day et al., 1997), para realizar una aproximación basada en reglas.

Por otro lado, en el trabajo llevado a cabo en Amitay et al. (2004), entre las técnicas más extendidas explicadas en la sección 2.1.1 para realizar dicha tarea, sus autores se decantaron por la utilización de *gazetteers*, ya que dicha aproximación no requería de un conjunto de entrenamiento, el cual resulta un tanto difícil de crear o encontrar en ocasiones. Además, debido a que los algoritmos de aprendizaje automático suelen ser más complejos, requieren de mucho más tiempo para procesar los datos.

Para la detección y desambiguación de los topónimos, los autores crearon un *gazetteer* obteniendo los datos de las siguientes fuentes: *GNIS*[2] para las localizaciones en EE.UU., *World-gazetteer.com*[3] para las localizaciones fuera

[1] http://annotation.exmaralda.org/index.php/Alembic_Workbench
[2] *GNIS - USGS Geographic Names Information System* http://geonames.usgs.gov.
[3] *World Gazetteer* http://www.world-gazetteer.com.

de los EE.UU., *UNSD*[4] para los continentes y los países, *ISO 3166-1*[5] para obtener las abreviaturas de los países, así como otras fuentes no especificadas para la obtención de las abreviaturas de los estados. Este *gazetteer* es la piedra angular de dicho sistema.

Para obtener los topónimos, en esta fase se cruzaron todos los términos pasados a minúsculas de la página web con los términos geográficos del *gazetteer*. Las abreviaturas no fueron tomadas en cuenta en esta fase.

En el trabajo llevado a cabo en Zong et al. (2005), para afrontar la tarea de detección de entidades geográficas, los autores utilizaron el *NER* del prestigioso proyecto *GATE*[6] (Cunningham et al., 2002). Se construyó un nuevo *gazetter* que contenía principalmente topónimos de EE.UU. para asistir a *GATE* en la extracción de los topónimos existentes en las páginas web de *DLESE*[7], es decir, este sistema combinó la búsqueda de topónimos basada en reglas con los *gazetteers*.

GATE está formado por un *tokenizador*[8], un separador de frases, un entiquetador gramatical (*POS tagger*) y un emparejador por ontología[9]. Los tipos de entidades nombradas que se pueden extraer mediante *GATE* son personas, lugares, organizaciones, fechas, identificadores alfanuméricos y monedas, aunque en esta implementación tan sólo se utilizaron las entidades de lugar.

Para la extracción de entidades de lugar en *GATE*, los autores incorporaron un diccionario geográfico. El diccionario geográfico predeterminado en *GATE* está compuesto por 6.713 nombres de lugar de diferentes países. Los topónimos que aparecen en cada página web del corpus fueron identificados directamente emparejándolos con el *gazetteer*. *GATE* también aplica *PLN* y patrones lingüísticos para identificar los nombres de lugares que pueden no aparecer en el *gazetteer*.

Como una gran parte de las páginas web referenciadas por el proyecto *DLESE* pertenecen a EE.UU., y el diccionario geográfico predeterminado de *GATE* no tiene suficiente información acerca de las ubicaciones de EE.UU., los autores decidieron incorporar a *GATE* el *gazetteer US Census 2000*[10] para que los topónimos de este país se pudieran extraer con mejor precisión. El nuevo *gazetteer* también facilita la construcción de una visión jerárquica de los topónimos estadounidenses basándose en las relaciones

[4] Departamento de Naciones Unidas para asuntos sociales y económicos http://unstats.un.org/unsd.

[5] Lista de códigos ISO 3166 http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=63545.

[6] http://www.gate.ac.uk

[7] Digital Library for Earth System Education. http://www.dlese.org.

[8] Separador de términos en un texto.

[9] La ontología se considera como el repositorio de conceptos que establecen conexiones entre los símbolos de una lengua y sus referentes en el mundo o submundo que se contempla.

[10] US Census Bureau. http://www.census.gov.

administrativas, las cuales se pueden utilizar en la desambiguación de los nombres de lugares extraídos y para la asignación del foco geográfico. Dado que a los topónimos en el *gazetteer US Census 2000* se les añade frecuentemente sufijos comunes (p. ej. "*city*", "*town*"), se crearon alias para algunos topónimos eliminando dichos sufijos y finalmente se añadieron también al *gazetteer*.

Los topónimos fueron clasificados en 4 grupos: estado, ciudad, condado y subdivisión de condado.

3.2. Desambiguación de topónimos

Como ya se ha comentado en la sección 2.1.2, la desambiguación de topónimos y la detección del foco geográfico están estrechamente ligadas debido a la necesidad imperiosa que se requiere para inducir el foco geográfico de un texto de partir de los inequívocos topónimos que aparecen en el mismo, es decir, si se quiere obtener correctamente el ámbito geográfico de un texto hay que saber previamente qué lugares concretos se mencionan en dicho texto resolviendo las posibles ambigüedades de los topónimos encontrados.

Según Buscaldi and Rosso (2008), existen dos aproximaciones fundamentales para la desambiguación de topónimos:

1. **Aproximación basada en mapa**. Esta aproximación se basa en el uso de información geográfica cuantitativa, empleando propiedades espaciales y geométricas de las localizaciones encontradas en el texto, como puede ser el cálculo de distancias entre lugares o el cálculo del centroide de un área geográfica (Leidner, 2008).

2. **Aproximación basada en conocimiento**. Esta aproximación se basa en la utilización de información geográfica cualitativa, empleando herramientas de (*PLN*) y conocimiento externo mediante el uso de diccionarios geográficos (*gazetteers*) y/o ontologías (Garbin and Mani, 2005).

La investigación en el campo de la detección del foco geográfico en textos se centra casi por completo en encontrar y resolver los topónimos (*toponym resolution*) (Serdyukov et al., 2009). Una detallada y completa descripción de la heurística para la resolución de topónimos la podemos ver en Leidner (2008), aunque la mayoría de aproximaciones para resolver dicha problemática se derivan de las siguientes tres ideas:

1. **Priorización de localizaciones**. Incluso sin ningún contexto, es posible hacer una conjetura inteligente sobre el referente más probable para un topónimo simplemente teniendo en cuenta las probabilidades previas de cada lugar.

Por ejemplo, los lugares con poblaciones más grandes, o los más frecuentemente menciones en los textos, son los candidatos más probables.

2. **Búsqueda de desambiguadores**. Se supone que cada lugar tiene una lista de desambiguadores, como pueden ser sus lugares vecinos en la jerarquía de un *gazetteer*, que resuelven la ambigüedad si se encuentran en la proximidad del topónimo mencionado.

   Por ejemplo, tanto Francia como el estado estadounidense de Texas son desambiguadores de la ciudad de París, que además de ser la capital del país europeo es una pequeña ciudad del estado estadounidense.

3. **Minimalidad espacial**. Si se mencionan varios topónimos en el texto sin que haya desambiguadores, estos lugares son seleccionados como referentes que minimizan el rectángulo delimitador mínimo que comprende dichos topónimos o la suma de sus distancias por pares entre sí.

   Por ejemplo, si Moscú y Helsinki aparecieran juntos en un texto, el Moscú de Rusia sería el seleccionado en vez del de Idaho (EE.UU.), ya que el ruso está miles de kilómetros más cerca de Helsinki, y Helsinki no es ambiguo (Serdyukov et al., 2009).

Otro problema adicional a afrontar a la hora de desambiguar topónimos son los alias, ya que frecuentemente nos podemos encontrar con distintos nombres que hacen referencia a la misma localidad. Por ejemplo, la ciudad de Nueva York puede ser referenciada en inglés como: *New York City*, *New York*, *Manhattan*, *NY City*, *NYC* o *Manhattanville*.

Siguiendo con el trabajo expuesto en Ding et al. (2000), en relación a la tarea de normalización y desambiguación de los topónimos, los autores tuvieron que identificar de manera inequívoca los nombres de las localidades mencionadas en los documentos web, para lo cual hicieron frente a la problemática de los alias creando una lista de los mismos obtenida del servicio postal de los Estados Unidos (*USPS*[11]), ya que dichos alias suelen estar bien delimitados.

Por otro lado, los autores realizaron la desambiguación de los topónimos teniendo en cuenta la ubicación del resto de topónimos que no eran ambiguos. Por ejemplo, si en el recurso web w se mencionan principalmente localidades del estado de Nueva York, se asume que la referencia a *Manhattan* es la de la Ciudad de Nueva York y no el *Manhattan* de Kansas.

Por otro lado, si miramos el trabajo expuesto en Amitay et al. (2004) en este mismo campo, la desambiguación, tal y como ya se ha dicho en la sección 2.1.2, debe hacer frente a distintos tipos de ambigüedades, tales como *geo/non-geo* y *geo/geo*. En cuanto a la desambiguación *geo/non-geo*, para

[11] http://www.usps.gov

identificar este tipo de nombres, los autores contaron el número de veces que cada uno de estos términos aparecía en un corpus de páginas web bien editadas[12], es decir, que estaban escritas en un lenguaje altamente formal. Se hicieron dos tipos de pruebas:

- Los nombres que aparecieron más de 100 veces pero mayoritariamente no empezaban por mayúsculas tal y como lo haría un nombre de lugar, fueron incluidos en una sección *non-geo* del *gazetteer* creado.

- Los nombres mencionados muchas más veces de lo que de su población se podría esperar fueron incluidos también en la sección *non-geo* del *gazetteer*.

En esta ocasión, los autores tuvieron que repasar manualmente los resultados para eliminar errores obvios y añadir algunos términos no detectados.

En cuanto a la desambiguación *geo/geo*, el algoritmo de desambiguación empleó las siguientes heurísticas para cada término obtenido en la fase anterior:

1. Se comprueba si los términos colindantes al supuesto topónimo encontrado en el texto pueden identificar inequívocamente a la localización.

   Por ejemplo, "*Chicago (IL)*", donde *IL* identificaría inequívocamente a *Chicago*, ya que es la abreviatura del estado de *Illinois* donde se encuentra dicha ciudad.

2. Cada candidato a topónimo que quede sin resolver se le asigna un valor acorde al número de habitantes de la localización a desambiguar.

3. En caso de que en la página aparezca en más de una ocasión el mismo topónimo y sólo uno esté claramente definido, el resto heredarán dicha inequívoca referencia.

4. Los topónimos aún no resueltos se desambiguarán por el contexto, es decir, si entre los topónimos no resueltos se encuentra una entidad geográfica superior común, estos topónimos serán los pertenecientes a dicha entidad.

   Por ejemplo, si en un texto en inglés nos topamos con los topónimos "*London*" y "*Hamilton*", se podría tener como posibles interpretaciones de *London* Inglaterra u Ontario (Canadá). Mientras que para *Hamilton* podrían ser Ohio (EE.UU.) u Ontario (Canadá), el cual es el único factor común entre los dos topónimos expuestos y por ende el que se utilizará para la desambiguación de ambos.

[12] El corpus consistió en alrededor de 1.200.000 páginas obtenidas bajo el dominio *.gov*.

Retomando el trabajo llevado a cabo en Zong et al. (2005), en el cual se utilizaba *GATE* para la detección de entidades geográficas, y dado que *GATE* lleva incorporados patrones para la extracción de ambigüedades *geo/non-geo*, pero no así para las *geo/geo*, los autores de este trabajo se centraron en resolver las ambigüedades de éste último tipo asumiendo que los topónimos encontrados en una misma web no tienen porqué pertenecer al mismo ámbito geográfico. Para ello realizaron una aproximación basada en reglas que hacía uso tanto de la información del contexto como de las distancias espaciales entre los topónimos.

Así pues, para desambiguar los topónimos obtenidos en la fase de detección de topónimos se procedió a realizar los siguientes 4 pasos:

1. Se buscaron patrones lingüísticos en el contexto próximo (las tres palabras posteriores) al topónimo encontrado. Los patrones eran del tipo:
   - Topónimo + "," + sentido del topónimo (p. ej. "*Chicago, an old city*").
   - Sentido del topónimo + "*of*" + topónimo (p. ej. "*state of California*").
   - Topónimo + sentido del topónimo (p. ej. "*Rio Grande County*").
   - Nombre de estado que aparecen en el contexto cercano (p. ej. "*Philadelphia is in PA*").
2. Emparejado perfecto con los nombres que aparecen en el *gazetteer*. Si encaja perfectamente el topónimo encontrado con un topónimo inequívoco existente en el *gazetteer*, entonces éste ya está desambiguado.
3. Propagación de topónimos desambiguados. Partiendo del topónimo no ambiguo más cercano al que se quiere desambiguar se comprueban los siguientes patrones lingüísticos:
   - Topónimo + "," + nombre de estado (p. ej. "*Denver, Colorado*").
   - Topónimo + "*of*" + nombre de estado (p. ej. "*Berkeley of California*").
   - Topónimo + nombre de estado (p. ej. "*Buffalo NY*").
   - Topónimo + "—" + nombre de estado o de ciudad (p. ej. "*Wisconsin — Minnesota*").
   - Nombre de estado o de ciudad + "—" + topónimo (p. ej. "*Washington — Aberdeen*").

   Una vez se ha buscado alguno de los anteriores patrones, se comprueba que el topónimo no sea aún ambiguo en el *gazetteer*.

4. Desambiguación basada en la distancia espacial. Si el topónimo encontrado todavía continúa siendo ambiguo, entonces se procede a desambiguarlo según la distancia al topónimo no ambiguo más cercano.

3.3. Cálculo del foco geográfico

Un documento puede tener dos tipos distintos de geografía asociada:

1. **El destino**. El destino geográfico viene determinado por el contenido del documento y está relacionado con el tema que se está tratando en dicho documento.

2. **La fuente**. La fuente geográfica tiene que ver con el origen del documento, especialmente cuando dicho documento es una página web, la localización de procedencia o localización física de del servidor donde está almacenado, la dirección de su autor o propietario, etc.

Una página web puede tener dos tipos distintos de geografía asociada:

1. **La fuente**. La fuente geográfica tiene que ver con el origen de la página web, la localización física del servidor donde está almacenada, la dirección de su autor o propietario, etc.

2. **El destino**. El destino geográfico viene determinado por el contenido de la página web y está relacionado con el tema que se está tratando en dicha página.

Si nos centramos en la geografía asociada al destino (que es la que nos atañe aquí, ya que es la que se puede inferir de su contenido textual), una vez solucionada la problemática de la obtención y desambiguación de topónimos en los textos, se conseguirá como resultado por cada documento del corpus una lista de topónimos que han sido mencionados, por lo que ya se puede proceder a deducir el foco geográfico de cada documento.

Así pues, cada mención del topónimo obtenido en estas fases previas añade cierto valor como candidato a ser el foco geográfico de dicha página, así como también lo hace en menor medida a cada miembro de su jerarquía geográfica superior e incluso inferior.

De esto se desprende que para la detección del foco geográfico de un texto, no solamente es necesaria la inequívoca localización de cada uno de los topónimos que aparecen en dicho texto, sino que también se hace necesaria la introducción de una jerarquía geográfica, es decir, el conocer las entidades geográficas superiores e inferiores de las explícitamente mencionadas en el texto.

Debido a la ineficiencia de los motores de búsqueda convencionales a la hora de identificar el foco geográfico de los recursos web, en Ding et al. (2000) se plantearon por primera vez la resolución de dicha problemática.

Los autores definieron el ámbito geográfico de un recurso web w como el área geográfica que el autor de w pretende abarcar.

En esta investigación, los autores se centraron en descubrir el foco geográfico dentro de un único país, Estados Unidos, con una jerarquía de tres niveles: país, estado y ciudad. En el algoritmo utilizado para la detección del foco geográfico, se definieron dos medidas: la potencia (*power*) para medir el interés, y la divulgación del recurso (*spread*) para medir la uniformidad, es decir, que los topónimos que aparecen en el texto estén uniformemente repartidos en un área geográfica concreta y no concentrados una subárea de ésta. Los lugares con un interés significativo (gran potencia) y una distribución uniforme (gran divulgación) se consideraban el ámbito geográfico del documento web. Esta tarea se afrontó mediante dos enfoques distintos: la *distribución geográfica de los enlaces (hyperlinks) a los recursos webs* y un enfoque basado meramente en el *contenido textual del recurso web*.

Para la primera aproximación, considérese un recurso web cuyo ámbito geográfico es todo Estados Unidos (por ejemplo, el periódico *USA Today*[13]). Tal recurso es probable que sea relevante para todo el país. La hipótesis de esta investigación en este apartado es que este interés se traduzca en las páginas web de todo el país que contendrán enlaces *HTML* a este recurso web. Por el contrario, un recurso con un ámbito geográfico mucho más limitado exhibirá un patrón de distribución de enlaces significativamente distinto para todo el país. Por lo tanto, un camino prometedor para estimar el ámbito geográfico de un recurso es el estudio de la distribución geográfica de los enlaces a los recursos. Más específicamente, un lugar l tendrá que satisfacer dos condiciones para estar en el ámbito geográfico de un recurso w:

1. *Power*: un porcentaje significativo de las webs de la localidad l contendrán enlaces a w.

$$Power(w, l) = \frac{Enlaces(w, l)}{Webs(l)} \tag{3.1}$$

2. *Spread*: las páginas web en la localidad l que contienen enlaces al recurso web w se distribuyen uniformemente a lo largo de l, es decir, que no haya un alto número de enlaces desde una web de la localidad l al recurso web w y en el resto de webs de la localidad l el número de enlaces hacia el recurso web w sea escaso.

   Es decir, para indicar que el ámbito geográfico de un recurso web es el estado de Nueva York, por ejemplo, es necesario que no sólo los enlaces (*Power*) procedan de recursos ubicados en la ciudad de Nueva York, sino en el resto de ciudades del estado de una manera uniforme, sin

[13] http://www.usatoday.com/

que haya una ciudad claramente predominante sobre el resto, en cuyo caso, dicha ciudad sería el foco geográfico.

Para centramos en la segunda aproximación, la que está más relacionada con esta tesis por basarse meramente en el contenido textual del recurso web, considérese un recurso web cuyo ámbito geográfico es el estado de Nueva York. La asunción de los autores en este punto es que dicho recurso debería mencionar con más frecuencia ciudades del estado de Nueva York que cualquier otra ciudad del mundo. Por lo tanto, para resolver el ámbito geográfico del recurso web se debe estudiar la distribución de las localidades mencionadas en dicho recurso. Análogamente a la anterior aproximación, existen dos condiciones que una localidad l debe satisfacer para ser el ámbito geográfico de un recurso web w:

1. *Power*: un porcentaje significativo de las localidades mencionadas en w deben ser la propia localidad l o una "sublocalidad" de ésta.

$$Power(w,l) = \frac{Referencias(w,l)}{Localidades(w) \times Ciudades(l)} \tag{3.2}$$

   Donde *Localidades(w)* es el número de referencias a localizaciones geográficas en el texto w. *Referencias(w,l)* es el número de referencias a ciudades del conjunto l mencionadas en el texto de w. Por último, *Ciudades(l)* es un factor de escala definido como el número de ciudades en la localización l, siendo $Ciudades(l) = 1$ si l es una ciudad en sí misma.

2. *Spread*: Las localidades mencionadas en el recurso web w están distribuidas uniformemente a lo largo de las "sublocalidades" de l.

Dado que la realización del cálculo del foco geográfico se basa principalmente en contar las veces que cada ciudad ha sido nombrada en el texto, y una localización puede ser en este ejemplo el país, un estado o una ciudad, se hace necesario el asignar valores a las ciudades que hay por debajo del país y/o estado mencionado, es decir, que si se menciona el estado de Nueva York, por ejemplo, todas las ciudades de ese estado obtendrán un valor por cada vez que se mencione dicho estado (Ding et al., 2000).

Al hilo de esta investigación, en Amitay et al. (2004) podemos ver otro interesante trabajo donde los autores una vez más afrontan la problemática de la detección del foco geográfico en los recursos web.

Para la detección del ámbito geográfico de las páginas web, los autores previamente tuvieron que realizar la pertinente detección (*geotagging*) y desambiguación (*toponym grounding*) de las entidades geográficas nombradas a nivel de ciudad, estado y país, tal y como se ha explicado en las secciones 3.1 y 3.2.

Así pues, los objetivos de este sistema son la identificación de todas las menciones geográficas dentro de las páginas web, la asignación de una

localización geográfica con su nivel de confianza, y derivar el/los foco/s geográfico de cada página.

Para la obtención del valor de los distintos miembros de la jerarquía geográfica se ignoraron los topónimos que formaban parte de o englobaban a un lugar más importante, así como los que no obtienen un valor lo suficientemente relevante, es decir, si ya se tiene un lugar con una puntuación significativa, por ejemplo la provincia de Alicante, los topónimos por encima de esta entidad geográfica (Comunidad Valenciana, España y Europa) que no hayan sido mentados o no tengan suficiente relevancia, al igual que los que hay por debajo en las mismas condiciones (Alicante ciudad, Elche, Elda, etc.), se ignorarán. La razón por la que los topónimos encontrados contribuyan menos a la valoración de sus entidades geográficas superiores que a la suya propia es porque esto permitiría que las entidades geográficas superiores siempre ganaran a no ser que las entidades inferiores fueran el único topónimo encontrado en el texto. En el ejemplo expuesto, si en un texto aparece 11 veces el topónimo *Elche* y 1 vez *Torrellano*, si se puntuase por igual las apariciones de los topónimos en sí como para sus entidades geográficas superiores (la provincia de Alicante en este caso), la provincia de Alicante superaría a la ciudad de *Elche* en la valoración como foco geográfico, puesto que obtendría además la puntuación de la aparición del topónimo *Torrellano* en el ejemplo expuesto.

Otro trabajo interesante sobre esta materia es el realizado por Zong et al. (2005). Una vez más, se intentó obtener el foco geográfico de un conjunto de páginas web utilizadas en el proyecto *DLESE*. Además, no sólo intentaron detectar el foco geográfico de la página web, sino que también lo hicieron con los párrafos de más de 200 términos.

Por cada uno de estos párrafos se obtenían todos sus topónimos y sus ascendientes geográficos, asignándoles a todos ellos una puntuación que indicara la relevancia que tenía cada uno en dicho párrafo. Esta puntuación se basaba en el número de apariciones de un topónimo dado entre el número total de términos en ese párrafo por un factor de peso establecido para los ascendentes geográficos y los propios topónimos.

Estos pesos hacen que cuanto más uniformemente repartidos estén las menciones de los topónimos de una entidad geográfica superior, mayor será el peso obtenido por dicha entidad y por ende más se contribuirá a que dicha entidad sea el foco geográfico.

3.4. Evaluación

La evaluación de los sistemas que intentan detectar el foco geográfico dista mucho de ser algo trivial, dado que es muy difícil encontrar corpus donde sus documentos tengan anotados el foco geográfico al que pertenecen.

Debido a esto, la aproximación más recurrida para dicha evaluación dentro de la comunidad científica suele ser una validación manual.

Aún así, dicha aproximación muestra un alto número de inconvenientes:

- Coste económico.
- Coste temporal.
- Imposibilidad de obtener un alto número de documentos evaluados.
- Distintos criterios según la persona y el momento en el que se evalúe.
- Distintos grados de validez.
- Fallos humanos a la hora de evaluar.

Pese a estos inconvenientes, tal y como se ha dicho al comienzo de esta sección, la evaluación manual es la más utilizada debido a la falta de recursos disponibles.

Estas evaluaciones manuales se suelen hacer obedeciendo a la granularidad del foco geográfico, es decir, a qué nivel se quiere localizar el ámbito del texto, pudiendo ir estas granularidades desde nivel de código postal o coordenadas, hasta a nivel de país o conjunto de países.

También resulta común el asignar distintos grados de acierto a la detección dada por los sistemas según el solapamiento que haya entre la solución dada y la correcta. Por ejemplo, imaginemos que tenemos un texto cuyo ámbito geográfico son las islas baleares. Podrían darse las siguientes situaciones:

- Que el sistema indique que el ámbito geográfico son las islas baleares, ante lo cual se obtendría la máxima puntuación.
- Que el sistema identifique el foco geográfico en una entidad geográfica superior, por ejemplo España, ante lo cual se obtendría cierta puntuación por el solapamiento de los dos ámbitos, pero no la máxima puntuación.
- Que el sistema identifique el ámbito geográfico en una entidad geográfica inferior, por ejemplo Menorca, ante lo cual, análogamente a la situación anterior, se obtendría cierta puntuación por el solapamiento de los dos ámbitos, pero no la máxima puntuación.
- Que el sistema identifique el ámbito geográfico solapando parte del ámbito real y parte de otro. Por ejemplo si el sistema dijera que el ámbito geográfico es Ibiza y Alicante. Análogamente a las situaciones anteriores, se obtendría cierta puntuación por el solapamiento de los dos ámbitos, pero no la máxima puntuación.

- Que el sistema identifique el ámbito geográfico completamente erróneo, por ejemplo la península ibérica. Pese a que la lógica pueda decir que no se le debe asignar ninguna puntuación, en ocasiones se le puede asignar puntuación acorde a la distancia existente entre el ámbito detectado y el real. Para este tipo de puntuaciones se suelen utilizar los mapas de rejillas, que permiten medir la distancia entre un ámbito y otro.

Debido a todos estos problemas resulta muy complicado el establecer una forma clara y objetiva de evaluar un sistema de detección del foco geográfico.

Entre los trabajos más relevantes en la materia, podemos ver como se suele recurrir a la evaluación manual, con granularidades a nivel de ciudad, estado (en el caso de EE.UU.) y país, tanto con evaluaciones binarias como graduales a la hora de decidir si se acertó o no en la detección del foco geográfico.

Para evaluar el trabajo realizado en Ding et al. (2000), se tenían que obtener un conjunto de páginas web que cumplieran los siguientes requisitos: que tuvieran un foco geográfico evidente e indiscutible, que dicho foco estuviese situado dentro de los EE.UU. y que fuera a nivel de país, estado o ciudad. Además, dichos recursos web debían tener un número suficientemente amplio de enlaces *HTML* que apuntasen a ellos para poder llevar a cabo los experimentos donde sólo trataban con los enlaces y no con la información textual, tal y como ya se comentó en la sección 3.3. Con estas premisas, los autores evaluaron el sistema seleccionando 150 páginas web obtenidas según el nivel geográfico jerárquico:

- **Nivel nacional**. Las 50 páginas más citadas de las web del gobierno federal, las cuales estaban listadas en la web *FedWorld*[14]. Para la obtención de estas 50 páginas se consultó al ya desaparecido motor de búsqueda de Internet *AltaVista* (comprado por *Yahoo!*).
- **Nivel de estado**. Las 50 páginas oficiales de cada uno de los estados de los EE.UU.
- **Nivel de ciudad**. Las 50 páginas oficiales más citadas de entre todas las ciudades de EE.UU. Se obtuvo dicha lista de páginas oficiales de la web *Piper Resources*[15].

Los resultado obtenidos mostraron que ambas aproximaciones obtuvieron unos buenos resultados, siendo mejor la utilización de cada aproximación dependiendo de una serie de características de los documentos webs analizados, aunque generalmente la aproximación basada meramente en el contenido textual obtuvo unos mejores resultados (82,5 % para el mejor de los casos para la aproximación basada en enlaces, y un 89,5 % para el mejor de los casos utilizando la aproximación basada en texto).

[14] http://fedworld.ntis.gov/
[15] http://www.statelocalgov.net/.

En el trabajo llevado a cabo en Amitay et al. (2004), para realizar las pruebas que comprobaban el buen funcionamiento del algoritmo de detección del foco geográfico se compararon los resultados obtenidos con la clasificación dada por los editores del *Open Directory Project*[16].

Según los autores de este trabajo, se consiguió un 80 % de aciertos en la etiquetación de los topónimos existentes en las páginas web analizadas, siendo la mayor fuente de errores la ambigüedades del tipo *geo/non-geo*.

Según los autores de este trabajo se consiguió una precisión del 91 % a nivel de país en la detección del foco/s geográfico/s.

Por último, en el trabajo desarrollado en Zong et al. (2005), para la evaluación del sistema de desambiguación, los autores obtuvieron aleatoriamente 50 páginas web que contenían más de 31 topónimos y menos de 200. La identificación inequívoca del topónimo se clasificó de manera binaria, correcta o incorrecta. Para aquellos clasificados como incorrectos se subclasificaron a su vez como incorrecto *geo/geo* o incorrecto *geo/non-geo*, siendo estos últimos errores provenientes únicamente de *GATE* ya que los autores no trataron este tipo de ambigüedad que, como ya se ha mencionado previamente, sí que trata intrínsecamente *GATE*.

Esta evaluación se realizó solamente con las entidades geográficas pertenecientes a EE.UU., obteniendo una precisión del 88,9 % según comprobaron manualmente, resultando la heurística del *emparejamiento perfecto con los nombres que aparecen en el gazetteer* la que más contribuyó a esta precisión con un 52,4 % de los aciertos.

Para la evaluación de esta tarea se emplearon dos métricas, según el nivel al que se estuviese comprobando la precisión:

1. A nivel de página. A este nivel los autores consideraron evaluar los resultados como correcto o incorrecto, obteniendo la precisión de dividir todos los resultados acertados entre todas las páginas evaluadas. La precisión obtenida fue del 66 %, fallando así pues en un 34 % de los casos, de los cuales el 24 % fue debido a errores en la desambiguación de entidades *geo/non-geo*.

2. A nivel de párrafo. A este nivel se hizo una evaluación con más grados:

   a) Completamente erróneo.
   b) Ámbito geográfico demasiado amplio.
   c) Ámbito geográfico demasiado pequeño.
   d) Ámbito geográfico correcto.

   En esta ocasión se dieron los resultados para este apartado obedeciendo a una precisión ajustada, en la que se contabilizaban todos los que

[16] *Open Directory Project* (*ODP*, actualmente conocido como *DMOZ*) es el directorio editado por humanos más extenso y más completo de la Web. Está construido y mantenido por una comunidad de editores voluntarios global. http://www.dmoz.org/.

obtuvieron el ámbito geográfico correcto dividiéndolo entre todos los párrafos que fueron evaluados, y obedeciendo a una precisión más relajada, en la que se contabilizaban todos los resultados que no tuvieran el ámbito geográfico completamente erróneo y se dividía entre todos los párrafos evaluados.

Para la precisión más ajustada se obtuvo una precisión del 30 %, mientras que para la más relajada se obtuvo una precisión del 42 %.

3.5. Detección del foco geográfico en textos informales

Hoy en día, no solamente nos encontramos con páginas web, artículos periodísticos, etc., con los que tratar para intentar identificar el foco geográfico de los mismos, sino que debido al auge de las redes sociales nos encontramos ante un abrumador volumen de textos, los cuales difieren por completos de los previamente mencionados debido a su brevedad e informalidad como norma general (ver figura 3.1).

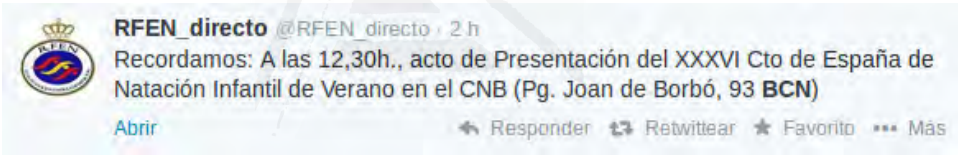


Figura 3.1: Ejemplo de texto emitido por un usuario de la red social *Twitter*.

El detectar la ubicación geográfica de este tipo de textos puede ser de gran utilidad para enviar información (Earle et al. (2012) y Sakaki et al. (2010)), recomendar servicios y comercios o enviar publicidad relevante a nivel local.

El problema subyace en que tan sólo alrededor del 0,42 % de los mensajes de texto que se envían a redes sociales como *Twitter* están geo-referenciados (Cheng et al., 2010), e incluso aquellos que están geo-referenciados pueden haber sido escritos en un lugar al que no se refieren en su contenido, por lo que se hacen necesarias otras técnicas que sean capaces de inferir la localización de los numerosos usuarios de dichas redes sociales.

La metodología para la detección del foco geográfico en textos informales claramente difiere de la de textos formales, principalmente debido a la dificultad existente a la hora de detectar los topónimos y realizar su desambiguación. Así pues, dicha metodología suele estar basada más en el uso de aproximaciones utilizando aprendizaje automático.

Dentro de los distintos tipos de algoritmos de aprendizaje automático, suelen destacar los algorimos de clasificación tales como *Naïve Bayes* y *Support Vector Machines*, y los modelos de lenguaje estadísticos, los cuales

asigna una probabilidad a una secuencia de m palabras $P(w_1, w_2, ..., w_m)$ mediante una distribución de probabilidad. De esta forma crean modelos de lenguaje a partir de un corpus de entrenamiento por cada una de las localizaciones o áreas que se quieren clasificar, para posteriormente comparar dicho modelo de lenguaje con el creado por el texto del que se quiere inducir su localización.

Si nos centramos en textos puramente informales, los trabajos científicos en esta área suelen basarse en corpus de fotografías de *Flickr* (ver sección 3.5.1) o en corpus de textos obtenidos de *Twitter* (ver sección 3.5.2).

3.5.1. Flickr

La inmensa mayoría de trabajos realizados sobre la detección de la ubicación geográfica de fotografías de *Flickr* desde el punto de vista de *PLN*, suele hacerse teniendo en cuenta únicamente las etiquetas que los usuarios asignan a las fotografías que suben a dicha plataforma.

Estas etiquetas se suelen agrupar por zonas geográficas tales como ciudades, puntos de interés, estados, paises, etc., para formar un conjunto de entrenamiento que sea capaz de clasificar dentro de dichas zonas. También es muy habitual la construcción de los denominados mapas de rejilla, que no son más que una representación del mundo (o del país o zona en la que se quieran ubicar las fotografías) en la que se divide por casillas cada una de las zonas de las que se quiere considerar como una ubicación, pudiendo estas casillas coincidir o no con una población (ver mapa de rejilla en la figura 3.2).

Sobre dichas etiquetas, como ya se ha comentado en esta misma sección, se suelen emplear técnicas de aprendizaje automático, las cuales podemos dividir entre algoritmos de clasificación (*Naïve Bayes* y *Support Vector Machine* entre los más utilizados) y modelos de lenguaje entre las técnicas más empleadas.

Si nos centramos en las aproximaciones basadas en modelos de lenguaje, hay que tener en cuenta que se requiere de la realización de un suavizado para solventar la problemática de los modelos que tienen un escaso vocabulario por falta de datos de entrenamiento.

En el trabajo llevado a cabo en Serdyukov et al. (2009), podemos ver una aproximación en la que se intenta localizar fotos de *Flickr*. Para ello, utilizaron como piedra angular de su investigación modelos de lenguaje creados a partir de las etiquetas que los propios usuarios ponían a las fotos que posteriormente se evaluarían.

El corpus de este sistema estaba compuesto por casi 400.000 fotos georreferenciadas de *Flickr* con su identificador, coordenadas y conjunto de etiquetas. De este conjunto de fotos se utilizó un filtro con el fin de eliminar todas aquellas fotos que fueron subidas en masa, es decir, aquellas fotos que un usuario subía a la vez y por ende eran más propensas a contener

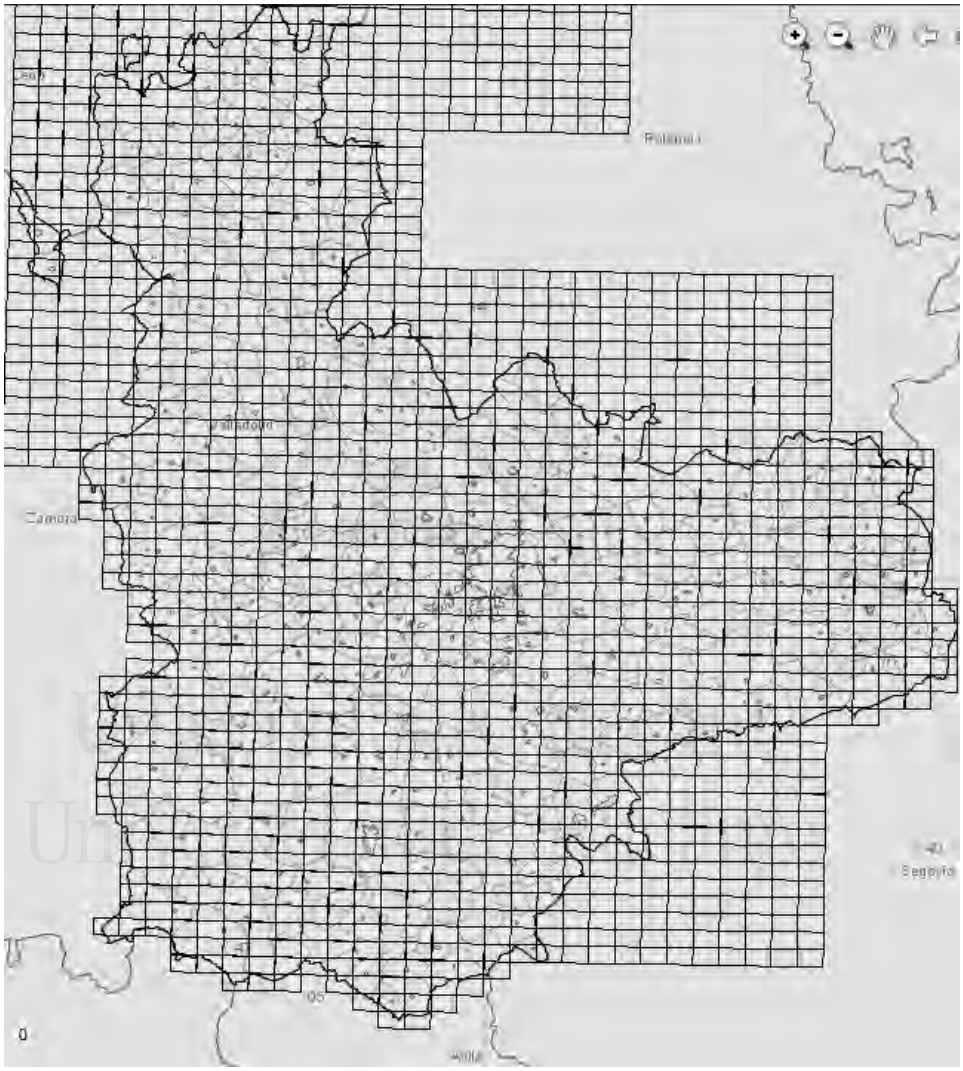


Figura 3.2: Ejemplo de un mapa de rejilla de la provincia española de Valladolid.

las mismas etiquetas pese a que fueran tomadas en lugares y situaciones diferentes. Tras este filtrado se trabajó con un corpus final de 140.000 fotografías pertenecientes a alrededor de 180 países distintos. Este conjunto final de fotografías se dividió en tres partes: una primera parte con 120.000 imágenes que fueron utilizadas para construir los modelos de lenguaje, una segunda con 10.000 fotografías utilizadas para afinar el sistema, y una última partición con otras 10.000 imágenes utilizadas para probar las distintas implementaciones realizadas.

La primera tarea de este trabajo fue asignar cada fotografía de la porción del corpus utilizada para la creación de los modelos de lenguaje a la casilla que le correspondía según sus coordenadas en un mapa de rejilla a nivel mundial construido para la ocasión, donde cada casilla fue definida como una localización. Se probó con distintos tamaños de celda (1, 5, 10, 50 y 100 kilómetros) para comprobar la precisión del sistema. Como línea de referencia (*baseline*), mediante un corpus de imágenes de *Flickr* georreferenciadas, se crearon modelos de lenguaje posicionando cada una de estas imágenes en su correspondiente casilla y obteniendo las etiquetas de todas las imágenes para cada una de las casillas. Las etiquetas fueron tratadas de forma individual (*bag-of-tags*), es decir, como si no tuvieran relación entre ellas, sin aplicar *stemming*[17] ni filtro de *stop-words*[18], aunque sí la normalización estándar de las etiquetas dado por el propio *Flickr*, la cual hacía que todas los términos de etiquetas compuestas fueran concatenados y todos los caracteres especiales eliminados.

Cada localización fue representada en un grafo no dirigido, donde el vínculo entre cualquier par de localizaciones sólo existe si está situado lo suficientemente cerca en la rejilla. Para calcular la distancia entre dos localizaciones se utilizó la distancia medida en número de casillas de distancia entre los dos lugares en cuestión. Al representar las localizaciones como *pseudo-documentos*, es decir, que había un conjunto de palabras (las etiquetas) por cada una de las localizaciones, se consiguió no sólo una semejanza espacial, sino también semántica.

Este *baseline* fue ampliado de las siguientes maneras:

- **Suavizado basado en las etiquetas**. La motivación para el suavizado basado en las etiquetas proviene de la necesidad de superar la escasez de datos y de la comprensión de que algunas etiquetas indican un área que supera los límites de la ubicación específica. La primera manera de utilizar el *vecindario espacial* es considerar que cada etiqueta encontrada dentro de una ubicación específica es

[17] *Stemming* es un método para reducir una palabra a su raíz (*stem* en inglés) o lema.

[18] Palabras vacías (*stop-words* en inglés) es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural (texto).

generada por medio de un modelo de lenguaje de la ubicación dada, o por modelos de lenguaje de localidades vecinas.

- **Suavizado basado en las casillas**. Es razonable suponer que los "buenos lugares" vienen de los "buenos vecinos". Esto significa que algún tipo de relevancia debe ser propagada a través de las casillas vecinas. Los autores aplicaron un enfoque simple ponderado gradualmente, es decir, que la probabilidad de generar el conjunto de etiquetas de un lugar determinado se ve aumentada por las probabilidades de localidades vecinas.

- **Suavizado basado en las casillas con propagación de la valoración**. Este enfoque es idéntico al anterior aunque dicha propagación tan sólo sucede entre todas las localidades vecinas que tienen una valoración menor a la que se está examinando. La razón que subyace en esta implementación es que tan sólo debe potenciar a aquellas localizaciones que tienen suficiente fuerza de por sí, sin que se aprovechen de que tienen un vecino con mucho peso.

- **Ponderación basada en topónimos**. Para esta implementación se introdujo un conocimiento preliminar sobre las etiquetas del *baseline* para potenciar o estimular algunas de ellas. Esta estimulación no consistía en otra cosa que en dar un mayor peso a aquellas etiquetas que aparecían dentro de la lista de lugares poblados de un *gazetteer*. En este caso, una vez más se utilizó *GeoNames*.

- **Suavizado basado en la ambigüedad de las etiquetas**. En esta ocasión, lo que se hizo fue dotar de menor peso a aquellas etiquetas que eran ambiguas, es decir, las etiquetas que aparecían en más de un lugar.

Empíricamente, los resultados fueron mejores cuando se utilizó como distancia una única celda. Como era de esperar, cuanto mayor fuera el tamaño de la celda (localización), mejor eran los resultados obtenidos. Los mejores resultados fueron dados con la utilización del suavizado basado en las casillas con propagación de la valoración, aunque mejoró ligeramente a la precisión del *baseline* (0,288 vs. 0,296) para un tamaño de celda de 100 km. Con todo, la mejor implementación vino dada cuando se juntaron con el *baseline* las tres últimas implementaciones expuestas anteriormente (0,317 de precisión). Las dos principales fuentes de error reportadas por los autores fueron las causadas por la escasez y el ruido de los modelos de localización, y los errores derivados de la ambigüedad y el carácter incompleto de los conjuntos de etiquetas utilizadas para su correlación.

Por otro lado, si nos centramos en trabajos que utilizan algoritmos de clasificación automática, cabe destacar la investigación llevada a cabo en Crandall et al. (2009), donde podemos ver otra aproximación en la que se

intenta clasificar geográficamente las fotos de *Flickr*. En esta ocasión, los autores realizaron esta tarea desde dos enfoques distintos, textual y visual, ayudándose también de la información temporal de las fotografías.

Si nos centramos en la parte meramente textual, por estar más relacionada con el trabajo llevado a cabo en esta tesis, al igual que en el trabajo anterior, los autores de esta investigación se basaron en las etiquetas que los usuarios de *Flickr* le habían asignado a sus fotos.

La resolución de este sistema se estableció en dos diferentes granularidades, a nivel de ciudad (alrededor de unos 100 km a la redonda) y a nivel de los puntos de interés (*landmarks*) (alrededor de 100 metros a la redonda).

El corpus de esta investigación estaba compuesto por casi 45 millones de fotografías pertenecientes a más de 300.000 usuarios distintos que llevaban asociadas los siguientes metadatos: etiquetas textuales, fecha y hora en la que fueron tomadas y su geolocalización. Algunas de estas fotografías recopiladas no tenían etiquetas textuales.

Para la clasificación de dichas imágenes entrenaron con las etiquetas de las propias fotografía tanto clasificadores Bayesianos como de *máquinas de vectores de soporte* (*SVM - Support Vector Machine*), dando éstos últimos clasificadores unos resultados ligeramente superiores. Se realizó una selección de las características basado en el filtrado por frecuencia de aparición. Concretamente, las etiquetas que aparecían menos de 3 veces fueron ignoradas por no ser de gran utilidad.

En sus experimentos intentaron hacer lo siguiente:

- Clasificar puntos de interés (*landmarks*) a nivel de ciudad, donde intentaron predecir el *landmark* dentro de una ciudad dada:
  - Entre 10 posibles candidatos. Obteniendo una precisión media entre las 10 ciudades más fotografiadas del mundo del 51,67 %.
  - Entre 25 posibles candidatos. Obteniendo una precisión media entre las 10 ciudades más fotografiadas del mundo del 44,64 %.
  - Entre 50 posibles candidatos. Obteniendo una precisión media entre las 10 ciudades más fotografiadas del mundo del 38,16 %.
- Clasificar *landmarks* a nivel mundial, donde intentaron predecir el *landmark* a nivel mundial entre los 10 *landmarks* más fotografiados del mundo, obteniendo una precisión media del 69,39 %.
- Clasificar a nivel de ciudad. Donde intentaron predecir en qué ciudad a nivel mundial fueron tomadas las fotografías, obteniendo una precisión media del 56,83 %.

3.5.2. Twitter

Si hay un claro exponente de textos informales, ese es *Twitter*. Una vez más, vemos como la inmensa mayoría de la comunidad científica, al

igual que sucediera con *Flickr*, ha decidido afrontar la tarea de la detección geográfica de los usuarios de *Twitter* basándose en algoritmos de aprendizaje automático, siendo los más recurridos los algoritmos de clasificación *Naïve Bayes* y *Support Vector Machine*, así como los modelos de lenguaje.

Estos algoritmos han servido para detectar la localización de los usuarios de *Twitter* a distintos niveles, siendo los más comunes los niveles de ciudad o estado (dentro de EE.UU.), aunque también se pueden encontrar algunas aproximaciones que van desde niveles más concretos como son los códigos postales, e incluso más genéricos como son los países (Cheng et al., 2010) (Kinsella et al., 2011) (Cranshaw et al., 2012) (Sadilek et al., 2012).

Otro factor común que se puede extraer de las aproximaciones más exitosas en este campo es la limpieza que se suele realizar en los corpus de tuits. Así pues, vemos como en la mayoría de casos se suelen eliminar los signos de puntuación, *URLs*, *stop-words* e incluso en ocasiones los nombres de usuarios de la propia red social, aunque también podemos encontrar algunas aproximaciones (especialmente las basadas en modelos de lenguaje) que prefieren mantener todos estos términos con el fin de captar las cuasi imperceptibles sutilezas existente en el lenguaje de los usuarios, según su región de procedencia. Debido a la dificultad por la informalidad del lenguaje, no se suele aplicar *stemming*.

Así pues, entre los trabajos más destacados y pioneros que podemos encontrar en este campo, que manejan la información meramente textual de los mensajes de *Twitter* para identificar el foco geográfico de los tuits de los usuarios, es decir, la ubicación de los usuarios de *Twitter* a partir de sus tuits emitidos, se encuentra el trabajo llevado a cabo en Cheng et al. (2010). En este trabajo sus autores propusieron y evaluaron un marco probabilístico para estimar la localización de los usuarios de *Twitter* a nivel de ciudad basándose meramente en el contenido de sus tuits. Esta investigación fue motivada por el hecho de que tan sólo el 0,42 % de los tuits emitidos tienen asociadas unas coordenadas geográficas, como los propios autores muestran en esta investigación.

Este trabajo se basa en tres características clave:

1. Sus datos de entrada basados puramente en el texto de los tuits, sin utilizar ningún tipo de datos externos provenientes de los usuarios o conocimiento obtenido de la web.

2. Un clasificador que identifica las palabras de los tuits que contienen un ámbito geográfico local.

3. Un modelo basado en un mapa de rejilla con suavizado entre casillas vecinas el cual es utilizado para refinar los resultados estimados.

Para la obtención del corpus que utilizaron en sus experimentos, accedieron al perfil de más de un millón de usuarios para obtener todos

sus tuits, previo filtro de su ubicación mediante los datos del perfil de los mismos. Concretamente, accedieron al campo de autoubicación existente en el perfil de los usuarios y desecharon a todos aquellos que no especificaran correcta y concisamente su ubicación. Para ello, únicamente conservaron los tuits de los usuarios que indicaban en su perfil de localización de usuario una ubicación dentro de los EE.UU. de alguna de las siguientes formas: "*nombre de ciudad*", "*nombre de ciudad, nombre de estado*" o "*nombre de ciudad, abreviatura de estado*", contrastando estas ubicaciones con las existentes en el *Census 2000 U.S. Gazetteer*, dando como válidas estas autoubicaciones. En caso de que alguna localidad fuera ambigua a nivel de país, tan sólo se mantuvieron si podían desambiguarla a través del nombre del estado o su abreviatura. Después de este filtrado el corpus final estaba compuesto por más de 130.000 usuarios con más de 4 millones de tuits.

Por otro lado, como conjunto de evaluación utilizaron los tuits de usuarios muy activos, aquellos de los que tenían mil o más tuits, y que habían facilitado su ubicación en la forma de coordenadas, evitando así que se solaparan estos usuarios con los que se utilizaron en el conjunto de entrenamiento. Así pues, el conjunto de evaluación estaba compuesto por más de 5000 usuarios y más de 5 millones de tuits distribuidos a lo largo de todo el territorio continental de los EE.UU.

Como *baseline* realizaron una aproximación donde calculaban la probabilidad de que el conjunto de palabras existente en el conjunto de tuits del usuario a evaluar perteneciera a una determinada localidad, asignándole así a dicho usuario la localidad que reportaba mayor probabilidad. Se eliminaron del conjunto de términos existente en cada uno de los usuarios a evaluar todas las *stop-words*, nombres de usuarios de *Twitter* (aquellos que empiezan por el símbolo '*@*'), enlaces y signos de puntuación. En lugar de utilizar *steamming* usaron *Jaccard Coefficient*[19] para verificar si una palabra que se encuentra por primera vez es una variación de alguna que haya sido encontrada previamente.

Partiendo del *baseline* anterior, los autores intentaron mejorarlo identificando las palabras que estaban más "localizadas" en algunas regiones. Para ello, construyeron un mapa de rejilla para observar la distribución geográfica de los términos que aparecían en los tuits obteniendo las que sobrepasaban cierto umbral. Estos términos geográficamente localizados fueron utilizados junto a los manualmente etiquetados existentes en *SCOWL* (*Spell Checker Oriented Word Lists*)[20] para construir un clasificador geográfico utilizando para ello algoritmos de clasificación como *Naïve Bayes*, *SVM* o *AdaBoost*, disponibles en *Weka toolkit* (Witten and Frank, 2005).

[19] El índice de *Jaccard*, también conocido como el coeficiente de similitud de *Jaccard*, es una estadística que se usa para comparar la similitud y la diversidad de conjuntos de muestra.

[20] http://wordlist.aspell.net/

Para tratar con los lugares donde había términos del conjunto de prueba que no aparecían se probó con distintos enfoques para realizar un suavizado, siendo el más exitoso el que realizaba un suavizado sobre los vecinos de celda basado en un mapa de rejilla.

Así pues, el *baseline* inicial se modificó incluyendo solamente los términos localizados y utilizando el proceso de suavizado descrito en el párrafo anterior. Mediante el proceso de eliminación de las palabras no localizadas se consiguió aumentar la precisión del sistema de un 10,12 % a un 49,8 % debido a la eliminación del ruido que proporcionaban el resto de términos del corpus de entrenamiento. El suavizado proporciono una leve mejora hasta llegar al 51 % de precisión para usuarios con 1000 o más tuits, o más de un 40 % utilizando sólo 100 tuits.

En Mahmud et al. (2012) se basaron en el trabajo anterior para construir un algoritmo capaz de inferir la localización de los usuarios de *Twitter* a diferentes niveles de granularidad. En esta ocasión, sus autores no sólo realizaron experimentos con el contenido textual de los tuits de los usuarios a evaluar, sino que también se basaron en el comportamiento (volumen de tuits por unidad de tiempo) de los mismos y utilizaron *gazetteers* con el nombre de las ciudades y estados de EE.UU.

El conjunto de datos utilizado en este trabajo estaba compuesto por tuits pertenecientes a las 100 ciudades más pobladas de los EE.UU., de las que obtuvieron hasta los 200 últimos tuits de 100 usuarios distintos por cada una de las ciudades. El conjunto final estaba compuesto por más de millón y medio de tuits pertenecientes a casi 10.000 usuarios.

Para los experimentos llevados a cabo en este trabajo se realizó una validación cruzada de 10 iteraciones (*10-fold cross-validation*). Cada usuario del corpus de entrenamiento correspondía a una muestra de entrenamiento. Los signos de puntuación del corpus, *URLs* y otros términos que contenía caracteres especiales fueron eliminados (excepto los *hashtags*). La salida final era un modelo entrenado con un número de clases igual al número total de localizaciones del conjunto de entrenamiento.

Para este sistema se construyen distintos tipos de clasificadores:

- **Clasificador estadístico basado en el contenido**. El contenido de los tuits fue utilizado para realizar la clasificación de los mismos, concretamente se utilizaron tres tipos distintos de contenidos:

  1. Palabras. Todas las palabras que había en cada tuit identificadas como nombre mediante un etiquetador gramatical[21] (*Apache Open NLP*[22]) y que no eran *stop-words*.
  2. *Hashtags*. Todas las etiquetas que había en cada tuit, es decir, todos los términos que comenzaban por el símbolo '#'.

[21] *PoS tagger* (*Part of Speech*) es un etiquetador de categorías gramaticales.
[22] http://opennlp.apache.org/

3. Topónimos. Todos los nombres de ciudad y de estado que había en cada tuit, identificados mediante emparejamiento con el *USGS gazetteer*. Para ello se tuvieron que crear no solamente unigramas, sino también bigramas y trigramas para obtener los nombres de ciudad o de estado que estaban compuestos por más de un término.

Se utilizaron varias heurísticas para la obtención de términos geográficamente localizados, tal y como se hizo en Cheng et al. (2010). Primero se computó la frecuencia de los términos seleccionados para cada una de las localizaciones, así como el número de personas en esa localización que habían utilizado esos términos en sus tuits, manteniendo así los términos que aparecían en los tuits de al menos K personas ($K = 5$ empíricamente). Posteriormente se calculaba la probabilidad condicionada media y máxima de cada localización para cada término, y si la diferencia de éstas sobrepasaban cierto umbral, se comprobaba si la máxima probabilidad condicionada sobrepasaba otro umbral para que los términos fueran finalmente seleccionados, ya que dichos términos estaban concentrados en ciertos lugares. Una vez que las características (los términos procedentes de la selección previamente mencionada) habían sido seleccionadas se procedía a construir modelos estadísticos utilizando para ello algoritmos de aprendizaje automático, concretamente *Naïve Bayes Multinomial*.

- **Clasificador heurístico basado en el contenido**.

1. Nivel local. La idea que subyace tras esto es que un usuario debe mencionar la ciudad o estado donde vive más a menudo que el resto de ciudades y estados. Así pues, por cada ciudad y estado del conjunto de entrenamiento se calcula la frecuencia con la que cada usuario los menciona en sus tuits y se usa como marcador de cada usuario con cada ciudad y estado mencionado. Para cada usurio, la ciudad o estado que haya obtenido un mayor marcador será seleccionada como su localización.

2. Historial de visitas. La idea es que un usuario suele visitar los lugares cercanos con más frecuencia que los más distantes. Para obtener el historial de visitas de cada usuario los autores buscaron las *URLs* generadas por *Foursquare*[23] con su servicio de registro (*check-in*) en los tuits, obteniendo así la información (ciudad y estado) del lugar visitado. La localización con mayor frecuencia es establecida como la del usuario.

[23]Foursquare es un servicio basado en localización web aplicada a las redes sociales. http://foursquare.com.

- **Clasificador basado en la actividad por zonas horarias**. Es un clasificador que intenta detectar la localización de los usuarios según en qué franja horaria envíen éstos sus tuits. Cada franja temporal representa una característica para el clasificador basado en la implementación de *Naïve Bayes* de *Weka*, siendo la duración de estas franjas de un minuto.

- **Conjunto de clasificadores agrupados**. En esta aproximación los autores utilizaron los clasificadores anteriormente mencionados asignándoles pesos a cada uno de ellos según el método expuesto en Jiménez (1998) para crear un ensamblado de clasificadores estadísticos y heurísticos.

- **Conjunto de clasificadores jerárquicos**. En esta ocasión los autores crearon un clasificador en dos pasos o dos jerarquías:

    1. Franja horaria. Se creó un ensamblado tal y como se ha explicado en el punto anterior utilizando los clasificadores estadísticos basados en contenido y el clasificador basado en la actividad por zonas horarias, pero esta vez se hizo entrenando previamente entre las ciudades que pertenecían a una determinada franja horaria de los EE.UU., por lo que sólo tenía que predecir entre esas ciudades.
    2. Estado. Igual que el anterior pero esta vez, en vez de agrupar por franjas horarias se hizo por estados de EE.UU. cuando existía más de una ciudad en dichos estados. El ensamblado, en esta ocasión tan sólo comprendía los clasificadores estadísticos basados en contenido.

Entre los clasificadores estadísticos, el que mejores resultados obtuvo fue el que estaba basado en topónimos con un 54 % de precisión, mientras que entre los clasificadores heurísticos a *nivel local* consiguió un 50 % de precisión. Aunque el mejor clasificador de esta investigación se encuentra entre los clasificadores jerárquicos, concretamente el de *franja horaria*, el cual obtuvo un 64 % de precisión.

Por último, otro trabajo interesante es el que podemos encontrar en Kinsella et al. (2011) donde sus autores una vez más crearon modelos de lenguaje de las localizaciones extrayendo las coordenadas de un corpus de tuits geo-referenciados, modelando las localizaciones a diferentes niveles de granularidad que iban desde el nivel de país al nivel de código postal. Para la construcción de los modelos de lenguaje utilizaron la aproximación descrita en Ponte and Croft (1998). En esta ocasión, sus autores no sólo intentaron predecir la localización del usuario, sino que también intentaron predecir la procedencia de los tuits individualmente.

Para obtener el ranking de las localidades más probables se podía hacer de dos maneras:

1. Clasificando mediante la probabilidad de que un modelo hubiera generado un tuit o conjunto de tuits (*QL*).

2. Clasificando comparando la divergencia que provocaba el modelo creado por cada localidad y el del propio tuit o conjunto de tuits, siendo elegida la localidad con menor divergencia. Para medir la divergencia entre un modelo y otro se utilizó la divergencia de *Kullback-Leibler* (*KL*) (Kullback, 1997).

El corpus utilizado fueron tuits que tenían asociadas las coordenadas facilitadas por los propios usuarios. Para obtener la verdadera localización en sus distintas granularidades a partir de éstas coordenadas, los autores utilizaron la herramienta *Yahoo! Geoplanet*[24], la cual no sólo obtenía el lugar especificado por las coordenadas, sino que también las entidades geográficas que estaban por encima de éste, es decir, su dirección postal, barrio, ciudad, estado, país y continente. No se realizó *stemming* ni se eliminaron las *stop-words*. Los nombres de usuario y los *hashtags* se incluyeron en los modelos de lenguaje. Cualquier tuit repetido o retuiteado fue eliminado, así como todos los enlaces web. Se obtuvieron dos conjuntos de datos distintos, uno para experimentos a una escala pequeña y el otro a escala global.

- Escala pequeña. Para obtener los tuits de este corpus se consultó el (*stream*) público de *Twitter* conocido como *Spritzer*[25], que por entonces permitía obtener un 5 % del total de tuits emitidos. Se filtraron los tuits para recuperar únicamente aquellos que pertenecieran a alguna de las 10 ciudades que más tuits emitía alrededor del mundo, que cuando se realizó este experimento (del 25 de mayo al 21 de junio de 2010) eran por orden decreciente: Jakarta (Indonesia), Nueva York (EE.UU.), Londres (Reino Unido), Chicago (EE.UU.), San Francisco (EE.UU.), Houston (EE.UU.), Toronto (Canadá), Amsterdam (Países bajos), Sydney (Australia) y Santiago de Chile (Chile). Por cada tuit se obtuvo la ciudad desde dónde se emitió el tuit y el barrio mediante el *reversing geo-coding* realizado con *Geoplanet*.

- Escala global. Esta vez los tuits se obtuvieron a través del *Twitter Firehose stream*, el cual permite obtener todos los tuits públicos de *Twitter*, con lo que se alcanzaron unos 7 millones de tuits, para los cuales se obtuvieron mediante *Geoplanet* el país, estado, ciudad y código postal.

[24] https://developer.yahoo.com/geo/geoplanet/
[25] https://dev.twitter.com/docs/streaming-apis/streams/public

Como *baseline* de esta investigación se utilizó *Yahoo! Placemaker* (ver sección 2.2.2). Para cada uno de estos conjuntos de datos se realizaron los siguientes experimentos:

- *Spritzer*. Para estos experimentos se particionó el corpus en 5 partes iguales para realizar validación cruzada. El corpus fue particionado a nivel de usuario con el fin de evitar textos altamente parecidos creados por el mismo usuario en distintas particiones. En los experimentos lanzados con este corpus cabe destacar los distintos idiomas, jergas y culturas que existen entre cada una de las 10 ciudades evaluadas. Los experimentos lanzados buscaron:

  1. *Detectar la procedencia de un tuit individual a nivel de ciudad entre las 10 ciudades del corpus*. Los mejores resultados fueron obtenidos al utilizar la aproximación *KL*, mediante la cual consiguieron un 65,7 % de precisión para las 10 posibles ciudades.
  2. *Predecir el barrio de procedencia de un tuit dentro de la ciudad de Nueva York*. Había tuits que pertenecían a alguno de los 502 barrios de Nueva York con los que se trabajó, para los cuales se obtuvo un precisión del 20,9 % utilizando, una vez más, las aproximación basada en *KL*.

- *Firehose*. Para estos experimentos se destinaron el 80 % de los tuits para construir los modelos de lenguaje, el 10 % como conjunto de validación, y el 10 % restante para evaluarlo. Los experimentos lanzados buscaron:

  1. *Detectar la procedencia de un tuit individual a nivel de país, estado, ciudad y código postal*. En esta ocasión, los mejores resultados para todos los niveles fueron obtenidos mediante la aproximación *QL*, dando una precisión a nivel de código postal, localidad y de estado del 13,9 %, 29,8 % y el 31,6 % respectivamente.
  2. *Predecir la localización de un usuario a nivel de país, estado, ciudad y código postal, a partir del conjunto de tuits que éste ha emitido*. Se tomó como norma el establecer como ubicación del usuario el lugar desde el que emitió tuits con más frecuencia. Una vez más la aproximación ganadora fue la *QL*, dando una precisión a nivel de código postal, localidad y de estado del 14,9 %, 31,9 % y el 44,9 % respectivamente.

Uno de los aspectos a destacar de esta investigación es que a pesar de que herramientas como *Yahoo! Placemaker* consiguieran detectar entidades geográficas en un 2,5 % de los tuits, los modelos de lenguaje pudieron ubicar correctamente a nivel de código postal un 13,9 % de ellos, lo que indica que

existen beneficios sustanciales en la utilización de otras características de estos textos a parte de las menciones geográficas explícitas.

3.6. Conclusiones

Como hemos podido ver en este capítulo, la detección del foco o ámbito geográfico en los textos es una de las tareas claves en los sistemas *GIR*, ya no sólo es importante para la búsqueda de documentos relacionados con ciertos requisitos demandados por los usuarios, sino que también resulta fundamental a la hora de visualizar dichos documentos, tal y como se vio en la sección 2.1.6. En dicha sección se explicaba cómo la búsqueda geográfica permite una interfaz de usuario única, el mapa interactivo, que puede ser utilizado no sólo para reducir el foco geográfico de la búsqueda del usuario, sino también para poner en relieve los eventos interesantes o documentos que puedan hallarse en la zona delimitada por el mapa, resaltando éstos de manera gráfica a través de la propia interfaz (Gey et al., 2011).

Las restricciones geográficas impuestas por una consulta lanzada a un sistema *GIR* deben ser satisfechas teniendo en cuenta la intersección geográfica de los documentos existentes en el corpus. Dicho de otro modo, se tiene que comprobar si el ámbito geográfico de la consulta y del documento analizado se solapan de alguna manera. Por ello se hace necesario el tener que explorar las entidades geográficas superiores de la consulta o de cada documento analizado, en busca de una intersección que le otorgue un mayor valor al documento analizado en función del grado de solapamiento geográfico existente entre ambos textos.

La detección de topónimos así como su correcta desambiguación son tareas previas claves cuando se intenta identificar con éxito el ámbito geográfico de los textos. Así pues, dichas tareas son factor común entre las aproximaciones más exitosas que han intentado detectar el foco geográfico de los textos.

Otro de los aspectos a tener en cuenta cuando se pretende detectar el ámbito geográfico de un texto es la granularidad con la que se quiere obtener éste, es decir, a nivel de país, estado o comunidad autónoma, condado o provincia, localidad, código postal, etc.

Así pues, esta granularidad ha de ser tenida en cuanta también cuando se quiere evaluar un sistema que intenta detectar el foco geográfico. Para dicho fin, se suelen hacer evaluaciones estadísticas, es decir, se examinan manualmente una serie de documentos seleccionados de manera aleatoria dentro de un corpus y se indica la precisión obtenida para éstos.

Las evaluaciones que se suelen hacer para este propósito pueden ser del tipo binario, correcto/incorrecto, o se pueden añadir un mayor número de grados en las mismas, tal y como vimos en Zong et al. (2005): completamente

erróneo, ámbito geográfico demasiado amplio, ámbito geográfico demasiado pequeño y ámbito geográfico correcto.

Respecto a los textos informales, como se ha mencionado en la sección 3.5, difiere de la de textos formales principalmente debido a la dificultad existente a la hora de detectar los topónimos y realizar su desambiguación.

Las investigaciones conducidas en esta materia se han centrado especialmente en dos redes sociales: *Flickr* y *Twitter*, teniendo en común las implementaciones más exitosas el uso de técnicas de aprendizaje automático tales como *Naïve Bayes*, *SVM*, *AdaBoost* o modelos de lenguaje.

Debido a la informalidad de dicho lenguaje, suele ser complicada la utilización de *stemmers* e incluso la eliminación de las *stop-words*, aunque sí que se suelen eliminar los caracteres especiales y de puntuación que suelen abundar en dichas redes sociales así como los numerosos enlaces que aparecen. Por otro lado, los nombres de usuario (aquellos términos que comienzan por el carácter '@') y etiquetas (aquellos términos que empiezan por el carácter '#') de *Twitter* suelen preservarse debido a que pueden ser de gran utilidad para la identificación geográfica de los usuarios.

Dentro de las aproximaciones más exitosas suele practicarse la selección de características con el objetivo de eliminar ruido, es decir, un decremento significativo de los términos que no aportan mucha información utilizados para la construcción de los clasificadores y/o modelos del lenguaje. Normalmente se suele trabajar con los topónimos e incluso cualquier otro tipo de sustantivo encontrado en los textos.

También suele ser bastante común la creación de mapas de rejilla que permiten crear modelos de lenguaje para cada una de las casillas del mapa que representan un lugar, haciéndose necesario algún tipo de suavizado que permita abordar la problemática de la escasez de algunos términos en determinadas casillas que no tienen un gran número de elementos. Los suavizados que mejor suelen funcionar son aquellos que se realizan a través de las casillas vecinas, es decir, los que comparten de algún modo términos con las ubicaciones colindantes.

La hipótesis seguida en esta tesis es que la información general del mundo asociada a las localizaciones geográficas puede mejorar la desambiguación de topónimos y la localización del foco geográfico en los documentos. La presencia en el texto de determinados eventos, nombres de personas, de organizaciones, fechas o incluso términos comunes, puede ser de gran utilidad para detectar de qué localidad concreta nos habla el texto (desambiguación de topónimos) y determinar su importancia con respecto al contenido del documento (detección del foco). Más aún, este tipo de información general podría servirnos para detectar el foco geográfico sin necesidad de que el nombre de la localización aparezca en el texto de forma explícita, infiriéndolo a partir de la aparición de determinados personajes, eventos, etc., relacionados con dicha localización.

Hasta donde alcanza nuestro conocimiento, el único sistema que ha empleado este tipo de información para la tarea de desambiguación de topónimos es el desarrollado en Roberts et al. (2010). En este trabajo incorporaban información de eventos, relacionando nombres de personas, organizaciones y otras localizaciones. En nuestro caso pretendemos ir más allá, incorporando también información relacionada con fechas y términos comunes que puedan ser representativos de un lugar (como pueden ser los nombres de determinadas comidas, expresiones artísticas, etc.).

El foco geográfico de un documento identifica las localidades relevantes que han sido mencionadas en el texto. En esta tesis se ha desarrollado una aproximación basada en aprendizaje automático trabajando con dos tipos distintos de clasificadores: *SVM* y modelos de lenguaje. A diferencia de otras aproximaciones que se centran únicamente la utilización de la información geográfica para llevar a cabo esta tarea, nuestra propuesta emplea toda la información textual incluida en el corpus bajo la asunción de que la presencia de ciertos términos, como los nombres de personas, eventos e incluso términos y expresiones comunes, pueden ser de gran ayuda para dilucidar el foco geográfico de un texto.

A pesar de que en la bibliografía sobre este tema podemos encontrarnos mucho ejemplos donde previamente a la identificación del foco geográfico se llevan a cabo una serie de pasos, la aproximación expuesta en esta tesis se centrará en los pequeños matices lingüísticos existentes en cada región para llevar a cabo dicha tarea, es decir, no se usará ninguna herramienta externa, sino solamente texto (aunque dicho texto puede proceder de diversas fuentes) para construir el sistema aquí expuesto que determinará el ámbito geográfico de los textos. Se ha decidido adoptar esta medida debido a que no sólo se va a tratar con textos formales como pueden ser los procedentes de artículos periodísticos, sino que también se hará con textos informales como lo son los mensajes emitidos por los usuarios de *Twitter*, en los cuales resulta muy complicado el llevar a cabo el tratado de entidades gramaticales con éxito, lo que haría arrastrar una serie de errores ajenos a la propia detección del foco geográfico.

Para llevar a cabo esta tarea se han utilizado técnicas de aprendizaje automático como *SVM* y modelos de lenguaje.

4

Corpus de trabajo

En este capítulo se van a detallar los recursos utilizados para la realización de los experimentos expuestos en el capítulo 5.

Uno de los mayores problemas a los que se enfrentan los sistemas *GIR*, o herramientas como las descritas en la sección 2.2, es el de los distintos tipos de textos que deben tratar. Estos textos, si se realiza una clasificación por su formalidad se podrían dividir en dos grupos:

1. **Formales.** En la actualidad se puede encontrar un amplio abanico de fuentes de textos formales, aunque, en cuanto a lo que esta tesis atañe, se centrará en dos de estas fuentes: artículos encontrados en periódicos, en concreto el diario *20Minutos*[1], y artículos de *Wikipedia*[2].

2. **Informales.** Debido al auge de las redes sociales, el número y volumen de corpus que están relacionados con éstas ha crecido exponencialmente en los últimos años. Debido a su lenguaje altamente informal es necesario un trato distinto a este tipo de textos. Concretamente, en esta tesis se trabajará con un corpus de *Twitter*[3] y otro de *Flickr*[4] como claros representantes de textos informales.

Con estos cuatro corpus se realizarán dos tareas de detección del foco geográfico en el texto, una primera en la que se intentará ubicar geográficamente las noticias existentes en el corpus del periódico *20 Minutos*, y otra en la que se intentará determinar la localidad donde los usuarios de *Twitter* han emitido un conjunto de tuits. Para estas dos tareas se hará uso de los corpus expuestos anteriormente con el fin de afrontar el principal cometido de esta tesis, conocer cómo afecta el uso de los distintos tipos de información y recursos a la hora de detectar el foco geográfico en textos formales e informales.

En las siguientes secciones se podrá ver cómo se obtuvieron y prepararon cada uno de los corpus con los que se ha trabajado en los experimentos.

[1] http://www.20minutos.es/
[2] http://www.wikipedia.org/
[3] https://twitter.com/
[4] https://www.flickr.com/

Figura 4.1: Portada del diario *20Minutos* donde se puede seleccionar el área geográfica de la que se quiere mostrar noticias relacionadas.

4.1. 20Minutos

20Minutos es un periódico español de información general y distribución gratuita que se publica de lunes a viernes. Perteneciente al *Grupo 20 Minutos*, con sede en Madrid y ámbito nacional.

Se ha seleccionado este medio y no otro debido a que tiene una sección donde las noticias están clasificadas geográficamente según a la ciudad a la que pertenecen (ver imagen 4.1). Concretamente, en los años que se recogieron datos para esta investigación, 2008-2011, las ciudades que se utilizaron fueron las capitales de provincia de cada una de las 50 provincias del estado español más las 2 ciudades autónomas del país. Así pues, se procedió a recoger todos los artículos publicados en cada una de dichas ciudades durante los 4 años indicados anteriormente, obteniendo finalmente un total de 519.563 artículos etiquetados geográficamente.

En la figura 4.2 se muestra un ejemplo de noticia publicada en la ciudad de Elche.

Así pues, el trabajo realizado para poder finalmente trabajar con este corpus es el detallado en los siguientes apartados.

Figura 4.2: Ejemplo de una noticia del periódico *20Minutos* clasificada geográficamente a nivel de ciudad.

Creación del corpus

Para la obtención del corpus del diario *20Minutos* utilizado en nuestros experimentos se creó una secuencia de comandos que recorrió el apartado de las noticias a nivel local del periódico (*crawler*), obteniendo y almacenando los artículos de cada una de las ciudades existentes, para cada uno de los días, en un servidor.

El formato original que tenían los ficheros de texto descargados que contenían dichas noticias fue *HTML*, el cual comprendía los siguientes campos:

- *<h1 class="article-title">*. Título del artículo.
- * *. Resumen de la noticia.
- * *. Texto completo de la noticia dividido en párrafos.
- *<p>*. Cada uno de los párrafos de la noticia que estaban incluidos dentro del campo * *. Este es el único campo del que podía haber más de uno por artículo, ya que había uno por cada párrafo que componía el texto de la noticia.

Se obtuvieron noticias desde el nacimiento del periódico en Internet, es decir, desde el 16 de enero de 2005, hasta el 31 de diciembre de 2011.

Durante los 2 primeros años de publicación del periódico en Internet, y parte del tercero, el número de ciudades existentes en formato digital no era muy extenso (Alicante, Barcelona, Bilbao, Córdoba, A Coruña, Granada, Madrid, Málaga, Murcia, Sevilla, Valencia, Valladolid, Vigo y Zaragoza, añadiéndose San Sebastián en el 2006), por lo que se decidió finalmente desechar dichos años y trabajar con los artículos comprendidos entre el 1 de enero del 2008 y el 31 de diciembre de 2011, ya que durante estos años existían al menos una ciudad por provincia española, más Ceuta y Melilla como ciudades autónomas, lo cual nos permitía tener una representación completa de todo el país, al menos a nivel provincial, al trabajar con estas ciudades como representantes de sus provincias.

La lista de ciudades existente para estas fechas eran inicialmente 60: A Coruña, Albacete, *Algeciras*, Alicante, Almería, Ávila, Badajoz, Barcelona, Bilbao, Burgos, Cáceres, Cádiz, *Cartagena*, Castellón de la Plana, Ceuta, *Elche*, *Gijón*, Girona, Granada, Guadalajara, Huelva, Huesca, Jaén, *Jerez de la Frontera*, Las Palmas de Gran Canaria (Las Palmas), León, Lleida, Logroño, Lugo, Madrid, Málaga, *Marbella*, Melilla, Murcia, Ourense, Oviedo, Palencia, Palma de Mallorca (Mallorca), Pamplona, Pontevedra, Salamanca, San Sebastián, Santa Cruz de Tenerife (Tenerife), Santander, *Santiago de Compostela*, Segovia, Sevilla, Soria, Tarragona, Teruel, Toledo, Valencia, Valladolid, *Vigo*, Vitoria, Zamora y Zaragoza. De ellas, se eligieron

finalmente para los experimentos como representantes de su provincia a las capitales de cada una de éstas, más Ceuta y Melilla como ciudades autónomas, siendo eliminadas las 8 ciudades puestas en cursiva, es decir, un total de 52 ciudades, que quedaron distribuidas tal y como muestra la tabla 4.1

Tabla 4.1: Número de artículos por ciudad y año del corpus del *20Minutos*.

Ciudad	2008		2009		2010		2011		**Media**
	Num	%	Num	%	Num	%	Num	%	
A Coruña	2140	4,2	740	2,7	1163	0,5	1463	0,6	1,0 %
Albacete	259	0,5	116	0,4	798	0,4	994	0,4	0,4 %
Alicante	2011	4,0	621	2,2	2803	1,3	2647	1,1	1,5 %
Almería	482	0,9	315	1,1	2059	1,0	2084	0,9	0,9 %
Ávila	370	0,7	261	0,9	997	0,5	915	0,4	0,5 %
Badajoz	305	0,6	384	1,4	10005	4,7	9981	4,3	4,0 %
Barcelona	3934	7,8	2632	9,5	1425	0,7	1656	0,7	1,9 %
Bilbao	2678	5,3	1115	4,0	8809	4,2	8343	3,6	4,0 %
Burgos	451	0,9	172	0,6	1453	0,7	1555	0,7	0,7 %
Cáceres	261	0,5	159	0,6	2536	1,2	2143	0,9	1,0 %
Cádiz	481	0,9	257	0,9	2665	1,2	3055	1,3	1,2 %
Castellón	336	0,7	164	0,6	1407	0,7	1333	0,6	0,6 %
Ceuta	350	0,7	114	0,4	258	0,1	245	0,1	0,2 %
Ciudad Real	303	0,6	155	0,6	1719	0,8	1858	0,8	0,8 %
Córdoba	2192	4,3	794	2,9	3192	1,5	2456	1,1	1,7 %
Cuenca	203	0,4	97	0,3	852	0,4	1171	0,5	0,4 %
Girona	398	0,8	278	1,0	1103	0,5	1183	0,5	0,6 %
Granada	2388	4,7	789	2,8	3342	1,6	2839	1,2	1,8 %
Guadalajara	250	0,5	113	0,4	801	0,4	1008	0,4	0,4 %
Huelva	813	1,6	352	1,3	3246	1,5	3417	1,5	1,5 %
Huesca	321	0,6	217	0,8	2416	1,1	2121	0,9	1,0 %
Jaén	399	0,8	227	0,8	2002	0,9	1817	0,8	0,9 %
Las Palmas	341	0,7	556	2,0	7475	3,5	4979	2,2	2,6 %
León	396	0,8	186	0,7	2073	1,0	1977	0,9	0,9 %
Lleida	393	0,8	299	1,1	1218	0,6	782	0,3	0,5 %
Logroño	301	0,6	443	1,6	6764	3,2	7439	3,2	2,9 %
Lugo	249	0,5	99	0,3	548	0,2	632	0,3	0,3 %
Madrid	3895	7,7	2588	9,4	1986	0,9	1994	0,9	2,0 %
Málaga	1985	3,9	812	2,9	5813	2,7	7966	3,5	3,2 %
Mallorca	469	0,9	394	1,4	8621	4,1	11285	4,9	4,0 %
Melilla	283	0,6	60	0,2	390	0,2	367	0,2	0,2 %

Continúa en la página siguiente...

Tabla 4.1 – *Continuación de la página anterior*

Ciudad	2008		2009		2010		2011		**Media**
	Num	%	Num	%	Num	%	Num	%	
Murcia	2665	5,3	1260	4,6	9522	4,5	14714	6,4	5,4 %
Ourense	323	0,6	128	0,5	573	0,3	632	0,3	0,3 %
Oviedo	1792	3,5	893	3,2	8676	4,1	10150	4,4	4,1 %
Palencia	346	0,7	76	0,3	832	0,4	823	0,3	0,4 %
Pamplona	789	1,6	501	1,8	5896	2,8	6320	2,8	2,6 %
Pontevedra	400	0,8	96	0,3	609	0,3	591	0,2	0,3 %
Salamanca	564	1,1	252	0,9	1659	0,8	1559	0,7	0,8 %
San Sebastián	480	0,9	314	1,1	2286	1,1	2244	1,0	1,0 %
Santander	827	1,6	886	3,2	15096	7,1	16048	7,0	6,3 %
Segovia	282	0,6	78	0,3	727	0,3	839	0,4	0,4 %
Sevilla	2221	4,4	1601	5,8	21101	9,9	25859	11,2	9,8 %
Soria	300	0,6	95	0,3	667	0,3	530	0,2	0,3 %
Tarragona	479	0,9	268	1,0	1159	0,5	780	0,3	0,5 %
Tenerife	547	1,1	475	1,7	6892	3,2	5840	2,5	2,6 %
Teruel	268	0,5	158	0,6	1771	0,8	1368	0,6	0,7 %
Toledo	396	0,8	419	1,5	8574	4,0	11179	4,9	4,0 %
Valencia	2696	5,3	1839	6,6	14217	6,7	13842	6,0	6,3 %
Valladolid	2770	5,5	1489	5,4	10030	4,7	13468	5,9	5,3 %
Vitoria	357	0,7	206	0,7	2594	1,2	2484	1,1	1,1 %
Zamora	273	0,5	107	0,4	1012	0,5	907	0,4	0,4 %
Zaragoza	1967	3,9	1013	3,7	8029	3,8	7778	3,4	3,6 %
TOTAL	**50.379**	**9,7**	**27.663**	**5,3**	**211.861**	**40,8**	**229.660**	**44,2**	
	519.563								

En dicha tabla se puede observar el número de artículos publicados en cada ciudad para cada año, así como el porcentaje en número de artículos que tenía cada una de estas ciudades por año. En la última columna se muestra el porcentaje en número de artículos que tenía cada una de estas ciudades en el global de todos los años. Por otro lado, en la penúltima fila se puede apreciar el número total de artículos que hay en cada año, así como el porcentaje de éstos en el total del corpus completo, es decir, el porcentaje en número de artículos que hay por cada año del corpus. Finalmente, en la última fila se muestra la suma total de artículos del corpus completo.

En la tabla 4.1 también se pueden ver las ciudades con más y menos artículos por cada uno de los años expuestos, así como en el global del corpus:

- **2008**. La que menos artículos publicó fue Cuenca con 203, un 0,4 % del total de artículos de ese año, mientras que la que más artículos publicó fue Barcelona con 3.934, un 7,8 % del total de artículos de ese

año. El total de artículos publicados para ese año fue de 50.379, lo que hace una media de 968,82 artículos por ciudad con una desviación estándar de 1.008,92.

- **2009**. La que menos artículos publicó fue Melilla con 60, un 0,2 % del total de artículos de ese año, mientras que la que más artículos publicó fue Barcelona con 2.632, un 9,5 % del total de artículos de ese año. El total de artículos publicados para este años fue de 27.663, lo que hace una media de 531,98 artículos por ciudad con una desviación estándar de 588,36.

- **2010**. La que menos artículos publicó fue Ceuta con 258, un 0,1 % del total de artículos de ese año, mientras que la que más artículos publicó fue Sevilla con 21.101, un 9,9 % del total de artículos de ese año. El total de artículos publicados para ese año fue de 211.861, lo que hace una media de 4.074,25 artículos por ciudad con una desviación estándar de 4.387,80.

- **2011**. La que menos artículos publicó fue Ceuta con 245, un 0,1 % del total de artículos de ese año, mientras que la que más artículos publicó fue Sevilla con 25.859, un 11,2 % del total de artículos de ese año. El total de artículos publicados para ese año fue de 229.660, lo que hace una media de 4.416,53 artículos por ciudad con una desviación estándar de 5.206,39.

- **Corpus completo**. La que menos artículos publicó fue Ceuta con 967, un 0,2 % del total de artículos del corpus, mientras que la que más artículos publicó fue Sevilla con 50.782, un 9,7 % del total de artículos del corpus. El total de artículos publicados para todo el corpus fue de 519.563, lo que hace una media de 9.991,59 artículos por ciudad con una desviación estándar de 10.337,97.

Preprocesado

De los documentos obtenidos en el apartado anterior se eliminaron todas las *URLs*, signos de puntuación y caracteres especiales tales como barras bajas, barras, etc., que había en los textos. También se pasaron todos los términos a minúsculas a excepción de la primera letra de cada uno de ellos que se mantuvo con el fin de poder identificar posibles nombres propios.

Una vez realizada esta "*limpieza*", partiendo del corpus obtenido se crearon cuatro corpus adicionales. Estos corpus estaban compuestos por distintas categorías gramaticales:

- *Corpus completo.* El corpus original.

- *Corpus completo menos los topónimos*. Al corpus original se le eliminaron los topónimos con el fin de poder apreciar qué aportaban dichos términos a la tarea en cuestión.
- *Adjetivos y sustantivos*. En esta ocasión, únicamente se utilizaron los adjetivos y sustantivos del corpus por ser los que teóricamente más aportan a la tarea de clasificación geográfica de textos.
- *Sustantivos*. Homónimo al anterior pero utilizando únicamente la categoría gramatical que más aporta a la detección del foco geográfico.
- *Topónimos*. Subcorpus del anterior donde los únicos sustantivos que hay son los topónimos.

Para la creación de cada uno de estos corpus adicionales se utilizó el etiquetador gramatical que hay dentro de la herramienta *FreeLing*[5], descrita con más detenimiento en la sección 5.2.1.

4.2. Twitter

Twitter es un servicio de microblogging (ver image 4.3). Desde que Jack Dorsey lo creó en marzo de 2006, y lo lanzó en julio del mismo año, la red ha ganado popularidad mundialmente y se estima que tiene más de 500 millones de usuarios, generando 65 millones de tweets al día y maneja más de 800.000 peticiones de búsqueda diarias, según comunico la propia compañía el 28 de marzo de 2008. Ha sido apodado como el "SMS de Internet". Entre sus usuarios se destacan grandes figuras públicas, como el presidente de los Estados Unidos Barack Obama (43,5 millones de seguidores), deportistas como Rafael Nadal (6,21 millones de seguidores), y músicos como Lady Gaga (41,5 millones de seguidores) o Katy Perry (53,6 millones de seguidores), entre otros.

Los mensajes que sus usuarios envían y leen (tuits) están limitados a 140 caracteres, que se muestran en la página principal del usuario. Los usuarios registrados pueden leer y crear tuits, pero los usuarios no registrados pueden simplemente leerlos. Los usuarios pueden suscribirse a los tuits de otros usuarios, a esto se le llama "*seguir*" y a los usuarios abonados se les llama "*seguidores*" (*followers*). Por defecto, los mensajes son públicos, pudiendo difundirse privadamente mostrándolos únicamente a unos seguidores determinados. Los usuarios pueden *twittear* desde la web del servicio, con aplicaciones oficiales externas (como para teléfonos inteligentes), o mediante el Servicio de mensajes cortos (*SMS*) disponible en ciertos países.

Los usuarios pueden agrupar mensajes sobre un mismo tema mediante el uso de etiquetas de almohadilla (*hashtag*), las cuales son palabras o frases

[5] http://nlp.lsi.upc.edu/freeling/node/1

Figura 4.3: Interfaz gráfica de Twitter.

iniciadas mediante el uso de una "#". De forma similar, la "@" seguida de un nombre de usuario se usa para mencionar o contestar a otros usuarios. Para volver a postear un mensaje de otro usuario, y compartirlo con los propios seguidores, la función de retuit se marca con un "*RT*" en el mensaje. A finales de 2009 se añadió la opción de listas, haciendo posible el seguir (así como mencionar y contestar) listas de usuarios en vez de usuarios individuales.

Un tuit o *status*, tiene asociados una serie de datos tales como: el usuario que lo emitió, datos de dicho usuario (entre ellos la localización que el usuario ha introducido en un campo de texto libre, es decir, puede escribir lo que desee), lugar de emisión (siempre y cuando tenga activada la geolocalización), etc.

Los mensajes fueron fijados a 140 caracteres máximo para la compatibilidad con los mensajes *SMS*, introduciendo la notación de la taquigrafía y el argot de Internet comúnmente usado en los *SMS*. El límite de 140 caracteres también ha llevado a la proliferación de servicios de reducción de *URLs*, como *bit.ly*, *goo.gl*, y *tr.im*, y web de alojamiento de material, como *Twitpic*, *memozu.com* y *NotePub* para subir material multimedia y textos superiores a 140 caracteres. *Twitter* usa *bit.ly* para acortar las *URLs* puestas en su servicio.

Según un estudio realizado por *Semiocast*[6] en 2012, analizando 383 millones de cuentas creadas antes de dicho año, los países con mayor número

[6] http://semiocast.com/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan

Figura 4.4: Mapa de actividad de los usuarios de Twitter.

Figura 4.5: Ejemplo de los textos emitidos en la red social *Twitter*.

de usuarios en *Twitter* son los Estados Unidos (107,7 millones), Brasil (33,3 millones), Japón (29,9 millones), Reino Unido (23 millones), Indonesia (19 millones), India (12 millones), México (10,5 millones), Filipinas (8 millones), España (7,9 millones) y Canadá (7,5 millones). En la figura 4.4 se puede observar un mapa con las zonas más activas de Twitter.

La gran expansión y asentamiento de esta red social, unido a su peculiar lenguaje de comunicación con un alto grado de informalidad repleto de contracciones y abreviaturas, ha hecho que la problemática de la detección del foco geográfico en estos textos adquieran una nueva dimensión. Si además le añadimos que la mayoría de los mensajes emitidos en *Twitter* son públicos y pueden estar geo-referenciados, esto ha hecho que nos hayamos decantado por esta red social con el fin detectar la ubicación geográfica de los usuarios de la misma a partir únicamente de los últimos mensajes emitidos por éstos. En la figura 4.5 se puede apreciar un ejemplo de tuit relacionado con la ciudad de Jaén.

Creación del corpus

Para la obtención de nuestro corpus de *Twitter* se utilizó la *SEARCH API*, ahora incluida en la *REST API v1.1*[7] de *Twitter* aunque con algunas limitaciones.

Mediante esta *API* se obtuvieron tuits georreferenciados pertenecientes a cada una de las 50 capitales de provincia españolas más sus 2 ciudades autónomas para conformar un total de 52 ciudades, que coinciden con las del corpus del *20Minutos* (ver sección 4.1), para así dar una cobertura similar a la expuesta para el corpus de dicha fuente a la hora de experimentar. Estos tuits fueron capturados desde el 20 de abril de 2013 al 10 de junio de ese mismo año.

Puesto que los usuarios pueden enviar tuits desde más de una ciudad del corpus, se decidió agrupar los tuits de los usuarios por ciudad, es decir, si hay tuits pertenecientes a un usuario procedentes desde la ciudad A y tuits de ese mismo usuarios desde la ciudad B, son tratados como conjuntos de tuits independientes. La lógica que subyace tras esto es que cuando un usuario tuitea desde otra ciudad, en muchas ocasiones lo hace haciendo referencia a algún lugar o entidad perteneciente al lugar de dónde se tuitea, por lo que puede ser de gran ayuda a la hora de extraer características de dicho lugar.

Los datos más relevantes extraídos de los tuits capturados son:

- El texto del tuit. Son los hasta 140 caracteres que cada usuario puede emitir por tuit.
- El lugar de procedencia del tuit. Es una de las 52 ciudades previamente mencionadas desde las cuales se puede haber emitido el tuit. Este dato ha sido obtenido mediante la aplicación de un filtro de tuits por ubicación que proporcionaba la propia *SEARCH API* de *Twitter*.
- El usuario que emite el tuit. Es el usuario que ha creado el tuit.

Los números de este corpus se pueden ver en la tabla 4.2.

Tabla 4.2: Número de tuits y usuarios de *Twitter* por ciudad y porcentaje de los mismos en el corpus recopilado.

Ciudad	**Tuits**		**Usuarios**	
	Número	Porcentaje	Número	Porcentaje
A Coruña	95497	2,01 %	2930	1,42 %
Albacete	53654	1,13 %	2197	1,06 %
Alicante	105869	2,22 %	4427	2,15 %
Almería	100953	2,12 %	3017	1,46 %

Continúa en la página siguiente...

[7] https://dev.twitter.com/docs/using-search.

Tabla 4.2 – *Continuación de la página anterior*

Ciudad	Tuits		Usuarios	
	Número	Porcentaje	Número	Porcentaje
Ávila	17037	0,35 %	867	0,42 %
Badajoz	81505	1,71 %	2530	1,22 %
Barcelona	259587	5,46 %	21295	10,34 %
Bilbao	112498	2,36 %	4490	2,18 %
Burgos	44043	0,92 %	1632	0,79 %
Cáceres	56438	1,18 %	2199	1,06 %
Cádiz	57117	1,20 %	2240	1,08 %
Castellón	44946	0,94 %	1978	0,96 %
Ceuta	11751	0,24 %	300	0,14 %
Ciudad Real	34151	0,71 %	1347	0,65 %
Córdoba	153445	3,23 %	5734	2,78 %
Cuenca	21309	0,44 %	864	0,41 %
Girona	14602	0,30 %	1170	0,56 %
Granada	120971	2,54 %	6439	3,12 %
Guadalajara	36055	0,75 %	1242	0,60 %
Huelva	83031	1,74 %	2697	1,30 %
Huesca	13198	0,27 %	506	0,24 %
Jaén	58117	1,22 %	2166	1,05 %
Las Palmas	45839	0,96 %	2006	0,97 %
León	52181	1,09 %	1870	0,90 %
Lleida	11844	0,24 %	904	0,43 %
Logroño	30640	0,64 %	1407	0,68 %
Lugo	23054	0,48 %	976	0,47 %
Madrid	857681	18,06 %	42127	20,46 %
Málaga	216369	4,55 %	8978	4,36 %
Mallorca	65937	1,38 %	3356	1,62 %
Melilla	9715	0,20 %	352	0,17 %
Murcia	170628	3,59 %	6458	3,13 %
Ourense	42276	0,89 %	1162	0,56 %
Oviedo	89750	1,89 %	3111	1,51 %
Palencia	30783	0,64 %	890	0,43 %
Pamplona	35261	0,74 %	1829	0,88 %
Pontevedra	25576	0,53 %	1204	0,58 %
Salamanca	77512	1,63 %	2913	1,41 %
San Sebastián	38182	0,80 %	2095	1,01 %
Santander	61992	1,30 %	2399	1,16 %
Segovia	20130	0,42 %	1133	0,55 %
Sevilla	365335	7,69 %	14439	7,01 %

Continúa en la página siguiente...

Tabla 4.2 – *Continuación de la página anterior*

Ciudad	**Tuits**		**Usuarios**	
	Número	Porcentaje	Número	Porcentaje
Soria	11530	0,24 %	432	0,20 %
Tarragona	11743	0,24 %	1248	0,60 %
Teruel	8120	0,17 %	399	0,19 %
Tenerife	26024	0,54 %	1475	0,71 %
Toledo	39375	0,82 %	2132	1,03 %
Valencia	498197	10,49 %	17043	8,27 %
Valladolid	99294	2,09 %	3713	1,80 %
Vitoria	37758	0,79 %	1547	0,75 %
Zamora	20045	0,42 %	794	0,38 %
Zaragoza	149487	3,14 %	5340	2,59 %
TOTAL	**4.748.032**			**205.895**

En dicha tabla se puede observar el número de tuits obtenidos en cada ciudad, así como el porcentaje en número de tuits por cada una de las ciudades del corpus. También se puede advertir en las dos últimas columnas el número de usuarios por ciudad existentes en el corpus y el porcentaje en dicho número de usuarios que tiene cada una de las ciudades del corpus respectivamente. La última fila muestra el número total de tuits del corpus y el de usuarios existentes en distintas ciudades. Nótese que si un usuario ha tuiteado en n ciudades, será contabilizado como n usuarios. La lógica que subyace bajo este procedimiento es que dado que se pretende averiguar dónde se encuentra un usuario según los tuits que ha emitido desde una ciudad dada, si nos centrásemos únicamente en la ciudad de procedencia del usuario y éste estuviera en cualquier otra ciudad, es muy probable que el usuario esté emitiendo tuits con términos pertenecientes a la ciudad que se encuentra visitando en vez de la de procedencia, lo que introduciría ruido a nuestro sistema además de no ser de gran utilidad práctica el averiguar la ciudad de procedencia cuando el usuario se encuentra en otra ciudad.

En la tabla 4.2 también se pueden ver las ciudades con más y menos tuits y usuarios de nuestro corpus:

- **Ciudad con mayor número de tuits**. Madrid es la ciudad desde la que se han recogido un mayor número de tuits con un total de 857.681 tuits emitidos en dicha ciudad, lo cual representa un 18,06 % del total de tuits del corpus.

- **Ciudad con menor número de tuits**. Teruel es la ciudad desde la que se han recogido un menor número de tuits, ya que tan sólo se han obtenido 8.120 tuits en dicha ciudad, lo cual representa un 0,17 % del total de tuits del corpus.

- **Ciudad con mayor número de usuarios**. Madrid es la ciudad desde la que más usuarios han tuiteado en nuestro corpus con un total de 42.127 usuarios distintos para esta ciudad, lo cual representa un 20,46 % del total de usuarios del corpus.

- **Ciudad con menor número de usuarios**. Ceuta es la ciudad desde la que menos usuarios han tuiteado en nuestro corpus, donde tan sólo hay 300 usuarios distintos para esta ciudad, lo cual representa un 0,14 % del total de usuarios del corpus.

- **Media y desviación estándar de tuits y usuarios por ciudad**. Según los datos que se reflejan en la tabla de nuestro copus de *Twitter*, se han obtenido una media de 91.308 tuits por ciudad con una desviación estándar de 139.659,33, mientras que si nos centramos en el número de usuarios podemos ver cómo la media es de 3.960 usuarios por ciudad con una desviación estándar de 6.686,25.

Preprocesado

Dada la gran informalidad del texto de los tuits, se procedió a eliminar todos los signos de puntuación de los mismos así como otros caracteres especiales (barras bajas, barras invertidas, etc.) y *URLs*, ya que éstos, a diferencia de en otras tareas de *PLN* como en el análisis de sentimiento, no aportaban prácticamente nada para identificar el foco geográfico del texto[8]. Por otro lado, los términos que comenzaban por los caracteres '#' y '@' se mantuvieron por representar etiquetas (*hashtags*) y menciones de usuarios de *Twitter* respectivamente, lo cual podría ser de gran ayuda a la hora de determinar la ubicación de un usuario.

También fueron eliminados los tuits repetidos (los que los usuarios copiaban y volvían a emitir, ya fueran originalmente suyos o de otros usuarios) así como los retuiteados, ya que no aportaban ninguna información adicional y sí que podían incluir ruido al poder ser emitidos desde distintas ciudades.

No se tuvieron en cuenta las relaciones de seguidor (*follower*) entre usuarios, ya que este estudio se basa meramente en *PLN* a través del texto procedente de los tuits, es decir, que tampoco se tuvo en cuenta otros campos de los perfiles de usuario tales como el campo '*ubicación*'.

4.3. Wikipedia

Wikipedia es una enciclopedia libre, políglota y editada colaborativamente. Es administrada por la *Fundación Wikimedia*, una organización sin ánimo

[8] Las *URLs* podrían ser útiles para detectar el foco geográfico, pero en esta tesis se va a hacer una aproximación puramente basada en *PLN*.

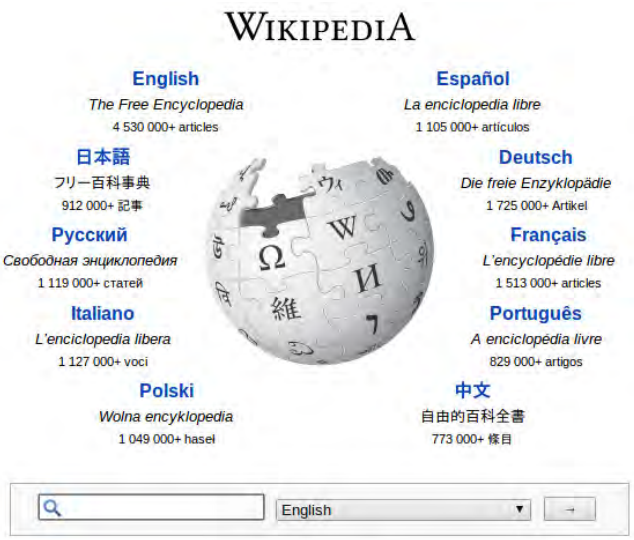


Figura 4.6: Página inicial de *Wikipedia*.

de lucro. Sus más de 37 millones de artículos en 284 idiomas (cantidad que incluye idiomas artificiales como el esperanto, lenguas indígenas o aborígenes como el náhuatl, el maya y las lenguas de las islas Andamán, y dialectos de muchos idiomas) han sido redactados conjuntamente por voluntarios de todo el mundo, y cualquier persona con acceso al proyecto puede editarlos. Iniciada en enero de 2001 por Jimmy Wales y Larry Sanger, es la mayor y más popular obra de consulta en internet (ver imagen de la página inicial de *Wikipedia* 4.6).

Un estudio realizado por la *Universidad Carnegie Mellon* y *Palo Alto Research Center* sobre las mayores categorías de artículos en la *Wikipedia* en inglés desde julio de 2006 hasta enero de 2008, mostró que el 14 % de sus artículos estaban relacionados con la geografía y lugares, teniendo éstos un porcentaje de crecimiento del 52 % desde julio de 2006 hasta enero de 2008 (Kittur et al., 2009).

La Wikipedia contiene una gran cantidad de nombres de lugares, que se pueden extraer de forma automática y ser así recopilado para formar recursos tan útiles como los *gazetteers* (Overell and Rüger, 2006).

La información geográfica de un artículo de la *Wikipedia* no viene dada sólo en su título cuando coincide con algún topónimo, sino que hay una serie de campos predeterminados que pueden aportar un gran soporte geográfico, sin tener que adentrarse en el texto del artículo en sí.

Así pues, si nos centramos en ciudades, los campos más relevantes que pueden aportar información para su correcta ubicación son:

- **Coordenadas**. Muestra las coordenadas *UTM* y es un campo que se puede encontrar para gran parte de los artículos de lugares geográficos,

a excepción de países y continentes. Es el campo más importante a la hora de ubicar correctamente el artículo en sí.

- **País, comunidad autónoma, etc**. Indica las entidades administrativas superiores a la ciudad dada.

- **Ubicación**. Es un campo que muestra cierta información que puede resultar de utilidad a la hora de identificar el emplazamiento del artículo en sí. Entre esta información se puede encontrar: altitud sobre el nivel del mar, distancia a las ciudades cercanas más importantes, etc.

- **Gentilicio**. Tanto en masculino como en femenino, así como en otras lenguas oficiales del lugar.

- **Código postal**.

- **Prefijo telefónico**.

- **Otros**: apodo, lema, superficie, año de fundación, altitud sobre el nivel del mar, alcalde, presupuesto, fiestas mayores, ciudades hermanadas, patrones, sitios webs, etc.

Creación del corpus

Para la recopilación del corpus de esta tesis se tuvieron únicamente en cuenta los artículos que trataban sobre cada una de las 52 ciudades previamente mencionadas en los corpus del *20Minutos* y *Twitter* (ver secciones 4.1 y 4.2), así como los artículos a los que se hacía referencia desde éstos 52 artículos iniciales.

La obtención de los 52 artículos pertenecientes a las capitales de provincia, Ceuta y Melilla, se realizó manualmente, dado que en muchas ocasiones existía algún tipo de ambigüedad con otros artículos de *Wikipedia* (provincias homónimas, ciudades con el mismo nombre en distintos lugares, etc.). Por otro lado, la obtención de los artículos mencionados en cada uno de los 52 artículos originales se hizo de forma automática mediante los enlaces que apuntaban desde estos 52 artículos a otros artículos existentes en *Wikipedia* (*outlinks*).

El objetivo que se buscaba añadiendo estos artículos referenciados era el de poder extender la cantidad de texto que se tenía con relación a cada localización con la que se trabajaba, ya que, pese a que los artículos de *Wikipedia* son mucho más extensos que los textos de las noticias, inicialmente tan sólo había un artículo por localidad.

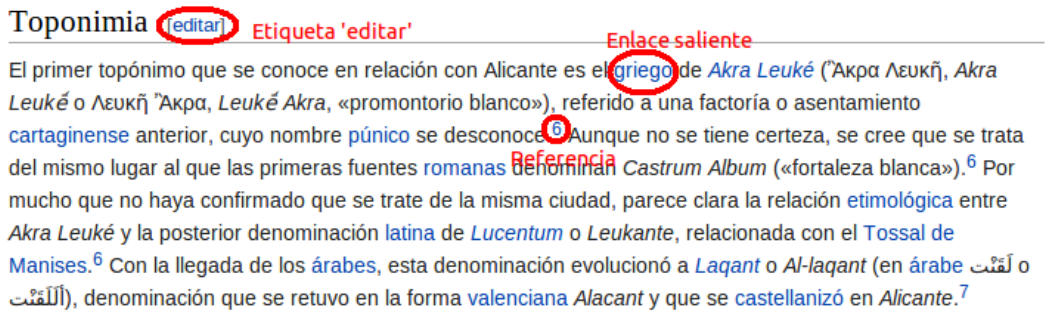


Figura 4.7: Extracto de un artículo de *Wikipedia*.

Preprocesado

De cada uno de estos artículos se utilizaron únicamente los términos que aparecían en el texto, sin tener en cuenta ningún metadato[9] al tratarse de una aproximación meramente textual.

De cada uno de los documentos obtenidos se eliminaron todos los signos de puntuación y caracteres especiales, así como todas las etiquetas '*editar*'(ver imagen 4.7) que aparecían en el texto. También se eliminaron los típicos números de las referencias que aparecen en mitad del texto de los artículos (ver imagen 4.7)

Pese a que en esta ocasión los textos del corpus de *Wikipedia* no se iba a intentar clasificar geográficamente, también se procedió a realizar la creación de cuatro corpus adicionales, al igual que se hizo con el corpus del diario *20Minutos* (ver sección 4.1), puesto que el corpus de *Wikipedia* iba a ser utilizado para la clasificación de textos del diario *20Minutos*, los cuales fueron a su vez divididos por categorías gramaticales.

Así pues, estos cuatro corpus adicionales contenían las mismas categorías gramaticales que los del diario *20Minutos*, es decir: todos los términos menos los topónimos, sustantivos y adjetivos, sólo sustantivos y sólo topónimos.

4.4. Flickr

Flickr es un sitio web gratuito que permite almacenar, ordenar, buscar, vender y compartir fotografías y vídeos en línea.

[9]Datos que describen otros datos, como por ejemplo la información asociada a cada documento de *Wikipedia* que trata sobre una localización, donde se puede obtener las coordenadas de dicha localización, el número de habitantes, etc.

Actualmente *Flickr* cuenta con una importante comunidad de usuarios que comparte las fotografías y vídeos creados por ellos mismos. Esta comunidad se rige por normas de comportamiento y condiciones de uso que favorecen la buena gestión de los contenidos.

La popularidad de *Flickr* se debe fundamentalmente a su capacidad para administrar imágenes mediante herramientas que permiten al autor etiquetar sus fotografías y explorar y comentar las imágenes de otros usuarios.

Para cada fotografía se tienen en cuenta 5 campos textuales distintos (ver figura 4.8):

- *Comentarios*. Son los comentarios que los usuarios de Flickr hacen de cada fotografía. Pueden haber múltiples comentarios pertenecientes a múltiples usuarios por cada fotografía.

- *Descripción*. Es la descripción que el usuario que ha subido la fotografía hace de la misma. Sólo puede haber una descripción por fotografía.

- *Etiquetas*. Son las etiquetas que el usuario que ha subido la fotografía le ha puesto a la misma. Puede que no haya ninguna etiqueta o que hayan múltiples etiquetas para una única fotografía. Las etiquetas se pueden repetir en distintas fotografías.

- *Notas*. Son las anotaciones que el usuario que ha subido la fotografía ha hecho de la misma. Inicialmente obedecían a los aspectos técnicos de la fotografía, tales como el objetivo o filtro utilizado, pero puesto que es un campo de texto libre, en numerosas ocasiones el contenido de este campo describe aspectos relevantes del lugar de la fotografía tal y como lo puede hacer el propio campo de descripción. Sólo puede haber una anotación por fotografía, la cual la realizará únicamente el usuario que la suba.

- *Título*. Es el título que el usuario que ha subido la fotografía le ha puesto a la misma. Este campo es el único de los 5 que es obligatorio en cada fotografía y del que sólo puede haber uno.

La localización de las fotografías utilizada en este corpus viene determinada por las coordenadas mostradas en el campo "*Location*" del fichero *XML* que describe dicha fotografía (ver figura 4.8).

En la tabla 4.3 se pueden observar los números del corpus de *Flickr*.

Figura 4.8: Ejemplo de los datos asociados a una fotografía de *Flickr* en formato *XML*.

Tabla 4.3: Números del corpus de *Flickr* por ciudad. En negrita las ciudades que no son capital de provincia pero son las más pobladas de su provincia.

Ciudad	**Comentarios**		**Descripción**		**Etiquetas**		**Notas**		**Títulos**	
	Num	%	Num	%	Num	%	Num	%	Num	%
A Coruña	11024	5,0	14063	4,5	14063	4,7	14063	4,5	14063	4,5
Albacete	2187	1,0	2693	0,9	0	0,0	2693	0,9	2693	0,9
Alicante	9386	4,3	11321	3,6	11321	3,8	11321	3,6	11321	3,6
Almería	3048	1,4	3439	1,1	3439	1,1	3439	1,1	3439	1,1
Ávila	1555	0,7	2252	0,7	2252	0,8	2252	0,7	2252	0,7
Badajoz	917	0,4	1170	0,4	1170	0,4	1170	0,4	1170	0,4
Barcelona	7790	3,5	14447	4,6	14447	4,9	14447	4,6	14447	4,6
Bilbao	6651	3,0	9846	3,2	9846	3,3	9846	3,2	9846	3,2
Burgos	4563	2,1	6717	2,2	6717	2,3	6717	2,2	6717	2,2
Cáceres	1819	0,8	2712	0,9	2712	0,9	2712	0,9	2712	0,9
Castellón	2784	1,3	4545	1,5	4545	1,5	4545	1,5	4545	1,5
Ceuta	249	0,1	290	0,1	0	0,0	290	0,1	290	0,1
Ciudad Real	748	0,3	1062	0,3	1062	0,3	1062	0,3	1062	0,3
Córdoba	3651	1,7	4434	1,4	4434	1,5	4434	1,4	4434	1,4
Cuenca	1320	0,6	1984	0,6	1984	0,7	1984	0,6	1984	0,6
Gijón	3345	1,5	5023	1,6	5023	1,7	5023	1,6	5023	1,6
Girona	5071	2,3	6672	2,1	6672	2,2	6672	2,1	6672	2,1
Granada	17072	7,8	22366	7,2	10532	3,5	22366	7,2	22366	7,2
Guadalajara	1262	0,6	1856	0,6	1856	0,6	1856	0,6	1856	0,6
Huelva	6564	3,0	7696	2,5	7696	2,6	7696	2,5	7696	2,5
Huesca	1883	0,9	3002	1,0	3002	1,0	3002	1,0	3002	1,0
Jaén	1887	0,9	2182	0,7	2182	0,7	2182	0,7	2182	0,7

Continúa en la página siguiente...

Tabla 4.3 – *Continuación de la página anterior*

Ciudad	Comentarios		Descripción		Etiquetas		Notas		Títulos	
	Num	%	Num	%	Num	%	Num	%	Num	%
Jerez	7966	3,6	9721	3,1	9721	3,3	9721	3,1	9721	3,1
Las Palmas	2951	1,3	4634	1,5	4634	1,6	4634	1,5	4634	1,5
Lleida	1810	0,8	2246	0,7	2246	0,8	2246	0,7	2246	0,7
León	7088	3,2	8776	2,8	8776	3,0	8776	2,8	8776	2,8
Logroño	752	0,3	1096	0,3	1096	0,4	1096	0,3	1096	0,3
Lugo	1485	0,7	1848	0,6	1848	0,6	1848	0,6	1848	0,6
Madrid	14567	6,6	25386	8,1	25386	8,6	25386	8,1	25386	8,1
Málaga	6318	2,9	7419	2,4	7419	2,5	7419	2,4	7419	2,4
Mallorca	11351	5,2	17157	5,5	17157	5,8	17157	5,5	17157	5,5
Melilla	2059	0,9	2526	0,8	2526	0,9	2526	0,8	2526	0,8
Murcia	4478	2,0	7438	2,4	7438	2,5	7438	2,4	7438	2,4
Ourense	680	0,3	1094	0,3	1094	0,4	1094	0,3	1094	0,3
Palencia	1578	0,7	1944	0,6	1944	0,6	1944	0,6	1944	0,6
Pamplona	515	0,2	622	0,2	622	0,2	622	0,2	622	0,2
Salamanca	3853	1,8	5697	1,8	5697	1,9	5697	1,8	5697	1,8
San Sebastian	751	0,3	1541	0,5	1541	0,5	1541	0,5	1541	0,5
Tenerife	700	0,3	1850	0,6	1850	0,6	1850	0,6	1850	0,6
Santander	567	0,2	836	0,3	0	0,0	836	0,3	836	0,3
Segovia	2579	1,2	5004	1,6	5004	1,7	5004	1,6	5004	1,6
Sevilla	9442	4,3	17139	5,5	17139	5,8	17139	5,5	17139	5,5
Soria	2792	1,3	3788	1,2	3788	1,3	3788	1,2	3788	1,2
Talavera	2562	1,2	4518	1,4	4518	1,5	4518	1,4	4518	1,4
Tarragona	2582	1,2	4289	1,4	4289	1,4	4289	1,4	4289	1,4
Teruel	983	0,4	1557	0,5	1557	0,5	1557	0,5	1557	0,5
Valencia	17076	7,8	23196	7,4	23196	7,8	23196	7,4	23196	7,4
Valladolid	6159	2,8	7256	2,3	7256	2,4	7256	2,3	7256	2,3
Vigo	2944	1,3	3511	1,1	3511	1,2	3511	1,1	3511	1,1
Vitoria	2686	1,2	3313	1,1	3313	1,1	3313	1,1	3313	1,1
Zamora	2221	1,0	2471	0,8	2471	0,8	2471	0,8	2471	0,8
Zaragoza	3332	1,5	4564	1,5	4564	1,5	4564	1,5	4564	1,5
TOTAL	**219.603**		**312.209**		**296.556**		**312.209**		**312.209**	

Al igual que ocurriera en los tres corpus previos (ver secciones 4.1, 4.2 y 4.3) se mantuvo una única ciudad como representante de cada una de las provincias de España, más Ceuta y Melilla como ciudades autónomas, pero a diferencia de éstos, en las provincias en las que existía una ciudad con mayor población que la capital de la propia provincia, ésta ciudad fue seleccionada debido al escaso número de fotografías georreferenciadas

que existían, especialmente en las localidades más pequeñas. Estas ciudades aparecen en negrita en la tabla 4.3.

Creación del corpus

En esta ocasión, para la obtención de un conjunto de fotografías suficientemente grande que nos permitiera trabajar con cada provincia de España más sus dos ciudades autónomas, se obtuvieron manualmente los grupos de usuarios más representativos de la ciudad más poblada de España por cada provincia más Ceuta y Melilla. A partir de estos grupos se obtuvieron los usuarios y fotografías que pertenecían a los mismos, así como todos sus datos asociados (título, etiquetas, coordenadas geográficas, etc.).

Preprocesado

Al igual que sucediera con el corpus de *Twitter*, los textos de *Flickr* contenían un alto grado de informalidad, por lo que se procedió a realizar un preprocesado que se encargará de "*limpiar*" el corpus de signos de puntuación y caracteres tales como barras bajas, barras invertidas, *URLs*.

4.5. Conclusiones

Con el objetivo de poder realizar experimentos que permitan conocer lo que pueden aportar recursos externos textuales no estructurados en la obtención del foco geográfico de otros recursos de la misma o distinta naturaleza, se han adquirido cuatro corpus de distinta naturaleza, dos de ellos se consideran textos formales, *20Minutos* y *Wikipedia*, y los otros dos informales, *Twitter* y *Flickr*.

En este capítulo se ha mostrado cómo se han obtenido los corpus mencionados así como el preprocesado aplicado para la posterior experimentación. También se han podido ver las cifras de cada uno de estos corpus con el fin de obtener una visión general del peso que cada una de las ciudades de nuestro corpus tiene.

Por el lado de los corpus de textos formales también se han filtrado dichos textos para obtener únicamente los términos que correspondían a las siguientes categorías gramaticales: topónimos, sustantivos (los cuales engloban a los anteriores), sustantivos más adjetivos, así como un corpus en el que fueron eliminados todos los topónimos. El objetivo de este filtrado no es otro que el comprobar el supuesto ruido que aportan el resto de términos que no entran dentro de estas categorías cuando se pretende determinar el foco geográfico de un texto.

En el próximo capítulo se experimentará sobre estos corpus para ver cómo afecta la agregación de recursos distintos del que se pretende obtener su foco geográfico. Así pues, se verá cómo clasificar geográficamente los textos

formales del corpus del periódico *20Minutos* utilizando como entrenamiento el propio corpus de *20Minutos*, un corpus de texto formal (*Wikipedia*), y otro corpus de naturaleza informal (*Twitter*), así como las diversas combinaciones de los mismos.

Por otro lado, y análogamente a lo que se va a hacer con el corpus de textos formales de *20Minutos*, se intentará obtener el foco geográfico de un corpus de textos informales, *Twitter*, utilizando para dicha tarea el propio corpus de *Twitter* expuesto en este capítulo, utilizando un corpus de textos formales (*Wikipedia*), y otro corpus de naturaleza informal (*Flickr*), así como las diversas combinaciones de dichos corpus.

5 Experimentación

En este capítulo se detallarán los experimentos que se han llevado a cabo en la detección del foco geográfico en los textos descritos en el capítulo anterior.

Para la detección del foco geográfico se han utilizado una serie de algoritmos de aprendizaje automático que se describirán en la sección 5.1.

Para la obtención de las características empleadas en los algoritmos de aprendizaje automático se han utilizado un conjunto de herramientas que se mostraran en la sección 5.2.

Una vez mostradas las técnicas con las que se pretende averiguar el foco geográfico de los textos tratados, se estudiará la aplicación de dichas técnicas obedeciendo al grado de formalidad que tengan los textos a clasificar. Así pues, en las secciones 5.3 y 5.4 se mostrarán las aproximaciones llevadas a cabo con corpus formales e informales respectivamente.

5.1. Algoritmos de aprendizaje

Dentro de las distintas técnicas utilizadas para la clasificación geográfica de textos, suelen predominar la utilización de clasificadores que utilizan técnicas de aprendizaje automático, tales como *SVM* y modelos de lenguaje, tal y como se comentó en el capítulo 3. En esta sección se va a detallar el funcionamiento de los algoritmos pertenecientes a dichas técnicas que se han utilizado en la experimentación realizada en este capítulo.

5.1.1. SVM

Para clasificar geográficamente los textos de los distintos corpus se ha utilizado un clasificador de múltiples clases (en nuestro caso han sido 52 clases, representadas por cada una de las ciudades del corpus) de las implementaciones del algoritmo de aprendizaje automático de *Máquinas de Vectores de Soporte* (*SVM: Support Vector Machine*) descrito en (Fan et al., 2008) que se encuentran dentro de la librería *LibLINEAR*[1]. Concretamente se ha utilizado la implementación *L2-SVM* que es un método de descenso

[1] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

de coordenadas (Hsieh et al., 2008). Para *L2-SVM LibLINEAR* se ha configurado un método *Newton de región de confianza* (Lin et al., 2008) para maximizar el logaritmo de la probabilidad del modelo de regresión logística.

Se ha optado por *LibLINEAR* ya que ofrece una implementación de *SVM* optimizada para trabajar en espacios de alta dimensionalidad, con un rendimiento superior en términos de velocidad y precisión para tareas de clasificación de textos (Hsieh et al., 2008).

LibLINEAR es particularmente útil dada su gran eficacia y fácil configuración. La manera más extendida en el procesamiento del lenguaje de representar los documentos de texto es como modelos de "*bolsa de palabras*" (*bag-of-words*) (Harris, 1954), es decir, cada término dentro del documento es tratado como un término independiente a los demás, el cual a su vez será tratado como una característica del clasificador.

Para crear los vectores de características se utiliza un esquema de ponderación de términos, es decir, se indica cuántas veces una característica concreta (término en nuestro caso) aparece en un documento dado (frecuencia de aparición). Los vectores son muy dispersos, con gran cantidad de dimensiones a 0, es decir, que la mayoría de términos no aparecen en todos los documentos, por lo que se utiliza un formato de representación compacto por eficiencia de espacio, tal y como se verá más adelante.

Puesto que *SVM* es un clasificador binario, la estrategia empleada para la clasificación con múltiples clases es la de *uno contra todos*, es decir, por cada una de las clases posibles (ciudades en nuestro caso), se calcula la probabilidad de que un texto pertenezca a ella. Esta ciudad sería la clase positiva, mientras que el resto de ciudades serían la clase negativa. Posteriormente se comprueba contra la siguiente ciudad, y así sucesivamente hasta haberlo comprobado contra todas las ciudades (clases) existentes, devolviendo como resultado la ciudad con mayor probabilidad.

El problema que trata de solucionar un algoritmo de *SVM* (Boser et al., 1992) es un problema de clasificación, el cual está expresado en la ecuación 5.1,

$$\min_{w} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} (\max(0, 1 - y_i w^T x_i))^2 \tag{5.1}$$

donde:

- w, representa la matriz de pesos.
- C, es un parámetro de penalización.
- l, es el número de características (términos de los documentos) distintas existentes.
- x, son el conjunto de características (términos de los documentos).

- y, indica la frecuencia de x en el documento.

5.1.2. Modelos de lenguaje

En esta sección se va a describir una aproximación distinta a *SVM* para clasificar geográficamente textos: los modelos de lenguajes probabilísticos (Ponte and Croft, 1998). Dicha aproximación se ha llevado a cabo con el objetivo de compararla con la de *SVM* detallada en la sección previa, ya que, como ya se comentó en el capítulo 3, es una de las aproximaciones que mejores resultados ha dado en tareas de clasificación de textos, especialmente cuando se trata de captar las sutilezas del lenguaje y no sólo un conjunto de topónimos que pueda identificar a la localidad en cuestión.

Para la creación de los modelos de lenguaje se han utilizado las herramientas incluidas en el proyecto *Lemur*[2], dentro del cual se han desarrollado motores de búsqueda, barras de herramientas para navegadores, herramientas de análisis de texto y recursos para la *IR* y la minería de textos.

El conjunto de herramientas de Lemur está diseñado para facilitar la investigación en modelado de lenguaje e *IR*, donde la *IR* es ampliamente interpretada para incluir tecnologías tales como recuperación distribuida con consultas estructuradas a la medida, *IR* entre distintos lenguajes, resumenes, filtrado y categorización. La arquitectura subyacente del sistema fue construida para soportar las tecnologías anteriormente descritas. Este conjunto de herramientas ha sido diseñado con el fin de permitir programar fácilmente sus propias personalizaciones y aplicaciones.

El motor de *IR* utilizado en el proyecto *Lemur* es conocido como *Indri*, el cual hace un cálculo de la relevancia basado en modelos de lenguaje. *Indri* es un motor de búsqueda que proporciona técnicas de búsqueda de texto y un lenguaje de consulta estructurado para colecciones de textos de hasta 50 millones de documentos (para una sola máquina) o 500 millones de documentos (en búsquedas distribuidas).

Indri está basado en una combinación de modelos de lenguaje y una red de inferencia (Zhai and Lafferty, 2004), creando una clasificación final utilizando para ello *KL-divergence* (Kinsella et al., 2011).

Concretamente, en los experimentos llevados a cabo en esta tesis con modelos de lenguaje, por cada localización existente en el sistema (ver tabla 4.1) se ha estimado la distribución de términos asociados con cada una de estas localidades. Una vez hecho esto, se estima la probabilidad de que el texto de un determinado documento haya podido ser creado por el modelo de lenguaje de una determinada ciudad.

Debido a que existe la posibilidad de que surjan términos en la noticia que se pretende clasificar que no estén en el modelo de lenguaje de alguna de las localidades del corpus, se realiza un suavizado mediante *Dirichlet*

[2]http://www.lemurproject.org/

smoothing (Zhai and Lafferty, 2004), el cual otorga una probabilidad muy baja a los términos que no han aparecido hasta ahora con el fin de poder realizar el calculo correctamente.

Por último, como se ha comentado previamente, se realiza una clasificación utilizando *KL-divergence* (Kinsella et al., 2011) según la probabilidad que haya devuelto cada uno de los modelos de lenguaje de haber generado el texto de la noticia.

5.2. Herramientas lingüísticas

En esta sección se van a describir las herramientas lingüísticas que se han empleado en los experimentos descritos en este capítulo.

5.2.1. *FreeLing*

Para obtener la categoría gramatical de los términos, o conjunto de términos, se utilizó el etiquetador gramatical incluido en el analizador lingüístico de código abierto *FreeLing* (Padró et al., 2010). Esta herramienta permite obtener las distintas categorías y subcategorías gramaticales de los términos, o conjunto de términos, que aparecen en el texto.

El analizador morfológico para el castellano utiliza un conjunto de etiquetas para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo *EAGLES*[3] para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Así pues, está pensado para recoger los accidentes gramaticales existentes en las lenguas europeas. Es por esto que dependiendo de la lengua hay atributos que pueden no especificarse. Si un atributo no se especifica significa que, o bien expresa un tipo de información que no existe en la lengua, o bien la información no se considera relevante. La infraespecificación de un atributo, como se verá a continuación, se marca con el caracter '0'.

Cada uno de los términos (o conjunto de términos) detectados por *FreeLing* serán etiquetado dentro de alguna de las categorías mostradas en la tabla 5.1. Tal y como se puede apreciar en dicha tabla, las etiquetas que el analizador morfológico utiliza para el castellano son: adjetivos, adverbios, artículos, determinantes, sustantivos, verbos, pronombres, conjunciones, numerales, interjecciones, abreviaturas, preposiciones y signos de puntuación.

De las 13 categorías gramaticales mencionadas, los experimentos llevados a cabo en este capítulo se han centrado en dos de ellas, adjetivos y sustantivos, ya que son las que mejor capturan la semántica del texto a la hora de detectar el foco geográfico en los documentos.

[3] http://www.ilc.cnr.it/EAGLES96/home.html

Tabla 5.1: *FreeLing*. Categorías gramaticales.

Categoría	Código	Ejemplo
Adjetivos	A	mallorquín
Adverbios	R	despacio
Artículos	T	el
Determinantes	D	aquella
Sustantivos	N	Benidorm
Verbos	V	viajar
Pronombres	P	nosotros
Conjunciones	C	pero
Numerales	M	ocho
Interjecciones	I	ah
Abreviaturas	Y	etc
Preposiciones	S	ante
Signos de puntuación	F	.

Dentro de la categoría de sustantivos existe un conjunto de subcategorías tal y como se puede apreciar en la tabla 5.2. De entre todas estas subcategorías, con el fin de poder obtener los topónimos existentes en el corpus, los experimentos se centrarán en la "*Clasificación semántica o NER*", y dentro de ésta, en la subsubcategoría de *Lugar*, indicada en las posiciones 5 y 6 con los caracteres '***G0***'.

La categoría de adjetivos (y el resto de categorías) tiene una tabla análoga, pero en esta ocasión no se va a discernir entre un tipo de adjetivo y otro, por lo que únicamente se tendrá en cuenta que el código devuelto por *FreeLing* para un término dado (o conjunto de términos) sea el que identifica a los adjetivos (los código que comienzan por 'A').

De este modo, si por ejemplo *FreeLing* devuelve el siguiente código: "*NP00G00*", significa que, según lo expuesto en la tabla 5.2, se trataría de un nombre propio de lugar.

La tabla 5.2, la cual representa las etiquetas "*sustantivos*", así como el resto de tablas de cada una de las otras doce categorías gramaticales, está compuesta de las siguientes columnas:

1. Número que hace referencia al orden y posición en que aparecen los atributos.

2. Se hace referencia a los atributos, el número de los cuales varía dependiendo de la categoría.

3. Se encuentran los valores que puede tomar cada atributo.

4. Muestra los códigos que se han establecido para su representación.

Tabla 5.2: *FreeLing*. Subcategorías gramaticales de los sustantivos.

SUSTANTIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	**Nombre**	**N**
2	Tipo	Común o neutro	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica o *NER*	Persona	SP
		Lugar	**G0**
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Las etiquetas en sí sólo son los códigos (columna 4) y se sabe a qué atributo pertenecen por la posición (columna 1) en la que se encuentran.

Así pues, si tenemos el siguiente extracto de noticia:

"*PP de Ceuta pone la ciudad como 'ejemplo de convivencia de todas las culturas' y rechaza alianzas de 'unos contra otros'.*

El portavoz del Partido Popular de Ceuta, Francisco Márquez, ha puesto a la ciudad autónoma como 'ejemplo de convivencia entre todas las culturas' y rechazó que se planteen alianzas de 'unos contra otros'."

El resultado tras pasar por el analizador gramatical de *FreeLing* sería el siguiente:

```
PP pp NP00000 1
de de SPS00 0.999919
Ceuta ceuta NP00G00 1
pone poner VMIP3S0 1
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
como como CS 0.998668
‘ ‘ Fe 1
ejemplo ejemplo NCMS000 1
```

```
de de SPS00 0.999919
convivencia convivencia NCFS000 1
de de SPS00 0.999919
todas todo DI0FP0 0.945312
las el DA0FP0 0.97051
culturas cultura NCFP000 1
'' Fe 1
y y CC 0.999812
rechaza rechazar VMIP3S0 0.993015
alianzas alianza NCFP000 1
de de SPS00 0.999919
‘ ‘ Fe 1
unos uno DI0MP0 0.879032
contra contra SPS00 0.976316
otros otro DI0MP0 0.56875
'' Fe 1

El el DA0MS0 1
portavoz portavoz NCCS000 1
de de SPS00 1
el el DA0MS0 1
Partido_Popular_de_Ceuta partido_popular_de_ceuta NP00O00 1
, , Fc 1
Francisco_Márquez francisco_márquez NP00SP0 1
, , Fc 1
ha haber VAIP3S0 0.998141
puesto poner VMP00SM 0.547619
a a SPS00 0.99585
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
autónoma autónomo AQ0FS0 1
como como CS 0.998668
‘ ‘ Fe 1
ejemplo ejemplo NCMS000 1
de de SPS00 0.999919
convivencia convivencia NCFS000 1
entre entre SPS00 0.995223
todas todo DI0FP0 0.945312
las el DA0FP0 0.97051
culturas cultura NCFP000 1
'' Fe 1
y y CC 0.999812
rechazó rechazar VMIS3S0 1
que que CS 0.4375
```

```
se se P0000000 0.465602
planteen plantear VMSP3P0 0.860063
alianzas alianza NCFP000 1
de de SPS00 0.999919
‘ ‘ Fe 1
unos uno DI0MP0 0.879032
contra contra SPS00 0.976316
otros otro DI0MP0 0.56875
’’ Fe 1
. . Fp 1
```

Del resultado devuelto por *FreeLing* se pueden apreciar 4 campos distintos para cada término:

1. *Forma*: el término analizado. Es el término tal cual aparece en el texto que se ha analizado, con la excepción de que cuando *FreeLing* detecta que se puede tratar de un nombre compuesto, trata cada término del nombre compuesto como un único término separándolo por ‘_’ (por ejemplo: Francisco_Márquez).

2. *Lema*: muestra el lema de los términos analizados. El lema es la forma en que aparece un término en la entrada de un diccionario: pasado a minúsculas, eliminando el género y número cuando corresponda o, si se trata de un verbo, devolviendo su forma en infinitivo. Cabe tener en cuenta que el corpus ha sido previamente limpiado que caracteres “extraños” (p. ej.: *#, $, !, (, ...*) y todos los términos pasados a minúsculas a excepción de la primera letra de cada uno de ellos, tal y como se comentó en la sección 4.1.

3. *Etiqueta*: la etiqueta gramatical del término. Indica el tipo de término que *FreeLing* ha identificado, tal y como se mostró en la tabla 5.1.

4. *Confianza*: muestra la seguridad que tiene *FreeLing* en su predicción de la categoría gramatical, estando dicha confianza en el rango comprendido entre 0 y 1, siendo 0 el nivel de confianza más bajo y 1 el más alto.

5.2.2. SRI Language Modeling Toolkit

SRI Language Model Toolkit[4] es un conjunto de herramientas para la construcción y aplicación de modelos estadísticos del lenguaje (*LMs*) para su uso en el reconocimiento de voz, el etiquetado y la segmentación estadística, y la traducción automática.

[4] http://www.speech.sri.com/projects/srilm/

En esta tesis, se utilizará para obtener la frecuencia de los términos que aparecen en cada uno de los documentos de los corpus, tal y como se puede observar en el siguiente ejemplo, donde se analiza el siguiente texto:
"*La Universidad de Alicante crea un analizador de textos el cual los clasifica geográficamente según su contenido.*"

De este texto se obtendría el siguiente resultado:

```
Alicante 1
La 1
Universidad 1
analizador 1
clasifica 1
contenido 1
crea 1
cual 1
de 2
el 1
geográficamente 1
los 1
según 1
su 1
textos 1
un 1
```

donde se puede apreciar el número de veces que aparece cada término en el documento.

En lugar de ser utilizado con los textos sin procesar de los corpus, *SRI Language Model Toolkit* se utilizará con los textos una vez hayan sido procesado con *FreeLing*, tal y como se explico en la subsección anterior.

5.3. Identificación del foco geográfico en textos formales

Tal y como se ha comentado en la introducción de este capítulo, los textos formales que se han utilizado en la experimentación han sido noticias georeferenciadas del diario *20Minutos*. Se ha utilizado este corpus de noticias frente a otros debido a que en dicho medio de información se puede realizar un filtrado de las noticias por la ciudad a la que pertenezcan, ya que dicho diario tiene clasificadas en su web las noticias según el origen geográfico de las mismas (foco geográfico), lo cual facilita las labores de entrenamiento y evaluación.

Estos textos, tal y como se explicó en la sección 4.1, estaban asociados a una ciudad concreta del territorio español. De entre todas las ciudades de las que se podían obtener noticias, se filtró para obtener finalmente las 50 capitales de provincia de España más sus dos ciudades autónomas, lo

que conforman un total de 52 ciudades que representan cada una de sus provincias, o territorios en el caso de Ceuta y Melilla. La elección de las capitales de provincia en lugar de cualquier otra ciudad viene determinada porque suelen ser las poblaciones que mejor representan a su provincia debido a que los gobiernos de las mismas se suelen alojar en ellas, y por ende producen un gran número de noticias, no sólo de la propia ciudad, sino que también del resto de la provincia. Además, dichas ciudades, salvo en contadas excepciones, son la ciudad más poblada de la provincia.

Las fechas de las noticias de dicho corpus van desde enero de 2008 hasta diciembre de 2011, ambos inclusive, es decir, 4 años de los que se han recogido todas las noticias de las citadas ciudades, pudiéndose ver las cifras de dicho corpus en la tabla 4.1 expuesta en el capítulo anterior.

Se han realizado distintos experimentos para ver lo que aportan los diversos términos del corpus dependiendo de su función gramatical (*PoS: Part Of Speech*). Para la obtención de la categoría gramatical de los términos se ha utilizado el etiquetador gramatical incluido en el analizador lingüístico de código abierto *FreeLing*, tal y como se explicó en la sección 5.2.1.

Los experimentos llevados a cabo en esta tesis se han centrado, como se ha comentado previamente, en las categorías gramaticales de los adjetivos y en los sustantivos, donde se ha utilizado el etiquetado devuelto por la función de *NER* de *FreeLing* para saber si dichos sustantivos identificaban un lugar.

Atendiendo a estas categorías gramaticales se han creado los siguientes corpus del diario *20Minutos*, para ver cómo se comporta el sistema dependiendo del tipo de categoría gramatical que se utiliza:

- *Topónimos*. Se guardaron únicamente en cada documento los términos detectados como topónimos por *FreeLing* por cada uno de los artículos que componían el corpus original. Es decir, los términos que tenían una '*N*' en la primera posición de la categoría gramatical identificada por *FreeLing*, y un '*G0*' en su quinta y sexta posiciones respectivamente. Si seguimos el ejemplo del estracto de texto expuesto en la sección 5.2.1, el nuevo fichero de texto estaría compuesto únicamente por el término:

  Ceuta

- *Sustantivos*. De manera análoga a la anterior, se guardó en cada documento del corpus no sólo los topónimos, sino que también cualquier sustantivo. Es decir, los términos que tenían una '*N*' en la primera posición de la categoría gramatical identificada por *FreeLing*. Si seguimos con el ejemplo expuesto en la sección 5.2.1, el nuevo fichero de texto estaría compuesto por los siguientes términos:

PP Ceuta ciudad ejemplo convivencia culturas alianzas portavoz Partido_Popular_de_Ceuta Francisco_Márquez ciudad ejemplo convivencia culturas alianzas

- *Sustantivos y adjetivos.* Además de los sustantivos, esta vez se añadieron también los adjetivos con el fin de ver la posible aportación de los mismos en la tarea de detectar el foco geográfico de los documentos. Es decir, los términos que tenían una '*N*' o una '*A*' en la primera posición de la categoría gramatical identificada por *FreeLing*. Si seguimos con el ejemplo expuesto en la sección 5.2.1, el nuevo fichero de texto estaría compuesto por los siguientes términos:

  PP Ceuta ciudad ejemplo convivencia culturas alianzas portavoz Partido_Popular_de_Ceuta Francisco_Márquez ciudad autónoma ejemplo convivencia culturas alianzas

- *Todo menos los topónimos.* Con el fin de poder discernir la aportación que el resto de términos que no son topónimos dan a la tarea de clasificación geográfica de textos, se creó un corpus sin topónimos para poder contrastarlo con un corpus donde sí se hayan incluido. Siguiendo con el texto de ejemplo mostrado en la sección 5.2.1, el nuevo texto obtenido sería el siguiente:

  PP de pone la ciudad como ' ejemplo de convivencia de todas las culturas ' y rechaza alianzas de ' unos contra otros '. El portavoz del Partido_Popular_de_Ceuta , Francisco_Márquez , ha puesto a la ciudad autónoma como ' ejemplo de convivencia entre todas las culturas ' y rechazó que se planteen alianzas de ' unos contra otros '.

- *Todo.* En esta ocasión no se utilizó *FreeLing*, ya que no se hizo ningún filtrado y se utilizaron todos los términos existentes en cada uno de los artículos del corpus. Es decir, los términos tal cual aparecían en la noticia original.

Puesto que la categoría gramatical detectada en cada término por *FreeLing* no es 100 % fiable, se han analizado manualmente 10 noticias del diario *20Minutos* para comprobar la fiabilidad de dicha herramienta a la hora de detectar topónimos. Los textos analizados se pueden ver en el anexo A.

Los resultados del análisis se muestran a continuación, pero previamente se muestran los criterios seguidos en dicho análisis:

1. Se considera *topónimo* todo aquel nombre de lugar (países, regiones, municipios, direcciones, etc.), cuando dicho nombre está haciendo referencia a una ubicación. Es decir, si un nombre de lugar coincide, por ejemplo, con el nombre de una persona, y el texto está haciendo referencia a la persona, entonces no se considerará como topónimo.

2. Si el nombre de una empresa, diario, etc., contiene el nombre de un lugar, este no será considerado como topónimo, ya que podría existir, por ejemplo, una "*Cafetería Viena*" sin estar en la capital austriaca.

3. Los nombres de empresas, pese a que tengan una ubicación física, no serán considerados topónimos.

4. Cuando un nombre de lugar venga a continuación de un edificio, comercio, etc., será considerado topónimo. Por ejemplo, si aparece el texto "*Teatro Principal de Alicante*", "*Alicante*" será considerado topónimo.

5. Los términos equivalentes que representen a un topónimo, por ejemplo, "*la capital cántabra*" para hacer referencia a "*Santander*", no serán considerados topónimos.

6. Las abreviaciones comunes de los topónimos, por ejemplo, "*El Estrecho*" para hacer referencia a "*El Estrecho de Gibraltar*", serán consideradas topónimos.

De los diez textos analizados manualmente se observa:

- *FreeLing* ha detectado 43 de los 72 topónimos, cometiendo frecuentemente fallos cuando tenía que detectar topónimos en conjuntos de términos como por ejemplo "*Teatro Principal de Alicante*". Esto da una cobertura del **60 %** de todos los topónimos existentes.

- Se han detectado 6 falsos positivos, siendo varios de ellos el mismo, "*Bioparc*", que es un zoológico ubicado en la ciudad de Valencia (entre otras). Tal y como se comentó en el punto 3 de la lista previa, los nombres de empresa no se consideran topónimos, más aún en el caso que nos atañe, puesto que existe más de un "*Bioparc*" ubicado en distintas localidades. Esto da una precisión a nivel de topónimo del **90 %**.

Hay que tener en cuenta los artículos que no contienen ningún topónimo, ya que entonces, no es posible clasificarlos con los algoritmos de aprendizaje automáticos expuestos en la sección anterior, por lo que son considerados como documentos mal clasificados. En el corpus hay por cada año:

- **Año 2008**, 50.170 artículos de los cuales 7.661 no se les ha detectado ningún topónimo con *FreeLing*.

- **Año 2009**, 27.400 artículos de los cuales 2.966 no se les ha detectado ningún topónimo con *FreeLing*.

- **Año 2010**, 211.610 artículos de los cuales 21.260 no se les ha detectado ningún topónimo con *FreeLing*.

- **Año 2011**, 229.410 artículos de los cuales 25.209 no se les ha detectado ningún topónimo con *FreeLing*.

Para los 5 corpus anteriores, tras el procesado de *FreeLing*, mediante la herramienta *SRI Language Model Toolkit* (*SRILM*) se obtuvo la frecuencia de cada uno de los términos que aparecían en cada uno de los artículos, tal y como se comentó en la sección 5.2.2.

Estas frecuencias se guardaron en ficheros de texto, donde cada uno de estos ficheros que almacenaban las frecuencias de los términos representaba una única y completa noticia del corpus original. Es decir, existía un fichero de frecuencia por cada noticia de cada uno de los cinco corpus (topónimos, sustantivos, sustantivos y adjetivos, sin topónimos y todos los términos) extraídos del original.

Una vez guardadas todas las frecuencias de los términos de cada artículo, dentro de cada uno de los distintos corpus, se pueden apreciar en la figura 5.1 los distintos años en los que están divididos cada corpus, los cuales abarcan desde el año 2008 hasta el año 2011, ambos inclusive. Esta separación temporal viene dada para poder estudiar con más detenimiento la evolución que tienen los terminos a lo largo de cada uno de los años indicados.

Figura 5.1: Estructura de directorios de los ficheros de frecuencia del corpus del periódico *20Minutos*.

En la figura 5.1 también se puede apreciar otro nodo (*2008-2011*) que representa todos los artículos sin distinción del año de publicación. Éste se ha hecho así para poder ver cómo funcionan los experimentos cuando se trabaja con un mayor rango temporal y número de ficheros para entrenar.

En dicha figura se ha representado la arquitectura que subyace bajo el corpus del diario *20Minutos* que utiliza todos los términos (*Todo*) en el año 2008, y por motivos de espacio y legibilidad se ha omitido la del resto

Tabla 5.3: Número de términos distintos incluyendo hápax legómenon del corpus del diario *20Minutos* según el año y su categoría gramatical.

PoS	2008	2009	2010	2011	2008-2011
Topónimos	28.831	18.139	85.128	84.766	172.283
Sustantivos	310.592	207.626	1.030.798	1.074.674	2.129.031
Adjetivos y sustantivos	326.645	221.048	1.060.241	1.104.299	2.174.606
Sin Topónimos	391.964	278.397	1.198.085	1.250.089	2.393.729
Todo	412.054	290.840	1.252.922	1.304.916	2.506.168

Tabla 5.4: Número de términos distintos eliminando hápax legómenon del corpus del diario *20Minutos* según el año y su categoría gramatical.

PoS	2008	2009	2010	2011	2008-2011
Topónimos	7.643	5.202	32.398	34.963	62.480
Sustantivos	88.626	64.237	384.463	434.519	766.905
Adjetivos y sustantivos	98.709	72.660	403.277	454.038	794.559
Sin Topónimos	139.897	106.384	488.429	548.136	932.827
Todo	145.853	110.540	510.580	571.874	975.067

de corpus (*Todo menos los topónimos*, *Nombres + Adjetivos*, *Nombres* y *Topónimos*) y años, aunque éstos mantienen una estructura idéntica, con la única excepción del contenido de los ficheros, los cuales están representados en el árbol de la figura 5.1 por sus nodos hoja.

Los términos que sólo aparecían una única vez en todo el corpus (*hápax legómenon*) fueron eliminados, dado que su aportación a la hora de clasificar textos iba a ser nula y permitiría agilizar la clasificación ocupando un menor tiempo y espacio. De este modo, se ha conseguido pasar del número de términos expuesto en la tabla 5.3, donde aún no se había eliminado los hápax legómenon, al expuesto en la tabla 5.4, donde ya se han eliminado los hápax legómenon.

Nótese que debido a que hay términos que en ocasiones son reconocidos como topónimos, y en otras ocasiones no, bien sea por un error en el etiquetado de *FreeLing*, bien porque en unas ocasiones el mismo término puede actuar como topónimo y en otras no, la suma del número de términos distintos existente en los corpus "*sin topónimos*" y "*topónimos*" es ligeramente superior a la suma total de todos los términos "*todo*".

Con el fin de poder realizar posteriormente una validación cruzada[5], se procedió a dividir cada uno de los corpus, los cuales ya estaban divididos en distintos años, tal y como se ha explicado en el párrafo anterior, en 10 partes iguales. Para ello se obtuvo aleatoriamente por cada una de las ciudades del corpus, una décima parte de sus artículos, es decir, se realizó una división por año, categoría gramatical, partición y ciudad, tal y como muestra la figura 5.1. Entre estas 10 particiones había la misma proporción de artículos en las localidades que la componían (*stratified cross-validation*).

En la figura 5.1 también se pueden apreciar las 10 particiones que se han realizado para posteriormente realizar la validación cruzada. Dichas particiones vienen indicadas por los recuadros *P1*, *P2*, ... y *P10*, representado el recuadro '...' las particiones que van desde la *P3* hasta la *P9*, que al igual que sucedia con el resto de años explicados en los párrafos anteriores, se han omitio por motivos de espacio y son idénticos a los de la partición *P1* mostrada en la imagen. En esta ocasión, los ficheros que hay en las hojas del árbol, los cuales representan los ficheros de frecuencia de cada noticia del corpus, son distintos para cada una de las particiones existentes.

En el siguiente nivel del árbol se pueden apreciar las distintas ciudades en las que está dividido el corpus y que ya fueron mencionadas en la tabla 4.1. Bajo cada una de estas ciudades se encuentra un número distinto de artículos para cada una de las ciudades, pudiendo estos artículos coincidir en nombre en distintas ciudades, aunque no en contenido.

En las hojas del árbol, tal y como se ha indicado anteriormente, se encuentran los ficheros de frecuencia de los términos de los artículos que representan las noticias originales. Los ficheros están nombrados con la fecha del día en la que se publicó la noticia que representan, siendo el último número de dicho fichero el ordinal numérico de la noticia para una ciudad y día dado.

A continuación se mostrarán los experimentos realizados según la naturaleza del corpus de entrenamiento utilizado, es decir, si este es el mismo que el utilizado para la evaluación, si pertenece a otra fuente pero sigue siendo de tipo formal, si pertenece a otra fuente y es de tipo informal o es un combinación de cualquiera de los tres tipos anteriores. Para todos estos experimentos, el conjunto de textos a evaluar será el de los corpus de noticias del diario *20Minutos*, el cual representa la fuente de textos formales que se quieren clasificar geográficamente.

[5]La validación cruzada o *cross-validation* es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica (Devijver y Kittler, 1982)

5.3.1. Entrenamiento con noticias del diario *20Minutos*

En esta sección se llevarán a cabo una serie de experimentos en los que se entrenará y evaluará el sistema única y exclusivamente con el corpus del diario *20Minutos*.

Se van a mostrar una serie de experimentos iniciales que ayudarán a determinar el camino a seguir para el resto de experimentos. Lo que se pretende conocer es la aproximación, entre *SVM* y modelos de lenguaje, que obtiene unos mejores resultados. Del mismo modo, dentro de la aproximación en la que se emplea *SVM*, se ha realizado una selección de características con el objetivo de conocer si era posible el reducir el coste espacial y temporal del algoritmo sin sufrir un decremento en el rendimiento del sistema.

SVM

Como se ha mostrado previamente, sobre el corpus inicial de noticias del diario *20Minutos* se han creado 5 corpus alternativos: topónimos, sustantivos, sustantivos y adjetivos, todos los términos de cada artículo menos los topónimos y todos los términos de cada artículo. El procedimiento y estructura empleados para la creación de cada uno de estos 5 corpus es idéntico (ver la sección 5.3 y la figura 5.1), diferenciándose únicamente en el contenido de los ficheros existentes en dichos corpus. Es decir, el número de ficheros que compone cada corpus es el mismo, pero el contenido de estos es distinto, ya que un corpus contendrá la frecuencia de cada uno de los términos que aparecen en el artículo en cuestión, otro solamente la frecuencia los términos identificados por *FreeLing* como topónimos, etc.

Así pues, por cada uno de los distintos grupos gramaticales y periodos de tiempos se realizó una validación cruzada evaluando con una partición de las 10 existentes y entrenando con las 9 restantes.

Para ello se creó un vector de características para cada una de las noticias del corpus de las 9 particiones de entrenamiento, mediante los cuales se entrenó *LibLINEAR* para construir el modelo, haciendo lo propio con la partición restante de evaluación.

Una vez que se han creado los ficheros de entrenamiento y de evaluación se crea un modelo y se evalúa el corpus de prueba creado, tal y como se explicó en la sección 5.1.1.

Con los resultados obtenidos en cada una de las 10 particiones se calcula la media aritmética para obtener el resultado final de dicho experimento, finalizando así la validación cruzada.

Los resultados obtenidos se muestran en la tabla 5.5, donde las columnas indican los años y las filas la categoría gramatical. En negrita se resaltan los mejores resultados para cada uno de los años.

El resultado mostrado refleja el tanto por cien de acierto a la hora de clasificar cada texto en su clase correspondiente, es decir, la ciudad de origen

de la noticia. Así pues, esta precisión fue medida según la formula expresada en la ecuación 5.2:

$$Precisión = \frac{número\ de\ textos\ clasificados\ correctamente}{número\ de\ textos\ totales} \tag{5.2}$$

Tabla 5.5: Resultados de la validación cruzada de *SVM* entrenando y probando con el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	58,20	60,03	63,36	62,69	63,33
Sustantivos	68,66	69,93	81,17	80,89	82,14
Adjetivos + Sustantivos	71,58	73,68	82,09	81,44	82,70
Sin topónimos	62,56	63,78	79,45	78,75	79,64
Todo	**73,55**	**74,82**	**84,12**	**83,33**	**84,35**

En los resultados obtenidos con los experimentos de *SVM* entrenando y evaluando con el corpus de noticias del diario *20Minutos*, tal y como muestra la tabla 5.5, se obtiene un funcionamiento mejor cuando se utiliza el corpus completo, es decir, cuando no se filtra por el tipo gramatical (*PoS*) del término del texto. Esta mejora es estadísticamente significativa ($p < 0,01$). Sin embargo, cabe destacar que para dichos resultados es necesario un mayor coste espacial y temporal, ya que el número de términos es mucho mayor en este caso, tal y como se puedo apreciar en la tabla 5.4.

Si centramos la atención en la utilización o no de topónimos, podemos observar cómo estos términos son de una mayor utilidad tanto en cuanto el corpus de entrenamiento es menor. En cuanto el corpus de entrenamiento tiene un volumen mayor, se pueden extraer rasgos geográficos del resto de términos, superando claramente al corpus que únicamente utiliza topónimos.

Por otro lado, si se realiza un filtrado para utilizar únicamente los términos que son adjetivos y sustantivos, ya sean estos últimos topónimos o no, el corpus se reduce considerablemente, lo que conlleva una reducción en el tiempo de procesamiento y los recursos necesarios para ello, sin que la precisión del sistema se vea muy afectada. Al trabajar con corpus de entrenamiento más extensos, incluso la eliminación de los adjetivos no conlleva una gran pérdida en la precisión del sistema y sí una ganancia en el tiempo de procesamiento y espacio requerido. Tal y como se observa en la tabla 5.5, los adjetivos aportan una mejora entre un 3 y un 4 por ciento en términos absolutos para los corpus más reducidos, y menos de un 1 % cuando se trabaja con corpus más amplios.

Hay que tener en cuenta que el número de topónimos existente en el corpus es escaso, tal y como se mostró en el análisis manual del extracto de

noticia que se hizo de los topónimos detectados por *FreeLing* en la sección 5.2.1, donde en todo el texto mostrado, tan sólo aparecía un topónimo. De media, según se puede ver en el anexo A, aparecen algo menos de 5 topónimos por documento, lo que propicia que sea muy difícil el poder detectar el foco geográfico de un texto con un número tan reducido de características. Además, hay que tener presente que muchos de los textos no contienen ningún topónimo. Concretamente, 57.096 documentos de los 518.590 de los que está compuesto el corpus de los 4 años, es decir, más de un 11 % de los documentos directamente no se pueden clasificar geográficamente por este criterio.

Pese a la gran importancia que tienen los topónimos a la hora de identificar el foco geográfico de los textos, obteniendo entre un 58 y un 63 por ciento de precisión con la utilización de únicamente estos términos, según los resultados obtenidos se pone de manifiesto la relevancia del resto de términos presentes en los documentos, ya que éstos son capaces de obtener una precisión mucho más alta que la lograda con la utilización única de los nombres de lugar. Incluso el corpus que no tiene topónimos logra unas resultados mejores que el que hace uso exclusivo de estos términos. Este hecho indica la relevancia de la información general del mundo a la hora de llevar a cabo esta tarea.

A la luz de los experimentos llevados a cabo en esta sección, se puede concluir que, si no existen restricciones temporales ni espaciales, la mejor aproximación es la resultante de la utilización del texto completo de las noticias que se quieren evaluar, mientras que si el tiempo o el espacio se deben tener en cuenta, una aproximación en la que se tenga en cuenta únicamente los sustantivos de los textos es la más acertada, siendo la aproximación implementada únicamente con topónimos una aproximación muy básica que dista mucho de la precisión obtenida por el resto de aproximaciones, poniéndose así de manifiesto la gran importancia de la información general del mundo que aportan todos los términos.

Modelos de lenguaje

Tal y como se detalló en la sección 5.1.2, se ha realizado un experimento en el que se pretende observar el comportamiento de un sistema capaz de clasificar geográficamente los artículos del diario *20Minutos* mediante la utilización de modelos de lenguaje, con el fin de compararlo con la aproximación que utiliza *SVM* y poder ver cuál es más eficaz a la hora de llevar a cabo esta tarea con un corpus de lenguaje formal.

Una vez más, al igual que se hiciera con en el primer experimento con *SVM*, se han creado 5 particiones gramaticales: todos los términos, todos los términos menos los topónimos, sustantivos y adjetivos, sustantivos, topónimos.

La configuración empleada en este experimento es análoga a la del experimento con *SVM*, la cual queda resumida en la figura 5.1, donde se muestra como el corpus está dividido según los distintos *PoS*, periodos temporales, particiones para realizar la pertinente validación cruzada, y ciudades del corpus. En esta ocasión todos los textos de entrenamiento fueron agrupados en un único fichero según la ciudad a la que pertenecían, a partir del cual se crearía el pertinente modelo de lenguaje de dicha ciudad.

En esta ocasión, la métrica utilizada para medir el sistema ha sido la *media de la precisión promedio* (*MAP: Mean Average Precision*), la cual se calcula de la siguiente manera:

$$MAP = \frac{\sum_{q=1}^{Q} \cdot PreP(q)}{Q} \tag{5.3}$$

donde Q es el número total de consultas lanzadas al sistema, q es una consulta concreta y $PreP(q)$ es la precisión promedio de la consulta q, calculada de la siguiente forma:

$$PreP(q) = \frac{s}{Posición\ en\ la\ que\ se\ ha\ devuelto} \tag{5.4}$$

donde s será 1 si es la localidad correcta y 0 en cualquier otro caso. En este caso, dado que las noticias del corpus pertenecen exclusivamente a una única ciudad:

$$PreP(q) = \frac{1}{Posición\ de\ la\ respuesta\ correcta} \tag{5.5}$$

Concretamente, se ha recogiendo única y exclusivamente la respuesta mejor valorada por el sistema y obteniendo la *MAP*, de tal manera que la utilización de *MAP* es equivalente a la medida de evaluación realizada en los experimentos mediante *SVM*, es decir, por cada texto de consulta enviado al sistema se ha considerado una única respuesta correcta, dividiendo el número total de aciertos entre el número total de consultas realizadas para obtener los resultados finales.

Así pues, los resultados obtenidos con estos experimentos se pueden observar en la tabla 5.6, donde las columnas indican los años y las filas la categoría gramatical que se utilizó para construir los modelos de lenguaje y las consultas.

En los resultados obtenidos con los experimentos llevados a cabo con modelos de lenguaje en esta sección, entrenando y evaluando con el corpus de noticias del diario *20Minutos*, tal y como se puede apreciar en la tabla 5.6, la utilización de todas las características (términos) del corpus no hace que el sistema clasifique geográficamente mejor los textos, ya que se introduce mucho más ruido, lo cual hace que el sistema vaya más lento y ocupe un mayor espacio.

Tabla 5.6: Resultados donde se muestra la precisión después de realizar la validación cruzada utilizando modelos de lenguaje con el corpus del diario *20Minutos*. Se resalta en negrita los mejores resultados para cada año.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	59,41	61,48	61,23	60,47	61,32
Sustantivos	67,96	67,89	77,25	76,79	75,68
Sustantivos + Adjetivos	**68,06**	**68,35**	**77,31**	**76,82**	**75,70**
Sin topónimos	58,50	59,71	70,89	70,47	67,94
Todo	64,73	65,18	74,93	74,59	72,76

Cuanto mayor es el corpus de entrenamiento, se puede apreciar una mayor diferencia entre el entrenamiento realizado utilizando únicamente topónimos y el resto de corpus en favor de estos últimos.

Observando la aproximación en la que solamente se utilizaban topónimos y la que se excluían éstos del entrenamiento, se aprecia como para corpus pequeños (años 2008 y 2009) obtienen unos resultados muy similares, mientras que cuanto mayor es el corpus de entrenamiento, más se acentúa la diferencia en favor del corpus que no tiene topónimos, lo cual apoya la hipótesis inicial sobre la utilidad de los matices lingüísticos para favorecer la detección del foco geográfico en textos.

Las aproximaciones en las que se utilizan solamente los sustantivos, así como la que además incluye los adjetivos, son las que mejores resultados obtienen, aportando estos últimos una mejora marginal. Esto da a entender que son principalmente los sustantivos, no solamente los topónimos, los que aportan mucho conocimiento a la hora de realizar esta tarea.

Comparando los resultados obtenidos utilizando *SVM* (tabla 5.5) y los de los modelos de lenguajes (tabla 5.6), se puede observar como los resultados son siempre mejores utilizando la primera aproximación, aunque cabe tener en cuenta que la aproximación de *SVM* tiene un mayor coste temporal a la hora de entrenar el modelo, especialmente cuando se utiliza un mayor número de características, como es el caso que obtiene los mejores resultados en *SVM*. Esto es debido a que *SVM* funciona bien con espacios de alta dimensionalidad. Esta comparación se puede apreciar de una manera más directa en la figura 5.2.

Así pues, viendo los resultados obtenidos hasta ahora, se puede concluir que salvo en casos concretos y cuando existan restricciones temporales, la mejor aproximación es la de *SVM* utilizando todas sus características, es decir, en la que se utiliza toda la información presente en los textos y no meramente la geográfica, tal y como se planteo en la hipótesis de partida. Si, por el contrario, hay que decantarse por una aproximación que utilice menos recursos espaciales, la realizada con *SVM* utilizando únicamente los

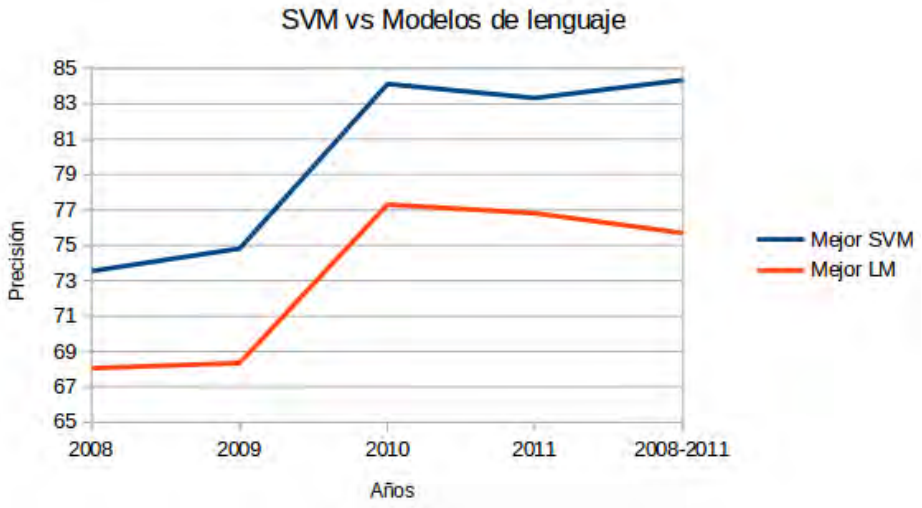


Figura 5.2: Precisión obtenida con la mejor aproximación de *SVM* y la mejor aproximación de modelos de lenguaje.

sustantivos parece ser la más adecuada cuando el corpus tiene un mayor tamaño, obteniendo unos resultados ligeramente superiores a los modelos de lenguaje cuando dicho corpus es más reducido. Si el coste temporal resulta crítico, la aproximación más acertada sería la de modelos de lenguaje utilizando sustantivos.

Selección de características

Tal y como se ha comentado en la introducción de esta sección, se ha llevado a cabo un experimento en el cual se ha procedido a reducir el número de características que se utilizan para la clasificación de los textos geográficamente mediante el algoritmo de *SVM* expuesto al comienzo de este capítulo, ya que *SVM* es el que mejores resultados ha dado en los experimentos realizados (ver figura 5.2).

En la tabla 5.3 se puede apreciar el número de características con las que tiene que trabajar el sistema por cada año. A este número de términos hay que recordar que ya se le han eliminado previamente las *stop-words* y los *hápax legómenon*, lo cual ha hecho decrecer drásticamente el número de características iniciales. Pese a ello, sigue siendo un número de características muy elevado, lo que ocasiona un alto coste computacional, lo cual se traduce en una gran demanda de tiempo y espacio cada vez que se quiere clasificar un texto geográficamente.

La inmensa mayoría de estas características tiene una aportación nula o mínima a la tarea de clasificación geográfica de textos, ya que no suelen ser representativas de ninguna región geográfica en concreto.

Si no se tiene en cuenta el tiempo y espacio necesarios para trabajar con este elevado número de características, los resultados, no solo no tiene por qué traducirse en una mejor clasificación, sino que en numerosas ocasiones introducen ruido[6] al sistema ocasionando así unos resultados más pobres.

En la práctica, el funcionamiento de los algoritmos de aprendizaje tales como *SVM* pueden ser mejorados frecuentemente mediante un proceso de preselección de características o *feature selection* (Cardie, 1996) (Wang and He, 2004) (Yang and Pedersen, 1997).

Así pues, al realizar una selección de características se reduce la dimensionalidad al descartar un gran número de características irrelevantes, convirtiéndose en un problema más asequible para los algoritmos empleados en la resolución de este tipo de tareas.

Otra ventaja de la reducción de características es que puede evitar un sobreajuste[7] del algoritmo, ya que cuanto mayor es el número de características en un algoritmo de aprendizaje automático, mayor es el ajuste que intenta realizar el sistema, produciéndose un sobreajuste cuando estas características no son lo suficientemente relevantes.

En la tarea que nos atañe, la de clasificar geográficamente un conjunto de textos, intuitiva y empíricamente, tal y como se ha podido comprobar en los experimentos llevados a cabo en esta sección, parece que si se seleccionan los topónimos existentes en el texto, puede ser una buena aproximación para reducir el número de características. El problema es que, como hemos visto en los experimentos previos, hay otros grupos gramaticales que pueden ayudar a realizar esta tarea de una manera más eficiente, ya que estos grupos gramaticales aportan un mayor conocimiento del mundo que nos rodea y por ende ayudan a ubicar geográficamente los textos.

Para ello existen técnicas basadas en la teoría de la información que permiten determinar estadísticamente cuáles son las características con un mayor peso a la hora de clasificar correctamente los textos, pudiendo descartar así aquellas que no resulten de utilidad.

Para llevar a cabo este proceso se emplea una función que mide la importancia de cada característica en la tarea de clasificación. En un estudio

[6] Una característica introduce ruido cuando al ser añadida al vector de aprendizaje incrementa el error de clasificación sobre nuevos datos.

[7] En aprendizaje automático, el sobreajuste (también es frecuente emplear el término en inglés *overfitting*) es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. El algoritmo de aprendizaje debe alcanzar un estado en el que será capaz de predecir el resultado en otros casos a partir de lo aprendido con los datos de entrenamiento, generalizando para poder resolver situaciones distintas a las acaecidas durante el entrenamiento. Sin embargo, cuando un sistema se entrena demasiado (se sobreentrena) o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo. Durante la fase de sobreajuste el éxito al responder las muestras de entrenamiento sigue incrementándose mientras que su actuación con muestras nuevas va empeorando.

llevado a cabo por Yang and Pedersen (1997) se demostró que técnicas sofisticadas de reducción de características tales como *Information Gain* (*IG*) y χ^2 pueden reducir la dimensionalidad del espacio de características por un factor de 100 sin pérdida (e incluso con pequeñas ganancias) en la efectividad en la tarea de la clasificación de textos.

En nuestro caso, debido a los buenos resultados obtenidos (ver trabajos de Zheng et al. (2004) y Moh'd A Mesleh (2007)), para la reducción del número de términos se ha utilizado el algoritmo de selección de características χ^2 que se encuentra dentro de *Weka toolkit* (Witten and Frank, 2005).

χ^2 es empleado en estadística con el fin de evaluar la independencia de dos eventos. En selección de características, estos eventos son la ocurrencia de la característica c y la ocurrencia de la clase i. Este algoritmo da una medida de cuánto se desvía la frecuencia esperada de la frecuencia observada. Así pues, un valor grande para una determinada característica indicaría que la hipótesis de independencia, que implicaría que los valores observados y esperados son similares, es incorrecta. Desde un punto de vista estadístico, este método puede ser problemático ya que, cuando un test estadístico se emplea en muchas ocasiones, como es el caso que nos atañe por cada característica, la probabilidad de obtener errores se incrementa. En cualquier caso, en la clasificación de textos, raramente tiene importancia que se añadan o eliminen unos cuantos términos que no son muy determinantes al conjunto de características final. Esta selección de características se efectúa ordenando las características en función del valor devuelto por χ^2 y escogiendo los n mejores valores. Una explicación más detallada de la distribución χ^2 se puede encontrar en Lancaster and Seneta (1969), así como un estudio de dicha distribución aplicada a la obtención de características para la clasificación de textos se puede observar en Yang and Pedersen (1997) y How and Narayanan (2004).

Por otro lado, debido a que en los próximos experimentos se iban a utilizar otros corpus para el entrenamiento del sistema, se ha optado por seleccionar el número de características del corpus que menos tenía, *Wikipedia*, con el fin de poder comparar estos experimentos con los que se mostrarán en la sección 5.3.2. Así pues, se ha experimentado con los siguientes números de características (palabras):

- 54.183. Son el número total de características que había en el corpus de la *Wikipedia*.
- 30.376. Son el número total de características que había en el corpus de la *Wikipedia* sin los hápax legómenon.
- 20.848. Son el número total de adjetivos y sustantivos que había en el corpus de la *Wikipedia* sin los hápax legómenon.

- 17.161. Son el número total de sustantivos que había en el corpus de la *Wikipedia* sin los hápax legómenon.
- 10.000. Debido a que había un gran salto entre el número de características utilizadas entre el experimento anterior y el próximo (más de 13 veces menor), se ha realizado un experimento con este número intermedio de características.
- 1.298. Son el número total de topónimos que había en el corpus de la *Wikipedia* sin los hápax legómenon.

Mediante el algoritmo de χ^2 se ha conseguido pasar del número de características expuesto en la tabla 5.3, al previamente indicado y expuesto en la primera columna de la tabla 5.7.

En esta ocasión, el experimento llevado a cabo se ha realizado con todos los términos del corpus, es decir, sin que haya ningún tipo de filtrado por la categoría gramatical a la que pertenece cada uno de los términos del corpus, tal y como se hizo en los experimentos previos. Esto es debido a que no se ha dado por hecho que términos que pertenecen a la categoría gramatical de adjetivos o sustantivos, incluso a topónimos, fueran a ser más determinantes a la hora de clasificar geográficamente los textos, y por ende se ha pretendido dejar que la función de χ^2 haga la selección de las características que considere más determinantes, sin tener en cuenta la categoría gramatical a la que éstas pertenecen.

Así pues, en este experimento, la primera división que se ha hecho del corpus ha sido siguiendo el número de características más determinantes a la hora de discernir la localidad de la que provienen los textos del corpus del diario *20Minutos*. El número de características seleccionado es el previamente descrito.

Los resultados obtenidos se pueden apreciar en la tabla 5.7, donde las columnas indicas los años y las filas el número de características que fueron empleadas para la clasificación geográfica de los textos. En la fila en la que se indica que se emplean todas las características (*Todas*) se muestran los resultados obtenidos sin aplicar la selección de características, es decir, los mismos que se podían apreciar en la columna '*Todo*' de la tabla 5.5 del experimento llevado a cabo con *SVM* en esta misma sección.

A la luz de los resultados mostrados en la tabla 5.7, cabe destacar cómo la reducción de características aporta una mejora significativa ($p < 0,01$) en términos de precisión, siendo ésta más notoria cuando el corpus con el que se trata es menos voluminoso, como se puede apreciar para los años 2008 y 2009.

Por último, si se compara los resultados obtenidos entre los distintos números de características reducidas, se puede ver como se obtienen unos resultados muy similares, por lo que se puede concluir que dado que se necesita menos tiempo y espacio, es mejor la utilización de un

Tabla 5.7: Resultados de la validación cruzada de *SVM* entrenando y probando con el corpus del diario *20Minutos* con reducción de características mediante χ^2.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Todas	73,55	74,82	84,12	83,33	84,35
54.183	**82,44**	**87,65**	88,11	87,34	87,76
30.376	81,96	87,52	87,96	87,32	87,74
20.848	81,70	86,88	87,94	87,34	87,78
17.161	81,79	86,81	87,95	87,36	87,98
10.000	81,95	86,24	87,82	**87,59**	**88,56**
1.298	82,09	86,84	**88,11**	87,19	86,71

número reducido de características (1.298 en nuestro caso) para el buen funcionamiento del sistema.

5.3.2. Entrenamiento con artículos de *Wikipedia*

En esta sección se va a detallar el procedimiento llevado a cabo para elaborar los experimentos conducidos con un entrenamiento con artículos de *Wikipedia* para clasificar geográficamente los artículos del diario *20Minutos*, tal y como se ha hecho en el apartado anterior. Con la utilización de los textos de *Wikipedia* como entrenamiento, se pretende comprobar si es posible realizar un entrenamiento con una fuente formal distinta a la que se pretende clasificar para un sistema que intenta detectar el foco geográfico de los textos.

La técnica de clasificación utilizada para detectar el foco geográfico de los artículos ha sido *SVM*, la cual se ha descrito en la sección 5.1.1 de este mismo capítulo. Se ha utilizado esta técnica, ya que en experimentos previos (ver figura 5.2) mostró mejores resultados que con los modelos de lenguaje.

Por otro lado, se han creado corpus obedeciendo a la categoría gramatical de los términos de los artículos originales, tal y como se hizo en los experimentos anteriores. Así pues, se ha trabajado con corpus que contienen todos los términos de los artículos, corpus donde solamente aparecen los adjetivos y sustantivos detectados por *FreeLing*, corpus donde sólo aparecen los sustantivos, corpus donde sólo aparecen los topónimos, y corpus donde aparecen todos los términos menos los topónimos. El número de términos de cada uno de estos corpus se puede observar en la tabla 5.8. Como se puede apreciar, el número de términos mostrado en esta tabla coincide con el que se utilizó en la reducción de características (ver tabla 5.7).

Esto se ha hecho así para poder comprobar cómo se comportaba *Wikipedia* como método de selección de características al utilizar su vocabulario únicamente. Es decir, con la utilización única y exclusiva de los textos de *Wikipedia* como entrenamiento, se pretende realizar una selección

de las características (palabras) directa utilizadas en la clasificación de los textos del diario *20Minutos*, que tan buenos resultados ha dado en el experimento anterior donde se seleccionaban las características mediante χ^2 (ver tabla 5.7).

Como ya se comentó en dicho experimento, el número de términos seleccionados mediante la técnica de reducción de características χ^2 se ha ajustado al número de éstas obtenidas en cada una de las distintas categorías gramaticales de los artículos de *Wikipedia*. De esta forma se puede realizar una comparación directa según el número de características utilizado para la clasificación geográfica.

Tabla 5.8: Número de términos distintos del corpus de artículos de ciudades de *Wikipedia* según la categoría gramatical del corpus obtenido con *FreeLing*.

CATEGORÍA	TÉRMINOS
Topónimos	1.298
Sustantivos	17.161
Adjetivos y Sustantivos	20.848
Sin Topónimos	29.078
Sin *Hápax Legómenon*	30.376
Todos	54.183

Estos artículos eran los de las ciudades sobre las que se pretendía etiquetar el corpus de noticias del diario *20Minutos*. Para ello se obtuvieron los artículos de las 50 capitales de provincia de España más sus 2 ciudades autónomas (ver ciudades en la tabla 4.1), de manera manual. De estos artículos únicamente se utilizó el texto que había en ellos, sin hacer uso de ningún metadato asociado a éstos, ya que el objetivo era el poder averiguar la aportación de una fuente diferente de lenguaje natural a la hora de clasificar geográficamente textos periodísticos. En otras palabras, se pretende averiguar la procedencia de una noticia simplemente con la aportación textual de un recurso lingüístico formal sin estructurar distinto a los propios textos que se quieren ubicar.

Al igual que se hizo con el corpus del diario *20Minutos*, se obtuvo la categoría gramatical de cada término que aparecía en los textos de *Wikipedia* mediante la herramienta de análisis lingüístico *FreeLing*, guardando nuevamente las 5 categorías descritas en el apartado anterior: topónimos, sustantivos, sustantivos+adjetivos, todos los términos menos los topónimos y todos los términos.

Por cada localidad presente en el sistema se tenía pues un único fichero de texto para entrenar, el correspondiente al artículo de *Wikipedia* que representaba a dicha localidad. Pese a ello, debido a la mayor extensión que suelen tener los artículos de *Wikipedia* en comparación con los artículos del diario *20Minutos*, aunque solamente se contaba con 52 textos para entrenar,

estos aportaban un gran número de términos (una media de 18.166 términos por artículo) que podían facilitar la identificación de la ubicación de una noticia dada.

Una vez creados los ficheros de texto para el entrenamiento se procedió a obtener la frecuencia de cada uno de estos términos con la herramienta *SRI Language Modeling*, tal y como se hizo con los textos del diario *20Minutos* en los experimentos previos.

Así pues, el vocabulario fue obtenido de los términos que aparecen en los artículos de *Wikipedia* utilizados, pudiéndose apreciar el número distinto de términos según la categoría gramatical de cada corpus en la tabla 5.8. Dicho número de términos es varias veces menor que el que se obtuvo del corpus del diario *20Minutos* y que se puede observar en la tabla 5.3.

Con estos corpus se realizaron los siguientes experimentos:

1. Clasificación geográfica de noticias del diario *20Minutos* utilizando como entrenamiento artículos de *Wikipedia* de las localidades de procedencia de las noticias.

2. Clasificación geográfica de noticias del diario *20Minutos* utilizando como selección de características los artículos de *Wikipedia* de las localidades de procedencia de las noticias, y para el entrenamiento un conjunto de textos del diario *20Minutos* distinto al que se evaluó.

3. Clasificación geográfica de noticias del diario *20Minutos* utilizando como entrenamiento artículos de *Wikipedia* de las localidades de procedencia de las noticias y los artículos que se citaban en los artículos de estas localidades (*outlinks*).

4. Clasificación geográfica de noticias del diario *20Minutos* utilizando como método de selección de características los artículos de *Wikipedia* de las localidades de procedencia de las noticias y los artículos que se citaban en los artículos de estas localidades, y para el entrenamiento un conjunto de textos del diario *20Minutos* distinto al que se evaluó.

A continuación se detallan los experimentos expuestos.

Entrenamiento procedente de artículos de localidades de *Wikipedia*

Para el entrenamiento del sistema, tal y como ya se ha indicado en la sección 4.3, se utilizaron única y exclusivamente artículos de *Wikipedia*.

Los ficheros de evaluación utilizados en este experimento coinciden con los utilizados en los experimentos previos, es decir, son las noticias del corpus del diario *20Minutos* utilizadas en la evaluación de los experimentos anteriormente descritos. Así pues, se entrena única y exclusivamente con ficheros de *Wikipedia* y se evalúa con los mismos ficheros del corpus del

diario *20Minutos* que se probaron en la ejecución entrenando únicamente con el corpus del diario, aunque, en esta ocasión, no se eliminaron los hápax legómenon ni del corpus de entrenamiento ni del corpus de evaluación, dado que en esta ocasión son corpus distintos y el que haya términos que sólo aparecen en una única ocasión en uno de los corpus, no quiere decir que en el otro no vaya a aparecer.

Una vez más, se ha vuelto a realizar una validación cruzada de 10 iteraciones para conocer los resultados del sistema.

Los resultados obtenidos se muestran en la tabla 5.9, donde las filas indican la categoría gramatical utilizada, mientras que las columnas señalan los años de los artículos que se clasificaron.

Tabla 5.9: Resultados de la validación cruzada de *SVM* entrenando con el corpus de artículos de las ciudades de *Wikipedia* y evaluando el corpus del diario *20Minutos*

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	**32,71**	**33,07**	**29,55**	**29,62**	**30,08**
Sustantivos	16,70	16,54	13,19	13,39	13,80
Sustantivos + Adjetivos	23,36	22,81	18,06	18,15	18,86
Sin topónimos	0,91	0,84	0,76	0,70	0,75
Todo	1,99	1,59	0,93	0,85	1,03

Como se puede apreciar en dicha tabla, los resultados son claramente peores que los obtenidos utilizando los propios textos del diario *20Minutos*. Pese a ello, estos resultados merecen un análisis más exhaustivo.

En primer lugar, llama la atención la baja precisión obtenida cuando se utilizan todos los términos y todos los términos menos los topónimos. Esto demuestra que el vocabulario existente en los artículos de *Wikipedia* claramente difiere del empleado en los artículos del diario *20Minutos*, siendo los adjetivos y sustantivos, y dentro de estos últimos los topónimos, los que aportan más información al sistema para que éste lleve a cabo la tarea con unos resultados mejores, mientras que los demás términos introducen ruido tal y como se puede apreciar en los resultados.

Por otro lado, en esta ocasión, se puede apreciar cómo haciendo un filtrado y utilizando bien sustantivos únicamente, o bien sustantivos y adjetivos, la precisión obtenida se incrementa considerablemente. Esto indica la cantidad de ruido que introducen el resto de categorías gramaticales de los términos utilizados en los artículos de *Wikipedia*.

Aunque, tal vez, lo más destacable en este experimento es la gran aportación que realizan los topónimos encontrados por *FreeLing* en los artículos de las localidades examinadas en *Wikipedia*. Esta mejora es resultado de la gran distancia que hay entre los textos de una fuente

y otra, pese a ser ambas fuentes consideradas como formales. Debido a esta diferencia entre los textos, parece que son única y exclusivamente los topónimos los que pueden aportar más información al sistema introduciendo una menor cantidad de ruido.

Otra posible causa sería el escaso número de artículos que se ha utilizado en el entrenamiento del sistema (simplemente los artículos de *Wikipedia* que representan a las localidades tratadas). Por ello es necesario llevar a cabo experimentos con un número mayor de textos para el entrenamiento, tal y como se va a mostrar en los próximos experimentos.

Selección de características mediante artículos de localidades de *Wikipedia* y entrenamiento procedente de *20Minutos*

De manera análoga a como se realizó el experimento anterior, en esta ocasión se utilizaron también los artículos de *Wikipedia* de las localidades en cuestión para obtener el vocabulario de términos existente en el entrenamiento. A diferencia del experimento anterior, esta vez, en lugar de utilizar también dichos artículos para el entrenamiento, se utilizaron los artículos del diario *20Minutos* correspondientes a la partición de entrenamiento. De este modo, por un lado se obtiene un conjunto de términos procedentes de una fuente distinta a la que se evalúa a modo de selección directa de características, tal y como se explicó al comienzo de la sección 5.3.2, y por otro se entrena un conjunto de textos mucho más amplio al utilizado en el experimento anterior con el fin de obtener unos resultados mejores. En otras palabras, se utilizaron únicamente los términos de *Wikipedia* como vocabulario, y los artículos del diario *20Minutos* como entrenamiento. De nuevo, se ha vuelto a realizar una validación cruzada de 10 iteraciones para conocer los resultados del sistema. Los resultados se pueden ver en la tabla 5.10.

Tabla 5.10: Resultados de la validación cruzada de *SVM* utilizando el vocabulario de los artículos de *Wikipedia*, entrenando con el corpus de artículos de noticias del diario *20Minutos* y evaluando el corpus del diario *20Minutos*

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	43,82	46,10	45,57	46,09	45,27
Sustantivos	58,38	60,54	68,99	68,86	70,57
Sustantivos + Adjetivos	63,80	66,28	71,98	71,65	73,25
Sin topónimos	55,63	57,99	70,59	69,92	70,16
Todo	**67,76**	**69,98**	**76,98**	**76,31**	**77,02**

En los resultados obtenidos, como era de esperar, se observa como mejoran considerablemente respecto al experimento anterior, ya que esta vez se ha entrenado con un corpus de textos de la misma naturaleza que el empleado en la evaluación.

En esta ocasión, se puede apreciar como los mejores resultados se obtienen cuando se emplean todos los términos del corpus, a diferencia del experimento anterior donde esta aproximación era una de las que peores resultados daba, lo que pone de manifiesto la importancia de los matices lingüísticos existentes en un texto para determinar su procedencia geográfica cuando se tiene un conjunto de entrenamiento no tan limitado como el del experimento previo y mucho más afín al texto que se pretende clasificar.

Por el contrario, la aproximación que se basa únicamente en los topónimos es, en esta ocasión, la que peores resultados da debido a la escasez de dichos términos en los textos, teniendo además que coincidir estos topónimos con los existentes en el vocabulario del sistema, es decir, con los existentes en los artículos de *Wikipedia*.

Es significativo el gran aumento de precisión obtenido con la aproximación que utilizaban todos los términos y la que utilizaba todos los términos menos los topónimos. Esto es debido a que los dos corpus, el de *Wikipedia* y el de *20Minutos*, pese a ser textos formales, difieren mucho en su vocabulario.

Por último, pese a obtener unos resultados ligeramente peores (alrededor de un 7 % peor) a los logrados en la aproximación en la que se utilizaba el vocabulario del diario *20Minutos* y entrenando con el mismo corpus, este enfoque resulta ser mucho más fácil de implementar al tener claramente desambiguados los artículos de cada localidad, mientras que el enfoque en el que se utilizan artículos del diario *20Minutos* se ha tenido que filtrar previamente las noticias según su procedencia geográfica para que únicamente se extraiga el vocabulario de aquellos artículos centrados en una o varias localidades concretas y no en un ámbito más genérico. El tiempo de ejecución de ambas aproximaciones también es un factor a tener en cuenta ya que con el vocabulario de *Wikipedia* se reduce drásticamente al haber menos características por las que clasificar (ver tabla 5.11).

Al final de esta sección se mostrará una comparativa de las distintas aproximaciones de selección de características.

Vocabulario y entrenamiento procedente de artículos de localidades de *Wikipedia* y los artículos referenciados

Otro experimento que se realizó con el entrenamiento único y exclusivo de textos de *Wikipedia* fue utilizando para el entrenamiento, aparte de los textos de los artículos de las propias ciudades, el texto de los artículos que se mencionaban en los artículos de *Wikipedia* de éstas localidades, es decir, el texto de los artículos a los que apuntaban los enlaces salientes o *outlinks*

(la recolección y filtrado de los textos de los artículos de estos enlaces se ha explicado con más profundidad en la sección 4.3).

El número total de artículos referenciados en cada uno de los artículos de las localidades analizadas en este experimento es 29.998, incluyendo los artículos de las propias ciudades, lo que da un promedio de unos 577 artículos por cada una de las 52 ciudades trabajadas, aunque, dicho promedio no está distribuido equitativamente como muestran las cifras de las ciudad con más artículos referenciados, Madrid con 1.296, y la ciudad con menos, Tarragona con 175.

El número de características que se utilizaron para cada categoría gramatical se muestra en la tabla 5.11, donde además, con el fin de poder comparar, aparecen el número de características empleado cuando únicamente se utilizaban los artículos de *Wikipedia* de las propias localidades y el número de características utilizado con el vocabulario del diario *20Minutos*.

Tabla 5.11: Número de términos distintos de los corpus *20Minutos*, únicamente artículos de las ciudades de *Wikipedia* y artículos de las ciudades de *Wikipedia* y los artículos referenciados en éstos según la categoría gramatical del corpus obtenido con *FreeLing*.

CATEGORÍA	***20Minutos***	***Wikipedia* ciudades**	***Wikipedia* enlaces**
Topónimos	62.480	1.298	50.711
Sustantivos	766.905	17.161	842.895
Adjetivos y Sustantivos	794.559	20.848	897.044
Sin Topónimos	932.827	29.078	1.043.129
Todos	975.067	54.183	1.074.174

Así pues, el procedimiento para el entrenamiento con los artículos de estos enlaces salientes ha sido el mismo que el realizado con los textos de únicamente los artículos de las ciudades de *Wikipedia* recién explicado, con la única salvedad de que en esta ocasión las ubicaciones vienen representadas no sólo por el artículo de su propia ciudad, sino que también por el de los artículos referenciados en éstos (*outlinks*), determinando así el vocabulario y la frecuencia de los términos para cada ciudad.

Los resultados obtenidos se muestran en la tabla 5.12.

Una vez más, se constata el hecho de que la utilización exclusiva de topónimos en el vocabulario y entrenamiento del sistema dada una fuente de texto de distinta naturaleza, resulta la aproximación más acertada para llevar a cabo la tarea de clasificación geográfica de textos formales.

Por un lado, los resultados obtenidos con el corpus en el que se incluían todos los términos y en el que se incluían todos los términos menos los topónimos, experimentan una sustancial mejora, aunque los resultados

Tabla 5.12: Resultados de la validación cruzada de *SVM* entrenando con el corpus de artículos de las ciudades de *Wikipedia* y el de los artículos referenciados en éstos para la evaluación del corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	**39,41**	**41,05**	**36,93**	**36,61**	**37,25**
Sustantivos	14,41	15,59	12,04	12,32	12,58
Sustantivos + Adjetivos	15,13	17,91	15,03	15,40	15,36
Sin topónimos	5,38	7,39	6,83	6,27	6,47
Todo	5,04	6,68	7,59	7,94	7,45

siguen siendo muy pobres debido a la gran diferencia de vocabulario entre las dos fuentes.

Si se centran la atención en las aproximaciones que utilizaban únicamente los sustantivos y la que utilizaba los sustantivos y los adjetivos, se puede apreciar como la precisión del sistema incluso decrece. Esto es debido a que los artículos referenciados contienen una gran cantidad de adjetivos y, sobre todo, sustantivos que no identifican las localidades desde las que se referencian estos artículos.

Por último, comparando los resultados obtenidos en este experimento con los obtenidos en el experimento en el que únicamente se tenían en cuenta los artículos de las ciudades en cuestión de *Wikipedia* (tabla 5.9), se aprecia una mejora de unos 7 puntos en términos absolutos (más de un 30 % en términos relativos) en la mejor aproximación, es decir, la que utiliza únicamente los topónimos de los textos. Esto es debido a que ahora se dispone un número mucho mayor de características para realizar la clasificación (ver tabla 5.11). Aún así, habría que considerar si compensa este incremento en la precisión teniendo en cuenta el coste computacional y espacial requerido para esta aproximación.

Selección de características mediante los artículos de localidades de *Wikipedia* y los artículos referenciados, y entrenamiento procedente de *20Minutos*

El experimento llevado a cabo en esta sección es una mezcla de los 3 previos, es decir, se ha utilizado para la obtención del vocabulario los términos de los artículos de *Wikipedia* de las ciudades en cuestión más el de sus *outlinks*, y para el entrenamiento el corpus de noticias del diario *20Minutos*. Los resultados obtenidos son los mostrados en la tabla 5.13.

A la luz de la precisión final obtenida en esta aproximación, se puede decir que es la más exitosa de las realizadas con el corpus de *Wikipedia*, aunque está seguida muy de cerca por la mostrada en los experimentos

Tabla 5.13: Resultados de la validación cruzada de *SVM* utilizando el vocabulario de los artículos de *Wikipedia* y sus enlaces salientes, entrenando con el corpus de artículos de noticias del diario *20Minutos* y evaluando el corpus del diario *20Minutos*

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	55,23	57,00	57,85	57,37	57,58
Sustantivos	65,74	67,22	75,48	74,96	76,10
Sustantivos + Adjetivos	69,13	71,38	77,46	76,63	77,52
Sin topónimos	60,58	62,07	76,04	75,12	75,28
Todo	**72,04**	**73,64**	**81,56**	**80,67**	**81,24**

llevados a cabo utilizando para vocabulario los artículos de las localidades de *Wikipedia* y entrenando con el corpus de noticias del diario *20Minutos* (ver tabla 5.10).

Pese a que esta aproximación es mucho más fácil de implementar al poderse obtener los artículos pertenecientes a las localidades tratadas de una manera más sencilla (simplemente indicando el artículo que representa a cada una de las localidades analizadas), se ha de tener presente que la aproximación en la que solamente se tenía en cuenta el texto de los artículos de las ciudades de *Wikipedia* resulta mucho más liviana que la expuesta en esta sección, por lo que puede que no se justifique la aproximación mostrada en este último apartado al lograr únicamente alrededor de un 4 % de mejora en valores absolutos.

Una vez más, al igual que se pudo apreciar en el anterior experimento donde se utilizaba el vocabulario de *Wikipedia* y se entrenaba con el corpus del diario *20Minutos*, se demuestra la importancia que aportan todos los términos y expresiones lingüísticas de los textos a la hora de determinar el foco geográfico de estos, obteniendo una mejora significativa ($p < 0,01$) con respecto al resto de aproximaciones que no hacen uso de todos los matices lingüísticos.

Comparación de aproximaciones con selección de características

En las figuras 5.3 y 5.4 se puede apreciar una comparativa de todas las aproximaciones realizadas con el corpus de *Wikipedia* expuestos en esta sección y los resultados logrados con la aproximación en la que se reducía el número de características expuesto en la sección 5.3.1, utilizando todos los términos y solamente los topónimos respectivamente. Para todos los experimentos se realizó una validación cruzada siendo completamente distinto el corpus de entrenamiento del de evaluación.

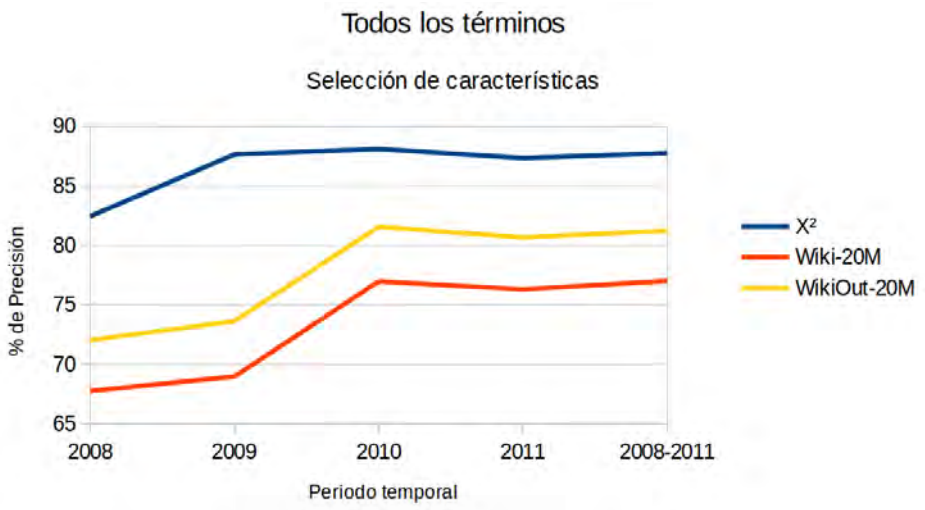


Figura 5.3: Precisión obtenida con las aproximaciones de *SVM* en las que se utilizaban como entrenamiento los 54.183 términos más discriminatorios obtenidos con χ^2 o los de los artículos de *Wikipedia*, entrenando con los artículos del diario *20Minutos*.

Las aproximaciones utilizadas son: χ^2 reducción de características (54.183), *Wiki-20M* vocabulario de ciudades de *Wikipedia* (54.183 características) y entrenamiento de *20Minutos*, y *WikiOut-20M* (1.074.174 características) vocabulario de ciudades y *outlinks* de *Wikipedia* y entrenamiento de *20Mintuos*.

La figura 5.3 muestra como la introducción de los artículos de *Wikipedia* referenciados (*outlinks*) aporta una mejora en la precisión, aunque, como se comentó en la sección previa, el coste computacional y temporal crece exponencialmente, tal y como se puede deducir del número de características utilizadas en estas aproximaciones (54.183 frente a 1.074.174).

χ^2 es claramente la mejor aproximación, aunque hay que tener en cuenta que para ello es necesaria una obtención previa de artículos del mismo diario para poder utilizarlos en la tarea de la extracción de estas características.

En cuanto a la figura 5.4, las aproximaciones utilizadas son: χ^2 reducción de características (1.298) con χ^2, *Wiki-20M* vocabulario creado con los topónimos (1.298 características) de ciudades de *Wikipedia* y entrenamiento de *20Minutos*, *WikiOut-20M* vocabulario creado con los topónimos (50.711 características) de ciudades y *outlinks* de *Wikipedia* y entrenamiento de *20Mintuos*.

Comparando estas dos figuras, se puede apreciar como se pueden extraer las mismas conclusiones en cuanto a los resultados de *Wikipedia* y *Wikipedia* con *outlinks*.

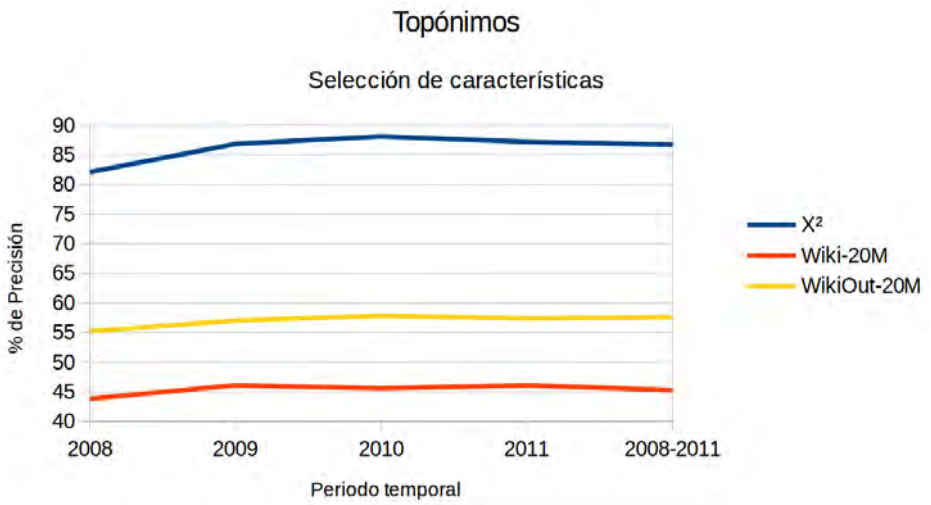


Figura 5.4: Precisión obtenida con las aproximaciones de *SVM* en las que se utilizaban como entrenamiento los 1.298 términos más discriminatorios obtenidos con χ^2 o los topónimos de los artículos de *Wikipedia*, entrenando con los artículos del diario *20Minutos*.

En cuanto a la selección de características con χ^2, en esta ocasión, la diferencia es mucho más marcada (un 20-35 % mejor). Hay que recordar que χ^2 no discrimina por categoría gramatical, lo que vuelve a poner de manifiesto que, pese a ser el mismo número de características, la diferencia de la utilización de todos los términos (figura 5.3) a crecido exponencialmente con la que utiliza solamente los topónimos (figura 5.4). Esto demuestra, una vez más, la gran importancia de todas las categorías gramaticales, y no sólo los topónimos, cuando se trata de realizar una clasificación de textos geográfica.

5.3.3. Entrenamiento con mensajes de *Twitter*

Siguiendo la metodología llevada a cabo en la sección anterior, en esta sección se va a mostrar cómo se ha llevado a cabo la experimentación utilizando una fuente de texto "*informal*", *Twitter*, para posteriormente evaluarlo sobre una fuente de texto "*formal*", el corpus de noticias del diario *20Minutos*.

Lo que se pretende averiguar con los experimentos llevados a cabo en esta sección es lo que puede llegar a aportar la información general del mundo, en este caso la fuente de textos informales *Twitter*, a la tarea de la clasificación geográfica de textos formales, intentando aprovechar así la abundancia de textos informales presentes en la web para esta tarea. La falta de formalidad puede ser compensada con el volumen de textos existentes.

En esta ocasión, a diferencia de lo que ocurriera en las dos secciones previas en las que se entrenaba un sistema *SVM* con textos que únicamente contenían términos englobados en unas específicas categorías gramaticales, se utilizará todo el texto ya que las herramientas disponibles para poder obtener dichas categorías gramaticales tienen un bajo índice de acierto debido a la informalidad de los textos tratados, tuits.

Por otro lado, debido a que los tuits proceden de un gran número de usuarios distintos, en esta ocasión cobra una mayor relevancia si cabe los matices lingüísticos existentes en los textos emitidos, ya que algunos términos pueden ser más frecuentes en ciertas áreas geográficas que en otras en las que en raras ocasiones se utilizan. Por ejemplo, si pensamos en términos como "*falla*", comprobaremos que éste será mucho más común en el área de Valencia que en cualquier otra debido a las populares fiestas que tienen lugar en la ciudad. O argot como "*chola*", el cual es comúnmente empleado en Canarias para referirse a las zapatillas.

El corpus utilizado para este experimento es el descrito en la sección 4.2. En dicha sección se puede apreciar que pese a que los tuits no son muy extensos, hasta 140 caracteres cada uno, el gran número que hay por cada una de las ciudades del corpus, las cuales coinciden con las del corpus del diario *20Minutos*, hace que sí que haya una gran cantidad de texto para realizar el entrenamiento del sistema.

Para los experimentos llevados a cabo en esta sección se realizaron dos aproximaciones obedeciendo al conjunto de tuits utilizados como entrenamiento del sistema, y una tercera aproximación donde se utilizó el vocabulario de *Twitter* como método de selección de características y los artículos del diario *20Minutos* para entrenar, tal y como se hiciera en los experimentos llevados a cabo en la sección anterior con *Wikipedia*.

Así pues, por un lado se realizaron experimentos agrupando todos los tuits emitidos en cada una de las ciudades dadas, dando como resultado un total de 52 textos de entrenamiento, uno por cada una de las ciudades del corpus.

Por otro lado, se realizó otro conjunto de experimentos unificando los tuits emitidos por cada usuario en cada ciudad. Es decir, si un usuario *U1* había emitido tuits en la ciudad *C1* esto haría que hubiera una muestra de entrenamiento para dicha ciudad. Si además, un segundo usuario *U2* emitió tuits desde la ciudad *C1* y desde la ciudad *C2*, en el conjunto de entrenamiento habría dos entradas más, una con los tuits pertenecientes a la ciudad *C1* y otra con los tuits pertenecientes a la ciudad *C2*. Es decir, para cada ciudad del corpus existen tantas muestras representándola como usuarios tuitearon desde esa ciudad.

Como en los experimentos anteriores, se procedió a crear los ficheros de frecuencia con los términos que aparecían en los tuits, obteniendo así el número de veces que dichos términos aparecían en cada una de las ciudades.

Para realizar dicho proceso, se utilizó la herramienta *SRILM* como en los experimentos anteriores.

Una vez obtenidos los ficheros de frecuencia de cada uno de los términos de las distintas ciudades del corpus, se procedió a obtener el vocabulario procedente de los términos existentes en dichos ficheros de frecuencia. Al igual que ocurriera con los experimentos de la sección anterior, donde se entrenaba única y exclusivamente con artículos de la *Wikipedia*, con los tuits tampoco existía una separación por años como ocurría con los experimentos de la sección 5.3.1 donde se entrenaba única y exclusivamente con noticias del propio corpus del *20Minutos*.

Una vez más, al igual que sucediera con los experimentos de las secciones anteriores, se reutilizaron los mismos conjuntos de evaluación del diario *20Minutos* con las mismas particiones descritas en dichas aproximaciones. Y al igual que entonces también, los ficheros de evaluación contenían únicamente las frecuencias de los términos que aparecían en el vocabulario correspondiente, que para los experimentos realizados con el corpus de *Twitter* fue creado únicamente con los términos del corpus de *Twitter* aquí utilizado.

Vocabulario y entrenamiento procedente del conjunto de tuits separados por usuario en cada ciudad

Como ya se ha comentado previamente, para el corpus de entrenamiento, *Twitter*, no se realizó ningún filtrado por categoría gramatical debido al escaso éxito que ofrecen las herramientas actuales en la realización de esta tarea para texto tan altamente informales, pero sí que se mantuvo el filtrado por categoría gramatical realizado en experimentos anteriores para el conjunto de evaluación, *20Minutos*. Así pues, la gran mayoría de los términos del conjunto de entrenamiento no aparecen en el conjunto de evaluación, especialmente cuando el conjunto de evaluación utiliza menos términos, como es el caso de cuando se utilizan únicamente los topónimos de los artículos. Esta aproximación se realiza de esta manera debido a que se pretende averiguar si los términos que aparecen son lo suficientemente discriminatorios y compensan el ruido introducido por el resto.

Con el vocabulario obtenido se procedió a crear los ficheros de entrenamiento, para lo cual, se unieron los conjuntos de tuits según el usuario que lo hubiera emitido en cada una de las distintas ciudades en las que éste emitió tuits, por lo que hubo tantos casos de entrenamiento como la suma de usuarios en todas las ciudades del corpus.

Con el fichero de entrenamiento y los de evaluación se crearon finalmente los modelos del sistema que permitieron evaluar dicho experimento.

Los resultados obtenidos se pueden ver en la tabla 5.14.

Según los resultado mostrados en la tabla 5.14, la aproximación que mejor ha funcionado ha sido la que intenta clasificar textos que contienen

Tabla 5.14: Resultados de la validación cruzada de *SVM* utilizando el vocabulario y textos de los tuits agrupado por usuario/ciudad para clasificar el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	6,21	7,27	0,78	0,75	1,64
Sustantivos	8,46	8,75	3,80	3,91	4,56
Sustantivos + Adjetivos	**11,42**	12,72	7,75	7,58	8,30
Sin topónimos	11,34	**13,86**	**8,52**	**8,59**	**9,11**
Todo	11,34	**13,86**	**8,52**	**8,59**	**9,11**

todas las categorías gramaticales del corpus. Aún así, debido a lo poco representativas que resultan unas muestras de entrenamiento con una longitud de texto tan reducida, los resultados distan mucho de los conseguidos con otras aproximaciones.

Vocabulario y entrenamiento procedente del conjunto de tuits agrupados por ciudad

En el experimento previo se ha entrenado con un gran número de ficheros de texto de un tamaño no demasiado extenso que representan a las 52 ciudades existentes en el corpus. Debido a esto, en muchas ocasiones, el sistema sufre un sobreajuste que puede llegar a provocar una clasificación errónea de los artículos del conjunto de evalución. Para salvar este sobreajuste, se ha realizado el mismo experimento agrupando los conjuntos de tuits según la ciudad de la que procedieran.

En esta aproximación, una vez se ha obtenido el vocabulario se procedió a crear los ficheros de entrenamiento del algoritmo de *SVM* implementado dentro de la librería *LibLINEAR*, tal y como se detalló en la sección 5.1.1 de este capítulo. El conjunto de entrenamiento estaba compuesto por 52 textos, uno por cada ciudad existente en el corpus. Cada uno de estos textos recogía todos los tuits emitidos desde cada una de las ciudades en cuestión.

Con el fichero de entrenamiento (*Twitter*) y los de evaluación (*20Minutos*) se crearon finalmente los modelos del sistema que permitieron evaluar dicho experimento.

Los resultados obtenidos se pueden ver en la tabla 5.15, donde en la primera fila se puede observar el año o años con los que se evaluó el sistema, mientras que en la primera columna se puede apreciar el tipo de términos que se utilizó en el conjunto de evaluación.

Los resultados muestran claramente como la mejor aproximación es la obtenida cuando se intentan clasificar geográficamente los artículos de noticias utilizando únicamente los topónimos existentes en éstas, lo cual

Tabla 5.15: Resultados de la validación cruzada de *SVM* utilizando el vocabulario y textos de los tuits agrupado por ciudad para entrenar y el corpus del diario *20Minutos* para evaluar.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	**37,64**	**38,86**	**36,17**	**36,59**	**36,64**
Sustantivos	31,02	31,22	28,76	28,98	29,21
Sustantivos + Adjetivos	30,78	31,07	28,63	28,72	29,01
Sin topónimos	2,11	2,10	3,34	3,72	3,32
Todo	3,18	3,29	4,51	4,93	4,50

pone de manifiesto como el factor común entre estos textos informales y los textos de las noticias son casi únicamente los topónimos, pese a no coincidir en numerosas ocasiones debido a las diferentes formas en las que éstos se escriben en lenguajes informales.

Si se eliminan los topónimos del corpus de noticias a evaluar, el rendimiento del sistema decae drásticamente, lo que reafirma la hipótesis expuesta en el párrafo anterior.

Algo similar sucede cuando se utilizan todos los términos de las noticias a evaluar. Aunque, en esta ocasión, es debido a la introducción de ruido en el sistema, dado que la proporción de topónimos con respecto al resto de categorías gramaticales es muy reducida.

En un término intermedio se encuentran las aproximaciones que hacen uso solamente de los sustantivos y los adjetivos. Éstas aproximaciones eliminan un gran número de características, y por ende de ruido, aunque no llegan a eliminar las suficientes como para obtener unos resultados parecidos a la aproximación que únicamente utiliza los topónimos de los textos a evaluar.

En la gráfica 5.5 se puedo observar una comparativa entre las aproximaciones que entrenaban con el conjunto de tuits agrupados por ciudad o lo separaban por usuario y ciudad, así como la que únicamente utilizaban los topónimos del conjunto de evaluación y la que utilizaba todos los términos.

Como se puede observar en dicha gráfica, los resultados son muy pobres cuando se utilizan las aproximaciones que separan los muestras de entrenamiento por usuario y ciudad. Además, separando las muestras de entrenamiento según los tuits emitidos por cada usuario en cada ciudad, hace que el sistema se ralentice considerablemente con respecto a la aproximación en la que se unificaban dichas muestras según la ciudad de procedencia de las mismas.

Por otro lado, la aproximación que hace uso de todos los términos en el conjunto de evaluación, debido a la disparidad de formato entre los textos formales (noticias del diario *20Minutos*) y los textos informales (tuits),

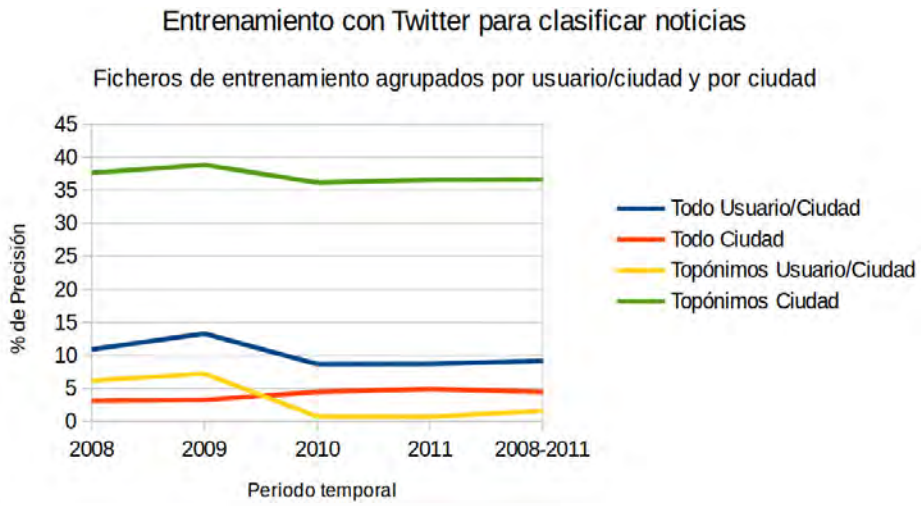


Figura 5.5: Precisión obtenida para los experimentos llevados a cabo en esta sección utilizando únicamente los topónimos o todos los términos, agrupando el conjunto de entrenamiento por ciudad o agrupándolo por ciudad y usuario.

tampoco logran unos buenos resultados. Esto conduce a la conclusión que cuando se intentan comparar dos corpus de distintas fuentes, siendo además estas fuentes de formalidad opuesta, las sutilezas del lenguaje no pueden ser utilizadas para poder realizar una clasificación geográfica correcta, siendo, por consiguiente, mucho más precisa la aproximación que hace uso únicamente de los topónimos del conjunto de evaluación.

En la figura 5.6 se comparan los resultados obtenidos en la mejor aproximación en la que se entrenaba mediante los textos de *Twitter* y en las que se entrenaba con textos de *Wikipedia*, con o sin artículos referenciados. Se puede apreciar cómo *Twitter* y *Wikipedia* en su versión más extensa, la que hace uso de los artículos referenciados, obtienen unos resultados casi idénticos pese a que sus textos contienen formalidades completamente distintas. Esto es debido a que tanto para los textos formales (*Wikipedia*), como para los informales (*Twitter*), los topónimos aparecen escritos de forma muy similar.

Selección de características mediante el conjunto de tuits, y entrenamiento procedente de los textos del diario *20Minutos*

Al igual que se realizó en los dos experimentos previos, con el vocabulario obtenido se procedió a crear los ficheros de entrenamiento, aunque en esta ocasión los textos utilizados para dicha tarea fueron los propios artículos del diario *20Minutos* mediante una validación cruzada.

El objetivo de este experimento es el de comprobar la utilidad de una fuente de textos informales (*Twitter*) como selector de características, al

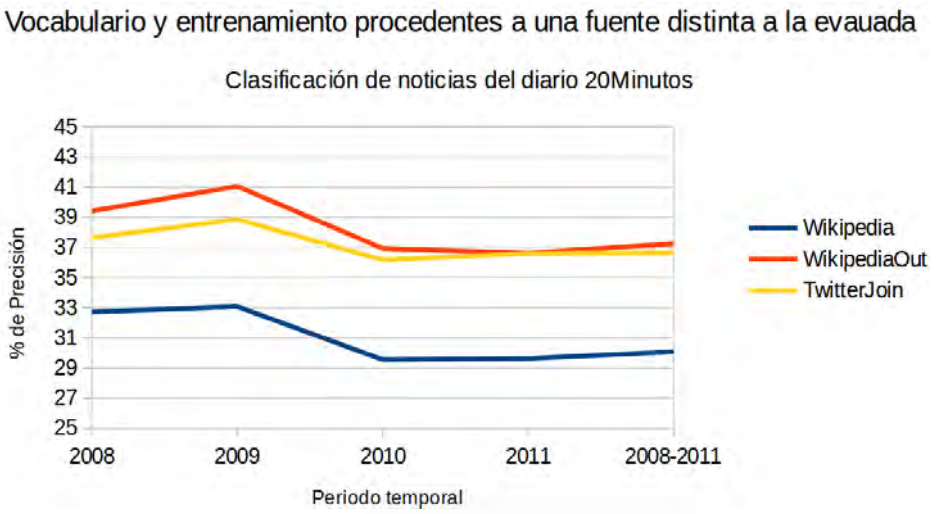


Figura 5.6: Comparación entre las ejecuciones de SVM llevadas a cabo con el vocabulario y el entrenamiento de textos formales (*Wikipedia*) o informales (*Twitter*) distintos a la fuente a clasificar (*20Minutos*).

igual que se hiciera con los textos formales (*Wikipedia*). En esta ocasión, el número de características utilizado fue de 2.080.857.

Con el fichero de entrenamiento y los de evaluación se crearon finalmente los modelos del sistema que permitieron evaluar dicho experimento.

Los resultados obtenidos se pueden ver en la tabla 5.16.

Tabla 5.16: Resultados de la validación cruzada de *SVM* utilizando el vocabulario y textos de los tuits y entrenando con los textos del diario *20Minutos*

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	51,58	52,92	52,53	52,52	52,46
Sustantivos	61,23	62,36	69,78	69,91	70,90
Sustantivos + Adjetivos	65,60	67,69	72,74	72,31	73,13
Sin topónimos	57,57	58,62	71,99	71,12	71,11
Todo	**69,03**	**70,03**	**77,83**	**77,12**	**77,49**

Los resultados muestran, al igual que sucediera con los experimentos análogos realizados con *Wikipedia*, cómo cuando se utiliza el corpus del propio *20Minutos*, los matices lingüísticos que aportan todos los términos ayudan a conseguir una mejora sustancial en la precisión, alrededor del 25 % más para los corpus más grandes. Este hecho se puede constatar al comprobar que incluso la aproximación que no hace uso de los topónimos,

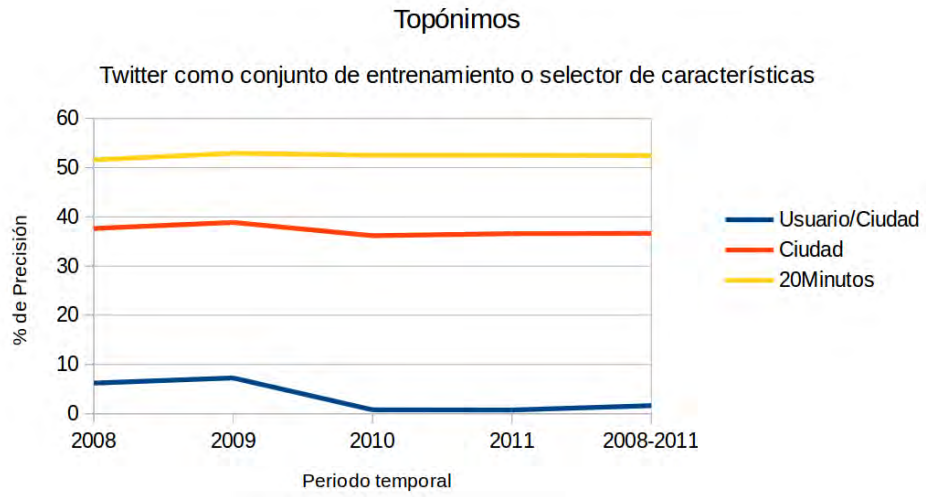


Figura 5.7: Precisión obtenida para los experimentos llevados a cabo en esta sección utilizando únicamente los topónimos del corpus de tuits.

vuelve a obtener mejores resultados que la que hace únicamente uso de estos topónimos, especialmente cuando los corpus son más extensos.

Al entrenar con corpus de entrenamiento más extensos, se consigue obtener un mayor número de dichos matices lingüísticos, haciendo que la clasificación mejore en más de 7 puntos absolutos (comparar los resultados de los corpus de los años con menor número de noticias, 2008 y 2009, con el de resto de años cuando se utilizan todos los términos).

Por el contrario, si únicamente se entrena con los topónimos, el ampliar el conjunto de entrenamiento no aporta ninguna mejora, incluso empeora ligeramente, debido a que no se aprovechan estos matices lingüísticos.

Comparación de aproximaciones

En las figuras 5.7 y 5.8 se pueden apreciar los resultados obtenidos en los 3 experimentos llevados a cabo en esta sección, donde se unificaron los tuits utilizados en el entrenamiento según el usuario y la ciudad de procedencia (*Usuario/Ciudad*), se unificaron los tuits utilizados en el conjunto de entrenamiento según la ciudad de procedencia sin importar el usuario emisor (*Ciudad*), se utilizó el conjunto de tuits como vocabulario que permitía seleccionar las características a tener en cuenta a la hora de entrenar con los propios textos del diario *20Minutos* para experimentar realizando una validación cruzada.

Así pues, en la figura 5.7 se muestran los 3 experimentos mentados utilizando sólamente los topónimos del corpus del diario *20Minutos*.

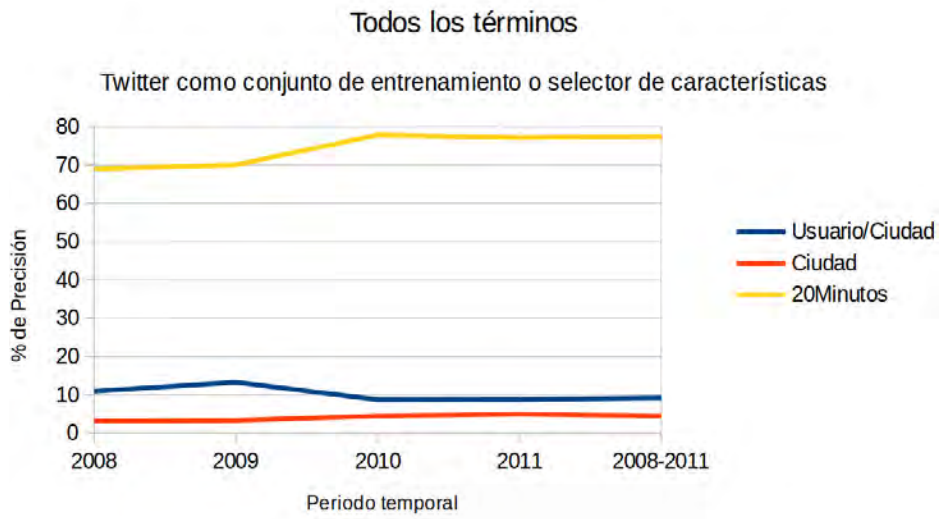


Figura 5.8: Precisión obtenida para los experimentos llevados a cabo en esta sección utilizando todos los términos del corpus de tuits.

En la figura 5.8 no se realizó ningún filtrado en los corpus por ninguna categoría gramatical, por lo que se utilizaron todos los términos que en ellos aparecían.

Lo primero que se aprecia en la figura es la mejora significativa obtenida cuando se utiliza el mismo conjunto de tuits para el entrenamiento pero agrupado por ciudad de procedencia (*Ciudad*).

Tal y como se mostró en sus respectivas tablas 5.14 y 5.15, se puede apreciar como la mejor aproximación, en cuanto a términos de precisión se refiere utilizando los textos de *Twitter* para entrenar, es la obtenida utilizando única y exclusivamente los topónimos de los textos que se pretenden clasificar, al igual que sucediera con los experimentos llevados a cabo con *Wikipedia* en la sección 5.3.2. Esto pone de manifiesto que cuando se utiliza una fuente de textos para el entrenamiento distinta a la que se pretende clasificar, bien sea esta fuente formal, bien informal, los términos más relevantes para realizar una clasificación geográfica son los topónimos, ya que se pierden los matices lingüísticos que se pueden encontrar dentro de textos de la misma fuente, los cuales ayudan enormemente a la detección del foco geográfico.

Por otro lado, cuando se utiliza la propia fuente que se pretende clasificar en el entrenamiento del sistema, *20Minutos* en este caso, los resultados son mejores cuanto mayor es el número de términos utilizados, tal y como se muestra en los resultados expuestos en la tabla 5.16, donde se aprecia cómo los resultados mejoran con corpus más extensos (2010, 2011 y 2008-2011) frente a los corpus más reducidos (2008 y 2009). Esto nos hace ver cuán

importantes son los matices lingüísticos existentes en los textos que son de la misma naturaleza pero de distinto origen geográfico.

Así pues, observando las dos figuras anteriores, se puede concluir que la mejor aproximación es aquella que utiliza para el entrenamiento todos los términos disponibles del conjunto de textos de la misma fuente que el que se pretende evaluar, *20Minutos* en este caso.

Comparación de aproximaciones de selección de características

En la figura 5.9 se muestran los resultados logrados en la selección de características con χ^2 con 54.183 características, artículos de las ciudades del corpus de *Wikipedia* con el mismo número de características, *Wikipedia* con los artículos de las ciudades más los referenciados en éstos con 1.074.174 características, y *Twitter* con 2.080.855 características.

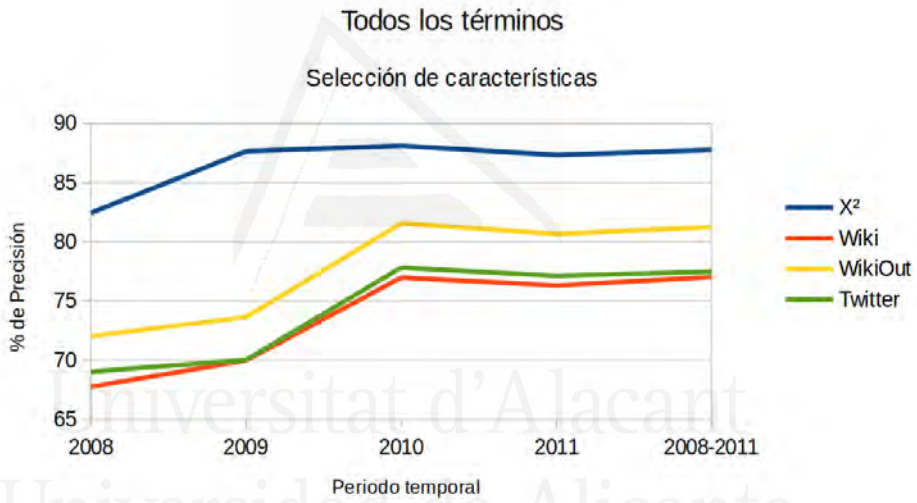


Figura 5.9: Precisión obtenida para los experimentos llevados a cabo con la selección de características mediante χ^2, artículos de las ciudades de *Wikipedia*, artículos de las ciudades de *Wikipedia* más los artículos referenciados en éstos y *Twitter*, teniendo en cuenta todas las categorías gramaticales y entrenando con el corpus del diario *20Minutos* haciendo una validación cruzada para su evaluación.

En esta figura se puede apreciar como χ^2 es cláramente la opción que mejores resultados obtiene, aunque, como ya se comentó previamente, hay que tener en cuenta que para esta aproximación hay que disponer previamente de un corpus los suficientemente extenso de textos de la misma fuente que se pretende clasificar, además del coste temporal que conlleva el cálculo de estas características mediante este algoritmo estadístico.

Centrando la atención en los recursos externos utilizados como selección de características, se puede apreciar como las 3 aproximaciones tienen

unos resultados muy similares, especialmente *Wikipedia* cuando utiliza únicamente los artículos de las ciudades del corpus, y *Twitter*, aunque cabe destacar que la implementación con *Twitter* resulta mucho más costosa en términos de tiempo y espacio, tal y como se puede deducir del número de características empleado en cada una de estas aproximaciones.

Así pues, se puede concluir que si se dispone de un corpus de noticias lo suficientemente extenso, la aproximación que hace uso de χ^2 es la más acertada.

Si, por el contrario, no se dispone de dicho corpus, queda reflejado como una aproximación en la que se utilizan artículos de otra fuente con una escritura formal como es la *Wikipedia*, da unos buenos resultados con un coste computacional bajo.

Por otro lado, la utilización de una fuente de texto completamente distinta a la que se quiere clasificar requiere de un gran volumen de estos textos para lograr unos resultados similares a los obtenidos con otra fuente con un nivel de formalidad más próximo a la que se está clasificando.

5.3.4. Combinación de corpus de entrenamiento

En esta sección se van a describir cómo se han combinado distintos corpus de entrenamiento para clasificar las noticias del corpus del diario *20Minutos*. El propósito de combinar dichos corpus es el de saber si se puede lograr una ayuda en la identificación del foco geográfico combinando textos para el entrenamiento del sistema de la propia fuente que se pretende clasificar con otras fuentes de distinta índole, tales como *Wikipedia* por el lado de las fuentes formales, o como *Twitter* por el lado de las informales, ya que ambos recursos son fácilmente accesibles y disponibles en numerosos idiomas.

Como se ha podido ver en las secciones previas, tal y como era de esperar, los textos que mejores resultados han dado al entrenar con ellos han sidos los del propio corpus que se pretendía clasificar geográficamente. En esta sección se mostrará lo que sucede cuando se le añaden otros textos de distintas fuentes para realizar el mismo cometido.

20Minutos y *Wikipedia*

El primero de los experimentos que se plantea en este apartado es el de la combinación de fuentes de textos formales, las cuales, en este caso se tratan de las propias noticias del corpus del diario *20Minutos* y los artículos de las ciudades de *Wikipedia* con los que se ha trabajado previamente.

La manera de proceder es análoga a la utilizada en las secciones 5.3.1 y 5.3.2, en las cuales se describía el uso de los corpus de *20Minutos* y *Wikipedia* para la clasificación de noticias procedentes del primero.

Una vez más, para los experimentos con la combinación de estas dos fuentes de texto formales se procedió a trabajar teniendo en cuenta las

distintas categorías gramaticales de los términos que aparecían en dichos textos. Al igual que en los experimentos previos, se trabajó con las siguientes cinco categorias: *topónimos*, *sustantivos*, *sustantivos+adjetivos*, *todos los términos menos los topónimos* y *todos los términos*.

Para la obtención de dichas categorías gramaticales se aprovecharon los ficheros creados en los experimentos que se llevaron a cabo en los entrenamientos con artículos del diario *20Minutos* y de *Wikipedia* por separado, tal y como se describió con anterioridad en sus respectivos experimentos (ver secciones 5.3.1 y 5.3.2, respectivamente).

Puesto que ya se tenía la frecuencia de los términos de cada uno de los artículos por los experimentos anteriores, se procedió a la obtención del vocabulario y la creación del conjunto de entrenamiento utilizando tanto las noticias del diario *20Minutos* como los artículos que representaban a cada una de las ciudades del corpus en *Wikipedia*. Estos artículos de *Wikipedia* fueron tratados como si fueran una noticia más del corpus de *20Minutos*, estando clasificado dentro de la ciudad a la que pertenecían y siendo un texto varias veces más extenso que el de las noticias del diario.

Pese a que los textos de los artículos de *Wikipedia* eran varias veces mayores que los artículos del diario, estos primeros quedaban muy diluidos entre el gran número de noticias obtenidas del diario. Por ello se decidió realizar otros experimentos que previnieran dicha dilución.

Una prueba que se realizó fue la de ponderar los textos según la fuente de procedencia de los mismos. Es decir, puesto que se pretende clasificar geográficamente textos del diario *20Minutos* utilizando textos del propio diario y de *Wikipedia*, se otorgó un peso a los conjuntos de entrenamiento dependiendo de la procedencia de los mismos. Hay que resaltar, una vez más, que dichas pruebas, con el fin de que fueran efectivas, fueron costosas temporalmente hablando al introducir una gran cantidad de textos. Por ello, tras probar con diversos pesos para las fuentes citadas y comprobar que no se conseguía mejorar los resultados obtenidos mediante la inclusión de los artículos de *Wikipedia* como una noticia más en el conjunto de entrenamiento, se desistió en esta aproximación.

También se probó con la incursión de textos de los enlaces salientes, tal y como se verá en la próxima sección.

Una vez obtenido el vocabulario del corpus se procedió a la creación de los ficheros de entrenamiento que se envían al algoritmo de *SVM*.

Por otro lado, como en cada ejecución de *SVM*, se crearon los pertinentes ficheros de evaluación que, una vez más, fueron los mismos que en las ejecuciones anteriores, ya que se sigue intentando ubicar geográficamente cada uno de los artículos del diario *20Minutos* que se utilizaron en los experimentos anteriores. Al igual que entonces, estos ficheros de evaluación estaban compuestos por los términos y frecuencias de la categoría gramatical, periodo temporal y partición de cada uno de los artículos del diario *20Minutos* que coincidían con los términos del vocabulario obtenido.

Con el modelo de entrenamiento y el fichero de evaluación se procedió a ejecutar el algoritmo de *SVM* para obtener la clasificación geográfica de cada una de las noticias del conjunto de evaluación, realizando de nuevo una validación cruzada con diez particiones para obtener la precisión final.

Los resultados obtenidos se muestran en la tabla 5.17. En dicha tabla se puede apreciar como, una vez más, tal y como sucediera en los experimentos en los que se utilizaba únicamente el corpus del diario *20Minutos* para el entrenamiento, los mejores resultados vienen cuando se emplean todos los términos del corpus.

Tabla 5.17: Resultados de la validación cruzada de *SVM* entrenando con el corpus de noticias del diario *20Minutos* y los artículos de las ciudades de *Wikipedia*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	58,18	59,91	63,36	62,72	63,34
Sustantivos	68,75	69,59	81,26	80,91	82,15
Adjetivos + Sustantivos	71,62	73,24	82,13	81,49	82,68
Sin topónimos	62,63	64,3	79,57	78,87	79,69
Todo	**73,46**	**74,97**	**84,09**	**83,36**	**84,42**

Si se comparan los resultados obtenidos con esta aproximación, con los obtenidos en la aproximación en la que únicamente se utilizaba el corpus del diario *20Minutos* (ver tabla 5.5), se puede apreciar con los resultados son prácticamente idénticos. Esto es debido a que, como ya se ha comentado, los 52 artículos de *Wikipedia* han quedado muy diluidos entre el gran número de textos del diario por cada ciudad.

Por esta razón, se ha realizado otra aproximación en la que además de los textos de los 52 artículos representativos de las 52 localidades del corpus, se añadieron los textos de los artículos referenciados en estos primeros (*outlinks*).

20Minutos y *Wikipedia* con enlaces salientes

En este experimento se ha realizado una extensión del anterior con la única salvedad que se une al conjunto de textos de entrenamiento los textos de los artículos de la *Wikipedia* procedentes de cada uno de los enlaces salientes de cada artículo de cada ciudad.

Una vez más se ha experimentado con las cinco distintas categorías gramaticales empleadas en los experimentos previos.

Dado que desde los experimentos previos ya se tenían las frecuencias de los términos de cada uno de los artículos, tanto del diario *20Minutos* como

de los artículos de *Wikipedia*, tanto de las ciudades como de los referenciados en éstos, se procedió a la obtención del vocabulario.

El conjunto de entrenamiento se creó con los artículos del diario *20Minutos* y los artículos de *Wikipedia* de los enlaces salientes mencionados juntos a los de las propias ciudades.

Dicho conjunto sirvió de entrenamiento a un algoritmo de *SVM*.

Una vez creado el modelo de entrenamiento con los nuevos artículos, junto a los ficheros de evaluación obtenidos de los experimentos previos, se procedió a ejecutar el algoritmo de *SVM* para obtener la clasificación geográfica de cada una de las noticias del conjunto de evaluación.

Los resultados de este experimentos se pueden observar en la tabla 5.18. Observando estos resultados, se puede apreciar un mejor funcionamiento cuando se utilizan únicamente los topónimos en los corpus más reducidos (2008 y 2009). Esto es debido a que el factor común más evidente entre los corpus del diario y *Wikipedia* son los nombres de lugar, y al tener un menor peso el corpus del diario *20Minutos* en los años 2008 y 2009, estos términos cobran una mayor relevancia.

Por otro lado, se puede apreciar cómo conforme se trabaja con corpus más voluminosos, los textos del diario *20Minutos* cobran más relevancia, y por ende los matices lingüísticos existentes en estos textos, llegando incluso a superar en los dos corpus más extensos (2011 y el que agrupa todos los años) la aproximación que no hace uso de los topónimos a la que hace un uso exclusivo de éstos.

Tabla 5.18: Resultados de la validación cruzada de *SVM* entrenando con el corpus de noticias del diario *20Minutos* y los artículos de las ciudades de *Wikipedia* y sus enlaces salientes, evaluando sobre el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	**57,11**	**57,91**	61,81	61,45	62,42
Sustantivos	47,07	43,20	67,84	69,52	73,01
Sustantivos + Adjetivos	48,48	44,58	**72,30**	**73,73**	**76,60**
Sin topónimos	36,19	36,23	62,52	65,27	67,03
Todo	40,72	40,35	69,03	70,13	74,09

En la figura 5.10 se muestra una gráfica, utilizando todos los términos de los corpus, comparando los resultados obtenidos con las aproximaciones realizadas utilizando únicamente el corpus del diario *20Minutos* como vocabulario y entrenamiento, *20Minutos* y los artículos de las ciudades del corpus de *Wikipedia* como vocabulario y entrenamiento, y *20Minutos* y los artículos de las ciudades del corpus de *Wikipedia* junto a los artículos referenciados en éstos como vocabulario y entrenamiento.

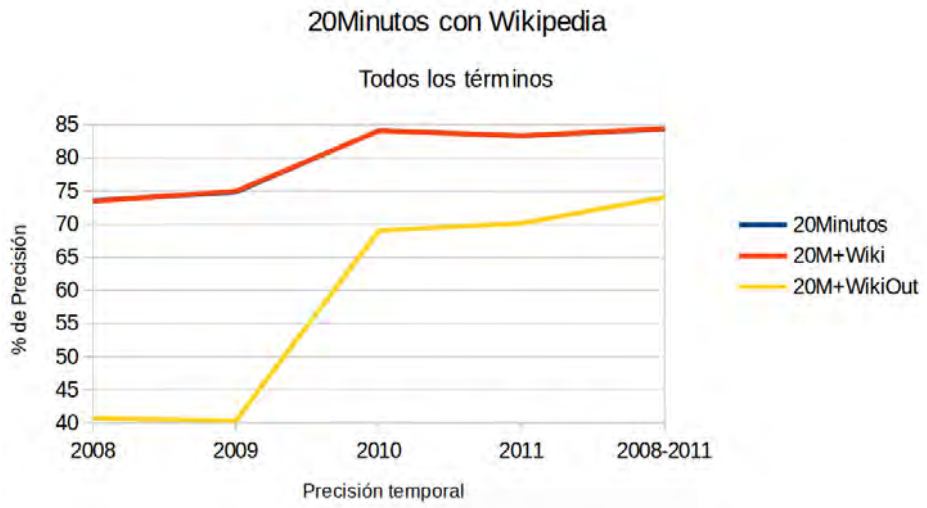


Figura 5.10: Precisión obtenida para los experimentos llevados a cabo utilizando el corpus del diario *20Minutos* únicamente o en combinación con el de *Wikipedia* (*20M+Wiki*) y *Wikipedia* con sus enlaces salientes (*20M+WikiOut*).

Observando la figura 5.10 se puede apreciar como la combinación de *20Minutos* más los artículos de las ciudades de *Wikipedia* hace que esta última se diluya ante el mayor volumen de los textos de diario, por lo que se obtienen unos resultados prácticamente idénticos a los logrados cuando no se añadían los textos de *Wikipedia*.

Al añadir los textos de los artículos referenciados en los artículos de las ciudades de *Wikipedia*, se observa como éstos introducen ruido, especialmente cuando el corpus del diario *20Miutos* no es tan voluminoso (años 2008 y 2009), e incluso cuando el diario adquiere mayor relevancia (años 2010, 2011 y todos los años), *Wikipedia* continúa penalizando la clasificación.

20Minutos y *Twitter*

En esta ocasión, el experimento llevado a cabo es similar al anterior pero en vez de utilizar junto con el propio corpus del *20Minutos* una fuente de texto formal como la de *Wikipedia*, se utilizó una informal como es *Twitter*, para complementar al anterior.

Para la combinación de ambos corpus se realizaron dos aproximaciones, una primera en la que se juntaban todos los tuits según la ciudad en la que fueron emitidos, es decir, que se unificaban todos los tuits en un único documento por ciudad dando como resultado 52 documentos, y una segunda en la que se hizo de manera análoga al experimento anterior con *Wikipedia*

y sus enlaces salientes, es decir, cada conjunto de tuits emitidos por cada usuario en cada ciudad fue tratado como si fuese una noticia más del corpus del diario *20Minutos* para el entrenamiento del sistema.

Tal y como se explicó en la sección en la que se realizaron los experimentos utilizando únicamente *Twitter* como corpus para clasificar noticias del diario *20Minutos* (sección 5.3.3), debido al alto índice de informalidad del lenguaje expresado en los tuits, las herramientas existentes hasta la fecha no son capaces de extraer con gran precisión las categorías gramaticales de los términos existentes en éstos. Por este motivo no se ha realizado dicha clasificación y por ende se utilizan todos los términos de los tuits que quedan tras realizar la limpieza explicada en la sección 4.2, en cada uno de los experimentos realizados con este corpus.

Los textos de evaluación vuelven a ser los mismos que en el resto de experimentos, así como las 10 particiones utilizadas para realizar la validación cruzada de los experimentos.

Con los ficheros de entrenamiento ya creados se procede a crear el modelo que permite identificar el foco geográfico de los distintos artículos de los conjuntos de evaluación, enviando así este modelo y el fichero de evaluación al sistema para que clasifique geográficamente los textos.

Los resultados obtenidos en la aproximación en la que se agrupaban todos los tuits según su localidad de procedencia, se pueden observar en la tabla 5.19.

Tabla 5.19: Resultados de la validación cruzada de *SVM* entrenando con el corpus de noticias del diario *20Minutos* y los mensajes de *Twitter* emitidos en las mismas ciudades de las noticias y agrupados como un único texto, evaluando el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	54,41	40,91	63,05	62,51	73,40
Sustantivos	56,02	55,02	82,74	82,15	83,96
Sustantivos + Adjetivos	57,60	53,84	82,53	82,33	**84,29**
Sin topónimos	44,99	43,09	73,04	75,45	73,62
Todo	**64,12**	**62,19**	**81,93**	**83,32**	82,89

Una vez más, los resultados muestran cómo al utilizar el propio corpus a evaluar como entrenamiento (*20Minutos*), la mayor precisión es obtenida cuando se utilizan todos los términos existentes en el corpus.

El sistema funciona mejor cuanto más grande es el corpus del propio diario *20Minutos*, ya que éste obtiene así un mayor peso en el entrenamiento del sistema.

La diferencia al utilizar todos los términos con respecto a la utilización única de topónimos es considerable. Incluso, cuando se utilizan todos los

términos menos los topónimos con corpus de mayor tamaño, el rendimiento del sistema es mucho mayor que con la utilización única de los topónimos.

En la tabla 5.20 se muestran los resultados obtenidos cuando se hace el mismo experimento pero separando los conjuntos de tuits por usuario y ciudad para pasarselos al sistema como muestras independientes de entrenamiento para cada ciudad.

Tabla 5.20: Resultados de la validación cruzada de *SVM* entrenando con el corpus de noticias del diario *20Minutos* y los mensajes de *Twitter* emitidos en las mismas ciudades de las noticias y separados por usuario/ciudad, evaluando el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	58,13	57,12	61,77	61,94	63,18
Sustantivos	71,03	68,93	82,57	83,00	83,95
Sustantivos + Adjetivos	**71,98**	**71,13**	83,05	81,95	83,77
Sin topónimos	61,40	60,51	78,69	78,87	80,42
Todo	70,57	69,83	**83,89**	**82,74**	**84,70**

Las aproximaciones que utilizan únicamente los sustantivos, los sustantivos y los adjetivos, y la que utiliza todos los términos obtienen unos resultados muy similares para los corpus de entrenamiento más reducidos (2008 y 2009), mientras que para los corpus de entrenamiento de mayor tamaño se aprecia cómo el sistema clasifica mejor con todos los términos.

Por otro lado, la aproximación que hace uso únicamente de los topónimos es la que peores resultados ha obtenido, viéndose incluso superada por la que omite estos términos, demostrándose una vez más la gran importancia de todos los matices lingüísticos a la hora de clasificar geográficamente los textos dados.

Si se separan por usuario/ciudad los conjuntos de entrenamiento de *Twitter*, hace que dicha fuente tenga mucho menos peso a la hora de entrenar el sistema debido a que la mayoría de muestras de entrenamiento carecen de los términos que ayudan a clasificar mejor geográficamente los textos, predominando así los textos más extensos del diario *20Minutos* y por ende imponiendo su mejor funcionamiento cuando se utilizan todos los términos disponibles en los artículos del diario.

En la figura 5.11 se puede apreciar visualmente el experimento llevado a cabo utilizando todos los términos de los corpus de entrenamiento. En dicha figura se comparan las aproximaciones en las que se utilizaba únicamente el corpus del diario *20Minutos* como vocabulario y entrenamiento (*20Minutos*), y los dos experimentos llevados a cabo en esta sección en los que se unía al corpus entrenamiento del diario *20Minutos* los textos de *Twitter*

unificados por ciudad (*20M+TwJ*) y unificados por usuario en cada ciudad (*20M+TwS*).

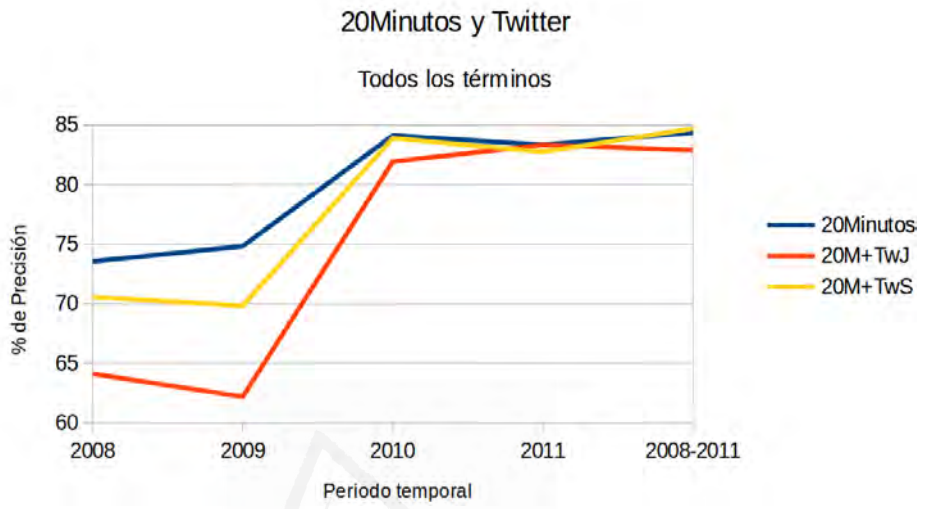


Figura 5.11: Precisión obtenida para los experimentos llevados a cabo utilizando el corpus de *Twitter* y *20Minutos* con todos los términos de ambos corpus tanto para el vocabulario como para el entrenamiento. *20Minutos* indica que se utiliza únicamente el corpus del diario, *20M+TJ* indica que se han utilizado los corpus del diario *20Minutos* y de *Twitter* agrupando los tuits por ciudad, y *20M+TS* indica que se han utilizado los corpus del diario *20Minutos* y de *Twitter* separando los conjuntos de tuits por usuario/ciudad.

Tal y como se puede apreciar en la figura 5.11, cuando se entrena con corpus de noticias no muy extensos (2008 y 2009), la influencia de los tuits es más significativa, lo que hace que se introduzca más ruido.

En la combinación del corpus de tuits precisamente con el de noticias con menor volumen es donde mejor se puede apreciar cómo el rendimiento recae cuando se agrupan todos los tuits por ciudad.

Cuando el volumen del corpus del diario aumenta, el corpus de *Twitter* pierde protagonismos, obteniéndose así unos resultados muy similares a cuando no se utiliza dicho corpus.

Por último, en la figura 5.12, se muestran las mejores aproximaciones expuestas en esta sección en donde se combinan el corpus del diario *20Minutos* con el de *Wikipedia*, con y sin los textos de los artículos referenciados en los artículos de las ciudades, y *Twitter* con los textos separados por usuario y ciudad.

Como se puede apreciar en la figura 5.12, se puede volver a realizar una división de los resultados acorde al volumen del corpus de entrenamiento del diario *20Minutos*, es decir, unos análisis de los resultados para un corpus más reducido, años 2008 y 2009, y otro para los más amplios.

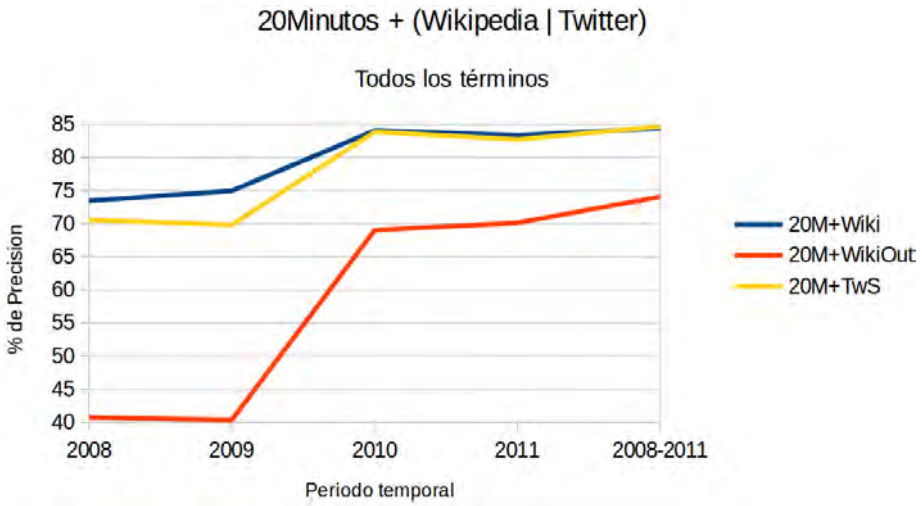


Figura 5.12: Precisión obtenida para los experimentos llevados a cabo utilizando la combinación del corpus del diario *20Minutos* con *Wikipedia* o *Twitter*. *20M+Wiki* es la aproximación en la que se utiliza *Wikipedia* y *20Minutos* para obtener el vocabulario y entrenar al sistema, *20M+WikiOut* es la aproximación que utiliza tanto *20Minutos* como *Wikipedia* con los enlaces referenciados para obtener el vocabulario y entrenar al sistema, y *20M+TwS* es la aproximación que utiliza tanto *20Minutos* como *Twitter* para obtener el vocabulario y entrena con ambos separando los tuits por usuario/ciudad.

20Minutos, *Wikipedia* y *Twitter*

Finalmente, como última prueba en este bloque de experimentos sobre textos formales, se va a proceder a realizar un experimento utilizando como conjunto de entrenamiento los corpus previamente descritos del diario *20Minutos*, *Wikipedia* y la red social *Twitter*, con el fin de dilucidar el foco geográfico de una serie de artículos periodísticos procedentes del diario *20Minutos*.

En esta ocasión, debido al gran volumen que tienen los corpus con la consiguiente demora temporal, se ha optado por utilizar únicamente el conjunto de artículos de *Wikipedia* de las ciudades del corpus, no utilizando así el resto de textos de los artículos de los enlaces salientes de estos artículos originales.

Por otro lado, dado que los resultados en los experimentos previos dieron mejores resultados con el corpus de tuits separados por usuario y ciudad (ver sección 5.3.4), se optó por dejarlos separados para estos experimentos.

Así pues, se ha utilizado la configuración de los experimentos previos para la combinación de los tres corpus, siguiendo la división de años y

evaluación previamente mencionadas, aplicando una validación cruzada de 10 particiones.

Los resultados obtenidos con la combinación de los tres corpus, en los que se mantenían separados los tuits por usuario/ciudad y se utilizaban solamente los artículos de *Wikipedia* pertenecientes a las ciudades evaluadas en el corpus, se pueden ver en la tabla 5.21.

Tabla 5.21: Resultados de la validación cruzada de *SVM* entrenando con el corpus de noticias del diario *20Minutos*, los artículos de las ciudades de *Wikipedia* y los mensajes de *Twitter* emitidos en las mismas ciudades de las noticias y separados por usuario/ciudad, evaluando el corpus del diario *20Minutos*.

	2008	**2009**	**2010**	**2011**	**2008-2011**
Topónimos	57,95	57,19	61,69	62,68	63,21
Sustantivos	71,90	68,75	82,58	82,03	83,74
Sustantivos + Adjetivos	**72,00**	**71,10**	83,00	82,12	83,78
Sin topónimos	61,62	60,73	78,77	78,76	80,38
Todo	70,53	69,73	**84,08**	**82,95**	**84,64**

Estos resultados son muy similares a los obtenidos en los experimentos previos donde se utilizaban bien el corpus de *Wikipedia*, bien el de *Twitter* en combinación con el del propio diario. Una vez más, los textos más numerosos o extensos del corpus del diario *20Minutos* provocan que el sistema clasifique de una manera similar a cuando se emplea únicamente esta fuente como entrenamiento, ya que en el experimento mostrado no se ha ponderado el peso de las otras dos fuentes.

Dicha ponderación resulta complicada de realizar debido al pobre rendimiento que obtenía el sistema cuando se utiliza estas otras fuentes. Pese a ello, tal y como ya se ha comentado previamente, se realizaron experimentos en los que se jugó con distintas ponderaciones para cada uno de los corpus. Dichos experimentos no lograron unos buenos resultados y por ello se decidió adoptar esta aproximación.

En la figura 5.13 se puede apreciar una comparativa entre la aproximación aquí expuesta y las que utilizaba *Wikipedia* para obtener el vocabulario y *20Minutos* para entrenar, y la que utilizaba tanto *Twitter* como *20Minutos* para obtener el vocabulario y entrenar.

En dicha gráfica también se valorar, una vez más, cómo los resultados se aproximan a medida que el corpus de noticias con el que se entrena es más grande, lo cual refleja el mayor peso que dicho corpus tiene a la hora de detectar el foco geográfico de las noticias al ser de la misma naturaleza.

En la gráfica se observa cómo los resultados se agrupan en dos para los años menos extensos del corpus del diario *20Minutos* (2008 y 2009). Por

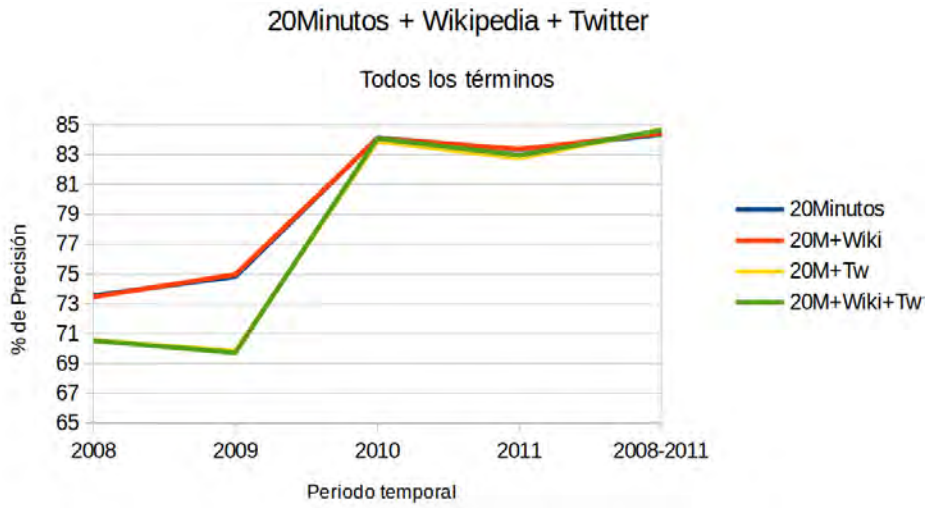


Figura 5.13: Precisión obtenida para los experimentos llevados a cabo utilizando la combinación del corpus del diario *20Minutos* con *Wikipedia* y con *Twitter*. *20M+Wiki* es la aproximación en la que se utiliza *Wikipedia* para obtener el vocabulario y *20Minutos* para entrenar al sistema, *20M+Tw* es la aproximación que utiliza tanto *20Minutos* como *Twitter* para obtener el vocabulario y entrena con ambos, separando los tuits por usuario/ciudad, y *20M+Wiki+Tw* es la aproximación en la que se emplean los corpus de las 3 fuentes utilizadas hasta hora para vocabulario y para entrenar, es decir, el último experimento llevado a cabo en esta sección.

un lado se obtienen unos resultados prácticamente calcados en la utilización de *20Minutos* en solitario y de la combinación de éste con los artículos de *Wikipedia* de las ciudades. Esto es debido, como ya se indicó previamente, al escaso peso que dicho medio tiene en este entrenamiento.

Por otro lado se aprecia cómo la aproximación que hace uso del corpus de *Twitter* en conjunto con el del diario, así como la que hace uso de todas las fuentes expuestas hasta ahora, obtienen unos resultados prácticamente idénticos. Una vez más, esto es debido a que los textos de la red social, aunque no muy extensos, sí que son muy numerosos, lo cual hace que adquieran una mayor relevancia especialmente cuando el corpus del diario no es muy extenso, tal y como se comentó en la sección anterior.

A la luz de estos resultados se puede concluir que:

1. El mejor de los escenarios es aquel que permite entrenar el sistema con textos de la misma fuente que se pretende clasificar.

2. La precisión del sistema se incrementa si se hace una selección previa de las características que ayudan a clasificar geográficamente mejor.

3. La utilización de todas las categorías gramaticales, y no simplemente los topónimos, es fundamental para obtener una precisión mucho mayor.

4. La utilización de otras fuentes de textos ajenas a las que se quieren clasificar, bien sean estas formales o informales, pueden dar buenos resultados cuando no se tienen textos previamente etiquetados de la fuente que se pretende clasificar.

5.4. Identificación del foco geográfico en textos informales

Tal y como se ha comentado en la introducción de este capítulo, los textos informales que se han utilizado en la experimentación han sido un conjunto de tuits georeferenciados procedentes de las 50 capitales de provincia del estado español más sus dos ciudades autónomas, Ceuta y Melilla (ver sección 4.2). La elección de esta red social ha sido debida a su gran número de usuarios y de mensajes (tuits) emitidos por éstos diariamente, pudiéndose obtener la localización a nivel de ciudad de cada uno de estos tuits, lo cual facilita la labor a la hora de entrenar y evaluar el sistema de geolocalización.

Las fechas de los tuits recogidos van desde 20 de abril de 2013 al 10 de junio de ese mismo año, pudiéndose ver los datos de dicho corpus en la tabla 4.2 del capítulo previo.

Así pues, el objetivo de los experimentos llevados a cabo en esta sección es el de geolocalizar a los usuarios de *Twitter* mediante, única y exclusivamente, los mensajes vertidos por éstos en la red social.

Debido a que los usuarios de *Twitter* de los cuales se han obtenido los tuits tienen un nivel de actividad muy dispar, se realizaron experimentos agrupando los usuarios según el número de tuits que habían emitido en cada localización, dando como resultado los siguientes grupos:

- Usuarios con una actividad baja en *Twitter*: entre 1 y 10 tuits. Se agruparon solamente los tuits de los usuarios que no emitieron más de 10 mensajes por ciudad, es decir, que en cada fichero del sistema había como mucho 10 tuits emitidos por un usuario dado en una ciudad.

- Usuarios con una actividad media en *Twitter*: entre 11 y 99 tuits. Se agruparon solamente los tuits de los usuarios que emitieron entre 11 y 99 mensajes por ciudad, es decir, que en cada fichero del sistema había como mínimo 11 tuits y como mucho 99 tuits emitidos por un usuario dado en una ciudad.

- Usuarios con una actividad alta en *Twitter*: cien o más tuits. Se agruparon solamente los tuits de los usuarios que emitieron al menos 100 mensajes por ciudad, es decir, que en cada fichero del sistema

había como mínimo 100 tuits emitidos por un usuario dado en una ciudad.

- Todo. No se discriminó por el número de tuits emitidos por cada usuario en cada ciudad, es decir, se utilizó el corpus de tuits entero.

Con esta división del corpus se puede analizar cómo afecta el nivel de actividad de los usuarios (la cantidad de texto disponible por cada usuario) a la hora de geolocalizar a dichos usuarios.

Esta división del corpus no sólo fue útil para estudiar la respuesta del sistema segúna el nivel de actividad de los usuarios en la red social. Puesto que el número de términos diferentes (características) del corpus de *Twitter* utilizado en esta tesis es varias veces mayor que el utilizado con los experimentos llevados a cabo en la sección 5.3.1 con los experimentos del diario *20Minutos* (2.080.857 frente a 426.200), el coste temporal de cada ejecución resultaba muy elevado. Mediante la utilización de las particiones según el nivel de actividad de los usuarios previamente mencionadas, se consiguió disminuir dicho coste drásticamente.

Así pues, el número de conjuntos de tuits (la suma de usuarios distintos en cada ciudad) del corpus utilizado según la actividad de los usuarios en cada ciudad es el mostrado en la tabla 5.22.

Tabla 5.22: Número de usuarios que han tuiteado desde las distintas ciudades analizadas con respecto al nivel de actividad de dichos usuarios.

	Número de usuarios
Baja	145.303
Media	49.424
Alta	10.162
Todo	204.889

Cada conjunto de tuits expuesto en la tabla 5.22 representa a un usuario distinto que ha tuiteado en una ciudad dada. Cada uno de dichos usuarios representará una instancia de entrenamiento o evaluación del sistema.

En esta sección se van a mostrar experimentos agrupados en los siguientes cuatro grupos:

1. Entrenamiento con textos de la misma fuente que se pretende clasificar, es decir, con textos del propio *Twitter*.

2. Entrenamiento con una fuente de textos formal, donde se utilizará de nuevo *Wikipedia* para acometer dicha tarea.

3. Entrenamiento con una fuente de textos informal distinta a la que se pretende clasificar, donde se utilizará en esta ocasión *Flickr* para llevar a cabo esta tarea.

4. Combinación de distintas fuentes, donde se utilizaran las fuentes de textos de los 3 puntos anteriores con distintas combinaciones.

5.4.1. Entrenamiento con textos de *Twitter*

En esta sección se van a realizar experimentos utilizando como corpus únicamente los textos recogidos en la red social *Twitter* (ver sección 4.2).

Puesto que la naturaleza de los textos que se pueden encontrar en los tuits es completamente distintas a los de una fuente formal, como pueden ser los de un periódico como el expuesto en la sección anterior, se han vuelto a realizar experimentos con *SVM* y modelos de lenguaje con el fin de comprobar cuál era la aproximación más efectiva para este tipo de textos.

Para dichas aproximaciones se utilizaron los conjuntos de tuits agrupados por usuario y ciudad, es decir, se unificaron todos los tuits que un usuario dado había realizado en una ciudad para crear un conjunto de tuits del cual se pretendía averiguar su procedencia geográfica.

A su vez, también se comprobó si era más efectiva la aproximación que utilizaba los conjuntos de tuits separados por usuario y ciudad, o la que agrupaba todos los tuits por cada ciudad del sistema.

Dado que en esta ocasión el corpus a tratar no tenía divisiones según la categoría gramatical ni el periodo temporal al que pertenecía, debido a la baja precisión de los sistemas de clasificación de entidades gramaticales en textos informales, tal y como ya se ha comentado, el primer experimento que se realizó fue una simple ejecución con los tres[8] "subcorpus" mencionados creados según el nivel de actividad de los usuarios, teniendo únicamente en cuenta si el entrenamiento se realizaba con los conjuntos de tuits agrupados por ciudad o separados por ciudad/usuario.

Para la corroboración de los experimentos se volvió a recurrir a la validación cruzada, por lo que se dividió el corpus aleatoriamente en partes iguales para dicho fin. Dichos experimentos se han realizado creando los conjuntos de entrenamiento con 9/10 partes del corpus evaluado, y el 100 % de los no evaluados. Es decir, si se quiere determinar el foco geográfico de los usuarios de *Twitter* que tienen una baja actividad, se dividiría el corpus de usuarios de baja actividad en 10 grupos para hacer una validación cruzada, y se utilizaría 1 parte para evaluar y 9 partes, junto al 100 % de los conjuntos de los grupos de media y alta actividad, para crear los modelos de lenguaje.

[8] no se realizaron experimentos con el corpus donde se utilizan todos los conjuntos de tuits indistintamente (sin filtrar por el nivel de actividad de los usuarios) debido al gran coste computacional y temporal que supone para *SVM* tal cantidad de características, y que con este experimento simplemente se pretendía saber si era mejor una aproximación que utilizara *SVM* con los textos de entrenamiento agrupados por ciudad o por usuario/ciudad para poder compararlo posteriormente con la aproximación que utilizaba modelos de lenguaje.

SVM

Para el entrenamiento necesario con esta técnica de aprendizaje automático, en el caso de la aproximación que entrenaba con los conjuntos de tuits agrupados por ciudad, se procedió a unir en un único fichero todos los textos de las particiones de entrenamiento (todas menos una), dejando la partición restante para efectuar la evaluación del sistema.

Por otro lado, cuando se trataba de la aproximación que separaba los tuits por usuario/ciudad, se entrenó el sistema con cada uno de estos conjuntos de tuits pertenecientes a las particiones de entrenamiento, dejando igualmente la partición restante para efectuar la evaluación del sistema.

Los resultados obtenidos mediante la validación cruzada se pueden observar en la tabla 5.23, donde las filas indican si los tuits estaban agrupados por usuario/ciudad o separados, mientras que las columnas señalan el grado de actividad de éstos usuarios. Se destacan los resultados con mayor precisión para cada grupo.

Tabla 5.23: Resultados de la validación cruzada de *SVM* entrenando con el corpus de tuits para evaluar los propios tuits.

	Baja	**Media**	**Alta**
Agrupado	8,65	3,36	1,80
Separado	**40,86**	**52,36**	**60,33**

Claramente, los resultados obtenidos muestran una amplia mejoría cuando se utiliza la aproximación separando los conjuntos de tuits por usuario/ciudad, aunque cabe destacar que esta aproximación requiere de un mayor coste temporal.

A la luz de los resultados obtenidos se procedió a realizar una aproximación utilizando únicamente todo el corpus de tuits separados por usuario/ciudad, sin discernir entre el nivel de actividad de los usuarios. La precisión obtenida una vez realizada la media de la validación cruzada es del **48,90 %**.

Otro aspecto a resaltar es que cuando se utiliza el conjunto de tuits agrupados, en contra de lo que era de esperar, los resultados empeoran cuanto mayor es el índice de actividad de los usuarios. Esto es debido a la disparidad que hay en el lenguaje de los distintos usuarios, lo cual, al unir todos los textos a nivel de ciudad como si hubiesen sido producidos por una única persona o fuente, hace que se introduzca una gran cantidad de ruido.

Por otro lado, según se puede observar en la tabla 5.23, cuando el conjunto de entrenamiento separa las muestras por usuario/ciudad, el sistema incremente considerablemente su precisión conforme dichas muestras son más extensas, ya que el sistema tiene más texto en la muestra de evaluación para poder discernir la procedencia geográfica de la mismas.

En esta ocasión, ya que en el entrenamiento se utilizaron muestras separadas por usuario/ciudad, y los sistemas de *SVM* funcionan mejor cuantas más muestras de entrenamiento tengan aunque éstas sean más reducidas, el sistema fue capaz de encontrar una mayor similitud entre algunas de estas muestras de entrenamiento al ser haber usuarios que empleaban un lenguaje muy similar al de la muestra que se pretendía clasificar, esquivando de esta forma el ruido introducido por otros usuarios.

Modelos de lenguaje

En esta sección se va a replicar el experimento anterior pero en lugar de utilizar *SVM* se van a utilizar los modelos de lenguaje descritos en la sección 5.1.2.

Para ello, se ha utilizado el mismo corpus con las mismas divisiones realizadas en el experimento anterior. Para la realización de este experimento también se tuvo en cuenta a la hora de crear el fichero de entrenamiento del sistema, el dejar separado o agrupados por ciudad los conjuntos de tuits de entrenamiento. Con dichos ficheros de entrenamiento se construyó el modelo de lenguaje oportuno del sistema.

Con el modelo de lenguaje ya creado se procedió a ejecutar el sistema y evaluar las respuestas dadas por éste.

Los resultados se pueden observar en la tabla 5.24, donde las filas muestran si los conjuntos de tuits fueron agrupados por ciudad o estaban separados para el conjunto de entrenamiento, mientras que las columnas indican el nivel de actividad de los usuario de *Twitter* evaluados (*Baja*, *Media* o *Alta*).

Tabla 5.24: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de tuits para evaluar los propios tuits.

	Baja	**Media**	**Alta**
Agrupados	**39,48**	**61,45**	**73,01**
Separados	36,06	43,57	48,42

Los resultados muestran sin lugar a dudas como la aproximación realizada con los conjuntos de tuits agrupados por ciudad, a diferencia de lo que ocurría con la aproximación de *SVM*, obtienen unos resultados mucho mejores que la aproximación que los mantenía separados por usuario/ciudad.

Esto parece ser debido a que el lenguaje empleado en cada ciudad procede de usuarios distintos con distintas expresiones y términos, pero al unificar todos los usuarios por ciudad, el modelo obtiene un factor común de los términos empleados por los usuarios de una misma ciudad, dando así unos mejores resultados.

Por otro lado, debido a que la implementación con modelos de lenguaje resulta mucho más rápida que la implementada con *SVM*, en esta ocasión

sí que fue factible realizar un experimento con ambos corpus, el agrupado y el separado, sin tener en cuenta el nivel de actividad de los usuarios. Así pues, la precisión obtenida cuando se utilizó el conjunto de tuits agrupados y separados fue de **45,55 %** y 37,66 % respectivamente.

Con respecto a la aproximación realizada en la sección previa con un sistema de *SVM*, los resultados obtenidos muestran como la aproximación realizada con modelos de lenguaje, a diferencia de lo que ocurría con los textos formales en la sección 5.3, además de tener una ejecución mucho más rápida, da unos resultados muy similares para usuarios con baja actividad o cuando no se tiene en cuenta el nivel de actividad de los mismos, y unos resultados estadísticamente mejores ($p < 0,01$) cuando se trata de usuarios un un nivel de actividad media o alta, tal y como se puede apreciar en la figura 5.14. En la gráfica se muestra una comparativa de la mejor aproximación de *SVM* para clasificar tuits, la que separaba los conjuntos de tuits por usuario y ciudad, comparada con la mejor aproximación que utilizaba modelos de lenguaje, la que agrupaba el conjunto de tuits del corpus de entrenamiento en una única muestra de texto por ciudad.

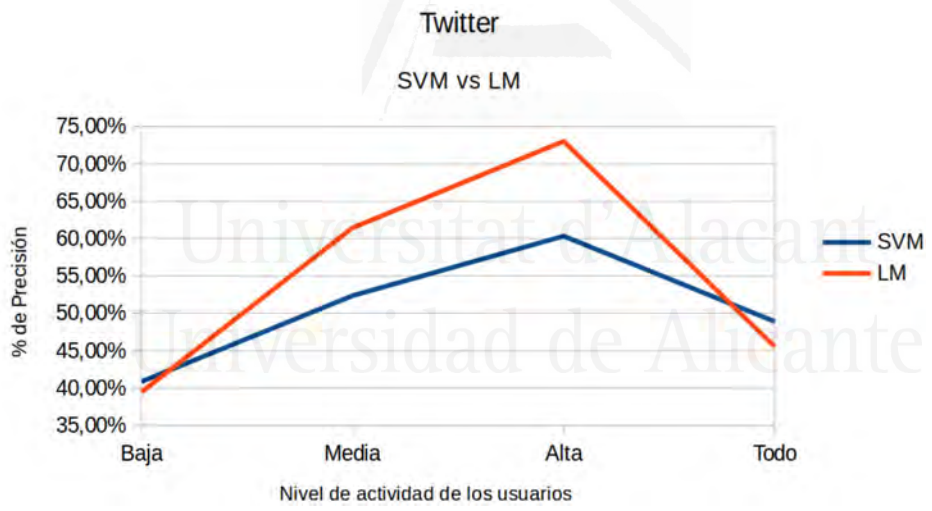


Figura 5.14: Precisión obtenida en la clasificación de conjuntos de tuits entrenando con el propio corpus de *Twitter* una aproximación basada en *SVM* (línea azul), y una aproximación basada en modelos de lenguaje (línea roja).

En esta gráfica se puede apreciar como cuando se intentan evaluar los conjuntos de tuits de usuarios con un bajo nivel de actividad, *SVM* ligeramente mejora, en términos de precisión, que no en tiempo, a los modelos de lenguaje (*LM*).

Conforme los conjuntos de tuits contienen más texto, se puede apreciar como es la aproximación implementada con modelos de lenguaje la que se impone claramente.

Si se observan los resultados cuando se clasifican conjuntos de tuits indistintamente (*Todo*), es *SVM* la que se impone por un ligero margen. Esto es debido a la existencia de un mayor número de textos procedentes de usuarios con un bajo nivel de actividad (ver tabla 5.22).

Si se calcula el promedio de los tres conjuntos de actividad (baja, media y alta), es decir, se suman las precisiones obtenidas para cada uno de estos grupos y se divide entre tres, los modelos de lenguaje claramente superan a la aproximación de *SVM*. Puesto que se va a trabajar con dichos conjuntos, se ha optado por la aproximación que hace uso de los modelos de lenguaje en detrimento de la de *SVM*. Además, cuando la aproximación de *SVM* se impone a la de los modelos de lenguaje lo hace por muy poco, y en cambio, cuando es la de modelos de lenguaje la que se impone lo hace por un margen muy significativo, requiriendo, además, un menor lapso de tiempo para su ejecución.

Por estos motivos, a partir de ahora, el resto de experimentos de esta sección se realizarán con modelos de lenguaje agrupando los conjuntos de tuits de entrenamiento por ciudad.

5.4.2. Entrenamiento con artículos de *Wikipedia*

Al igual que se hizo en los experimentos llevados a cabos en la sección de los textos formales cuando se entrenaba con *Wikipedia* (sección 5.3.2), se ha vuelto a utilizar este corpus para entrenar un sistema capaz de obtener el foco geográfico de los textos, que en esta ocasión eran conjuntos de tuits.

Una vez más, los artículos utilizados para el entrenamiento han sido los artículos de *Wikipedia* que representaban a las ciudades del corpus, el cual coincidía con las ciudades utilizadas en el corpus de *Twitter*, y el de los enlaces salientes o *outlinks* utilizados en estos artículos.

Así pues, se han realizado dos aproximaciones distintas, una en la que solamente se utilizaban los artículos de las ciudades, y otra en la que además se utilizaba el texto existente en los artículos salientes.

Entrenamiento procedente de artículos de localidades de *Wikipedia*

Tal y como se ha descrito en el apartado anterior, la aproximación utilizada va a crear un modelo de lenguaje con los términos, en esta ocasión del corpus de *Wikipedia*, del que se utilizarán los artículos de las 52 ciudades existentes en el sistema. Puesto que en dicho corpus sólo existía un artículo por cada ciudad, no ha sido necesario el agrupar todos los textos.

Así pues, con el el sistema entrenado como se ha comentado en el párrafo anterior, se procedió a lanzar cada uno de los distintos conjuntos de evaluación, que como se comentó al comienzo de esta sección, estaban divididos en el nivel de actividad de los usuarios y en particiones para poder realizar la posterior validación cruzada.

Los resultados obtenidos en esta aproximación se pueden ver en la tabla 5.25, donde las filas indican la categoría gramatical de los términos de *Wikipedia* utilizados para el entrenamiento, y las columnas muestran el nivel de actividad de los usuarios evaluados.

Tabla 5.25: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de ciudades de *Wikipedia* para determinar el foco geográfico de los conjuntos de tuits.

	Baja	**Media**	**Alta**	**Todo**
Topónimos	15,06	**9,05**	**3,90**	**12,96**
Sustantivos	9,59	3,69	1,03	7,71
Sustantivos + Adjetivos	9,70	3,74	1,08	7,81
Sin topónimos	**15,61**	6,68	2,31	12,76
Todo	15,57	6,59	2,32	12,71

Como se puede apreciar en los resultados, normalmente, el mejor rendimiento es obtenido cuando únicamente se utilizan los topónimos, ya que estos se supone que son los que añaden un mayor valor semántico desde el punto de vista geográfico cuando se intentan clasificar textos informales entrenando con una fuente de texto formal. Pero hay que resaltar como el mejor resultado es obtenido precisamente cuando se omiten estos términos en el corpus de usuarios de baja actividad. Los matices lingüísticos vuelven a ser importantes cuando se trata de geolocalizar textos de una extensión tan limitada, pese a que los conjuntos de entrenamiento y de evaluación sean tan dispares como *Wikipedia* y *Twitter*.

Por otro lado, curiosamente, cuanto más pequeño es el conjunto de tuits a evaluar, mejor funciona esta aproximación. Esto es debido a que los términos que aparecen en fuentes formales como *Wikipedia* que son capaces de ubicar geográficamente un texto, se encuentran más dispersos dentro de los grandes conjuntos de tuits. Pese a ello, estos resultados distan mucho de los obtenidos con *Twitter* como conjunto de entrenamiento.

Entrenamiento procedente de artículos de localidades de *Wikipedia* y los artículos referenciados

En esta ocasión, se ha realizado un experimento análogo al anterior, aunque esta vez se ha utilizado también los textos de los artículos que se citaban en los propios artículos de las localidades del corpus.

Puesto que en el experimento efectuado para detectar el foco geográfico de los usuarios de *Twitter* con *SVM* los mejores resultados se obtenían cuando se separaba el conjunto de entrenamiento (tuits) por usuario/ciudad, y el efectuado con modelos de lenguaje pasaba justo al contrario, se ha decido realizar la misma prueba, aunque esta vez agrupando o separando los artículos referenciados de *Wikipiedia*, para dilucidar cuál obtiene una precisión mayor.

Así pues, en la tabla 5.26 se pueden apreciar tanto los resultados obtenidos para la aproximación que utilizaba todos los artículos de *Wikipedia*, previamente mencionados, agrupados por ciudad (columnas encabezadas con '*Agr*'), como los que entrenaban al sistema como artículos distintos que representan a una misma ciudad (columnas encabezadas con '*Sep*'). Todo ello, una vez más, teniendo en cuenta el nivel de actividad de los usuarios en *Twitter*, y la categoría gramatical de los términos de los artículos de *Wikipedia* utilizados.

Tabla 5.26: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de artículos de ciudades de *Wikipedia* y el de los artículos referenciados en éstos, tanto agrupados por ciudad (*Agr*) como separados por artículo/ciudad (*Sep*), para determinar el foco geográfico de los conjuntos de tuits.

	Baja		Media		Alta		Todo	
	Agr	Sep	Agr	Sep	Agr	Sep	Agr	Sep
Topónimos	12,35	9,70	8,39	10,31	3,67	7,55	10,94	13,02
Sustantivos	15,14	13,95	18,04	10,70	17,29	9,25	15,95	12,89
Adjetivos + Sustantivos	15,32	14,00	**18,44**	9,51	**17,69**	6,48	16,20	12,53
Sin topónimos	15,51	13,90	17,66	9,28	17,52	5,99	16,13	12,41
Todo	**15,79**	14,34	17,50	4,52	17,16	1,90	**16,27**	8,04

Observando los resultados obtenidos, se puede ver claramente como la aproximación en la que se agrupan los artículos consigue mejores resultados.

También se puede apreciar como con un conjunto de entrenamiento pequeño como es el de topónimos, la precisión del sistema cae drásticamente. Esto es debido principalmente a que los artículos referenciados en los artículos de las ciudades, en muchas ocasiones no contienen topónimos relacionados con el artículo de la ciudad desde el que se refenciaba. Tal es así que varios de los artículos referenciados por una ciudad, son también referenciados por otra localidad del corpus, haciendo así que el sistema no consiga realizar un entrenamiento muy preciso.

Otra razón es el escaso número de términos (topónimos) que existe para entrenar el sistema y la aún más escasa correlación de estos términos formales con los plasmados en los tuits. Por ejemplo, si en un artículo se habla de Barcelona, en los tuits, normalmente se escribiría como *bcn*, tal y como se puede apreciar en la figura 5.15 donde se muestra un tuit procedente de una fuente formal como es la cuenta de la policía nacional de España.

Figura 5.15: Ejemplo de tuit donde se menta la ciudad de Barcelona mediante la abreviatura *BCN*.

Con respecto al resto de categorías gramaticales utilizadas para la clasificación de tuits, cabe destacar como obtienen un resultado muy parejo, el cual, pese a ser mucho mejor que el obtenido utilizando solamente los topónimos, sigue siendo muy bajo.

Si comparamos esta aproximación con la llevada a cabo sin los artículos referenciados, se puede apreciar como la aproximación realizada con los artículos referenciados generalmente mejora con respecto a *Wikipedia* sin los artículos referenciados, especialmente cuando la cantidad de texto a clasificar es mayor (usuarios más activos).

En la figura 5.16 se puede observar una comparación de las aproximaciones utilizando los artículos de las localidades del corpus de *Wikipedia* únicamente (líneas azul y roja), y la utilización de las localidades y los artículos mencionados en éstas (líneas amarilla y verde), tanto utilizando todos los términos de dichos corpus para el entrenamiento, como utilizando solamente los topónimos.

En la gráfica se puedo apreciar como a medida que los conjuntos de tuits de los que se pretende obtener el foco geográfico son más grandes, la precisión decrece drásticamente debido a la gran abundancia de términos existentes en estos conjuntos de tuits, los cuales no se encuentran en el corpus de *Wikipedia*.

La única excepción viene cuando se utiliza el corpus de *Wikipedia* con los enlaces salientes utilizando todos los términos de dicho corpus. En esta aproximación, la precisión se mantiene estable sin importar mucho el nivel de actividad de los usuarios. Esto es debido a que a medida que se introducen nuevos artículos que no versan puramente sobre las propias localidades, el léxico empleado se enriquece pudiendo así coincidir con el existente en los conjuntos de tuits evaluados.

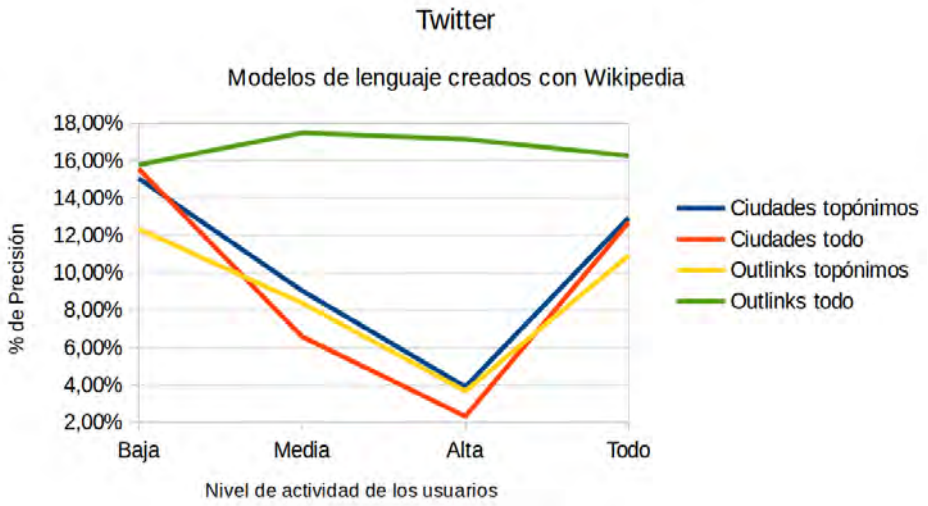


Figura 5.16: Precisión obtenida en la clasificación de conjuntos de tuits creando los modelos de lenguaje con *Wikipedia*. '*Ciudades*' indica que se han utilizado únicamente los artículos de las ciudades del corpus. '*Outlinks*' indica que se han utilizado los artículos de las ciudades del corpus y a los que se hacía referencia en éstos. '*topónimos*' indica que únicamente se han utilizado los topónimos detectados por *FreeLing* en los artículos. '*todo*' indica que se han utilizado todos los términos de los artículos.

La gráfica nos permite también ver con claridad cómo al utilizar un corpus de entrenamiento más extenso, el uso de topónimos pierde relevancia, ya que el resto de categorías gramaticales y matices lingüísticos aportan un gran peso a la hora de determinar el foco geográfico de los usuarios de *Twitter*.

Si se comprueban los resultados de la tabla 5.26, se puede ver cómo la aproximación que no hace uso de topónimos obtiene unos resultados muy similares a los logrados con todos lo términos, lo cual refuerza la hipótesis del peso que tienen las demás categorías gramaticales a la hora de determinar el foco geográfico incluso de textos de diversa formalidad.

5.4.3. Entrenamiento con textos de *Flickr*

Siguiendo la metodología llevada a cabo en los experimentos previos, en esta sección se va a mostrar cómo se ha llevado a cabo la experimentación utilizando una fuente de texto informal, *Flickr* (sección 4.4), para preparar un modelo de lenguaje que sea capaz de detectar el foco geográfico de textos procedentes de otra fuente de textos informales, *Twitter*.

Lo que se pretende averiguar con este experimento es cuan útil puede resultar una fuente de texto informal para determinar el ámbito geográfico de los textos de otra fuente informal.

En esta ocasión, puesto que los textos de *Flickr* también son informales, sucede algo parecido a lo que ocurría con los textos de *Twitter*, es decir, que las herramientas existentes para determinar la categoría gramatical de los términos que componen los textos de dicha fuente no llegan a obtener unos buenos resultados. Por este motivo, en lugar de separar los experimentos según la categoría gramatical de los términos del conjunto de entrenamiento, se ha hecho según la procedencia de dichos textos dentro de *Flickr*, es decir, dependiendo de si los textos provenían de: los comentarios, la descripción, las notas de los usuarios, las etiquetas, el título o todos ellos juntos, tal y como se describió en la sección 4.4 donde se detalló el corpus de entrenamiento aquí utilizado.

El procedimiento de los experimentos expuestos en esta sección es idéntico al explicado en las secciones previas, es decir, se han creado modelos de lenguaje obedeciendo a la procedencia mencionada en el párrafo anterior de los textos de *Flickr*, para posteriormente poder detectar el foco geográfico de los conjuntos de tuits según el nivel de actividad de los usuarios de *Twitter*.

A la luz de los resultados mostrados en secciones previas, la aproximación utilizada en este experimento ha sido la que agrupaba todos los textos según su ciudad de procedencia.

Los resultados de este experimento se pueden ver en la tabla 5.27, donde una vez más las columnas muestran el nivel de actividad de los usuarios y las filas indican el campo de procedencia de los textos de *Flickr*.

Tabla 5.27: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de *Flickr* para determinar el foco geográfico de los conjuntos de tuits.

	Baja	**Media**	**Alta**	**Todo**
Comentarios	13,01	7,75	4,29	11,29
Descripción	18,55	**13,77**	**8,08**	16,86
Notas	13,37	6,31	2,15	11,08
Etiquetas	18,43	11,56	4,64	16,06
Título	15,23	5,75	2,35	12,27
Todo	**19,87**	12,77	6,40	**17,48**

Según se puede apreciar en los resultados mostrados, la mejor precisión se obtiene cuando se utiliza para la creación de los modelos de lenguajes los textos procedente de la descripción de las fotografías, así como juntando los textos de todos los campos. Esto es debido a que el volumen de texto del corpus de *Flickr* utilizado en estos experimentos no es muy extenso, y en los

campos citados es donde más texto se pude hallar, de ahí esa ligera mejora a la hora de clasificar los conjuntos de tuits.

Por otro lado, una vez más se puede observar como los mejores resultados se obtienen cuando se intenta obtener el foco geográfico de los usuarios con un bajo nivel de actividad. Esto es debido a que ambos corpus, pese a ser de carácter informal, son muy dispares y, por ende, cuanto mayor es el conjunto de términos del corpus a evaluar, mayor es la diferencia existente con el modelo de lenguaje de cada ciudad y más difícil de clasificar correctamente.

Comparación de aproximaciones con selección de características

En la figura 5.17 se muestra los mejores resultados obtenidos al crear los modelos de lenguaje con cada uno de los 3 corpus expuestos en esta sección: *Twitter*, *Wikipedia* (con todos los términos de los artículos de las localidades y sus enlaces salientes) y *Flickr* (con todos los textos).

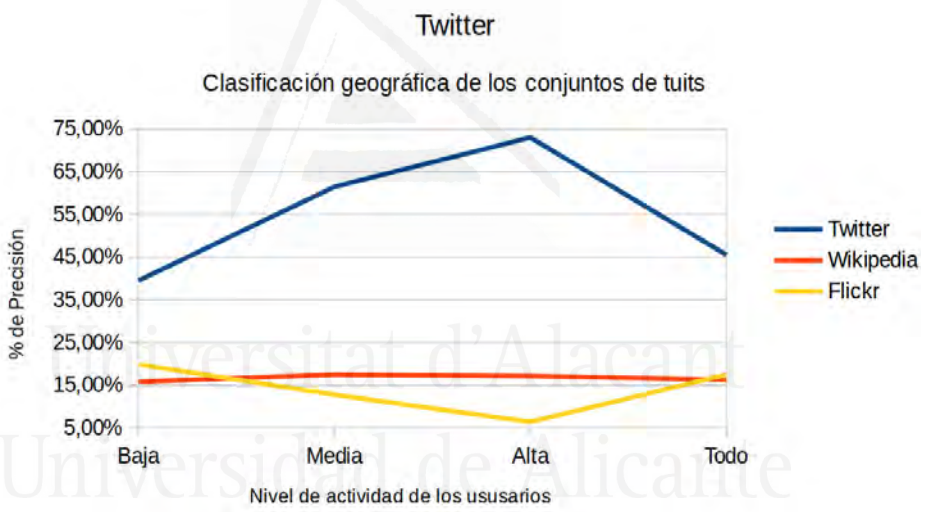


Figura 5.17: Precisión obtenida en la clasificación de conjuntos de tuits creando los modelos de lenguaje con textos procedentes de distintas fuentes.

Como era de esperar, creando los modelos de lenguaje con tuits, para posteriormente obtener el foco geográfico de otros conjuntos de tuits, es la aproximación que mejores resultados obtiene, aunque la más costosa de implementar en términos de tiempo y espacio.

Al intentar evaluar conjuntos de tuits con más textos, la precisión aumenta linealmente cuando se crean los modelos de lenguaje con los propios tuits, puesto que a mayor cantidad de texto, más fácil resulta ubicarlo.

Por otro lado, con los corpus de *Wikipedia* y *Flickr* no sucede lo mismo debido a, como ya se ha comentado anteriormente, la gran disparidad que existe entre estos corpus y el de *Twitter*, viéndose los términos más relevantes

a la hora de clasificar los conjuntos de tuits geográficamente diluidos entre esta mayor cantidad de texto.

5.4.4. Combinación de corpus de entrenamiento

En esta sección se mostrarán aproximaciones en las que se ha combinado algunos de los corpus de entrenamiento, de manera análoga a como se hizo con los textos formales (ver sección 5.3.4) con los que se ha trabajado en las secciones previas, con el fin de detectar el foco geográfico de los conjuntos de tuits.

El propósito de esta combinación es el de conocer bajo qué circunstancias se podría lograr una ayuda en la identificación del foco geográfico de una fuente de texto informal como es *Twitter*, mediante el agregado de recursos de distinta índole como son *Wikipedia* y *Flickr*, puesto que estos recursos son fácilmente accesibles y están disponibles en numerosos idiomas.

Twitter y *Wikipedia*

El primero de los experimentos que se ha planteado ha sido el de la combinación del corpus de *Twitter* con el de *Wikipedia*, es decir, una fuente de texto informal con una formal. En esta ocasión se ha optado por utilizar únicamente los artículos de *Wikipedia* de las propias ciudades, sin tener en cuenta los de los artículos referenciados en estos.

La manera de proceder con este experimento ha sido análoga a como se hizo con el experimento llevado a cabo utilizando únicamente *Twitter* para generar los modelos de lenguaje (ver sección 5.4.1), es decir, se ha dividido el corpus en 3 particiones obedeciendo al nivel de actividad de los usuarios, y por otro lado se han dejado todos los tuits existentes en el corpus. Dentro de cada una de estas particiones, a su vez se han creado otras particiones adicionales para realizar la oportuna validación cruzada. Una de estas últimas particiones es utilizada para la evaluación del sistema, mientras que el resto de particiones del mismo nivel de actividad de los usuarios, el resto del corpus de tuits con otro nivel de actividad de los usuarios (ver sección 5.4.1), y los artículos de las localidades de *Wikipedia*, se utilizan para la creación de los modelos de lenguaje. En la siguiente iteración se utilizará una de las particiones aún no utilizada del mismo nivel de actividad de los usuarios para la evaluación del sistema, y el resto, incluida la partición utilizada para la evaluación en la iteración anterior, se utilizan para crear los modelos de lenguaje.

Por ejemplo, si se está pretendiendo clasificar geográficamente un conjunto de tuits de usuarios con una actividad baja, el corpus donde se encuentran todos los tuits de usuarios que han tuiteado entre 1 y 10 veces en alguna ciudad se dividirá en partes iguales. Una de esas partes se utilizará para la posterior evaluación del sistema, mientras que el resto de

particiones de usuarios con actividad baja, se utilizará para crear el modelo de lenguaje junto al resto de particiones de usuarios con actividad media y alta, así como el conjunto de artículos de *Wikipedia*. En la siguente iteración de la validación cruzada, se utilizará otra partición distinta de usuarios con un nivel de actividad bajo para la evaluación y el resto, junto con la que en la primera se había utilizado como evaluación, se utilizarán para crear el modelo de lenguaje. Y así sucesivamente hasta haber probado con todas las particiones de los usuarios con actividad baja. Una vez completado esto, se obtendrá la media de los resultados, completando así la validación cruzada.

Puesto que los artículos de *Wikipedia* estaban divididos según la categoría gramatical de los términos, se realizaron experimentos probando todas estas categorías contra todos los niveles de actividad de los usuarios de *Twitter* en los que se había dividido dicho corpus, de la misma manera que se hizo con los experimentos llevados a cabo en la sección 5.4.2, en la que se utilizaba únicamente *Wikipedia* para crear los modelos de lenguaje.

Así pues, los resultados obtenidos se pueden observar en la tabla 5.28.

Tabla 5.28: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de *Twitter* y *Wikipedia* para determinar el foco geográfico de los conjuntos de tuits.

	Baja	**Media**	**Alta**	**Todo**
Topónimos	**39,49**	61,45	**73,00**	45,55
Sustantivos	39,45	61,45	72,90	45,56
Sustantivos + Adjetivos	39,46	61,47	72,88	**45,58**
Sin topónimos	39,44	**61,50**	72,81	**45,58**
Todo	39,45	**61,50**	72,81	**45,58**

En la figura 5.18 se puede apreciar una comparativa entre la aprosimación que utilizaba únicamente *Twitter* y la actual que ha utilizado *Twitter* junto al corpus de artículos de las ciudades de *Wikipedia*.

Estos experimentos ponen de manifiesto el escaso peso que tienen los artículos de *Wikipedia* entre el océano de tuits utilizados para la creación de los modelos de lenguaje, siendo prácticamente insignificante la incursión de esta fuente de textos. Por ello, tal y como sucedió en la sección 5.3.4 en la que se realizó la clasificación de noticias del diario *20Minutos* con 52 artículos de *Wikipedia*, se realizaron otros experimentos para poder dotar de una mayor relevancia al corpus de *Wikipedia* para que ésta no perdiera notoriedad.

Se probó con la ponderación de ambos corpus, otorgándole diversos pesos a éstos. Pero, una vez más, este experimento obtuvo unos resultados muy inferiores a los mostrados. Por ello, se decidió añadir más texto al corpus de *Wikipedia*. Para ello se utilizaron los textos de los artículos citados en los propios artículos de las ciudades analizadas.

Figura 5.18: Precisión obtenida en la clasificación de conjuntos de tuits creando los modelos de lenguaje con textos procedentes de *Twitter* y *Wikipedia*.

Twitter y *Wikipedia* con enlaces salientes

Por los motivos expuestos en el experimento anterior, se ha decidido implementar una aproximación donde, aparte de incluir los artículos de *Wikipedia* referentes a las localidades analizadas, se utilizaron también los artículos referenciados en éstos, obteniendo así un volumen de texto mucho mayor y más cercano al empleado en el corpus de *Twitter*.

Puesto que en los experimentos previos llevados a cabo en la sección 5.4.2 se demostró que la precisión mejoraba cuando se unificaban los artículos de *Wikipedia* en un único fichero, como si se tratase de un conjunto de tuits más, la aproximación llevada a cabo en esta sección se enfocó de la misma manera, es decir, agrupando todos los textos que aparecían en los artículos referenciados en el artículo de un ciudad dada, en un único texto que representaba a dicha ciudad.

El resto del procedimiento para llevar a cabo este experimento es análogo al anterior, donde se lanzaron ejecuciones donde se medían todos los tipos de categorías gramaticales de *Wikipedia* contra todos los corpus de *Twitter* clasificados según el nivel de actividad de los usuarios por cada una de las localidades del corpus.

Para asegurar la corrección de los experimentos, una vez más se realizó una validación cruzada. Los resultados de este experimentos se pueden ver en la tabla 5.29.

Una vez más se puede apreciar como *Wikipedia* introduce ruido a la hora de generar modelos de lenguajes capaces de obtener el foco geográfico

Tabla 5.29: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de *Twitter* y *Wikipedia* con sus enlaces salientes para determinar el foco geográfico de los conjuntos de tuits.

	Baja	**Media**	**Alta**	**Todo**
Topónimos	38,78	**60,98**	**72,36**	**44,86**
Sustantivos	**40,35**	50,26	51,30	42,21
Sustantivos + Adjetivos	40,25	49,03	48,96	41,70
Sin topónimos	40,16	49,37	49,53	41,82
Todo	40,19	49,56	49,88	41,89

de los conjuntos de tuits evaluados. Esto se puede apreciar al observar la precisión obtenida cuando los modelos de lenguaje obtenidos tienen más términos (*Todo*), mejorando claramente dicha precisión cuando los modelos de lenguaje contienen menos términos (*Topónimos*), donde se obtienen unos resultados similares a los conseguidos con la utilización única del corpus de *Twitter* para generar dichos modelos de lenguaje (ver tabla 5.24) al haber un ratio mucho menor de términos procedentes de *Wikipedia* que de *Twitter* cuando se generan estos modelos.

En la figura 5.19 se muestra una comparativa de las aproximaciones en las que sólo se utilizaba *Twitter* para crear el sistema, y las dos aproximaciones expuestas en esta sección donde se utilizaba *Wikipedia* y *Wikipedia* con el texto de los artículos referenciados en los artículos de las ciudades del corpus, junto al propio corpus de *Twitter*, para generar los modelos de lenguaje.

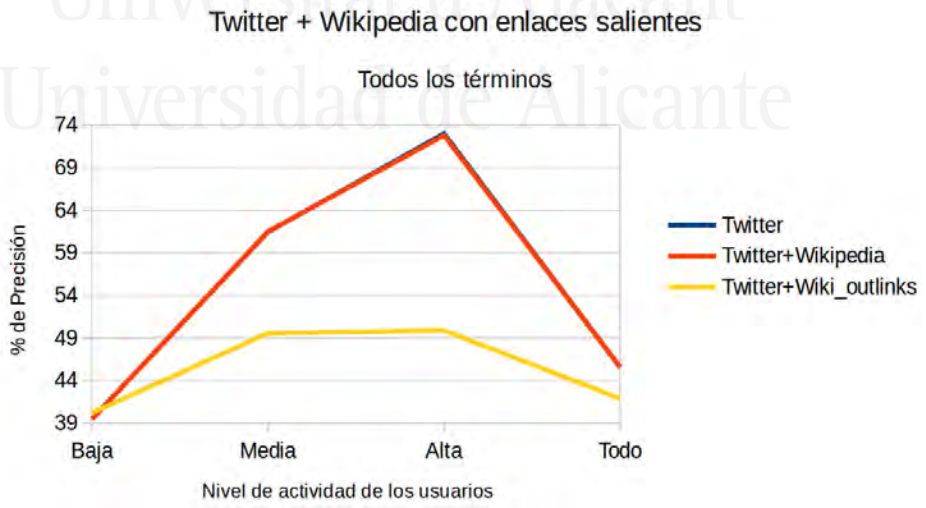


Figura 5.19: Precisión obtenida en la clasificación de conjuntos de tuits creando los modelos de lenguaje con textos procedentes de *Twitter* y *Wikipedia*.

En esta ocasión, se puede apreciar como cuanto mayor es el volumen de los textos a clasificar, mayor es la cantidad de ruido que introduce el entrenamiento de *Wikipedia* con sus enlaces salientes con respecto a la aproximación que no hace uso de estos artículos. Aunque, cabe destacar que cuando se intenta clasificar conjuntos de tuits con un escaso volumen, la incorporación de los textos de *Wikipedia* con los enlaces salientes, aporta una ligera mejoría.

Twitter y *Flickr*

En esta sección se van a realizar experimentos para intentar determinar el foco geográfico de una fuente de textos informales, *Twitter*, entrenando con esa misma fuente más otra fuente de texto informal como lo es *Flickr*.

Así pues, estos experimentos se realizan de manera análoga a como se han hecho en los experimentos llevados a cabo en la sección 5.4.3, donde se intentaba determinar el foco geográfico de los mismos conjuntos de tuits creando modelos de lenguajes únicamente con textos del corpus de *Flickr*, aunque, en esta ocasión, se utilizará también el propio corpus de *Twitter* para crear los modelos de lenguaje.

En esta ocasión, de igual manera a como se realizó en los experimentos en los que se creaban los modelos de lenguaje con *Flickr*, se han creado los modelos de lenguaje atendiendo al campo de procedencia de los textos dentro de *Flickr*, es decir, comentarios, descripción, notas, etiquetas, títulos y todos estos textos juntos.

Tal y como se ha realizado en todos los experimentos llevados a cabo bajo la aproximación de los modelos de lenguaje, y como se describió en la sección 5.1.2, por cada una de las ciudades del corpus se creó un modelo de lenguaje, para posteriormente obtener la probabilidad de que cada uno de estos modelos hubiera generado cada uno de los conjuntos de evaluación. Posteriormente se adjudicó la localidad de la cual se había generado el modelo de lenguaje que mayor probabilidad tenía de haber generado el conjunto de tuits de evaluación.

La creación de los modelos de lenguaje de las localidades se realizó de manera análoga a la de la secciones previas, es decir, unificando los textos de *Flickr* como si se tratara de un conjunto de tuits más por cada una de las ciudades del corpus. Los resultados obtenidos de la validación cruzada se pueden ver en la tabla 5.30.

Al igual que sucedió con los experimentos realizados en la sección anterior con la combinación de *Twitter* y *Wikipedia*, se realizaron pruebas ponderando el peso de cada uno de los corpus. Pero, al igual que entonces, los resultados empeoraron también cuanto más peso se le quitaba a *Twitter* frente a *Flickr*, siendo claramente la mejor aproximación la que otorgaba todo el peso al conjunto de tuits.

Tabla 5.30: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de *Twitter* y el de *Flickr* para determinar el foco geográfico de los conjuntos de tuits.

	Baja	**Media**	**Alta**	**Todo**
Comentarios	36,70	47,23	49,69	38,11
Descripción	39,60	60,83	71,47	45,25
Notas	39,48	**61,49**	**73,03**	45,55
Etiquetas	**40,48**	60,48	70,46	**45,86**
Título	39,76	61,37	72,87	45,70
Todo	37,19	46,05	46,97	37,91

Pese ello, se pueden extraer unas conclusiones interesantes del experimento llevado a cabo en esta sección. Una de ellas es que el uso de los textos procedentes de los comentarios vertidos sobre cada foto en *Flickr* introduce una cantidad ingente de ruido que, incluso con el escaso peso que tienen estos textos en el sistema, hace que la detección del foco geográfico de los conjuntos de tuits decaiga drásticamente.

Algo similar sucede con la aproximación en la que se han empleado todos los textos disponibles del corpus de *Flickr*, sin distinción de su campo de procedencia, ya que se puede apreciar como la caída en la precisión es aún más notable que utilizando únicamente los textos de los comentarios.

Una vez más, debido a la gran desproporción que existe entre el tamaño del corpus de *Twitter* y el de *Flickr* (4.748.032 de tuits procedentes de 205.895 usuarios frente a las 312.209 fotografías obtenidas para el corpus de *Flickr*), los mejores resultados mostrados en la tabla 5.30 no difieren mucho de los mostrados en 5.24, donde se creaban los modelos de lenguaje únicamente con los tuits, tal y como se puede apreciar en la figura 5.20. En dicha figura se ha optado por mostrar los resultados de la combinación de *Twitter* y *Flickr* que hacía uso únicamente del texto de las etiquetas por ser junto a las notas el que mejores resultados ofrecía.

Una vez más, se observa cómo a mayor volumen de textos ajenos a la fuente informal que se pretende clasificar, peores resultados se obtienen si el conjunto de entrenamiento no procede de la misma fuente que la clasificada.

En cambio, al igual que sucedía en el experimento anterior con *Wikipedia*, si los textos a clasificar contienen una cantidad mínima de datos, el aporte de estas otras fuentes no sólo no es perjudicial, sino que puede llegar a ser beneficioso.

Twitter, *Wikipedia* y *Flickr*

Por último, en esta sección se ha realizado un experimento en el que se van a utilizar las tres fuentes de textos con las que se ha trabajado en la

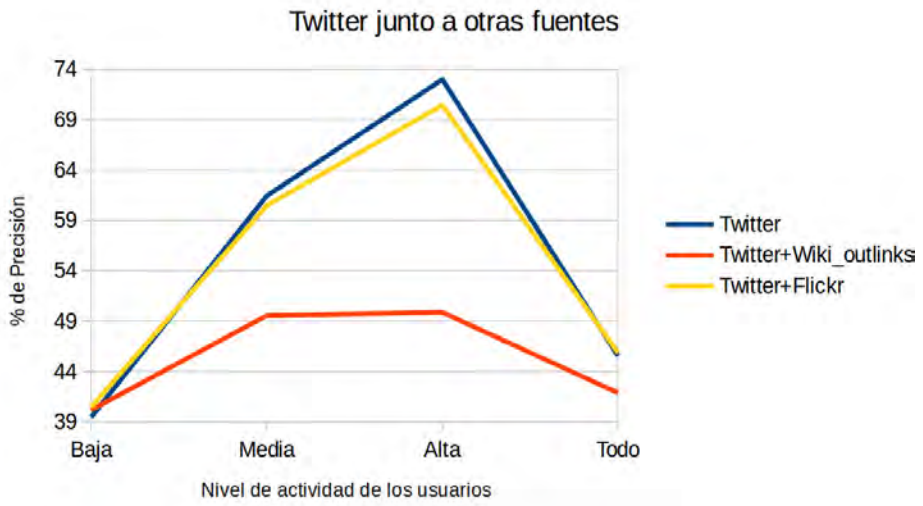


Figura 5.20: Precisión obtenida en la clasificación de conjuntos de tuits creando los modelos de lenguaje con textos procedentes de *Twitter*, *Twitter* y *Wikipedia* con los enlaces salientes o *Twitter* y *Flickr* con el texto de las etiquetas.

sección de clasificación de textos informales (5.4), para crear los modelos de lenguaje oportunos que obtengan el foco geográfico de los conjuntos de tuits evaluados en el resto de secciones.

Puesto que la envergadura de los corpus aquí utilizados impedía el poder realizar pruebas con todas las combinaciones expuestas en sus respectivos experimentos, se ha optado por acotar dichos experimentos.

Para ello, se ha realizado una aproximación usando los campos procedentes de *Flickr* que mejores resultados dieron en la aproximación en la que se utilizaba únicamente esta fuente de textos para crear los modelos de las localidades (5.4.3), es decir, *descripción* y *todo*. Los textos, una vez más, fueron agrupados por ciudad como si se tratara de un conjunto de tuits más.

Por otro lado, por los mismos motivos, únicamente se ha utilizado el corpus de artículos de ciudades de *Wikipedia* con los *topónimos* y *todos* los términos como categorías gramaticales. Cada texto de cada ciudad fue tratado como un conjunto de tuits de esa ciudad.

Así pues, las diversas combinaciones de los corpus de *Twitter*, *Wikipedia* y *Flickr* fueron lanzadas para clasificar geográficamente los conjuntos de tuits según los cuatro distintos niveles de actividad de los usuarios que se han ido utilizando en todos los experimentos de esta sección. Para obtener la precisión final del sistema se realizó una validación cruzada, cuyos resultados se pueden ver en la tabla 5.31.

Tabla 5.31: Resultados de la validación cruzada de *modelos de lenguaje* entrenando con el corpus de *Twitter*, artículos de ciudades de *Wikipedia* y el de *Flickr* agrupado por ciudad, para determinar el foco geográfico de los conjuntos de tuits.

		Baja	**Media**	**Alta**	**Todo**
Wikipedia Topónimos	**Flickr Descripción**	39,67	**60,82**	**71,43**	45,31
	Flickr Todo	38,55	49,13	51,88	40,37
Wikipedia Todo	**Flickr Descripción**	**39,76**	60,78	70,98	**45,39**
	Flickr Todo	38,59	49,31	52,02	40,49

Una vez más cabe destacar la escasa aportación de *Wikipedia* al experimento debido a su bajo volumen de texto comparado con los otros dos corpus.

Respecto a *Flickr*, se ratifica como los resultados utilizando todos los campos claramente introducen una gran cantidad de ruido en los modelos de lenguaje, manteniéndose éste estable cuando se utiliza el campo de descripción de las fotos, el cual logra unos resultados muy similares a cuando se utilizan exclusivamente los conjuntos de tuits para generar los modelos de lenguaje.

Una visión más global de los experimentos expuestos en esta sección se muestra en la figura 5.21, donde se pueden apreciar los resultados obtenidos mediante todas las aproximaciones que se realizan con combinaciones, así como la aproximación que mejores resultados ha cosechado, en la que se utilizaba únicamente el corpus de *Twitter* para generar los modelos de lenguaje.

En dicha figura, la aproximación de *Twitter* (línea azul) hace referencia a la que empleaba únicamente el corpus de *Twitter* para crear los modelos de lenguaje. La aproximación llevada a cabo con *Wikipedia* (línea roja) fue la que se utilizaban todos los términos de los artículos referenciados y artículos de las ciudades del corpus juntos a los textos de *Twitter*. La aproximación de *Flickr* (línea amarilla) fue la empleada con todos los términos de las fotografías del corpus junto al corpus de *Twitter*. Por último, la aproximación en la que se empleaban los 3 corpus (*Todo*, línea verde), es la que utilizaba todos los términos de los artículos de *Wikipedia* de las ciudades del corpus, más los términos de todo el corpus de *Flickr* junto al de los tuits.

En el gráfico se puede apreciar claramente como los mejores resultados son obtenidos con la aproximación que utilizaba únicamente los tuits para generar los distintos modelos de lenguaje de las ciudades del corpus. Esto es así debido a que el lenguaje que se puede encontrar en *Twitter* es muy específico, y está repleto de abreviaturas, contracciones, etc., lo cual hace muy difícil de equiparar a cualquier lenguaje empleado en cualquier otra fuente de textos.

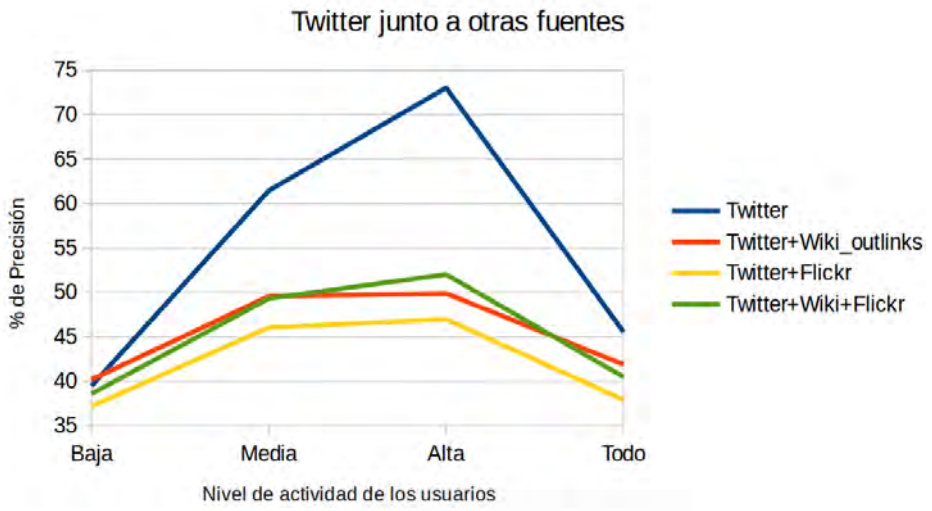


Figura 5.21: Precisión obtenida en la obtención del foco geográfico de conjuntos de tuits creando los modelos de lenguaje con la combinación de textos procedentes de distintas fuentes.

También es curioso apreciar como pese a que *Flickr* contiene un lenguaje más próximo en la informalidad al de *Twitter*, introduce mucho más ruido que *Wikipedia*, la cual se supone que es una fuente de texto con un lenguaje formal. Aunque, cuando se utilizan las etiquetas de las fotografías en lugar de todo el texto disponible, se consigue una mejora para usuarios de baja actividad, tal y como se pudo observar en la tabla 5.30 y la figura 5.20.

Así pues, la agregación de recursos textuales supone una acumulación de ruido que hace que el sistema sea menos preciso al detectar el foco geográfico de los conjuntos de tuits.

5.5. Conclusiones

A lo largo de este capítulo se han realizado experimentos que mostraban aproximaciones para obtener el foco geográfico en textos de diversa formalidad. Para ello, se ha han utilizado técnicas de aprendizaje automático como *SVM* (*Support Vector Machines*: Máquinas de Vectores de Soporte) o Modelos de Lenguaje probabilísticos (*Probabilistic Language Models*).

5.5.1. Identificación del foco geográfico en textos formales

Se han mostrado diversas aproximaciones para la identificación del foco geográfico en textos formales.

Entrenamiento con noticias del diario *20Minutos*

Con el propio corpus del diario *20Minutos* se hizo una calibración previa para ver qué algoritmo era el que mejores resultados daba a la hora de determinar el foco geográficos de las noticias del propio diario.

El sistema *SVM* demostró obtener una mayor precisión, entre 5 y 9 puntos porcentuales absolutos por encima del de modelos de lenguaje, siendo, además, la mejor aproximación la que hacía uso de todos los términos existentes en el corpus. Esta diferencia en la precisión se acentuó aún más cuanto más grande era el corpus de entrenamiento utilizado, al ser de este corpus de donde se podían extraer un mayor número de sutilezas lingüísticas que ayudaran a esta tarea.

Pese a que los topónimos son una parte fundamental a la hora de llevar a cabo un sistema que identifique geográficamente los textos, la precisión obtenida por el corpus formado únicamente por éstos resulta ser la más baja de todas. Esto es debido a la información que se pierde al desechar el resto de términos donde residen matices lingüísticos importantes a la hora de hacer una clasificación geográfica. Además, hay que tener en cuenta que alrededor de un 11 % de las noticias del corpus trabajado no contienen topónimos, lo que imposibilita su geolocalización utilizando únicamente esta categoría.

Los resultados obtenidos mediante la reducción de características mejoraron entre 9 y un 3 puntos porcentuales absolutos a los logrados en la mejor implementación realizada sin esta reducción. Esta mejora fue más significativa tanto en cuanto el corpus era más pequeño (años 2008 y 2009).

Entrenamiento con artículos de *Wikipedia*

En los experimentos que utilizaron el corpus de *Wikipedia*, con o sin los artículos referenciados en los 52 previos, para entrenar el sistema, los mejores resultados coincidieron con las aproximaciones que utilizaban solamente los topónimos. Esto es debido a la disparidad existente entre ambos corpus, lo que hace que categorías como la de topónimos sean las que mejor discriminen geográficamente los textos.

Cabe destacar que dentro de estos experimentos que utilizaron el corpus de *Wikipedia*, para entrenar el sistema, la aproximación que hacía uso de todos los sustantivos y adjetivos, funcionó mejor que la que hizo uso solamente de los sustantivos. De ahí se puede deducir que el uso de otras categorías gramaticales tales como los adjetivos, puede ser de gran ayuda a la hora de determinar el foco geográfico de este tipo de texto.

En cuanto a los experimentos que utilizaron *Wikipedia* como mero selector de características, al utilizar el propio corpus del diario *20Minutos* como conjunto de entrenamiento los resultados mejoraron claramente, obteniéndose los mejores resultados cuando se hacía uso de todos los

términos existentes en ambos corpus, mostrando una vez más la importancia de todos los términos a la hora de hacer este tipo de clasificaciones.

La aproximación que hizo uso de los artículos referenciados obtuvo mejores resultados que la que hizo uso únicamente de los 52 artículos de las localidades existentes. Esto fue debido a que esta primera aproximación tenía un mayor número de características que eran a su vez relevantes para esta tarea.

Entrenamiento con mensajes de *Twitter*

Los resultados obtenidos con la aproximación que utilizaba los tuits agrupados por usuario y ciudad no fueron muy buenos debido a lo poco representativas que eran unas muestras con tan poco texto.

Al igual que sucediera con la aproximación que entrenaba únicamente con muestras de *Wikipedia* (textos formales), al entrenar con muestras más extensas, la que agrupa los tuits por ciudad de procedencia, los mejores resultados se obtienen cuando se clasifican textos que contienen únicamente topónimos. Esto es debido a que son prácticamente los únicos términos que comparten estos corpus tan dispares.

Por otro lado, si se utilizan los términos de los textos de *Twitter* únicamente como selector de características, y se entrena con los propios textos del diario *20Minutos*, la precisión lograda supera en varias veces a las anteriores, siendo la mejor aproximación la que entrena y clasifica textos que contienen todos sus términos.

En cuanto a aproximaciones de selección de características, se ha comprobado que la utilización de χ^2, al utilizar el propio corpus de noticias para obtener el vocabulario y entrenar al sistema, es la que mejores resultados ha dado.

El problema que tiene dicha aproximación es que se ha de disponer previamente de un corpus lo suficientemente extenso ya etiquetado de las mismas características al que se pretende clasificar geográficamente, con la dificultad que ello atañe por la escasez de los mismos. Además, dicha aproximación requiere de un proceso de selección de característica temporalmente costoso.

Por estos motivos, si se pretende clasificar textos geográficamente, de los cuales no se tiene ningún corpus previamente etiquetado, se puede afirmar que la mejor aproximación es la que hace uso de otro medio de la misma formalidad que hace uso de un recurso que ya está geográficamente etiquetado, es extenso y está libremente disponible: *Wikipedia* con los enlaces que apuntan a otros artículos.

Combinación de corpus de entrenamiento

Los experimentos realizados con la combinación de *20Minutos* y *Wikipedia* dan unos resultados prácticamente idénticos a los realizados únicamente con *20Minutos*. Esto es debido a que el volumen de los textos del diario es varias veces superior a los de los 52 artículos de *Wikipedia*. Se realizó una ponderación de los corpus, la cual no consiguió mejorar los resultados. Por ello se decidió añadir más textos a *Wikipedia* mediante la incursión de los textos de los artículos referenciados.

Los experimentos realizados con la combinación de *20Minutos* y *Twitter* cuando se agrupaba el conjuntos de tuits por ciudad de procedencia muestra que tanto en cuanto mayor es el conjunto de entrenamiento del corpus del diario *20Minutos*, mejores son los resultados, siendo siempre la mejor aproximación la que hace uso de todas las características del lenguaje. Así pues, esta aproximación puede ser válida para cuando se quiere apoyar a un conjunto escaso de textos etiquetados de la fuente a clasificar.

Los experimentos realizados con la combinación de los corpus *20Minutos*, *Wikipedia* y *Twitter* vuelven a mostrar unos resultados similares a los anteriores, donde los mejores resultados se lograban cuando el corpus del diario *20Minutos* era más representativo (años 2010, 2011 y 2008-2011).

Así pues, se puede concluir que la combinación de corpus puede ser interesante cuando no se disponga de una gran cantidad de textos etiquetados del mismo corpus que se quiere clasificar geográficamente, ya que puede obtener unos buenos resultados aunque siempre lejos de los logrados por aproximaciones que sí que dispongan de un número extenso de textos etiquetados de la misma fuente a clasificar.

5.5.2. Identificación del foco geográfico en textos informales

Estos experimentos perseguían el conocer cómo pueden afectar la incursión de otras fuentes de texto en la tarea de georreferenciar textos informales.

Entrenamientos con textos de *Twitter*

En el experimento efectuado para detectar el foco geográfico de los usuarios de *Twitter* con *SVM* los mejores resultados se obtenía cuando se separaba el conjunto de entrenamiento (tuits) por usuario/ciudad, y el efectuado con modelos de lenguaje pasaba justo al contrario, por lo que parece que intuitivamente, para una aproximación basada en modelos de lenguaje, agrupar los textos por ciudad puede dar unos resultados mejores, tal y como mostraron finalmente los experimentos.

Los resultados obtenidos muestran como la aproximación realizada con modelos de lenguaje, a diferencia de lo que ocurría con los textos formales, además de tener una ejecución mucho más rápida, da unos resultados muy

similares para usuarios con baja actividad, y unos resultados claramente mejores cuando se trata de usuarios de un nivel de actividad media o alta. Si se calcula el promedio de los tres conjuntos de actividad (baja, media y alta), los modelos de lenguaje claramente superan a la aproximación de *SVM*.

Entrenamiento con artículos de *Wikipedia*

En lo que respecta a la aproximación que crea los modelos de lenguaje utilizando únicamente los 52 artículos representativos de las 52 ciudades del corpus, normalmente, el mejor rendimiento es obtenido cuando únicamente se utilizan los topónimos, ya que estos se supone que son los que añaden un mayor valor semántico desde el punto de vista geográfico cuando se intentan clasificar textos informales entrenando con una fuente de texto formal. Pero hay que resaltar como el mejor resultado es obtenido precisamente cuando se omiten estos términos en el corpus de usuarios de baja actividad. Los matices lingüísticos vuelven a ser importantes cuando se trata de geolocalizar textos de una extensión tan limitada, pese a que los conjuntos de entrenamiento y de evaluación sean tan dispares como *Wikipedia* y *Twitter*.

Por otro lado, curiosamente, cuanto más pequeño es el conjunto de tuits a evaluar, mejor funciona esta aproximación. Esto es debido a que los términos que aparecen en fuentes formales como *Wikipedia* que son capaces de ubicar geográficamente un texto, se encuentran más dispersos dentro de los grandes conjuntos de tuits.

Otra conclusión interesante que se puede extraer de esta aproximación es que en un conjunto de entrenamiento pequeño como es el de topónimos, la precisión del sistema cae drásticamente. Esto es debido principalmente a que los artículos referenciados en los artículos de las ciudades, en muchas ocasiones no contienen topónimos relacionados con el artículo de la ciudad desde el que se referenciaba. Tal es así que varios de los artículos referenciados por una ciudad, son también referenciados por otra localidad del corpus, haciendo así que el sistema no consiga realizar un entrenamiento muy preciso.

Otra razón es el escaso número de términos (topónimos) que existe para entrenar el sistema y la aún más escasa correlación de estos términos formales con los plasmados en los tuits.

Si se compara la aproximación con los artículos referenciados con la llevada a cabo sin éstos, se puede apreciar como la aproximación realizada con los artículos referenciados generalmente mejora con respecto a *Wikipedia* sin los artículos referenciados, especialmente cuando la cantidad de texto a clasificar es mayor (usuarios más activos). Esta mejora es mucho más significativa cuando se hace uso de todos los términos que aparecen en el corpus de los artículos.

Al utilizar un corpus de entrenamiento más extenso, el uso de topónimos pierde relevancia, ya que el resto de categorías gramaticales y matices lingüísticos aportan un gran peso a la hora de determinar el foco geográfico de los usuarios de *Twitter*.

La aproximación que no hace uso de topónimos obtiene unos resultados muy similares a los logrados con todos lo términos, lo cual refuerza la hipótesis del peso que tienen las demás categorías gramaticales a la hora de determinar el foco geográfico incluso de textos de diversa formalidad.

Entrenamiento con textos de *Flickr*

La mejor precisión se obtuvo cuando se utilizó para la creación de los modelos de lenguajes los textos procedente de la descripción de las fotografías, así como juntando los textos de todos los campos. Esto es debido a que el volumen de texto del corpus de *Flickr* utilizado en estos experimentos no es muy extenso, y en los campos citados es donde más texto se pude hallar, de ahí esa ligera mejora a la hora de clasificar los conjuntos de tuits.

Por otro lado, una vez más se puede observar como los mejores resultados se obtienen cuando se intenta obtener el foco geográfico de los usuarios con un bajo nivel de actividad (entre 1 y 10 tuits). Esto es debido a que ambos corpus, pese a ser de carácter informal, son muy dispares y, por ende, cuanto mayor es el conjunto de términos del corpus a evaluar, mayor es la diferencia existente con el modelo de lenguaje de cada ciudad y más difícil de clasificar correctamente.

Como era de esperar, creando los modelos de lenguaje con tuits, para posteriormente obtener el foco geográfico de otros conjuntos de tuits, es la aproximación que mejores resultados obtiene, aunque la más costosa de implementar en términos de tiempo y espacio.

Respecto a las aproximaciones que hacían uso de *Wikipedia* o *Flickr*, hay que destacar que *Wikipedia* obtiene unos resultados lineales, sin verse alterados éstos por el nivel de actividad de los usuarios, mientras que *Flickr* empeora conforme más activos son los usuarios debido a la gran diferencia existente entre ambos corpus pese a ser ambos informales.

Combinación de corpus de entrenamiento

Al igual que sucediera con los experimentos con textos formales, se ha experimentado ponderando los distintos corpus, pero dicha ponderación no ha sido efectiva ya que siempre funcionaba mejor la implementación que otorgaba todo el peso al corpus de la fuente a clasificar, *Twitter*.

Wikipedia introduce ruido a la hora de generar modelos de lenguajes capaces de obtener el foco geográfico de los conjuntos de tuits evaluados.

Esto se puede apreciar al observar la precisión obtenida cuando los modelos de lenguaje obtenidos tienen más términos (*Todo*), mejorando claramente dicha precisión cuando los modelos de lenguaje contienen menos términos (*Topónimos*), donde se obtienen unos resultados similares a los conseguidos con la utilización única del corpus de *Twitter* para generar dichos modelos de lenguaje al haber un ratio mucho menor de términos procedentes de *Wikipedia* que de *Twitter* cuando se generan estos modelos.

Comparando la aproximación de *Twitter* y *Wikipedia*, con la de *Twitter* y *Wikipedia* con los artículos referenciados, se observa que cuanto mayor es el volumen de los textos a clasificar, mayor es la cantidad de ruido que introduce el entrenamiento de *Wikipedia* con sus enlaces salientes con respecto a la aproximación que no hace uso de estos artículos. Aunque, cabe destacar que cuando se intenta clasificar conjuntos de tuits con un escaso volumen, la incorporación de los textos de *Wikipedia* con los enlaces salientes, aporta una ligera mejoría.

Las conclusiones que se pueden extraer de la aproximación que genera los modelos de lenguaje a partir de *Twitter* y *Flickr*, muestra que el uso de los textos procedentes de los comentarios vertidos sobre cada foto en *Flickr* introduce una cantidad ingente de ruido que, incluso con el escaso peso que tienen estos textos en el sistema, hace que la detección del foco geográfico de los conjuntos de tuits decaiga drásticamente. Lo mismo sucede cuando se utilizan todos los campos de textos disponibles en *Flickr*.

En cambio, si los textos a clasificar contienen una cantidad mínima de datos, el aporte de estas otras fuentes no sólo no es perjudicial, sino que puede llegar a ser beneficioso.

Respecto a los experimentos llevados a cabo utilizando los 3 corpus conjuntamente, una vez más hay que decir lo diluidos que quedan los textos de *Wikipedia* con respecto a los otros corpus debido a su reducido tamaño.

En cuanto a *Flickr*, se ratifica como los resultados utilizando todos los campos claramente introducen una gran cantidad de ruido en los modelos de lenguaje, manteniéndose éste estable cuando se utiliza el campo de descripción de las fotos, el cual logra unos resultados muy similares a cuando se utilizan exclusivamente los conjuntos de tuits para generar los modelos de lenguaje.

Si se comparan las distintas combinaciones expuestas en este apartado con la utilización única y exclusiva de *Twitter* para llevar a cabo esta tarea, se observa como los mejores resultados son obtenidos con la aproximación que utilizaba únicamente los tuits para generar los distintos modelos de lenguaje de las ciudades del corpus. Esto es así debido a que el lenguaje que se puede encontrar en *Twitter* es muy específico, y está repleto de abreviaturas, contracciones, etc., lo cual hace muy difícil de equiparar a cualquier lenguaje empleado en cualquier otra fuente de textos.

También cabe resaltar que pese a que *Flickr* contiene un lenguaje más próximo en la informalidad al de *Twitter*, introduce mucho más ruido que *Wikipedia*, la cual se supone que es una fuente de texto con un lenguaje formal. Aunque, cuando se utilizan las etiquetas de las fotografías en lugar de todo el texto disponible, se consigue una mejora para usuarios de baja actividad.

6

Análisis geográfico del lenguaje

En este capítulo se va a mostrar el estudio realizado sobre los corpus del diario *20Minutos* y *Twitter* empleados en esta tesis. En dicho estudio se mostrará la evolución de los términos que componen los textos según su disposición geográfica y temporal así como su medio de publicación. El objetivo de dicho estudio es aumentar el conocimiento sobre la relación que existe con la terminología empleada para, de esta manera, mejorar futuros sistemas de recuperación de información geográfica (*GIR*) en la detección del foco geográfico.

Para dicho análisis, se han abordado las siguientes tareas:

1. Análisis de la correlación existente entre los términos del corpus del diario *20Minutos* y los de *Twitter*.
2. Análisis de la evolución temporal de los términos en ambos corpus.
3. Análisis de la distribución geográfica de los términos de ambos corpus.
4. Análisis de la similitud existente del lenguaje empleado entre diferentes localizaciones.
5. Análisis de los términos más representativos de cada lugar.

6.1. Correlación entre corpus

En esta sección se va a mostrar la correlación existente entre los términos del corpus del diario *20Minutos* y el de los mensajes de *Twitter*. Para ello se medirá cuán similar son los textos obtenidos de la red social con los del diario. Esta similitud se medirá entre el corpus de tuits y el de noticias que comprenden tres distintos periodos de tiempo: el que va antes de cuando se obtuvieron los tuits (*antes*), el que abarca el mismo tiempo que cuando se obtuvieron los tuits (*durante*) y el posterior a la recogida de los mensajes de *Twitter* (*después*).

Los tuits del corpus se recogieron desde el 20 de abril de 2013 hasta el 10 de junio de ese mismo año, ambos inclusive, lo que conforma un total de 52 días, aunque el primer y el último día no sean días completos. Puesto

que se pretende comprobar lo similar que resulta el corpus de *Twitter* con el de *20Minutos*, para este análisis se han obtenido las noticias del diario *20Minutos*, publicadas en las ciudades de las que se recogieron los tuits, pertenecientes a los 52 días previos al 20 de abril de 2013 (*antes*), las pertenecientes a los 52 días posteriores al 10 de junio del 2013 (*después*), y las que tuvieron lugar en los mismos días que se recogieron los tuits (*durante*).

El motivo de esta distinción es debido a la hipótesis que se quiere corroborar: los textos vertidos en la red social *Twitter* suelen ser reflejo de lo que sucede en la sociedad, por lo que debería de existir una mayor correlación entre textos de las fuentes analizadas cuando éstas sean coetáneas.

Así pues, se compararon los términos de cada uno de estos tres conjuntos de noticias con el conjunto de tuits con el que se ha trabajado a lo largo de esta tesis. La comparación se realizó a nivel de localidad, es decir, se comparó la similitud de los términos del corpus de noticias de una ciudad dada para cada uno de los periodos temporales previamente descritos, con la de los textos obtenidos en *Twitter* para esa misma localidad.

Para comprobar la correlación existente entre cada uno de estos textos se ha utilizado la divergencia de *Kullback-Leibler* (*KL*) (Kullback and Leibler, 1951) (Kullback, 1997) (Kullback, 1987).

En teoría de la probabilidad y teoría de la información, la divergencia de *Kullback-Leibler* (*KL*) (también conocida como divergencia de la información, ganancia de la información, entropía relativa o *KLIC* por sus siglas en inglés) es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad P y Q. En otras palabras, es la cantidad de información perdida cuando Q se utiliza para aproximarse a P. La divergencia *KL* también mide el número esperado de bits adicionales necesarios para codificar ejemplos de P utilizando un código optimizado para Q en lugar del código optimizado para P. Generalmente P representa la "verdadera" distribución de los datos, observaciones, o cualquier distribución teórica. La medida Q generalmente representa una teoría, modelo, descripción o aproximación de P (Burnham and Anderson, 2003).

La divergencia *KL* para un corpus P con respecto de un corpus Q, cuyo vocabulario se encuentra en el conjunto finito de términos χ, se define en la ecuación 6.1.

$$D(P||Q) = \sum_{x \epsilon \chi} P(x) log \frac{P(x)}{Q(x)} \tag{6.1}$$

Puesto que la divergencia *KL* es una medida no simétrica, se ha procedido a calcular la divergencia *KL* simétrica (la distancia *KL*) basándonos en la aproximación descrita en Bigi (2003).

Así pues, el cálculo de la distancia *KL* que se ha efectuado entre cada uno de los tres corpus del diario *20Minutos* (*antes*, *durante* y *después*) y

el corpus de *Twitter* por cada una de las localidades de dichos corpus, se muestra en la ecuación 6.2.

$$D(P||Q) = \sum_{x \epsilon \chi} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} \tag{6.2}$$

De este modo, para cada ciudad analizada se obtienen todos los términos existentes en *Twitter* y la frecuencia de estos en dicho corpus. Por otro lado, se realiza la misma operación para el corpus de noticias del periodo temporal contra el que se quiere medir la distancia de la misma ciudad que se está analizando en *Twitter*.

La unión de todos los términos existentes en ambos corpus, es decir, el vocabulario existente en ambos corpus, queda recogido en la ecuación 6.2 por el símbolo χ. Por otro lado, P y Q representan los vectores normalizados de la frecuencia de los términos que aparecen en los corpus que se están analizando de *Twitter* y *20Minutos* respectivamente.

Para esta normalización se tiene en cuenta el número total de términos que aparecen en el corpus de la ciudad dada y sus frecuencias. A esta normalización también se le ha aplicado un suavizado (expresado en la ecuación 6.5) que tiene en cuenta los términos que no aparecen en el corpus de *Twitter* de la ciudad tratada pero sí en en vocabulario existente en χ.

Una vez se tienen todos los términos normalizados de la ciudad tratada de *Twitter*, se crea un vector con el peso de cada término que aparece en χ. Para los términos que aparecen en el vocabulario (χ) y no lo hacen en el corpus de *Twitter* de la ciudad analizada, se le da un valor obtenido de dividir el valor de suavizado ϵ obtenido en la ecuación 6.5 entre el número total de términos que aparecen en χ y no lo hacen en el vocabulario de la ciudad tratada de *Twitter*, teniendo en cuenta también la frecuencia de estos términos. Así pues, la normalización se efectúa según se muestra en 6.3:

$$P(t_i, c_j) = \begin{cases} \frac{freq(t_i)}{d}, si\ t_i\ está\ en\ el\ corpus \\ \frac{\epsilon}{d}, si\ el\ t_i\ no\ está\ en\ el\ corpus \end{cases} \tag{6.3}$$

En esta definición, $P(t_i, c_j)$ representa la probabilidad de un término i en una ciudad j, $freq(t_i)$ es la frecuencia del término i en la ciudad j, d es el divisor que distribuye el peso de ϵ entre los términos que no aparecen en el corpus, y ϵ es el suavizado aplicado para los términos que no aparecen en el corpus.

Como se ha mencionado previamente, el vector resultante se denota en la ecuación 6.2 por el símbolo P, mientras que Q representa el vector de términos obtenido del corpus de noticias de la ciudad analizada para un periodo de tiempo concreto. Dicho vector se ha construido de manera análoga al vector P.

El cálculo del divisor d mostrado en la definición 6.3 se realiza según lo expuesto en la ecuación 6.4.

$$d = np * \epsilon + fp \tag{6.4}$$

Donde np es el número de términos de χ que no aparecen el un corpus analizado, ϵ es el valor de suavizado y fp es el número total de términos y sus frecuencias que aparecen en el corpus analizado.

El cálculo del valor de suavizado ϵ se realizó según se muestra en la ecuación 6.5.

$$\epsilon = \frac{\frac{1}{1+(|vp-vq|)}}{1+p+q} \tag{6.5}$$

En esta ecuación, basada en la empleada en Bigi (2003), $|vp - vq|$ es el número total de términos que no son comunes en los dos corpus analizados, p representa al número total de términos del corpus de la ciudad analizada de *Twitter* multiplicando cada término por su frecuencia en el corpus, y q representa al número total de términos de la ciudad analizada para el periodo de tiempo con el que se está comparando del corpus del diario *20Minutos* multiplicado por la frecuencia de cada término en dicho corpus.

Así pues, se ha procedido a realizar el cálculo de la distancia *KL* previamente descrito, a nivel de ciudad entre el corpus de dicha ciudad de *Twitter* y el de esa misma ciudad en el diario *20Minutos* para cada uno de sus periodos de tiempo comentados.

Los resultados se pueden observar en la tabla 6.1. Nótese que a menor distancia *KL*, mayor correlación entre corpus.

Tabla 6.1: Resultados de la distancia *KL* de los corpus de las ciudades en *Twitter* con respecto a los corpus de las ciudades en el diario *20Minutos* por intervalo de tiempo. En negrita se resaltan los mejores resultados y las ciudades que lograron obtener dichos resultados cuando los textos de *20Minutos* y *Twitter* coincidían en el tiempo (*Durante*).

Localidad	**Intervalo temporal**		
	Antes	*Durante*	*Después*
A Coruña	**12,94**	**12,94**	13,84
Albacete	15,08	**14,31**	14,42
Alicante	13,00	13,06	**12,97**
Almería	13,48	**13,20**	13,49
Ávila	17,01	17,25	**16,15**
Badajoz	**11,70**	11,75	12,01
Barcelona	15,07	14,83	**14,69**
Bilbao	11,85	**11,79**	11,94

Continúa en la página siguiente...

Tabla 6.1 – *Continuación de la página anterior*

Localidad	**Intervalo temporal**		
	Antes	*Durante*	*Después*
Burgos	**16,09**	16,80	16,19
Cáceres	**12,88**	12,92	13,17
Cádiz	13,05	**12,96**	13,32
Castellón	14,27	**13,67**	14,47
Ceuta	**20,27**	20,65	20,49
Ciudad Real	**13,51**	13,65	13,57
Córdoba	12,27	12,11	**11,65**
Cuenca	18,70	**18,02**	18,47
Girona	21,57	20,93	**20,85**
Granada	12,06	12,14	**11,97**
Guadalajara	**15,31**	15,58	15,84
Huelva	13,66	13,22	**13,13**
Huesca	15,26	14,84	**14,72**
Jaén	13,48	13,18	**13,15**
Las Palmas	12,79	**12,55**	12,90
León	**13,86**	14,43	14,61
Lleida	19,47	**19,21**	20,15
Logroño	12,71	**12,45**	12,82
Lugo	15,26	15,08	**14,62**
Madrid	10,35	**10,01**	10,46
Málaga	10,75	**10,38**	10,65
Mallorca	14,00	**13,69**	13,78
Melilla	17,86	**17,83**	18,03
Murcia	10,50	**10,27**	10,58
Ourense	14,73	**14,12**	15,84
Oviedo	11,89	**11,82**	11,87
Palencia	**15,95**	16,38	16,25
Pamplona	**13,60**	13,72	13,62
Pontevedra	**15,81**	15,98	15,97
Salamanca	13,50	**13,04**	13,19
San Sebastián	**16,41**	16,64	16,59
Santa Cruz de Tenerife	**13,79**	14,04	14,25
Santander	11,90	11,51	**11,33**
Segovia	15,20	**15,09**	**15,09**
Sevilla	9,82	**9,64**	9,84
Soria	18,34	**18,04**	18,35
Tarragona	21,48	21,10	**20,95**
Teruel	16,14	**15,71**	16,56

Continúa en la página siguiente...

Tabla 6.1 – *Continuación de la página anterior*

Localidad	**Intervalo temporal**		
	Antes	*Durante*	*Después*
Toledo	12,71	**12,33**	12,52
Valencia	10,26	**10,13**	10,23
Valladolid	10,88	**10,77**	11,08
Vitoria	**14,57**	14,59	14,95
Zamora	16,77	17,20	**16,62**
Zaragoza	11,18	**11,05**	11,13

Como se puede observar en la tablac 6.1, 26 de las 52 poblaciones obtienen la mínima distancia, es decir, que hay una correlación mayor entre los corpus de *Twitter* y el de *20Minutos* para el periodo temporal que coincide con las fechas de emisión de los tuits, lo cual muestra una relevancia mucho más alta de lo que cabría esperar por puro azar (50 % frente al 33 %). Esta relevancia es mucho más significativa cuando se trabaja con ciudades más grandes, es decir, con corpus más extensos. Así pues, si se observan únicamente las 10 ciudades más pobladas de España, es el 90 % de las ciudades las que ofrecen los resultados esperados, tal y como se puede ver en la tabla 6.2.

Tabla 6.2: Resultados de la distancia *KL* de los corpus de las ciudades en *Twitter* con respecto a los corpus de las ciudades en el diario *20Minutos* por intervalo de tiempo entre las 10 ciudades más pobladas.

Localidad	**Intervalo temporal**		
	Antes	*Durante*	*Después*
Barcelona	15,07	14,83	**14,69**
Bilbao	11,85	**11,79**	11,94
Las Palmas	12,79	**12,55**	12,90
Madrid	10,35	**10,01**	10,46
Málaga	10,75	**10,38**	10,65
Mallorca	14,00	**13,69**	13,78
Murcia	10,50	**10,27**	10,58
Sevilla	9,82	**9,64**	9,84
Valencia	10,26	**10,13**	10,23
Zaragoza	11,18	**11,05**	11,13

Cabe destacar la naturaleza distinta de los corpus, donde el corpus del diario *20Minutos* muestra temáticas relacionadas con la actualidad que ocurrió en una zona geográfica concreta durante las fechas descritas, mientras que en la red social *Twitter* se pueden encontrar temáticas muy dispares, sin que tengan éstas por qué estar relacionadas con la localidad desde la que se emite el mensaje. De hecho, en muchas ocasiones se habla

sobre temas que son específicos de otras localidades, por ejemplo: "*4-1 le ha metido el Borussia al Madrid!!!*", en un tuit emitido desde Lugo.

También hay que tener presente, como ya se ha comentado en capítulos anteriores, la escasa o nula información geográfica de la inmensa mayoría de los mensajes emitidos desde *Twitter*.

Por último, otro aspecto a tener presente es que las fechas de los tres corpus de *20Minutos* empleados en este análisis, pese a ser distintas, son contiguas en el tiempo, es decir, que pese a no intersectar fechas sí que están muy próximas, por lo que las temáticas tratadas en estos 3 periodos pueden solaparse, especialmente con el periodo temporal que está en medio que es el que coincide con el del corpus de *Twitter*.

Así pues, teniendo en cuenta las dificultades previamente descritas, se puede concluir que los resultados obtenidos alcanzan una precisión más que notable, sobre todo si se tiene en cuenta el tamaño de las ciudades. De este modo, como ya se ha comentado previamente, se es capaz de lograr un 90 % de precisión para las 10 ciudades más grandes.

También se realizaron pruebas reduciendo el número de términos a 100, 1.000 y 10.000, seleccionando éstos mediante el cálculo de la unidad tipificada (*z-score*) (Benzécri and Bellier, 1976) comentada en la siguiente sección, pero se obtuvieron unos resultados similares aplicando dicha selección de características (términos).

6.2. Evolución temporal del lenguaje

En esta sección se va a mostrar cómo evolucionan los términos a nivel de localidad de ambos corpus, *20Minutos* y *Twitter*, por separado a lo largo del tiempo. De este modo, se podrá comprobar cómo la relevancia de los términos va variando con el tiempo en cada una de las localidades mostradas, así como en el conjunto del país.

Para analizar cómo evolucionan los términos en el tiempo, se dividieron los corpus en periodos temporales (ver secciones 6.2.1 y 6.2.2). Para cada uno de estos periodos se obtuvieron los términos y las frecuencias que aparecían en cada una de las ciudades por separado. Con estos términos y sus frecuencias se procedió a calcular la unidad tipificada de cada término en comparación con el resto de periodos temporales por cada ciudad dada.

Las unidades tipificadas muestran el número de desviaciones típicas en que un valor dado se sitúa por encima o debajo de la media de su muestra o población. Se usan también para comparar valores de diferentes muestras o poblaciones.

La unidad tipificada se calcula según se muestra en la ecuación 6.6,

$$z = \frac{x - \bar{a}}{\sigma} \tag{6.6}$$

donde x es el valor analizado (la frecuencia normalizada del término analizado para un periodo de tiempo concreto en nuestro caso), $\bar{a}$ es la media de valores (la media normalizada de dicho término para una ciudad dada en todos sus periodos temporales) y σ es la desviación típica de dichos valores.

Así pues, se procedió a observar la evolución del *z-score* o unidad tipificada de un conjunto de términos que representaban algunas de las principales preocupaciones de los españoles según el *Centro de Investigaciones Sociológicas* (*CIS*)[1].

Para la detección de las preocupaciones de los españoles, el *CIS* mensualmente realiza encuestas a los ciudadanos pidiéndoles que muestren qué tres temas les preocupan más de entre una lista de ellos[2]. Los encuestados deben seleccionar hasta tres temas y el *CIS* mostrará finalmente el tanto por ciento de ciudadanos que ha seleccionado cada tema.

Los temas seleccionados han sido: paro, corrupción y educación por ser a su vez temas actuales y tres de las principales preocupaciones sociales de los españoles por ambas épocas. De este modo, se pretende saber si en periodos determinados, la preocupación sobre estos temas se puede detectar o predecir, pudiéndose así obtener información social relevante de un modo rápido y económico, sin tener que estar realizando constantes y costosas encuestas.

Para captar en los textos de *20Minutos* y *Twitter* las tres preocupaciones expuestas en el párrafo anterior, se decidió observar las apariciones de los términos que representan dichas temáticas: paro, corrupción y educación. Puesto que estos términos podían aparecer tanto en mayúsculas como en minúsculas, con o sin tilde, especialmente en *Twitter* debido a su informalidad, se decidió agrupar estos términos sin importar si estaban escritos con o sin mayúsculas y con o sin tilde.

En las secciones 6.2.1 y 6.2.2 se muestran las evoluciones temporales de estos términos para un conjunto de ciudades de ejemplo: Alicante, Madrid y Sevilla, así como la evolución del conjunto del país agrupando el resultado de todas las localidades estudiadas (las 50 capitales de provincia más las dos ciudades autónomas) y obteniendo la media. En dicha evolución también se muestra el valor obtenido en la encuesta del *CIS*. Dicho valor representa el tanto por cien de personas encuestadas que mostraba preocupación por dichas temáticas, pudiéndose éstas solapar, es decir, que los encuestados podían mostrar su preocupación por más de una temática.

Para comprobar si un término ganaba o perdía peso en cada uno de los periodos temporales de cada ciudad analizada, se ordenaron todos los términos que aparecían en cada uno de los periodos de cada una de estas

[1] http://www.cis.es/cis/opencms/ES/index.html

[2] http://www.cis.es/opencms/-Archivos/Indicadores/documentos_html/TresProblemas.html

ciudades de acuerdo a su *z-score*, de mayor a menor, es decir, los términos que mostraban un mayor índice de preocupación para un periodo de tiempo dado en una ciudad dada, aparecían al principio de la lista, teniendo en cuenta la posición de la lista que ocupaban el término analizado y el total de términos existentes en el periodo temporal analizado para dicha ciudad, tal y como se muestra en la ecuación 6.7.

$$\textit{Índice de preocupación} = 100 - 100 \times \frac{\textit{Posición del término}}{\textit{Número total de términos}} \quad (6.7)$$

Mediante esta ecuación se obtenían valores en el rango 0-100, lo que permitía compararlos directamente con los mostrados en las encuestas del *CIS*.

Una vez obtenido el índice de preocupación del término dentro de su periodo temporal, se comparaba éste con la obtenida en el resto de periodos temporales.

Nótese que cuanto más próximo a 100 estén los resultados, significa que mayor índice de preocupación tuvo ese término en el año dado.

6.2.1. *20Minutos*

El corpus del diario *20Minutos* fue dividido por años, dando como resultado 4 corpus distintos por cada ciudad: 2008, 2009, 2010 y 2011.

En las siguientes subsecciones se van a mostrar los resultados obtenidos para cada uno de los tres terminos analizados en las ciudades comentadas previamente, así como en el conjunto del país y lo reflejado en el *CIS*.

Evolución temporal de la '*Corrupción*'

En la figura 6.1 se muestra la evolución de la presencia del término *corrupción* en el diario *20Minutos* entre los años 2008 y 2011. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

En lo que respecta al término '*corrupción*' se puede apreciar como difiere mucho el índice de preocupación obtenido en cada ciudad entre sí y entre la que le otorga el *CIS*.

Se puede apreciar también como ciudades como Alicante o Sevilla se sitúan en la media de localidades de España a lo largo del periodo temporal analizado, mientras que Madrid registró una tendencia opuesta a éstas. A nivel de ciudad suceden grandes variaciones en la percepción de la corrupción dependiendo de si ha surgido algún caso de ésta en la ciudad o provincia, y de cuando está en apogeo dicho caso.

Así pues, si analizamos una ciudad como Alicante, casos como Gürtel[3] que estalló a finales de 2007 en la comunidad Valenciana, hacen que durante

[3] https://es.wikipedia.org/wiki/Caso_G%C3%BCrtel

Figura 6.1: Evolución temporal de la preocupación por la corrupción según el índice de preocupación del término 'corrupción' en las noticias del diario *20Minutos* para las ciudades indicadas.

los años siguientes fuera aumentando la percepción de corrupción, llegando a su apogeo cuando el expresidente de la comunidad, Francisco Camps, presentó su dimisión a mediados de 2011.

Si centramos más el foco en la ciudad de Alicante, el caso Brugal[4] que se destapó en el año 2006, vio como en el 2010 comenzó su segunda etapa, lo que hace que durante ese año y los siguientes, la sensación de corrupción en dicha ciudad aumentara.

Otro caso que contribuyó a la percepción de una mayor corrupción en Alicante fue el de la Caja de Ahorros del Mediterráneo (*CAM*)[5], que fue intervenida en julio de 2011.

En otras ciudades como Sevilla sucedieron casos parecidos (casos de los ERE [6]), lo cual hace que la percepción de la corrupción, como se ha indicado previamente, varíe en función de los casos que haya en cada momento en cada lugar y del momento en el que éstos se encuentren.

En lo que si que coinciden las muestras expuestas, menos Madrid, y la del *CIS* es en la tendencia alcista que hay en este tema, con lo que se puede decir que la prensa reflejaba lo que estaba ocurriendo.

Otra cosa a resaltar de este estudio es que según el *CIS*, la corrupción no parecía ser una gran preocupación para los españoles durante los años analizados (preocupaba entre al 1 y el 5 % de los españoles). En los años posteriores a los mostrados en el estudio, hubo un incremento exponencial

[4] https://es.wikipedia.org/wiki/Caso_Brugal
[5] https://es.wikipedia.org/wiki/Caja_Mediterr%C3%A1neo
[6] https://es.wikipedia.org/wiki/Caso_ERE_en_Andaluc%C3%ADa

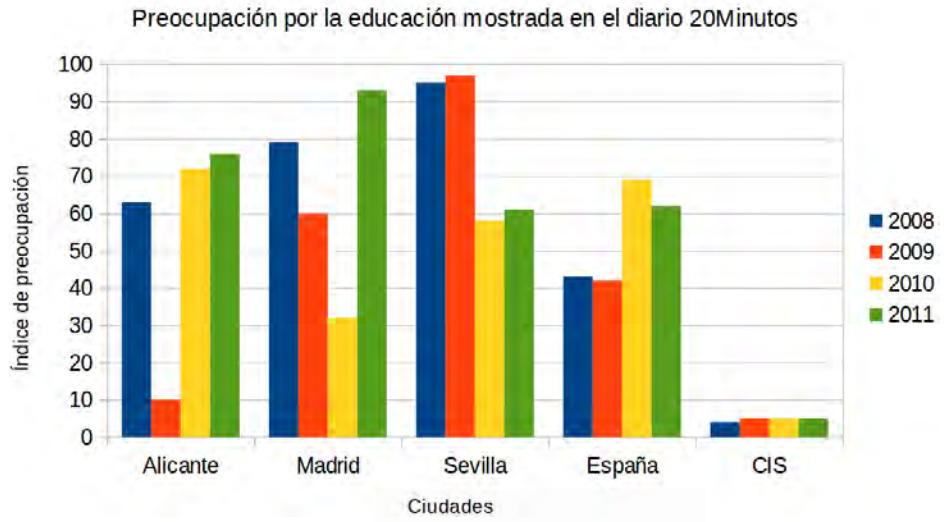


Figura 6.2: Evolución temporal de la preocupación por la educación según el índice de preocupación del término 'educación' en las noticias del diario *20Minutos* para las ciudades indicadas.

de ésta, llegando a tener niveles por encima del 60 % en 2014, y alrededor del 50 % en 2016. Habría que analizar si este aumento en la percepción de la corrupción es debido a un aumento de la misma, o a una mayor presencia de ésta en los medios de comunicación, pudiéndose de esta manera anticipar por dónde van a ir las futuras preocupaciones de los ciudadanos.

Evolución temporal de la '*Educación*'

En la figura 6.2 se muestra la evolución de la presencia del término *educación* en el diario *20Minutos* entre los años 2008 y 2011. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

Una vez más, al pertenecer cada una de las ciudades analizadas a distintas comunidades autónomas, y al tener políticas distintas en educación cada una de estas comunidades debido a que tienen derivada dicha competencia, se pueden observar grandes variaciones entre unas ciudades y otras, dependiendo de las medidas y recortes adoptados en cada una de sus respectivas comunidades. Aún así, se puede apreciar como la preocupación por la educación va creciendo a nivel nacional en los textos del diario a medida que va avanzando la crisis económica, y por ende los recortes. Finalmente, entre las ciudades mostradas se pude apreciar un grado final de preocupación por la educación muy similar entre ellas, y muy por encima de lo que muestra el *CIS*.

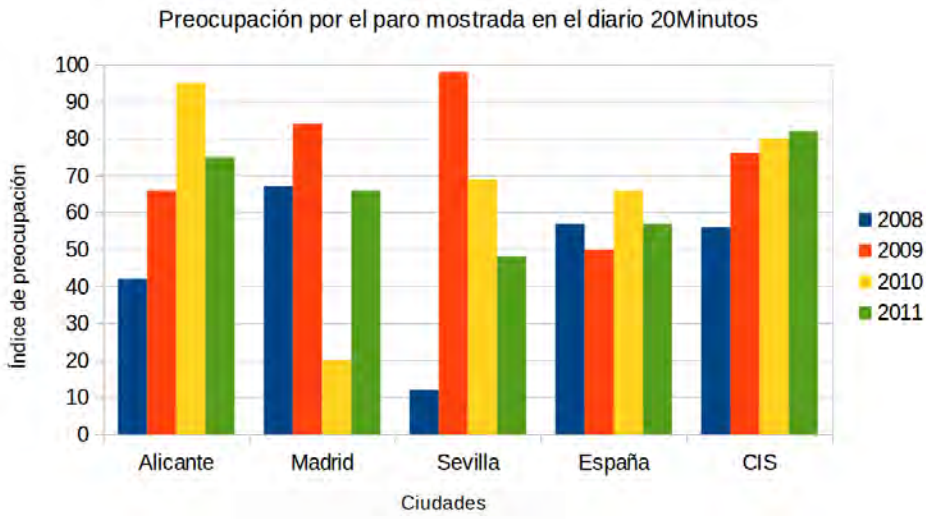


Figura 6.3: Evolución temporal de la preocupación por el paro según el índice de preocupación del término 'paro' en las noticias del diario *20Minutos* para las ciudades indicadas.

Según el *CIS*, el grado de preocupación de la ciudadanía por la educación en los años estudiados estaba alrededor del 5 %. Este grado ha ido creciendo paulatinamente, llegando a duplicar esa cifra inicial en el 2016, por lo que, una vez más, si nos fijamos en el índice a nivel nacional, parece que lo mostrado en la prensa se adelanta a lo que posteriormente recogen las encuestas del *CIS*. Esto puede ser debido también a que las medidas que adopta el gobierno, generalmente recortes en plena crisis financiera, de las cuales se hace eco la prensa, no tienen una repercusión final hasta pasado cierto tiempo. Pasado este tiempo, es cuando la ciudadanía empieza a notar el efecto de las medidas adoptadas, y por tanto a expresar su preocupación.

Evolución temporal del '*Paro*'

En la figura 6.3 se muestra la evolución de la presencia del término *paro* en el diario *20Minutos* entre los años 2008 y 2011. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

Durante los 4 años analizados, cabe destacar que el paro fue la mayor preocupación entre los españoles según el *CIS*. En esta ocasión, la preocupación mostrada por el *CIS* coincide con la mostrada en el diario *20Minutos* en las ciudades dadas y la media del país.

Según el *CIS*, la preocupación por el desempleo fue creciente durante estos 4 años analizados. Algo parecido muestra el ejemplo de Alicante, el cual es el más parecido de los expuestos en la gráfica a la media del país,

con la excepción del último año mostrado (2011) que comenzó a decrecer, al igual que a nivel nacional. Según el *CIS*, pese a que el desempleo continúa siendo la principal preocupación de los españoles en el año 2016, ésta ha decrecido hasta niveles que se sitúan entre los mostrados para los años 2009 y 2010, lo cual ya se pudo observar en los resultados obtenidos del diario *20Minutos* a nivel nacional.

Más allá de que los resultados mostrados por los medios de comunicación puedan condicionar la opinión de los ciudadanos, según el estudio realizado, estos medios de comunicación parecen ser un termómetro de lo que opinarán los ciudadanos en un futuro.

6.2.2. *Twitter*

Puesto que el corpus de *Twitter* estaba compuesto por tuits recogidos en 50 días completos distintos, más dos días, el inicial el final, no completos, con el fin de poder observar una evolución temporal se decidió dividir el corpus en cinco partes, comprendiendo cada una de estas divisiones periodos de 10 días (el primer y el último periodo tiene un poco más de 10 días, aunque menos de 11): 20/04/2013-30/04/2013, 01/05/2013-10/05/2013, 11/05/2013-20/05/2013, 21/05/2013-30/05/2013 y 31/05/2013-10/06/2013.

En las siguientes secciones se muestra la evolución de las preocupaciones mentadas previamente en las localidades de muestra indicadas en las figuras.

En esta ocasión, hay que resaltar que puesto que el *CIS* muestra los resultados de su encuesta mensualmente, y dado que cada periodo analizado engloba únicamente 10 días, los valores del *CIS* no varían durante varios de los periodos mostrados en las gráficas (los que comprenden el mes de abril).

También hay que destacar que debido a que los periodos de tiempo abarcan únicamente 10 días, los resultados de la unidad tipificada obtenidos pueden oscilar considerablemente.

Evolución temporal de la '*Corrupción*'

En la figura 6.4 se muestra la evolución de la presencia del término *corrupción* en *Twitter* entre las fechas indicadas. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

En las redes sociales, el índice de preocupación medio del término '*corrupción*' para el periodo de tuits indicado ha sido menor que lo mostrado en el diario *20Minutos* para los años expuestos, pese a que como se puede apreciar en la comparación de los resultados del *CIS* en ambas gráficas(6.1 y 6.4), la preocupación de los españoles por la corrupción ha aumentado considerablemente (unos 30 puntos más) según los estudios del *CIS*.

Entre las ciudades mostradas, Alicante y Sevilla se aprecia como siguen una tendencia para el índice de preocupación del término '*corrupción*'

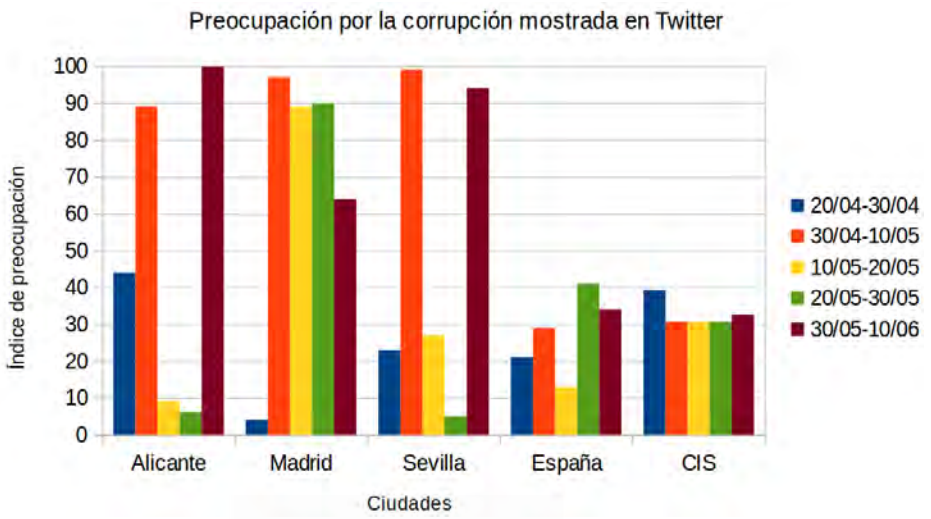


Figura 6.4: Evolución temporal de la preocupación por la corrupción según el índice de preocupación del término 'corrupción' en los mensajes emitidos en la red social *Twitter* para las ciudades indicadas.

prácticamente idénticas, mientras que en Madrid, la preocupación por la corrupción no descendió en el tercer y cuarto periodo tal y como lo hicieron éstas dos ciudades.

Hay que resaltar cómo la media obtenida para el conjunto de las ciudades obtiene una puntuación muy similar a la lograda en la encuesta del *CIS*, especialmente si se hiciera la media de lo ocurrido en el mes de abril, que es lo que finalmente muestra la encuesta del *CIS*, por lo que en este aspecto, *Twitter* parece reflejar fielmente lo que muestra la encuesta del *CIS*.

Evolución temporal de la '*Educación*'

En la figura 6.5 se muestra la evolución de la presencia del término *educación* en *Twitter* entre las fechas indicadas. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

En esta ocasión, el término '*educación*' ha obtenido unos valores más diferenciados que el anterior. La fluctuación que ha sufrido el índice de preocupación de este término en *Twitter* ha sido muy alta, oscilando entre los 20 y 100 puntos.

Esto es debido a que el término seleccionado como representación de la educación, en el sentido de la enseñanza, tal y como muestra la figura 6.6, tiene un uso muy extendido en otros sentidos como se puede apreciar en la figura 6.7. Tal vez, sería más interesante filtrar tuits que utilicen dicho término dentro de un *hashtag*, tal y como se muestra en la figura 6.8, aunque

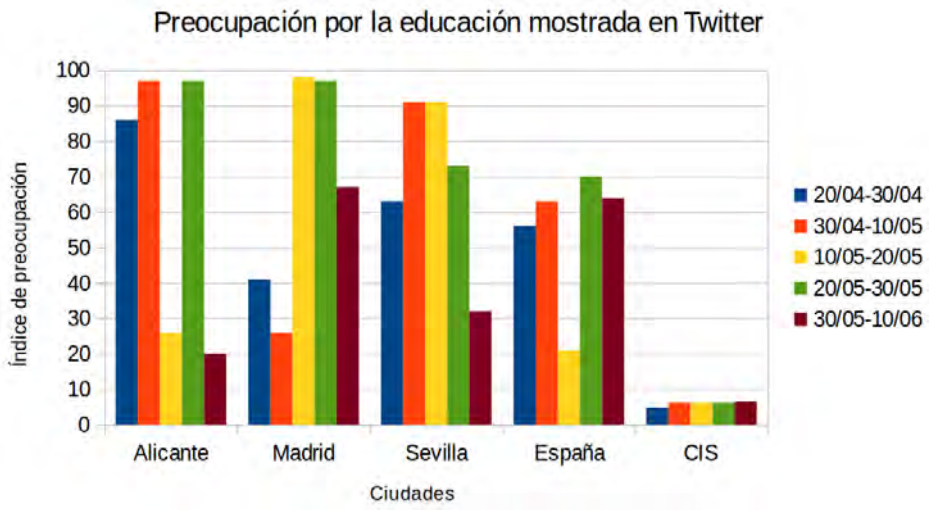


Figura 6.5: Evolución temporal de la preocupación por la educación según el índice de preocupación del término 'educación' en los mensajes emitidos en la red social *Twitter* para las ciudades indicadas.

Figura 6.6: Tuit que muestra preocupación por la*educación*' en el sentido que ha sido preguntado en las encuestas del *CIS*.

ello suponga que se pierdan tuits como el mencionado en la figura 6.6, pero el problema es que en un espacio de 10 días no existe una gran cantidad de tuits que contengan dichos *hashtags*. Incluso hay ciudades que no los contienen.

Según se aprecia en la gráfica de la figura 6.5, Alicante parece ser una muestra significativa de cómo se ha percibido la educación en el resto de España a través de *Twitter*.

Por otro lado, según muestran los datos del *CIS* y de *Twitter*, la preocupación por la educación es mucho mayor en la red social que en la encuesta publicada, en la que la educación, al igual que sucediera con lo mostrado en los resultados del *CIS* en la gráfica 6.2, no parecía ser una de las principales preocupaciones de los españoles, aunque, comparando los resultados del *CIS* de ambas gráficas, sí que se percibe que ha aumentado considerablemente de una fecha a otra.

Figura 6.7: Tuit que muestra otra acepción del término '*educación*'.

Figura 6.8: Tuit que muestra preocupación por la educación incluyendo el término '*educación*' en un *hashtag*.

Evolución temporal del '*Paro*'

En la figura 6.9 se muestra la evolución de la presencia del término *paro* en *Twitter* entre las fechas indicadas. Este índice de preocupación se compara con el mostrado en las encuestas llevadas a cabo por el *CIS* para este mismo periodo temporal.

Respecto al término '*paro*', los datos muestra un índice de preocupación desigual entre las distintas ciudades.

En esta ocasión, el paro es más relevante en la encuesta publicada por el *CIS* (es la primera preocupación de los españoles, con diferencia, según dicha encuesta) que en la media obtenida en la red social.

Por otro lado, como ya se comentó previamente, al analizarse periodos temporales tan cortos encontramos grandes oscilaciones en el *z-score* obtenido, por lo que si obviamos el cuarto periodo temporal, se puede apreciar como en la ciudad de Alicante el paro consigue una preocupación muy similar a la mostrada en la encuesta del *CIS*, así como Sevilla si no se tiene en cuenta el segundo periodo temporal.

6.3. Distribución geográfica de términos

En esta sección se seguirá trabajando con las preocupaciones del *CIS* con las que se trabajó en la sección anterior para mostrar cuán relevante es cada

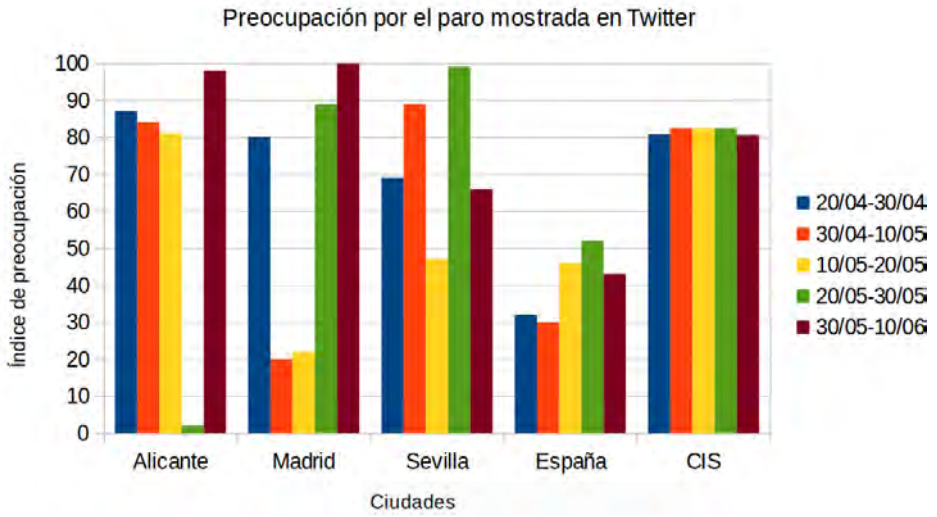


Figura 6.9: Evolución temporal de la preocupación por el paro según el índice de preocupación del término 'paro' en los mensajes emitidos en la red social *Twitter* para las ciudades indicadas.

una de estas preocupaciones en cada una de las 52 localidades tratadas con respecto a las demás. Estas localidades, como ya se ha comentado previamente, se han utilizado como las representantes de cada una de sus respectivas provincias. De esta forma, se han podido crear mapas coropléticos[7] que muestren el nivel de relevancia de cada uno de los términos que representan las citadas preocupaciones ciudadanas obtenidas del *CIS* expuestas en la sección anterior.

Este estudio pretende mostrar la conexión geográfica que hay por cada una de las preocupaciones comentadas previamente. Es decir, si el desempleo aumenta en un promedio más alto en una zona geográfica concreta, se debería de reflejar dicho aumento en los medios de comunicación y redes sociales al expresar estos un creciente aumento en los textos que comentan dicha problemática.

Las zonas geográficas tratadas en este estudio comprenden las capitales de provincia en representación de estas mismas provincias, pero hay que tener en cuenta que cuando surge alguna alarma social, como puede ser el caso del desempleo, no se suele ver afectada de modo aislado una única provincia, sino que, al menos, toda su comunidad autónoma lo hace, y en muchas ocasiones las limítrofes a ésta.

[7]Un mapa coroplético, mapa coropleto o mapa de coropletas, es un mapa temático en el que las regiones se colorean de un motivo que muestra una medida estadística, como puede ser la densidad de población o el ingreso por habitante. Este tipo de mapa facilita la comparación de una medida estadística de una región con la de otra o muestra la variabilidad de esta para una región dada.

Para la obtención de la relevancia de cada uno de los términos se volvió a utilizar la medida de la unidad tipificada explicada en la sección 6.2, y expresada en la ecuación 6.6, aunque, en esta ocasión, en lugar de obtener el *z-score* de cada uno de los términos comparando dichos términos entre distintos periodos temporales del mismo corpus de la misma ciudad, se ha obtenido comparando los términos mencionados contra cada una de las 52 localidades existentes en los corpus del diario *20Minutos* y *Twitter* por separado.

Con el valor de la unidad tipificada de cada ciudad, la cual representa a su provincia, de cada uno de los términos seleccionados, se han creado los mapas coropléticos con la herramienta *QGIS*[8] que se muestran en las secciones 6.3.1 y 6.3.2, pertenecientes a los corpus de *20Minutos* y de *Twitter* respectivamente.

En dichos mapas, los colores más cálidos (los más próximos al rojo) representan una mayor relevancia de los términos en dicha ciudad, mientras que los más fríos (los más próximos al azul), quieren decir que el término analizado tuvo una menor repercusión en la localidad en cuestión con respecto al resto de localidades del corpus. Los rangos de valores están separados en 5 grupos por cuantiles, es decir, introduciendo el mismo número de clases en cada grupo (+/-1), o lo que es lo mismo, 10 u 11 localidades por cada tonalidad mostrada en el mapa. De esta forma, se clasificaron las provincias en los siguientes 5 grupos: muy relevante (rojo), relevante (naranja), medianamente relevante (amarillo), poco relevante (azul claro) y muy poco relevante (azul oscuro).

6.3.1. *20Minutos*

En esta sección se va a mostrar la distribución geográfica de las tres preocupaciones estudiadas en la sección anterior, según la relevancia que se le dio en las noticias publicadas en el diario *20Minutos*, de las 52 ciudades del corpus, para los años comprendidos entre el 2008 y 2011.

Distribución geográfica de la '*Corrupción*'

La figura 6.10 muestra la relevancia del término '*corrupción*' según se ha recogido en el diario *20Minutos* durante el periodo de tiempo indicado.

Se puede apreciar cómo dicho término tomó una especial relevancia en el levante español (Comunidad Valenciana, Baleares, Murcia y Almería), así como en Castilla la Mancha o las Islas Canarias, todos ellos, sitios muy azotados por tramas de corrupción que afloraron en el periodo temporal mencionado.

[8] www.qgis.org

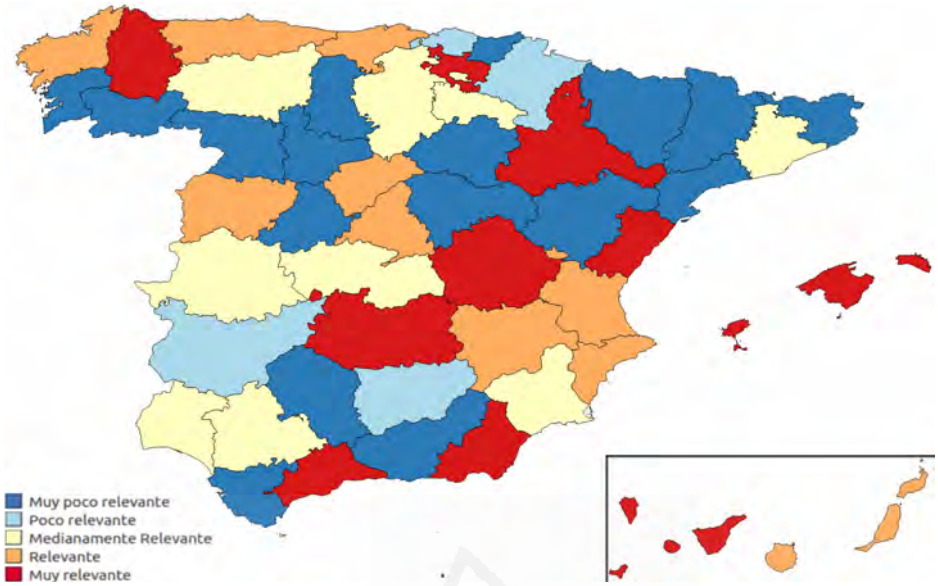


Figura 6.10: Mapa coroplético que muestra la relevancia del término '*corrupción*', medido con la unidad tipificada, en el corpus del diario *20Minutos* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

Distribución geográfica de la '*Educación*'

La figura 6.11 representa la preocupación por la 'educación' emitida en el diario *20Minutos* durante las fechas indicadas.

Se puede apreciar como se le otorga una mayor importancia a este término en la zona central de la península, especialmente en comunidades autónomas como Madrid, Castilla la Mancha, Castilla León, La Rioja y Aragón.

Los motivos que otorgan mayor importancia al término educación en estas comunidades son diversos, pero se pueden clasificar como motivos con connotaciones positivas y motivos con connotaciones negativas.

Por un lado, hay artículos relacionados con recortes o imposiciones, tales como la mostrado en la figura 6.12. Este tipo de artículos han sido publicados principalmente en provincias con gobiernos más conservadores (Madrid, Valencia, etc.) por poner impedimentos a medidas adoptadas por un gobierno central más progresista y alineado con las ideas del diario *20Minutos*.

Por otro lado, se han publicado otro tipo de noticias que ensalzan aspectos positivos de medidas adoptadas en el ámbito de educación, como la mostrada en la figura 6.13.

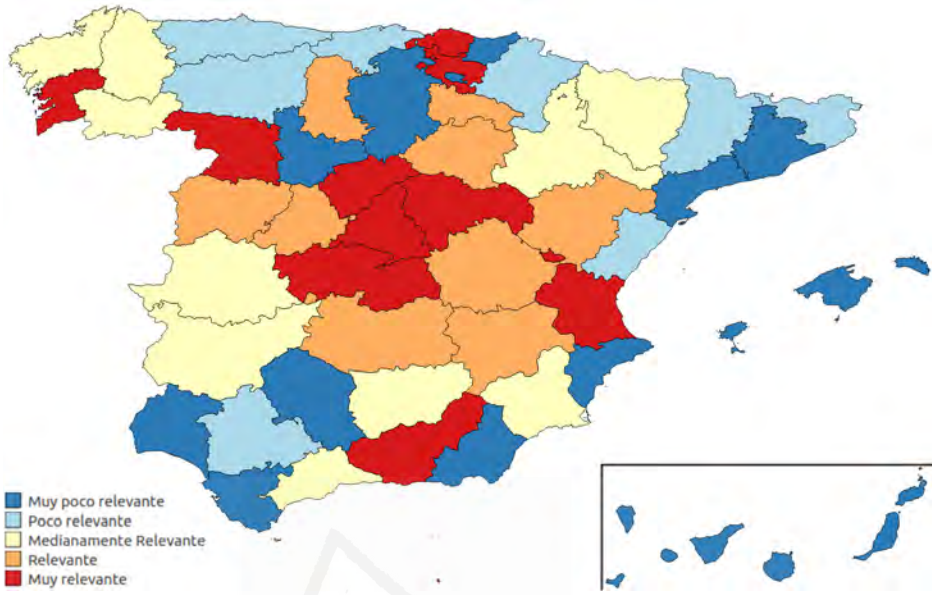


Figura 6.11: Mapa coroplético que muestra la relevancia del término '*educación*', medido con la unidad tipificada, en el corpus del diario *20Minutos* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

Figura 6.12: Noticia del diario *20Minutos* relacionada con educación con connotación negativa.

Figura 6.13: Noticia del diario *20Minutos* relacionada con educación con connotación positiva.

Distribución geográfica del '*Paro*'

La figura 6.14 representa la preocupación por el '*paro*' mostrada en las distintas noticias del diario *20Minutos* publicadas en las ciudades del corpus estudiado.

Se puede observar como en comunidades autónomas como Galicia, Cataluña y Castilla la Mancha mostraron una especial preocupación en el periodo analizado.

Si se consultan los resultados de la Encuesta de Población Activa (EPA)[9] en el Instituto Nacional de Estadística (INE) para el periodo de tiempo analizado (desde el año 2008 al 2011), se observa que el desempleo aumento un 282 % en el conjunto del país desde el último trimestre del año 2007 al primer trimestre del año 2012. Comunidades como Cataluña y Castilla la Mancha estuvieron muy por encima de esa media con un 338 % de aumento para estas dos comunidades, lo que coincide con lo mostrado la figura 6.14.

6.3.2. *Twitter*

En esta sección se va a mostrar un estudio análogo al anterior pero en esta ocasión se va a utilizar el corpus de *Twitter* para mostrar la representación geográfica de los temas previamente indicados.

El conjunto de tuits, como ya se ha visto previamente, estaba comprendido entre el 20 de abril y el 10 de junio de 2013.

[9] http://www.ine.es/jaxiT3/Tabla.htm?t=4247&L=0

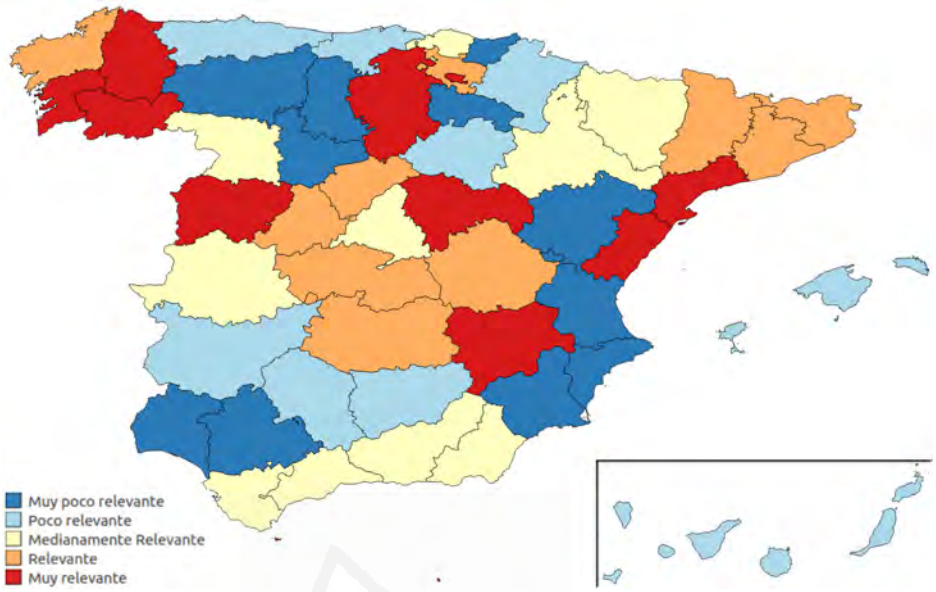


Figura 6.14: Mapa coroplético que muestra la relevancia del término '*paro*', medido con la unidad tipificada, en el corpus del diario *20Minutos* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

Distribución geográfica de la '*Corrupción*'

La figura 6.15 muestra la distribución geográfica del término *corrupción* en *Twitter*.

En esta figura se muestra como en *Twitter* se percibe una mayor preocupación por la 'corrupción' en el levante del país, así como en Madrid y Canarias. Esta preocupación coincide con la mostrada en la figura 6.10 del diario *20Minutos*, en la que pese a comprender un intervalo de tiempo distinto, también se mostraba una mayor preocupación en la zona este del país, extendiéndose a Cataluña en corpus de *Twitter* en el periodo que comprendían los tuits analizados.

Distribución geográfica de la '*Educación*'

La figura 6.16 muestra la distribución geográfica del término *educación* en *Twitter*.

En esta ocasión, a diferencia de lo que ocurría con la corrupción, los usuarios de *Twitter* mostraron una menor preocupación por la 'educación' en el levante español con respecto al resto del territorio, lo cual vuelve a coincidir con lo mostrado en la figura 6.11 que analizaba el mismo término para en el diario *20Minutos* para los años comprendidos entre el 2008 y el 2011.

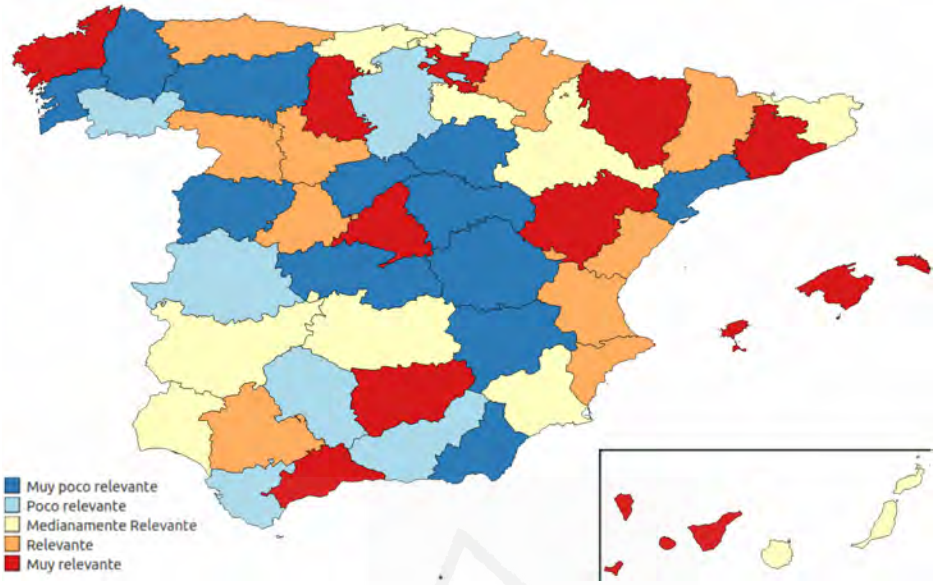


Figura 6.15: Mapa coroplético que muestra la relevancia del término '*corrupción*', medido con la unidad tipificada, en el corpus de *Twitter* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

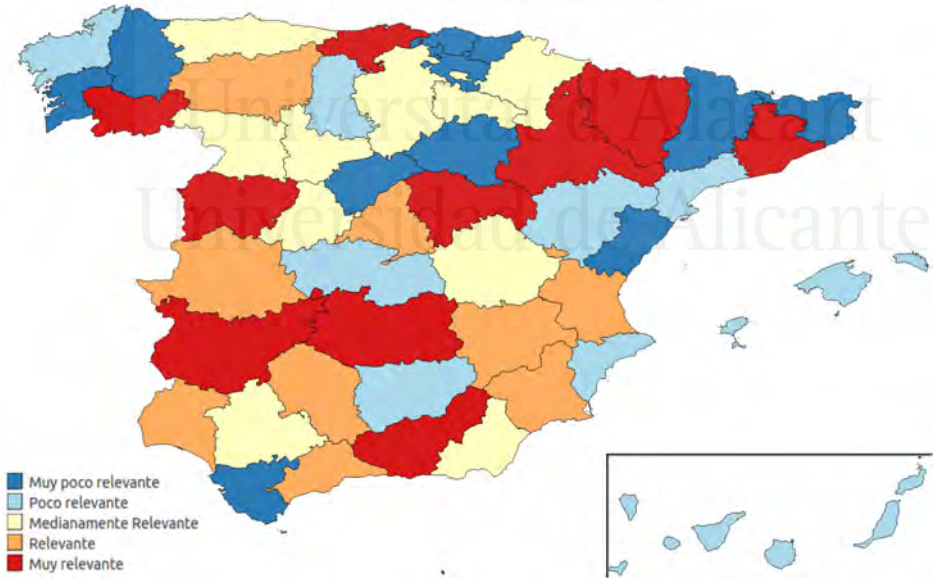


Figura 6.16: Mapa coroplético que muestra la relevancia del término '*educación*', medido con la unidad tipificada, en el corpus de *Twitter* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

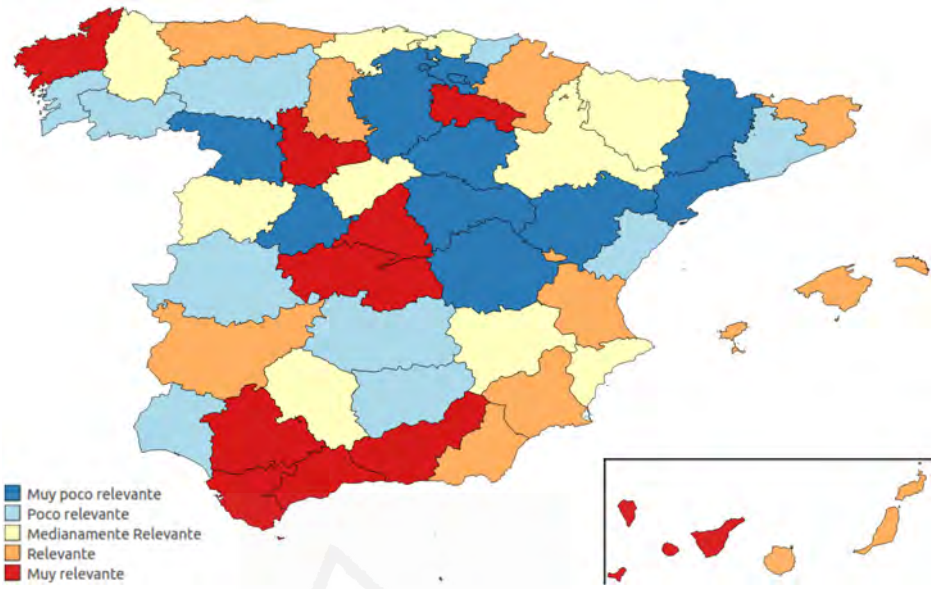


Figura 6.17: Mapa coroplético que muestra la relevancia del término '*paro*', medido con la unidad tipificada, en el corpus de *Twitter* de cada una de las ciudades del corpus que representa a su provincia. Cuanto más cálido es el color, mayor relevancia tiene el término en dicha provincia.

Distribución geográfica del '*Paro*'

La figura 6.17 muestra la distribución geográfica del término *paro* en *Twitter*. En esta figura se aprecia como existe una mayor preocupación por el desempleo en el sur en general, y en las comunidades de Madrid, Andalucía y Canarias en particular.

En esta ocasión, el centro y sur de la península, al igual que sucediera con el diario *20Minutos* (ver figura 6.14) siguen siendo unas de las principales zonas en cuanto a la preocupación por el paro. Por el contrario, Cataluña y Galicia, al contrario de lo que pasó con lo mostrado en el diario, no resaltaron tal preocupación.

6.4. Agrupamiento de ciudades por terminología

En esta sección se va a mostrar la similitud entre el lenguaje empleado entre las distintas ciudades del corpus del diario *20Minutos* y la red social *Twitter* por separado.

El objetivo de este experimento es el de comprobar si existe una similitud entre los distintos matices lingüísticos existentes, expresados tanto en textos formales como informales, entre áreas geográficas próximas unas a las otras.

Para comprobar la similitud existente entre los términos empleados entre cada ciudad, se ha utilizado un método de agrupamiento (*clustering*). Concretamente, se ha utilizado la implementación del algoritmo *k-means* (MacQueen et al., 1967) implementada en el proyecto *Carrot*^{2}[10] (Osiński and Weiss, 2005).

En esta implementación, el algoritmo puede definir automáticamente el número de *clusters* existente, o también se le puede indicar el número máximo de *clusters* deseado, siendo estos *clusters* excluyentes, es decir, ninguna localidad puede aparecer al mismo tiempo en más de un *cluster*. Para comprobar qué número de clusters aportaba una mayor coherencia, se ha probado con un números distinto de los mismos.

Para la utilización de dicho algoritmo se han utilizado todos los términos de los documentos, incluidas sus frecuencias. Los documentos fueron agrupados previamente según la ciudad de procedencia de los mismos, por lo que finalmente se clasificaron 52 documentos, uno por cada ciudad existente en el corpus.

Se ha de tener en cuenta que, a diferencia de los mapas coropléticos mostrados en la sección 6.3, donde la proximidad entre colores tenía un significado, en esta ocasión los mapas son categorizados, es decir, cada color representa una categoría (*cluster*) distinta, sin que signifique que dos *clusters* en los que los colores que los representan tengan una mayor proximidad cromática sean más similares entre sí que un tercer *cluster* representado por un color completamente diferente.

6.4.1. *20Minutos*

Como se ha comentado previamente, se ha decidido probar el sistema con distintos números de *clusters*. Concretamente, se han ido probando un número incremental de *clusters*, empezando por 2 hasta 17 (uno por cada comunidad autónoma).

Para cada número de *clusters* se han obtenido unos resultados similares desde perspectiva de coherencia con respecto a las CCAA, es decir, desde el punto de vista que se esperaba que el lenguaje perteneciente a cada una de las CCAA o provincia colindante fuera más parecido entre sí que entre el resto de ciudades de otras CCAA o provincias más distantes.

A modo de ejemplo, en las figuras 6.18, 6.19 y 6.20 se muestran los resultados obtenidos con el corpus del diario *20Minutos* con 17 *clusters* (uno por cada comunidad autónoma), 10 *clusters* y 5 *clusters* respectivamente.

En la figura 6.18 se ve como al forzar al algoritmo a realizar 17 *clusters* distintos, los resultados obtenidos no llegan a mostrar una gran similitud con lo esperado (provincias categorizadas por comunidad autónoma). Sea cual sea el dominio, siempre es difícil encontrar una cohesión fuerte en los *clusters*

[10] http://project.carrot2.org/

Figura 6.18: Mapa categorizado donde se agrupan las provincias en alguno de sus 17 *clusters* según la similitud que tenga el lenguaje empleado en los artículos del diario *20Minutos* publicados en sus respectivas capitales.

con un número tan grande de ellos. Aún así, sí que se pueden observar unos resultados coherentes como ocurre con las provincias de Extremadura, País Vasco o Canarias. Incluso Castilla León o Ceuta y Melilla (ciudades autónomas distintas pero con muchas características comunes) obtienen unos buenos resultados.

Hay que tener presente la enorme dificultad que entraña para un algoritmo de *clustering* el hacer unos *clusters* que sean coherentes en general, y particularmente desde la tarea que nos atañe, la de detectar similitudes en el lenguaje desde el punto de vista geográfico (Bezdek et al., 1984).

La figura 6.19 muestra como al reducir el número de *clusters* los resultados son mucho más parecidos a lo que cabría esperar, agrupando provincias según la proximidad geográfica de las mismas, lo que demuestra una mayor relación entre el lenguaje empleado entre localidades limítrofes, existiendo, como era de esperar, algunos valores atípicos.

Así pues, se puede apreciar cómo la Comunidad Valenciana está toda incluida en un único *cluster*, aunque, como se ha comentado previamente, en dicho *cluster* se ha añadido un valor atípico, Ávila.

Otro *cluster* interesante que se puede observar es el que forman todas las provincias de las comunidades autónomas de Aragón, País Vasco y Navarra, las cuales son comunidades limítrofes y por tanto tienen muchos matices lingüísticos y culturales en común.

Figura 6.19: Mapa categorizado donde se agrupan las provincias en alguno de sus 10 *clusters* según la similitud que tenga el lenguaje empleado en los artículos del diario *20Minutos* publicados en sus respectivas capitales.

Por lo demás, se observa una gran coherencia en el resto de los *clusters* al existir agrupamientos por provincias del sur en Andalucía, Cataluña y Baleares, las cuales tienen cultura y lengua en común, zona central de la península, ambas Castillas, Extremadura, etc.

Por último, la figura 6.20, en la que se ha limitado a 5 el número de *clusters*, corrobora la teoría de una mayor similitud existente en el lenguaje empleado entre localidades próximas entre sí en comparación con las más lejanas.

En dicha figura, se puede observar *clusters* compactos, como el que engloba todas las provincias que conforman las comunidades autónomas de Aragón, Cataluña, Comunidad Valenciana y Baleares, las cuales tienen muchas similitudes lingüísticas y culturales. En dicho *cluster*, también están incluidas todas las provincias de Galicia. Pese a que dicha comunidad no es limítrofe, hay que tener en cuenta que se ha forzado al algoritmo a crear únicamente 5 *clusters*, y ha agrupado todas las provincias de esta comunidad autónoma, lo cual ya de por sí es muy complicado, aunque haya unido estas provincias con otras más lejanas.

También se pueden apreciar *clusters* como el formado por las provincias del País Vasco y Navarra, las cuales están fuertemente unidas culturalmente, el que forma prácticamente toda Castilla León, el de Castilla la Mancha al completo, o el sur con casi toda Andalucía incluida y Canarias.

Figura 6.20: Mapa categorizado donde se agrupan las provincias en alguno de sus 5 *clusters* según la similitud que tenga el lenguaje empleado en los artículos del diario *20Minutos* publicados en sus respectivas capitales.

Por otro lado Huelva y Salamanca han quedado fuera de sus respectivos *clusters* de CCAA, y tal vez otras CCAA uniprovinciales como Madrid, Asturias, Cantabria o La Rioja deberían haber quedado agrupadas en *clusters* más próximos geográficamente a éstas, aún así, son un número reducido y los resultados muestran una gran coherencia con la hipótesis inicial planteada para este estudio.

6.4.2. *Twitter*

A continuación se muestran mapas con el número de *clusters* igual al mostrado en la subsección anterior, donde, a diferencia de lo que ocurría con el corpus del diario *20Minutos*, en esta ocasión el lenguaje empleado en cada región no guardaba la correlación esperada según la proximidad geográfica.

En la figura 6.21 se muestran los resultados obtenidos cuando se fuerza al algoritmo a crear 17 *clusters*. En dicha imagen, salvo en contadas excepciones, cuesta encontrar similitudes por proximidad geográfica.

En la figura 6.22 se muestra la clasificación obtenida forzando al algoritmo a crear 10 *clusters*. Los resultados mejoran ligeramente a los obtenidos cuando se obtenían 17 grupos distintos, tal y como se puede observan en el *cluster* obtenido para las provincias del norte del país: A Coruña, Lugo, Asturias, Cantabria. Pese a ello, los resultados distan mucho

Figura 6.21: Mapa categorizado donde se agrupan las provincias en alguno de sus 17 *clusters* según la similitud que tenga el lenguaje empleado en *Twitter* emitido desde cada una de las capitales de provincia españolas.

de los obtenidos con el mismo número de *clusters* con los textos del diario *20Minutos* (ver figura 6.19).

En la figura 6.23, que representa la agrupación en 5 *clusters*, se muestra una mayor coherencia entre la correlación del lenguaje empleado entre las distintas localidades del corpus de *Twitter*, aunque, eso sí, sin llegar a la precisión obtenida con los textos del diario *20Minutos*. Aún así se puede apreciar como en la zona este del país predomina el *cluster* de color verde, en la centro el azul celeste y en la oeste el morado.

En esta ocasión, los resultados obtenidos con el corpus de *Twitter* tienen menor congruencia que los obtenidos con el corpus del diario *20Minutos*. Esto es debido a que el corpus de *Twitter* está compuesto por unos textos más heterogéneos, debido, principalmente, a un número mucho mayor de usuarios que escriben los distintos textos procedentes de cada una de las localidades tratadas, así como un lenguaje mucho más genérico en los textos informales que en los formales.

Esto puede ser debido también a que los textos del diario *20Minutos* publicados dentro de las secciones de un ámbito geográfico concreto, tratan específicamente sobre temas que afectan a dicho ámbito, mientras que los tuits emitidos por los usuarios de la red social pueden además tratar temas más genéricos, los cuales pueden afectar a más de una provincia, comunidad o país.

Figura 6.22: Mapa categorizado donde se agrupan las provincias en alguno de sus 10 *clusters* según la similitud que tenga el lenguaje empleado en *Twitter* emitido desde cada una de las capitales de provincia españolas.

Figura 6.23: Mapa categorizado donde se agrupan las provincias en alguno de sus 5 *clusters* según la similitud que tenga el lenguaje empleado en *Twitter* emitido desde cada una de las capitales de provincia españolas.

Otra conclusión que se puede extraer del estudio realizado es que con un número alto de *clusters* es complicado obtener unos resultados coherentes en dicha tarea, especialmente en textos procedentes de *Twitter*. Por el contrario, al reducir el número de *clusters* se puede obtener una mayor congruencia en cuanto a la correlación entre textos de áreas geográficas más próximas, lo cual podría ser utilizado para dirigir mensajes adaptados a dichos grupos que, aunque el ámbito geográfico en el que se centren los mensajes sea más amplio que el de una única provincia o localidad, siempre será mucho más preciso que cuando se dirigen dichos mensajes a nivel nacional.

Como ya se comentó previamente, los algoritmos de *clustering* no suelen tener unos grandes resultados cuando intentan clasificar geográficamente textos basándose meramente en los términos de los propios textos, pero, en los experimentos llevados a cabo en esta sección, se ha logrado una gran coherencia en sus resultados, especialmente cuando se reduce el número de *clusters* (ver figuras 6.19, 6.20 y 6.23).

6.5. Términos más representativos de cada ciudad

En esta última sección se van a mostrar los términos más significativos de las 10 ciudades con más población de España para los corpus del diario *20Minutos*, en el periodo de tiempo que abarca desde el año 2008 al 2011, ambos inclusive, y la red social *Twitter*, en el periodo de tiempo con el que se ha ido trabajando a lo largo de esta tesis.

El objetivo de este experimento es el de poder observar qué términos han sido los más relevantes a la hora de identificar cada una de estas ciudades de muestra, para realizar futuras investigaciones teniendo en cuenta estos términos o patrones que se puedan extraer de los resultados aquí obtenidos.

Para la obtención de estos términos se ha vuelto a utilizar la medida de la unidad tipificada, explicada en la sección 6.2 mediante la ecuación 6.6. De manera análoga a como se hizo en la sección 6.3, se obtuvo la unidad tipificada de cada uno de los términos de cada ciudad del corpus, comparándolos contra cada una de las 52 ciudades existentes en el corpus del diario *20Minutos* y *Twitter* por separado, es decir, se realizó un ranking de los términos más discriminatorios para cada una de las ciudades del corpus. Así pues, una vez obtenido el valor de la unidad tipificada de cada término por cada localidad, se ordenaron de mayor a menor *z-score*, seleccionando los 10 primeros de la lista para ser mostrados.

6.5.1. *20Minutos*

En la tabla 6.3 se muestran los 10 términos más relevantes (con mayor *z-score*) de las 10 ciudades más pobladas del país. Se puede apreciar cómo la inmensa mayoría de los términos son sustantivos, de los cuales, la mayoría son bien topónimos o bien nombres o siglas de empresas o entidades.

Tabla 6.3: Términos más representativos en el corpus del diario *20Minutos* de las 10 ciudades más pobladas de España.

Ciudad	Términos
Madrid	Moratalaz, Hortaleza, SerMaS, Samur-protección, Fuencarral-el, Valdemingómez, Honour, Summa-112, Vicálvaro, Samur
Barcelona	vehicles, cop, Però, carrers, alumnes, majoria, dilluns, sha, trànsit, famílies
Valencia	Benicalap, FEHV, Russafa, Dipu, Emarsa, Benimàmet, l'horta, Betoret, Bromera, Tarongers
Sevilla	Tussam, Mellet, Salteras, Parasol, Aussa, Emasesa, Lipasam, CSIF-A, UCE-A, Tamarguillo
Zaragoza	Lambán, Arcosur, DPZ, ACTUR, Almozara, Meriva, Valdespartera, Juslibol, Tuzsa, Falo
Málaga	MUPAM, AEHCOS, GEMAC, Caneda, Cohard, POTAUM, Moclinejo, Alcazabilla, CPB, Martiricos
Murcia	Sangonera, Limusa, UPCT, Portmán, Lorquí, Belluga, Argem, Murciana, Albudeite, TSJRM
Palma de Mallorca	Munar, Ibatur, PSIB, PSM-IV-ExM, IB-Salut, Cabrer, Inestur, Andratx, Artà, Calvià
Las Palmas de Gran Canaria	Emalsa, Alcaravaneras, Cuyás, Cofete, Jinámar, Inalsa, Tasarte, Pamparacuatro, GC-1, Vegueta
Bilbao	Santutxu, Bilboko, Alhóndigabilbao, Artxanda, Otxarkoaga, Konpartsak, Olabeaga, baracaldeses, Miribilla, Zorrozaurre

Por ejemplo, uno de los términos con un valor más alto para Valencia es '*FEHV*' (Federación Empresarial de Hostelería de Valencia). Este término se puede encontrar frecuentemente en noticias de la propia ciudad. Concretamente en 68 noticias publicadas en esta ciudad. Un ejemplo de noticia con dicho término es la mostrada en la figura 6.24.

El término *FEHV* no aparece en ninguna noticia que no haya sido publicada en Valencia. Otros ejemplos similares al anterior son términos como '*Samur*' (Servicio de Asistencia Municipal de Urgencia y Rescate) en Madrid o el TSJRM (Tribunal Superior de Justicia de la Región de Murcia) en Murcia.

Por otro lado, en las ciudades en las que existe una lengua cooficial (Barcelona, Valencia, Palma de Mallorca y Bilbao), se observa un gran número de términos escritos en dicha lengua, ya que son términos que suelen identificar claramente, o al menos acotar, el origen del texto. En este aspecto cabe destacar que puesto que se está hablando de una fuente de texto formal, el diario *20Minutos*, estos términos no suelen ser muy frecuentes, ya que la lengua en la que se escriben los artículos en dicha fuente es la misma para todas las localidades presentes en el corpus tratado, el castellano, pero sí que existen múltiples citas entrecomilladas escritas en la lengua cooficial de dichas localidades que permiten identificar el origen de estos textos.

En lo que respecta a los sustantivos, muchos de estos pueden ser localizados en ciertas regiones con una mayor facilidad cuando en la citada región existe otra lengua cooficial, ya que suelen ser términos existentes en el vocabulario de esta otra lengua y no existentes en castellano, además de prefijos y/o, sobre todo, sufijos más comunes en la lengua cooficial.

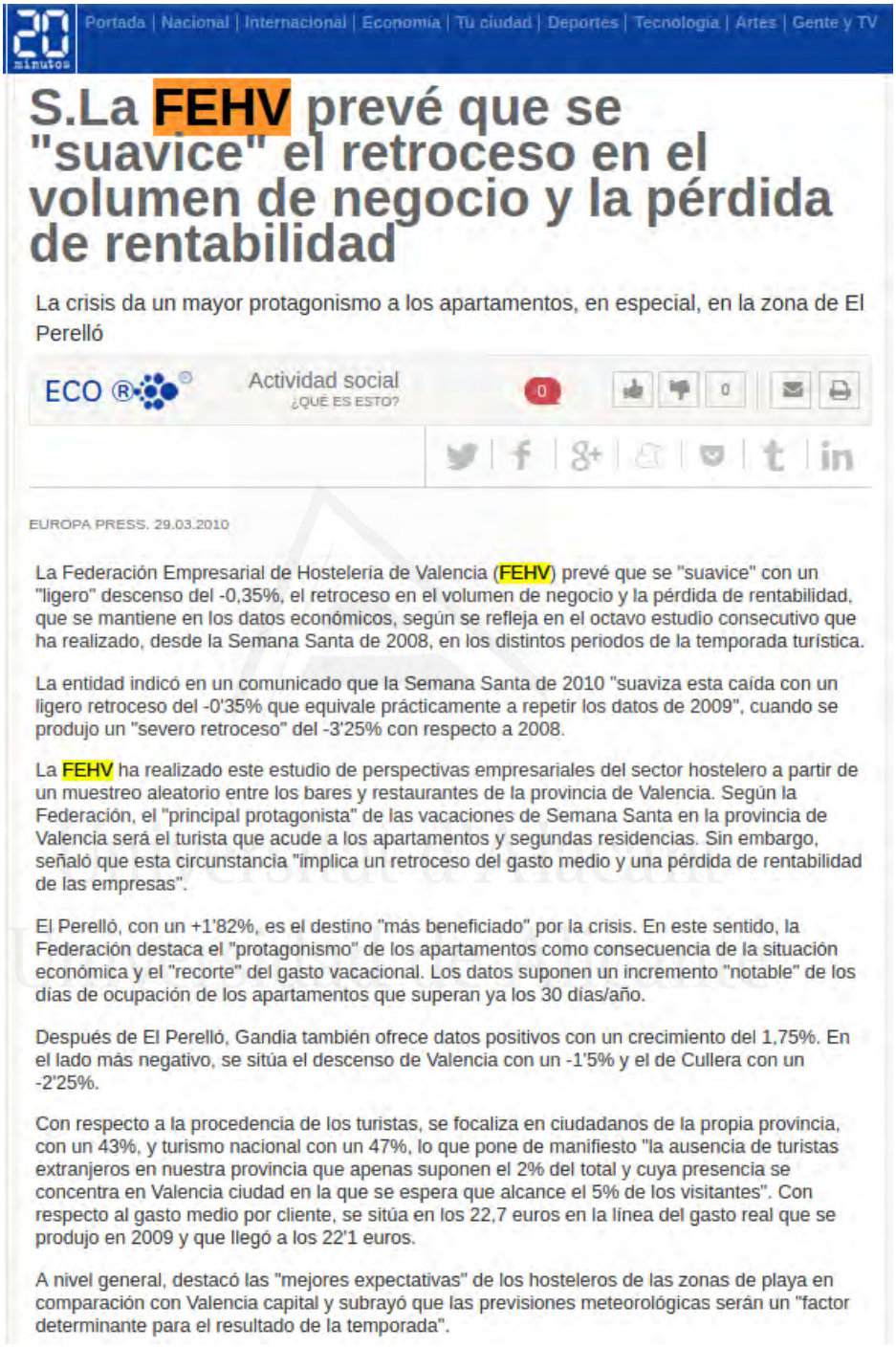
6.5.2. *Twitter*

En la tabla 6.4 se pueden observar los 10 términos más representativos del corpus de mensajes de *Twitter* de las 10 ciudades más pobladas de España.

En esta ocasión, puesto que, como era de esperar, entre los términos más relevantes de cada ciudad existen muchos nombres de usuario (los términos que empieza por '@' en *Twitter*), se ha decidido no tener en cuenta dichos términos por no aportar nada al estudio.

La lista está ordenada de la ciudad más poblada a la que menos y del término más representativo al que menos.

En esta ocasión, se puede observar como el número de topónimos que aparecen en los 10 primeros puestos del ranking de cada ciudad no es muy significativo, en contra de lo que cabría esperar, por suponerse que son los términos que mejor clasifican geográficamente un texto. Tan sólo Madrid con 4 topónimos (Fuencarral, Atocha, Barajas y Callao) es la que hace uso de más nombres de lugar para poderse identificar geográficamente. Esto vuelve a poner de manifiesto la hipótesis inicial de esta tesis en la que se defendía



20 minutos | Portada | Nacional | Internacional | Economía | Tu ciudad | Deportes | Tecnología | Artes | Gente y TV

S.La FEHV prevé que se "suavice" el retroceso en el volumen de negocio y la pérdida de rentabilidad

La crisis da un mayor protagonismo a los apartamentos, en especial, en la zona de El Perelló

ECO ® Actividad social ¿QUÉ ES ESTO? 0 0

EUROPA PRESS. 29.03.2010

La Federación Empresarial de Hostelería de Valencia (FEHV) prevé que se "suavice" con un "ligero" descenso del -0,35%, el retroceso en el volumen de negocio y la pérdida de rentabilidad, que se mantiene en los datos económicos, según se refleja en el octavo estudio consecutivo que ha realizado, desde la Semana Santa de 2008, en los distintos periodos de la temporada turística.

La entidad indicó en un comunicado que la Semana Santa de 2010 "suaviza esta caída con un ligero retroceso del -0'35% que equivale prácticamente a repetir los datos de 2009", cuando se produjo un "severo retroceso" del -3'25% con respecto a 2008.

La FEHV ha realizado este estudio de perspectivas empresariales del sector hostelero a partir de un muestreo aleatorio entre los bares y restaurantes de la provincia de Valencia. Según la Federación, el "principal protagonista" de las vacaciones de Semana Santa en la provincia de Valencia será el turista que acude a los apartamentos y segundas residencias. Sin embargo, señaló que esta circunstancia "implica un retroceso del gasto medio y una pérdida de rentabilidad de las empresas".

El Perelló, con un +1'82%, es el destino "más beneficiado" por la crisis. En este sentido, la Federación destaca el "protagonismo" de los apartamentos como consecuencia de la situación económica y el "recorte" del gasto vacacional. Los datos suponen un incremento "notable" de los días de ocupación de los apartamentos que superan ya los 30 días/año.

Después de El Perelló, Gandia también ofrece datos positivos con un crecimiento del 1,75%. En el lado más negativo, se sitúa el descenso de Valencia con un -1'5% y el de Cullera con un -2'25%.

Con respecto a la procedencia de los turistas, se focaliza en ciudadanos de la propia provincia, con un 43%, y turismo nacional con un 47%, lo que pone de manifiesto "la ausencia de turistas extranjeros en nuestra provincia que apenas suponen el 2% del total y cuya presencia se concentra en Valencia ciudad en la que se espera que alcance el 5% de los visitantes". Con respecto al gasto medio por cliente, se sitúa en los 22,7 euros en la línea del gasto real que se produjo en 2009 y que llegó a los 22'1 euros.

A nivel general, destacó las "mejores expectativas" de los hosteleros de las zonas de playa en comparación con Valencia capital y subrayó que las previsiones meteorológicas serán un "factor determinante para el resultado de la temporada".

Figura 6.24: Noticia del diario *20Minutos* relacionada con *FEHV* y publicada en la ciudad de Valencia.

Tabla 6.4: Términos más representativos en el corpus del diario *20Minutos* de las 10 ciudades más pobladas de España.

Ciudad	Términos
Madrid	Term, Ilustracion, #LaPosada, ALQUILA, Fuencarral, Weasley9, Atocha, Barajas, Estanque, #Madrid, Callao
Barcelona	Barceloneta, ocupant, Gràcia, aparèixer, apareix, Boqueria, endirecte, OFFF, #igersbcn, Batlló, DIV
Valencia	màx, Olivereta, #Valencia, #Amunt, bonaire, #AMUNT, vax, Guerreiro, s86, Saler, #Noticias
Sevilla	#Seville, kmh, LLuvia, SSO, #Sevilla, Racha, ST, #sevillahoy, #TDSActualidad, #3galablogosur, #seville
Zaragoza	#atecaeschocolate, #CW13, #agapitofueraya, bilba, #cw13, #VamosCAI, #SiempreBajoAragón, #internetforum, StC, zaragoza3, #huesitos
Malaga	#ValiolapenaJG, Huelin, Malagueta, blondx, #málaga, Nove, nobe, AGP, Bst, #seLía, huelin
Murcia	01h, Presion, Sangonera, Fica, condomina, CS3, 02h, atalayas, #Murcia, Condomina, Región
Palma	Illes, #Palma, Balears, 1FM, #palmademallorca, 94euros, PMI, #IllesBalears, #elsheroisdemestalla, #VMB, #xbox360
Palmas de Gran Canaria	#Adria, #CanaryIslandsNeedsA1DConcert, Canteras, maanso, #Dime, #grancanaria, Vegueta, #yosoydeiker, Vira, #laspalmas, ULPGC
Bilbao	Lleo, #Bizkaia, Euskalduna, #sherpasummit, Guggenheim, #a8, Barria, Ibilaldi, BilbaoBasket, ariane, Pilepic

Figura 6.25: Tuit con el *hashtag* '*#agapitofueraya*'que identifica claramente a la ciudad de Zaragora por estar vinculado a su club de fútbol.

que el resto de la información que acompaña a los topónimos, puede ser de gran ayuda en la tarea de clasificar geográficamente los textos.

También se pueden encontrar otros términos relacionados con la lengua cooficial que se emplea en las ciudades que la tienen, la cual, es esta ocasión está mucho más presente de lo que lo estaba en los textos del diario *20Minutos* al haber un gran número de usuarios de la red social que utilizan estas otras lenguas.

Por otro lado, cabe destacar el uso de los *hashtags* empleados en cada ciudad, ya que en numerosas ocasiones pueden identificar claramente a la localidad en cuestión. Por ejemplo, en el tuit mostrado en la figura 6.25 con el *hashtag* '*#agapitofueraya*'se identifica claramente a la ciudad de Zaragora por estar vinculado a su principal club de fútbol a través de su expresidente Agapito Iglesias.

6.6. Conclusiones

En este capítulo se han realizado cinco estudios diferentes que pretendían aportar información sobre los corpus de *20Minutos* y *Twitter* utilizados a lo largo de esta tesis.

En el experimento llevado a cabo para mostrar la correlación entre estos corpus, los resultados logrados muestran que hay una correlación temporal en el 50 % de las ciudades, llegando al 90 % si se consideran únicamente las ciudades más pobladas del país, las cuales son a su vez de las que tuits se ha recogido en el corpus. Es decir, dado un número suficientemente alto de mensajes procedentes de *Twitter*, se puede observar una correlación temporal entre ambos corpus del 90 %.

Este es un gran resultado teniendo en cuenta las dificultades de la tarea, tales como la naturaleza distinta de los corpus, que los mensajes vertidos en *Twitter* suelen hacer referencia a otros lugares, la escasa información geográfica encontrada en los tuits o que las fechas del corpus de noticias eran contiguas.

En cuanto a los experimentos que mostraban la evolución temporal del lenguaje, según se ha podido observar en los resultados obtenidos, parece que lo que publica la prensa, transcurrido un tiempo, se refleja en las encuestas

del *CIS*, por lo que esto puede servir para poder predecir tendencias en las preocupaciones que sufrirán los ciudadanos en un futuro, bien sea porque la prensa se adelanta a dichas preocupaciones, o bien por ésta influye en los ciudadanos.

Más allá de que los resultados mostrados por los medios de comunicación puedan condicionar la opinión de los ciudadanos, según el estudio realizado, estos medios de comunicación parecen ser un termómetro de lo que opinarán los ciudadanos en un futuro.

Con respecto a *Twitter* y la evolución temporal de los términos habría que realizar un seguimiento en un periodo más amplio de tiempo, ya que si se dividen dichos periodos temporales en 10 días puede ocurrir grandes oscilaciones en la evolución de los términos, especialmente cuando se habla a nivel de ciudad.

Si se centran la atención en los experimentos llevados a cabo para comprobar la distribución geográfica de los términos, en los resultados obtenidos con el corpus del diario *20Minutos*, así como con el corpus de tuits, se aprecia una gran correlación con lo que ocurrió en cada área geográfica con respecto a los términos analizados.

Respecto al agrupamiento de áreas geográficas según la terminología empleada, como ya se comentó previamente, los algoritmos de *clustering* no suelen tener unos grandes resultados cuando intentan clasificar geográficamente textos basándose meramente en los términos de los propios textos, pero, en los experimentos llevados a cabo en esta sección, se ha logrado una gran coherencia en sus resultados, especialmente cuando se reduce el número de *clusters*.

Analizando el lenguaje empleado en cada una de las distintas localidades del corpus del diario *20Minutos*, se puede apreciar cómo se han logrado *clusters* compactos que engloba todas las provincias que conforman comunidades autónomas con grandes similitudes lingüísticas y culturales. Los resultados muestran una gran coherencia con lo hipótesis inicial planteada para este estudio.

El último de los experimentos que mostraba los términos más representativos de las 10 ciudades más pobladas de España, por un lado, en los experimentos llevados a cabo con el corpus del diario *20Minutos*, se ha puesto de manifiesto que la mayoría de estos términos pertenecían a la categoría gramatical de sustantivos (topónimos, nombres o siglas de empresas o entidades, etc.).

En las comunidades en las que existe una lengua cooficial se pueden encontrar una gran cantidad de términos que ayudan a ubicar, o al menos a acotar, la procedencia geográfica de los textos que provienen tanto de *20Minutos* como de *Twitter*. Estos términos siempre se encuentran entre los que mayores valores obtienen de la unidad tipificada.

En cuanto a *Twitter*, se ha observado cómo en esta ocasión el número de topónimos no es muy significativo entre los términos más representativos

de cada localidad, justo al contrario de lo que sucede con los '*hashtags*', los cuales ayudan a identificar inequívocamente las localidades del corpus de esta red social.

7

Conclusiones y trabajo futuro

Para finalizar, en este capítulo se mostrarán las conclusiones de las propuestas y experimentos realizados.

Con el objetivo de poder realizar experimentos que comprueben el principal propósito de esta tesis, el conocer lo que pueden aportar recursos externos textuales no estructurados en la obtención del foco geográfico de otros recursos de la misma o distinta naturaleza, se han adquirido cuatro corpus de distinta naturaleza, dos de ellos se consideran textos formales, *20Minutos* y *Wikipedia*, y los otros dos informales, *Twitter* y *Flickr*.

Por el lado de los corpus de textos formales también se han filtrado dichos textos para obtener únicamente los términos que correspondían a las siguientes categorías gramaticales: topónimos, sustantivos (los cuales engloban a los anteriores), sustantivos más adjetivos, así como un corpus en el que fueron eliminados todos los topónimos. El objetivo de este filtrado no es otro que el comprobar la aportación que realizan cada una de las categorías mencionadas cuando se pretende determinar el foco geográfico de un texto.

7.1. Identificación del foco geográfico en textos formales

Para la clasificación geográfica de los textos formales se ha dividido el corpus del diario *20Minutos* de forma jerárquica según su categoría gramatical, periodo temporal, partición para la posterior validación cruzada, localidad y artículo.

Con el propio corpus del diario *20Minutos* se hizo una calibración previa para ver qué algoritmo era el que mejores resultados daba a la hora de determinar el foco geográficos de las noticias del propio diario.

El sistema *SVM* demostró obtener una mayor precisión, entre 5 y 9 puntos porcentuales absolutos por encima del de modelos de lenguaje.

7.1.1. Clasificación geográfica de textos formales mediante el propio corpus de textos formales

La mejor aproximación ha sido la que ha hecho uso de todos los términos existentes en el corpus. Esta mejora se acentuó aún más cuanto más grande era el corpus de entrenamiento utilizado, al ser de este corpus de donde se podían extraer un mayor número de sutilezas lingüísticas que ayudaran a esta tarea.

Pese a que los topónimos son una parte fundamental a la hora de llevar a cabo un sistema que identifique geográficamente los textos, la precisión obtenida por el corpus formado únicamente por éstos resulta ser la más baja de todas. Esto es debido a la información que se pierde al desechar el resto de términos donde residen matices lingüísticos importantes a la hora de hacer una clasificación geográfica. Además, hay que tener en cuenta que alrededor de un 11 % de las noticias del corpus trabajado no contienen topónimos, lo que imposibilita su geolocalización utilizando únicamente esta categoría.

Los resultados obtenidos mediante la reducción de características mejoraron entre 3 y 9 puntos porcentuales absolutos a los logrados en la mejor implementación realizada sin esta reducción. Esta mejora fue más significativa tanto en cuanto el corpus era más pequeño (años 2008 y 2009).

La utilización de χ^2, al utilizar el propio corpus de noticias para obtener el vocabulario y entrenar al sistema, es la que mejores resultados ha dado.

El problema que tiene dicha aproximación es que se ha de disponer previamente de un corpus lo suficientemente extenso ya etiquetado de las mismas características al que se pretende clasificar geográficamente, con la dificultad que ello atañe por la escasez de los mismos. Además, dicha aproximación requiere de un proceso de selección de característica temporalmente costoso.

7.1.2. Clasificación geográfica de textos formales mediante otro corpus formal distinto al que se pretende clasificar

En los experimentos que utilizaron el corpus de *Wikipedia*, con o sin los artículos referenciados en los 52 artículos iniciales, para entrenar el sistema, los mejores resultados coincidieron con las aproximaciones que utilizaban solamente los topónimos. Esto es debido a la disparidad existente entre ambos corpus, lo que hace que categorías como la de topónimos sean las que mejor discriminen geográficamente los textos.

Cabe destacar que dentro de estos experimentos que utilizaron el corpus de *Wikipedia* para entrenar el sistema, la aproximación que hacía uso de todos los sustantivos y adjetivos, funcionó mejor que la que hizo uso solamente de los sustantivos. De ahí se puede deducir que el uso de otras categorías gramaticales tales como los adjetivos, pueden ser de gran ayuda a la hora de determinar el foco geográfico de este tipo de texto.

En cuanto a los experimentos que utilizaron *Wikipedia* como mero selector de características, al utilizar el propio corpus del diario *20Minutos* como conjunto de entrenamiento, los resultados mejoraron claramente a los logrados con cualquier aproximación donde se entrenaba únicamente con el corpus de *Wikipedia*, obteniéndose los mejores resultados cuando se hacía uso de todos los términos existentes en ambos corpus, mostrando una vez más la importancia de todos los términos a la hora de hacer este tipo de clasificaciones.

La aproximación que hizo uso de los artículos referenciados obtuvo mejores resultados que la que hizo uso únicamente de los 52 artículos de las localidades existentes. Esto fue debido a que esta primera aproximación tenía un mayor número de características que eran a su vez relevantes para esta tarea.

Así pues, dada la escasez de recursos textuales formales desestructurados y georeferenciados, se demuestra que fuentes como *Wikipedia* pueden ser de gran ayuda cuando el número de textos georeferenciados procedentes de la propia fuente a clasificar sea escaso o nulo.

7.1.3. Clasificación geográfica de textos formales mediante un corpus de textos informales

Los resultados obtenidos con la aproximación que utilizaba los tuits agrupados por usuario y ciudad no fueron muy buenos debido a lo poco representativas que eran unas muestras con tan poco texto.

Al igual que sucediera con la aproximación que entrenaba únicamente con muestras de *Wikipedia* (textos formales), al entrenar con muestras más extensas, la que agrupa los tuits por ciudad de procedencia, los mejores resultados se obtienen cuando se clasifican textos que contienen únicamente topónimos. Esto es debido a que son prácticamente los únicos términos que comparten estos corpus tan dispares.

Por otro lado, si se utilizan los términos de los textos de *Twitter* únicamente como selector de características, y se entrena con los propios textos del diario *20Minutos*, la precisión lograda supera en varias veces a las que solamente se trabajaba con el corpus de *Twitter* para el entrenamiento, siendo la mejor aproximación la que entrena y clasifica textos que contienen todos sus términos.

7.1.4. Clasificación geográfica de textos formales mediante la combinación de diversos corpus con distinta formalidad

Los experimentos realizados con la combinación de *20Minutos* y *Wikipedia* dan unos resultados prácticamente idénticos a los realizados únicamente con *20Minutos*. Esto es debido a que el volumen de los textos del diario es

varias veces superior a los de los 52 artículos de *Wikipedia*. Se realizó una ponderación de los corpus, la cual no consiguió mejorar los resultados. Por ello se decidió añadir más textos a *Wikipedia* mediante la incursión de los textos de los artículos referenciados. Esta extensión de los textos de *Wikipedia* acabó introduciendo una gran cantidad de ruido.

Los experimentos realizados con la combinación de *20Minutos* y *Twitter* cuando se agrupaba el conjuntos de tuits por ciudad de procedencia muestra que tanto en cuanto mayor es el conjunto de entrenamiento del corpus del diario *20Minutos*, mejores son los resultados, siendo siempre la mejor aproximación la que hace uso de todas las características del lenguaje. Así pues, esta aproximación puede ser válida para cuando se quiere apoyar a un conjunto escaso de textos etiquetados de la fuente a clasificar.

Los experimentos realizados con la combinación de los corpus *20Minutos*, *Wikipedia* y *Twitter* vuelven a mostrar unos resultados similares a los anteriores, donde los mejores resultados se lograban cuando el corpus del diario *20Minutos* era más representativo (años 2010, 2011 y 2008-2011).

De estas combinaciones se pueden extraer las siguientes conclusiones generales:

- La combinación de corpus puede ser interesante cuando no se disponga de una gran cantidad de textos etiquetados del mismo corpus que se quiere clasificar geográficamente, ya que puede obtener unos buenos resultados aunque siempre lejos de los logrados por aproximaciones que sí que dispongan de un número extenso de textos etiquetados de la misma fuente a clasificar. Del mismo modo, bajo estas mismas circunstancias de escasez de textos etiquetados geográficamente, se puede afirmar que la mejor aproximación es la que hace uso de otro medio de la misma formalidad que hace uso de un recurso que ya está geográficamente etiquetado, es extenso y está libremente disponible: *Wikipedia* con los enlaces que apuntan a otros artículos.

- No se justifica la introducción de los artículos referenciados en los artículos de *Wikipedia* de las ciudades del corpus en combinación con el corpus del diario *20Minutos*, ya que estos hacen que el espacio y el tiempo requerido para procesarlos, así como los resultados obtenidos, empeore drásticamente, entre un 34 % y un 13 % peor.

- Si se analizan los resultados obtenidos, se puede comprobar cómo la precisión del sistema mejora cuando el corpus del diario *20Minutos* es mayor, dado que los textos del propio diario tienen una mayor presencia con respecto al de las otras fuentes y dicha fuente es la que mejor puede clasificar sus propios textos.

- *Wikipedia* aporta una ligera mejora cuando el corpus del diario es menos extenso (año 2009), aunque ésta no es estadísticamente

significativa (74,82 % frente a 74,97 %). Por el contrario, la fuente de textos informales, *Twitter*, introduce más ruido (alrededor de un 4 % de perdida de precisión) cuando el corpus de *20Minutos* tiene menos peso debido a que los textos de *Twitter* adquieren una mayor relevancia.

7.2. Identificación del foco geográfico en textos informales

Para la clasificación geográfica de textos informales se intentó obtener el foco geográfico de mensajes de *Twitter* emitidos por usuarios dentro de alguna de las 52 ciudades expuestas en los experimentos con textos formales, es decir, las 50 capitales de provincia españolas más Ceuta y Melilla.

En esta ocasión, puesto que los textos de *Twitter* son altamente informales, y las herramientas disponibles para obtener las categorías gramaticales de cada término no obtienen unos resultados buenos, no se realizó una división del corpus por dichas categorías. En su lugar, se realizó una división obedeciendo al nivel de actividad de los usuarios en las distintas ubicaciones.

De este modo, se procedió a realizar los experimentos análogos a los realizados con los textos formales.

7.2.1. Clasificación geográfica de textos informales mediante el propio corpus de textos informales

Puesto que la naturaleza de los textos de *Twitter* difería mucho con la de las noticias del diario *20Minutos*, se volvió a comprobar qué aproximación, entre *SVM* y los modelos de lenguaje, era la más adecuada para clasificar geográficamente los textos de *Twitter*.

También se comprobó lo que era más adecuado, si una aproximación en la que se agruparan todos los tuits por ciudad a la hora de realizar el entrenamiento, u otra en la que se separada por usuario y ciudad, obteniendo una precisión que oscila entre un 40 % y un 60 % dependiendo del nivel de actividad de los usuarios.

Así pues, los resultados obtenidos en la aproximación realizada mediante *SVM* mostraron un funcionamiento del sistema mucho mejor cuando se separaron los tuits por usuario y ciudad, logrando una precisión entre un 40 % y un 73 % según el nivel de actividad de los usuarios.

Por otro lado, los experimentos llevados a cabo con modelos de lenguaje, a diferencia de lo ocurrido con *SVM*, mostraron un mejor funcionamiento cuando se agruparon los conjuntos de tuits por ciudad.

Únicamente, *SVM* supera a los modelos de lenguaje cuando el conjunto de tuits evaluado pertenece a un usuario con poca actividad para una ciudad dada, y esta mejora no es significativa. También se produjo una ligera mejora

de *SVM* respecto a los modelos de lenguaje cuando no se diferenció entre los conjuntos de tuits por su nivel de actividad, pero, una vez más, esta mejora no es significativa y fue debida a que hay un número mucho mayor de usuarios con un nivel de actividad bajo que del resto (más de 145.000 usuarios de los poco más de 200.000 totales tienen un nivel de actividad baja). Tanto en cuanto el nivel de actividad de los usuarios era mayor, mayor era también la diferencia entre los modelos de lenguaje y *SVM* a favor de la primera aproximación. Además, los modelos de lenguaje tenían una ejecución mucho más rápida que *SVM*, por lo que claramente se puede afirmar que los modelos de lenguaje superaron a *SVM* en estos experimentos, de ahí que hayan sido la aproximación utilizada en el resto de experimentos con textos informales.

7.2.2. Clasificación geográfica de textos informales mediante un corpus de textos formal

Se ha realizado una aproximación para detectar el foco geográfico en los conjuntos de tuits expuestos en el experimento anterior utilizando un corpus de textos formales, *Wikipedia*. Con los textos de los artículos de *Wikipedia* utilizados previamente se crearon los oportunos modelos de lenguaje utilizados en esta aproximación.

En un primer momento, se realizó un entrenamiento procedente de artículos de localidades de *Wikipedia*. En esta aproximación se utilizaron únicamente los textos procedentes de los artículos de *Wikipedia* que representaban a las ciudades del corpus. Normalmente, el mejor rendimiento es obtenido cuando únicamente se utilizan los topónimos, alrededor de un 13 % de precisión. Esto es debido a que estos términos son los que añaden un mayor valor semántico desde el punto de vista geográfico cuando se intentan clasificar textos informales entrenando con una fuente de texto formal. Pero hay que resaltar como el mejor resultado es obtenido precisamente cuando se omiten estos términos en el corpus de usuarios de baja actividad, obteniendo un 15,61 % de precisión sin dicha categoría gramatical frente a un 15,06 %. Los matices lingüísticos vuelven a ser importantes cuando se trata de geolocalizar textos de una extensión tan limitada, pese a que los conjuntos de entrenamiento y de evaluación sean tan dispares como *Wikipedia* y *Twitter*. También se ha podido apreciar como cuanto más pequeño eran los conjuntos de tuits a evaluar, mejores eran los resultados, 15 % para los usuarios de baja intensidad por 4 % de los de alta. Esto es debido a que los términos que aparecen en fuentes formales como *Wikipedia* que son capaces de ubicar geográficamente un texto, se encuentran más dispersos dentro de los grandes conjuntos de tuits.

Con el fin de poder ampliar el corpus de entrenamiento que crea los modelos de lenguaje se añadieron también los textos de los artículos referenciados en los artículos de *Wikipedia* previos. Se probó agrupando y

separando los textos por ciudad, dando claramente unos mejores resultados al agruparlos (8 % de precisión por separado por un 16 % al agruparlos). Cuando se utilizaron únicamente los topónimos para crear los modelos de lenguaje, la precisión del sistema, en esta ocasión, cayó dos puntos debido a que los artículos referenciados, en muchas ocasiones no tenían topónimos o tenían un número muy reducido de éstos que a su vez estuvieran presentes en los tuits escritos de igual forma, es decir, sin abreviaciones, con tildes, etc.

La precisión dista mucho de la obtenida utilizando únicamente *Twitter* para crear los modelos de lenguaje (casi 20 veces peor para los usuarios activos y tres veces peor de media), lo que denota la gran diferencia entre el lenguaje empleado en un medio y en el otro.

La aproximación llevada a cabo con los textos de los artículos referenciados mejora a la realizada únicamente con el texto de los artículos de *Wikipedia* de las ciudades del corpus, especialmente con los usuarios más activos. Con la utilización de los artículos referenciados la precisión se mantiene estable independientemente del nivel de actividad de los usuarios, 16 % (actividad baja = entre 1 y 10 tuits), 17 % (actividad media = entre 11 y 99 tuits) y 17 % (actividad alta = más de 100 tuits), mientras que cuando no se utilizan estos artículos referenciados, dicha precisión sufre grandes variaciones, 15 %, 9 % y 4 % respectivamente.

Al utilizar un corpus de entrenamiento más extenso, el que hace uso de los textos de los artículos referenciados, el uso de topónimos pierde relevancia, ya que el resto de categorías gramaticales y matices lingüísticos aportan un gran peso a la hora de determinar el foco geográfico de los usuarios de *Twitter*. Los resultados con la utilización única de los topónimos está en torno a la mitad de la obtenida con la utilización de todos los términos.

La aproximación que no hace uso de topónimos obtiene unos resultados muy similares a los logrados con todos lo términos, lo cual refuerza la hipótesis del peso que tienen las demás categorías gramaticales a la hora de determinar el foco geográfico incluso de textos de diversa formalidad.

7.2.3. Clasificación geográfica de textos informales mediante otro corpus informal distinto al que se pretende clasificar

Para los experimentos llevados a cabo para la detección del foco geográfico de textos informales con el entrenamiento de una fuente distinta de textos también informales, se utilizaron los textos procedentes de *Flickr*.

Los textos de *Flickr* fueron obtenidos de los cinco campos textuales distintos que hay en cada fotografía existente en esta plataforma: comentarios, descripción, notas, etiquetas y título. Además, se creo un conjunto de textos adicionales unificando todos los campos de textos anteriores.

Se continuó utilizando la aproximación realizada con modelos de lenguaje debido a que se había demostrado una mejor precisión en un menor tiempo de ejecución para textos informales.

La aproximación que mejores resultados dio fue la que utilizaba los textos procedentes de la descripción de cada fotografía, con un 17 % de media de precisión. En dicho campo se solía incluir una descripción del lugar donde se tomó. Los resultados obtenidos con la utilización de todos los campos de texto eran muy parecidos a los mejores.

Por otro lado, una vez más se puede observar como los mejores resultados se obtienen cuando se intenta obtener el foco geográfico de los usuarios con un bajo nivel de actividad (entre 1 y 10 tuits), logrando un 20 % de precisión para éstos. Esto es debido a que ambos corpus, pese a ser de carácter informal, son muy dispares y, por ende, cuanto mayor es el conjunto de términos del corpus a evaluar, mayor es la diferencia existente con el modelo de lenguaje de cada ciudad y más difícil de clasificar correctamente.

Si se resumen los resultados logrados con las tres aproximaciones expuestas hasta ahora ubicando textos informales, se puede concluir que la aproximación que utiliza los propios tuits para crear los modelos de lenguaje es la que, como era de esperar, obtiene los mejores resultados.

Al intentar evaluar conjuntos de tuits con más texto, es decir, conjuntos de tuits de usuarios más activos, la precisión aumenta linealmente utilizando *Twitter* para crear los modelos de lenguaje.

Si, por el contrario, se utiliza otra fuente de textos para crear los modelos de lenguaje, el efecto logrado es justo el contrario, es decir, cuanto mayor es el volumen del texto (nivel de actividad) a evaluar, menor es la precisión, pasando de un 16 % de precisión a un 4 % en el caso de *Wikipedia* y de un 20 % a un 6 % en el caso de *Flickr*. Esto es debido a la gran disparidad que existe entre estos corpus y el de *Twitter*, viéndose los términos más relevantes a la hora de clasificar los conjuntos de tuits geográficamente, diluidos entre esta mayor cantidad de texto.

7.2.4. Clasificación geográfica de textos informales mediante la combinación de diversos corpus con distinta formalidad

También se realizaron experimentos combinando las distintas fuentes de texto utilizadas hasta ahora para detectar el foco geográfico de los textos de *Twitter*.

Los mejores resultados se obtienen cuando se utiliza únicamente textos de la propia fuente que se va a evaluar. Esto es así debido a que el lenguaje que se puede encontrar en *Twitter* es muy específico, y está repleto de abreviaturas, contracciones, etc., lo cual hace muy difícil de equiparar a cualquier lenguaje empleado en cualquier otra fuente de textos.

Pese a que *Flickr* contiene un lenguaje más próximo en la informalidad al de *Twitter*, introduce mucho más ruido que *Wikipedia*, la cual es una fuente de texto con un lenguaje formal.

La agregación de recursos textuales supone una acumulación de ruido que hace que el sistema sea menos preciso al detectar el foco geográfico de textos tan singulares como los procedentes de la red social *Twitter*.

Se puede lograr una ligera mejora en conjuntos con un escaso número de tuits cuando se añaden estas fuentes externas al propio corpus de *Twitter* para generar los modelos de lenguaje de las localidades.

7.3. Análisis de los corpus de noticias y de *Twitter*

Por último, en el capítulo 6 se ha mostrado un estudio realizado sobre los corpus con los que se ha trabajado a lo largo de esta tesis en el que se han analizados diversos aspectos.

7.3.1. Correlación entre corpus

Para comprobar la correlación existente entre el corpus de noticias del diario *20Minutos* y el de tuits se ha calculado la distancia *Kullback-Leibler* (*KL*) en el vocabulario de cada ciudad entre el corpus de tuits y tres corpus de *20Minutos*, el que comprendía un periodo temporal inmediatamente anterior al de la emisión de los tuits, el que comprendía el mismo periodo temporal de los tuits y el que comprendía un periodo temporal inmediatamente posterior.

Se ha obtenido una precisión media de un 50 % en la correlación entre los corpus del diario *20Minutos* y *Twitter* cuando ambos corpus coincidían en el tiempo. Esta precisión creció hasta un 90 % para las 10 ciudades más pobladas del país, es decir, las 10 ciudades de las que se tenían unos textos más extensos. Estas cifras indican claramente la correlación entre ambos corpus pese a la dificultad que implica aspectos tales como la disparidad entre las temáticas de ambos corpus, la escasa o nula información geográfica existente en los tuits y que los periodos temporales que se compararon eran contiguos, con el consiguiente solapamiento entre temas tratados en los textos.

7.3.2. Evolución de la relevancia de los términos

Se ha podido constatar cómo la evolución sufrida por los términos analizados en el diario *20Minutos* acaba plasmándose en las encuestas llevadas a cabo por el *CIS*. Más allá de que los resultados mostrados por los medios de comunicación puedan condicionar la opinión de los ciudadanos, según el estudio realizado, estos medios de comunicación parecen ser un termómetro de lo que opinarán los ciudadanos en un futuro.

Por un lado, pese a que se puede apreciar cierta coincidencia entre la evolución de los términos estudiados en *Twitter* y las encuestas del *CIS*, debido al escaso lapso temporal que se cubre con los tuits recogidos (apenas 50 días), es necesario un estudio más extenso para poder sacar unas conclusiones más sólidas.

7.3.3. Relevancia geográfica de los términos

En la distribución geográfica de las preocupaciones de los españoles analizadas en esta tesis, se ha podido observar cómo los términos analizados tanto en el diario como en *Twitter* adquirían una mayor relevancia en las zonas en las que manifiestamente existía una mayor preocupación por dichos temas, tal y como sucede con el paro en el sur del país o la corrupción en el este.

En ocasiones, en temas como la educación, esta mayor notoriedad no solamente tiene que ver con aspectos negativos, sino que también se ha podido comprobar cómo se le daba una mayor importancia por aspectos positivos relacionados con dicho asunto.

7.3.4. Relación de la terminología por área geográfica

Se ha analizado la similitud del lenguaje empleado entre las distintas ciudades del corpus del diario *20Minutos* y de la red social *Twitter* por separado. Para dicho análisis se han creado *clusters* que agrupaban las distintas áreas geográficas según la similitud de los términos empleados en los textos emitidos en dichas áreas.

Para este estudio hay que tener en cuenta que los textos del diario *20Minutos* publicados dentro de las secciones de un ámbito geográfico concreto tratan específicamente sobre temas que afectaban a dicho ámbito, mientras que los tuits emitidos por los usuarios de la red social pueden además tratar temas más genéricos, los cuales pueden afectar a más de una provincia, comunidad o país. De ahí que los resultados obtenidos trabajando con el corpus de tuits están por debajo de los obtenidos con el corpus del diario *20Minutos*

Con un número elevado de *clusters* es complicado obtener unos resultados coherentes en esta tarea, especialmente en textos procedentes de *Twitter*. En cambio, si se reduce dicho número, se puede obtener una mayor congruencia en cuanto a la correlación entre textos de áreas geográficas más próximas, tal y como se mostró en los resultados obtenidos con diez y cinco *clusters* donde se observaron agrupaciones de áreas tales como Extremadura, País Vasco y Navarra, Galicía, o las provincias que comprendían la antigua corona de Aragón. Esto podría ser utilizado para dirigir mensajes adaptados a dichos grupos que, aunque más amplios, son mucho más precisos.

7.3.5. Términos más representativos de las distintas áreas geográficas

Se han mostrado los términos más representativos de las 10 ciudades más pobladas de España con el fin de poder observar qué términos han sido los más relevantes a la hora de identificar cada una de estas ciudades.

La inmensa mayoría de los términos más relevantes que aparecen en ambos corpus son sustantivos, de los cuales, la mayoría son nombres o siglas de empresas o entidades, tales como "*SAMUR*" (Servicio de Asistencia Municial y Urgencia y Rescate de Madrid). También se pueden encontrar adjetivos como "*herculano*" que hace alusión a los hinchas del principal equipo de fútbol de la ciudad de Alicante.

En ciudades en las que existen lenguas cooficiales, se observa un gran número de términos escritos en dicha lengua, ya que son términos que suelen identificar claramente, o al menos acotar, el origen del texto. Estos términos existen en el vocabulario de esta otra lengua y no existentes en castellano, o se trata de prefijos y/o, sobre todo, sufijos más comunes en la lengua cooficial que en la lengua castellana.

Además de estos aspectos, en *Twitter*, entre los términos más relevantes de cada ciudad se aprecian muchos nombres de usuario y *hashtags*. Cabe destacar el uso de estos últimos, ya que en numerosas ocasiones pueden identificar claramente a la localidad en cuestión.

7.4. Trabajo futuro

Los sistemas *GIR* siguen siendo un campo de investigación con un amplio margen de mejora. Tal y como se ha visto en el capítulo 2 de esta tesis, una de las principales claves del éxito de estos sistemas radica en la correcta detección del ámbito geográfico de los textos a recuperar. Por ello se ha realizado el trabajo descrito en esta tesis, donde se pretendía aportar una aproximación diferente para esta problemática.

Como es de esperar, los enfoques mostrados aquí son susceptibles de ser ampliados para poder obtener así unos resultados aún más óptimos. Entre estas ampliaciones caben destacar las siguientes:

- Dado que tanto la simple agregación como la ponderación de corpus de entrenamiento empeoran en la mayoría de los casos los resultados obtenidos al utilizar únicamente un corpus de entrenamiento de la misma fuente que se pretende clasificar geográficamente, resulta crítico el encontrar nuevas técnicas de combinación de corpus de entrenamiento que sean capaces de aprovechar las fortalezas de cada uno de estos corpus de entrenamiento, desechando así el ruido introducido al agregar recursos de origen distinto al del corpus que se pretende clasificar geográficamente.

- Puesto que el número de corpus geográficamente etiquetados es escaso, será interesante abordar la clasificación de textos utilizando estas escasas fuentes de texto existentes que ya están clasificadas geográficamente, tales como *Wikipedia*, conjuntos de tuits, etc., de una manera más eficiente para poder clasificar otras fuentes de las que no se dispongan grandes volúmenes de textos previamente clasificados capaces de entrenar a los sistemas.

- Al hilo de lo expuesto en el punto anterior, habría que comprobar cómo funcionaría el sistema si se entrenase con un número escaso de muestras etiquetadas del mismo corpus a evaluar, tal y cómo sucede cuando se pretende clasificar un corpus del que no se tiene ningún número de documentos geográficamente clasificados y hay que anotar a mano unos cuantos para entrenar al sistema. El objetivo de este experimento sería comprobar si con otros corpus genéricos tales como los de *Wikipedia* o *Twitter*, los cuales se pueden encontrar fácilmente geolocalizados, el sistema funciona mejor que con un conjunto reducido de muestras etiquetadas a mano.

- Otra continuación interesante sería la de realizar experimentos con otra lengua distinta al castellano. Pese a que el sistema descrito en esta tesis no tiene en cuenta la lengua utilizada, habría que comprobar si se obtienen unos resultados similares a los logrados con el castellano.

- Utilizar otras técnicas para la clasificación geográfica de textos. En trabajos como el mostrado en Iyyer et al. (2014) se puede ver cómo un sistema basado en un modelo de redes neuronales recursivas (*RNN: recursive neural network*) en combinación con métodos de *IR*, es capaz de lograr unos grandes resultados en tareas de búsqueda de respuestas (*QA: Question Answering*). Dicho sistema está pensado para situaciones en las que apenas existen entidades nombradas en los textos, por lo que parece que una adaptación a la clasificación geográfica de textos utilizando algoritmos de clasificación como *SVM* puede dar grandes resultados.

- Extender el conjunto de localizaciones posibles en la clasificación geográfica. Tal y como se a visto en esta tesis, se ha trabajado con un número cerrado de localizaciones. Éstas pertenecían todas al mismo país, España, por lo que habría que ampliar dicho ámbito incluyendo otros países y/o continentes.

- Clasificar tuits individualmente. En el capítulo 5 se clasificaron conjuntos de tuits agrupados por usuario y ciudad. En un futuro se pretende clasificar tuits a nivel individual, es decir, sin tener que agrupar éstos con otros tuits emitidos por el mismo usuario. Dado que esta tarea resulta muy compleja ya que, como se indicó en capítulos

anteriores, la inmensa mayoría de tuits no hace ninguna referencia a la ubicación dónde se encuentra el usuario, se podrían agrupar estos conjuntos de tuits sin saber previamente si han sido emitidos en la misma ciudad.

- Trabajar con ámbitos geográficos más reducidos. A lo largo de esta tesis se han utilizado textos pertenecientes a las principales ciudades de cada una de las provincias españolas como representantes de las mismas. En un futuro trabajo se pretende ir más allá, siendo capaces de detectar el foco geográfico incluso a nivel de barrio de cada localidad, reduciendo así la granularidad.

- Otro de los aspectos interesantes revelados en esta tesis ha sido la influencia que tienen los medios de comunicación en la opinión que posteriormente muestra la población, o cómo los medios recogen temas que acaban preocupando a los ciudadanos a posteriori. Un trabajo interesante sería el de comprobar si esto sigue cumpliéndose a lo largo del tiempo y con qué dilación.

- Finalmente, habría que insertar el sistema de detección del foco geográfico aquí descrito dentro de un sistema *GIR* completo para ver su comportamiento en conjunto.

7.5. Principales aportaciones

El trabajo desarrollado en esta tesis ha sido parcialmente publicado en diversos congresos y revistas. A continuación se enumeran las principales contribuciones.

7.5.1. Sistemas de recuperación de información geográfica

- *University of Alicante at NTCIR-9 GeoTime* (Peregrino et al., 2011b). Trabajo desarrollado para la conferencia *NTCIR-9* bajo la tarea de *GeoTime* (geotemporal). Para ello se desarrolló un sistema *GIR* que se basaba en motores de búsqueda y aplicaciones como *Yahoo! PlaceMaker* y un sistema de búsqueda de respuestas.

- *Estudio sobre el impacto de los componentes de un sistema de recuperación de información geográfica y temporal* (Peregrino et al., 2012a). En este trabajo se desarrollo un completo sistema de *IR* geotemporal, donde la parte geográfica se afrontaba utilizando una aproximación basada en múltiples motores de búsqueda de respuesta, mientras que la temporal se basaba en un sistema de búsqueda de respuestas.

- *Question Answering and Multi-search Engines in Geo-Temporal Information Retrieval* (Peregrino et al., 2012d). Trabajo que desarrolló un sistema con una arquitectura modular para poder integrar o eliminar partes capaces de recuperar información tanto geográfica como textual. El sistema estaba compuesto por un conjunto de motores de búsqueda y un subsistema de búsqueda de respuestas.

- *Map-based Filters for Fuzzy Entities in Geographical Information Retrieval* (Peregrino et al., 2011a). Aproximación que abordaba la problemática entidades geográficas difusas en los sistemas *GIR*. Para ello se utilizaban imágenes raster como filtros geográficos para determinar la relevancia de los documentos dependiendo de la localización de los lugares mencionados en dichos documentos.

7.5.2. Detección del foco geográfico en textos

- *Clasificación geográfica de textos informales* (Peregrino Torregrosa et al., 2014). En este artículo se realizó un estudio sobre las técnicas más empleadas en la clasificación de textos informales combinando dichas técnicas con recursos de distinta índole y formalidad.

- *Every Move You Make I'll Be Watching You: Geographical Focus Detection on Twitter* (Peregrino et al., 2013). En este trabajo se detectó el foco geográfico de un conjunto de usuarios de *Twitter* basándose meramente en la información textual de los tuits. Para ello se emplearon modelos de lenguaje y otras fuentes de información textual externas tales como *Wikipedia*.

- *Una aproximación basada en corpus para la detección del foco geográfico en el texto* (Peregrino et al., 2012c). Trabajo que diseñó un sistema que clasificaba geográficamente noticias basándose en aspectos del propio corpus de noticias a clasificar, tales como la aparición de determinados personajes, eventos, fechas e incluso términos comunes.

- *Tratamiento de la dimensión espacial en el texto y su aplicación a la recuperación de información* (Tomás et al., 2012). Trabajo en el que se analizaba la información espacial en el texto para la desambiguación de topónimos y la detección del foco geográfico. Para dicho fin se hizo uso de herramientas de *PLN* y recursos como nomenglátores y ontologías. La desambiguación de topónimos y la detección del foco geográfico se llevó a cabo mediante la incorporación de conocimiento general del mundo (como entidades, roles, fechas y eventos).

- *Tratamiento inteligente de la información para ayuda a la toma de decisiones* (Vázquez et al., 2014). Trabajo centrado en el tratamiento inteligente de información procedente de diversas fuentes tales como

blogs, foros, portales especializados, etc. La finalidad era generar conocimiento a partir de la información semántica recuperada, para lo que fue necesario determinar el área geográfica de procedencia de los textos analizados.

7.5.3. Extracción de información

- *Mapping Routes of Sentiments* (Peregrino et al., 2012b). Propuesta para la creación de un sistema que recogiera la información procedente de distintas redes sociales tales como *Twitter*, *FourSquare* o *Flickr* para poder trazar futuras rutas turísticas adaptadas a los perfiles de los usuarios. La información extraída debe estar asociada a un área geográfica.

- *Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario* (Lloret et al., 2015). En este proyecto se creó un marco tecnológico basado en tecnologías del lenguaje humano para poder procesar y anotar semánticamente la información procedente de Internet.

A

Anexo 1: Documentos analizados por *FreeLing*

Ficheros de noticias del diario *20Minutos* analizados manualmente para determinar el grado de acierto de *FreeLing* clasificando términos en su correspondiente categoría gramatical.

Se han seleccionado una noticia de 10 ciudades del corpus del diario *20Minutos* para realizar la tarea.

Se presenta en primer lugar el texto de la noticia limpio (ver sección 4.1), tal cual se le pasa al analizador léxico *FreeLing*, y a continuación la salida producida por *FreeLing*.

Alicante

Buscan a un hombre que intentó violar a una mujer en Elda y la hirió con una navaja La Policía Nacional busca a un hombre que presuntamente intentó violar a una mujer en Elda Alicante y la hirió con un arma blanca en el forcejeo según confirmó hoy la subdelegada del Gobierno en la provincia de Alicante Encarna Llinares Momentos antes de la reunión del Consejo de Administración de acuaJúcar en la Subdelegación de Gobierno en Alicante Encarna Llinares se refirió así preguntada por los medios de comunicación a la agresión sufrida en Elda por una mujer ocurrida el pasado martes cuando un hombre de mediana edad supuestamente intentó violarla y le clavó una navaja en el abdomen En este sentido Llinares indicó que se están haciendo averiguaciones sobre la identidad del presunto autor de los hechos al tiempo que expresó su deseo de que en el menor tiempo posible se le pueda detener Los hechos según publica hoy el diario Información ocurrieron en la noche del pasado martes 22 de enero en un bar situado en la calle Magallanes de Elda donde la víctima trabaja Al parecer ambos se quedaron solos y la mujer le pidió que se marchara porque era la hora de cierre Al salir para cerrar la puerta el hombre supuestamente se acercó a ella con el fin de agredirla sexualmente Al defenderse se produjo un forcejeo entre el presunto agresor y la víctima quien recibió un navajazo en el abdomen Después el hombre la dejó herida y tirada en la calle hasta que fue hallada e ingresada en el

Hospital de Elda

```
 Buscan buscar VMIP3P0 1
a a SPS00 0.99585
un uno DI0MS0 0.986987
hombre hombre NCMS000 0.95679
que que PR0CN000 0.5625
intentó intentar VMIS3S0 1
violar violar VMN0000 0.801435
a a SPS00 0.99585
una uno DI0FS0 0.951241
mujer mujer NCFS000 0.959459
en en SPS00 1
Elda elda NP00G00 1
y y CC 0.999812
la lo PP3FSA00 0.0277626
hirió herir VMIS3S0 1
con con SPS00 1
una uno DI0FS0 0.951241
navaja navaja NCFS000 1
La_Policía_Nacional la_policía_nacional NP00O00 1
busca buscar VMIP3S0 0.190476
a a SPS00 0.99585
un uno DI0MS0 0.986987
hombre hombre NCMS000 0.95679
que que PR0CN000 0.5625
presuntamente presuntamente RG 1
intentó intentar VMIS3S0 1
violar violar VMN0000 0.801435
a a SPS00 0.99585
una uno DI0FS0 0.951241
mujer mujer NCFS000 0.959459
en en SPS00 1
Elda_Alicante elda_alicante NP00V00 1
y y CC 0.999812
la lo PP3FSA00 0.0277626
hirió herir VMIS3S0 1
con con SPS00 1
un uno DI0MS0 0.986987
arma arma NCFS000 0.944444
blanca blanco AQ0FS0 0.875
en en SPS00 1
el el DA0MS0 1
forcejeo forcejeo NCMS000 0.75
```

según según SPS00 0.986667
confirmó confirmar VMIS3S0 1
hoy hoy RG 0.877358
la el DA0FS0 0.972146
subdelegada subdelegado NCFS000 0.442539
de de SPS00 1
el el DA0MS0 1
Gobierno gobierno NP00O00 1
en en SPS00 1
la el DA0FS0 0.972146
provincia provincia NCFS000 1
de de SPS00 0.999919
Alicante_Encarna_Llinares_Momentos alicante_encarna_llinares_momentos
NP00G00 1
antes_de antes_de SPS00 1
la el DA0FS0 0.972146
reunión reunión NCFS000 1
de de SPS00 1
el el DA0MS0 1
Consejo_de_Administración consejo_de_administración NP00O00 1
de de SPS00 0.999919
acuaJúcar acuajúcar NCCS000 1
en en SPS00 1
la el DA0FS0 0.972146
Subdelegación_de_Gobierno subdelegación_de_gobierno NP00O00 1
en en SPS00 1
Alicante_Encarna_Llinares alicante_encarna_llinares NP00V00 1
se se P0000000 0.465602
refirió referir VMIS3S0 1
así así RG 0.994118
preguntada preguntar VMP00SF 1
por por SPS00 1
los el DA0MP0 0.97623
medios medio NCMP000 0.95
de de SPS00 0.999919
comunicación comunicación NCFS000 1
a a SPS00 0.99585
la el DA0FS0 0.972146
agresión agresión NCFS000 1
sufrida sufrir VMP00SF 0.557461
en en SPS00 1
Elda elda NP00G00 1
por por SPS00 1
una uno DI0FS0 0.951241

```
mujer mujer NCFS000 0.959459
ocurrida ocurrir VMP00SF 1
el el DA0MS0 1
pasado_martes [M:??/??/??:??.??:??] W 1
cuando cuando CS 0.983796
un uno DI0MS0 0.986987
hombre hombre NCMS000 0.95679
de de SPS00 0.999919
mediana mediano AQ0FS0 0.509558
edad edad NCFS000 1
supuestamente supuestamente RG 1
intentó intentar VMIS3S0 1
violar violar VMN0000 1
la lo PP3FSA00 1
y y CC 0.999812
le le PP3CSD00 1
clavó clavar VMIS3S0 1
una uno DI0FS0 0.951241
navaja navaja NCFS000 1
en en SPS00 1
el el DA0MS0 1
abdomen abdomen NCMS000 1
En en SPS00 1
este este DD0MS0 0.956743
sentido sentido NCMS000 0.948276
Llinares llinares NP00SP0 1
indicó indicar VMIS3S0 1
que que CS 0.4375
se se P0000000 0.465602
están estar VAIP3P0 1
haciendo hacer VMG0000 0.958333
averiguaciones averiguación NCFP000 1
sobre sobre SPS00 0.994203
la el DA0FS0 0.972146
identidad identidad NCFS000 1
de de SPS00 1
el el DA0MS0 1
presunto presunto AQ0MS0 1
autor autor NCMS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
hechos hecho NCMP000 0.375
a_el_tiempo_que al_tiempo_que CS 1
expresó expresar VMIS3S0 1
```

```
su su DP3CS0 1
deseo deseo NCMS000 0.5
de de SPS00 0.999919
que que CS 0.4375
en en SPS00 1
el el DA0MS0 1
menor menor AQ0CS0 0.978261
tiempo tiempo NCMS000 1
posible posible AQ0CS0 1
se se P0000000 0.465602
le le PP3CSD00 1
pueda poder VMSP1S0 0.490196
detener detener VMN0000 1
Los el DA0MP0 0.97623
hechos hecho NCMP000 0.375
según según SPS00 0.986667
publica publicar VMIP3S0 0.993015
hoy hoy RG 0.877358
el el DA0MS0 1
diario diario NCMS000 0.5
Información información NP00000 1
ocurrieron ocurrir VMIS3P0 1
en en SPS00 1
la el DA0FS0 0.972146
noche noche NCFS000 1
de de SPS00 1
el el DA0MS0 1
pasado_martes_22_de_enero [M:22/1/??:??.??:??] W 1
en en SPS00 1
un uno DI0MS0 0.986987
bar bar NCMS000 1
situado situar VMP00SM 0.9
en en SPS00 1
la el DA0FS0 0.972146
calle_Magallanes_de_Elda calle_magallanes_de_elda NP00G00 1
donde donde PR000000 0.963542
la el DA0FS0 0.972146
víctima víctima NCFS000 0.961538
trabaja trabajar VMIP3S0 0.875
A a SPS00 1
el el DA0MS0 1
parecer parecer VMN0000 0.875
ambos ambos PI0MP000 0.638889
se se P0000000 0.465602
```

```
quedaron quedar VMIS3P0 1
solos solo AQ0MP0 0.75
y y CC 0.999812
la el DA0FS0 0.972146
mujer mujer NCFS000 0.959459
le le PP3CSD00 1
pidió pedir VMIS3S0 1
que que CS 0.4375
se se P0000000 0.465602
marchara marchar VMSI1S0 0.5
porque porque CS 1
era ser VSII1S0 0.490859
la el DA0FS0 0.972146
hora hora NCFS000 1
de de SPS00 0.999919
cierre cierre NCMS000 0.889286
A a SPS00 1
el el DA0MS0 1
salir salir VMN0000 1
para para SPS00 0.998507
cerrar cerrar VMN0000 1
la el DA0FS0 0.972146
puerta puerta NCFS000 1
el el DA0MS0 1
hombre hombre NCMS000 0.95679
supuestamente supuestamente RG 1
se se P0000000 0.465602
acercó acercar VMIS3S0 1
a a SPS00 0.99585
ella él PP3FS000 1
con con SPS00 1
el el DA0MS0 1
fin fin NCMS000 1
de de SPS00 0.999919
agredir agredir VMN0000 1
la lo PP3FSA00 1
sexualmente sexualmente RG 1
A a SPS00 1
el el DA0MS0 1
defender defender VMN0000 1
se se PP3CN000 1
se se P0000000 0.465602
produjo producir VMIS3S0 1
un uno DI0MS0 0.986987
```

```
forcejeo forcejeo NCMS000 0.75
entre entre SPS00 0.995223
el el DA0MS0 1
presunto presunto AQ0MS0 1
agresor agresor NCMS000 0.833333
y y CC 0.999812
la el DA0FS0 0.972146
víctima víctima NCFS000 0.961538
quien quien PR0CS000 1
recibió recibir VMIS3S0 1
un uno DI0MS0 0.986987
navajazo navajazo NCMS000 1
en en SPS00 1
el el DA0MS0 1
abdomen abdomen NCMS000 1
Después después RG 1
el el DA0MS0 1
hombre hombre NCMS000 0.95679
la lo PP3FSA00 0.0277626
dejó dejar VMIS3S0 1
herida herir VMP00SF 0.111111
y y CC 0.999812
tirada tirar VMP00SF 0.75
en en SPS00 1
la el DA0FS0 0.972146
calle calle NCFS000 0.96875
hasta hasta SPS00 0.955172
que que CS 0.4375
fue ser VSIS3S0 0.932292
hallada hallar VMP00SF 1
e e CC 0.973684
ingresada ingresar VMP00SF 1
en en SPS00 1
el el DA0MS0 1
Hospital hospital NCMS000 1
de de SPS00 0.999919
Elda elda NCFS000 1
```

Barcelona

La Fiscalía archiva las diligencias por la proclama muera el Borbón de Joan Tardà Joan Tardà Joan Tardà en la redacción de 20minutos Miquel Taverna Ampliar La Fiscalía del Tribunal Superior de Cataluña ha archivado las diligencias abiertas al diputado de ERC Joan Tardà por lanzar la proclama muera el Borbón tras un acto de las juventudes de Esquerra al estimar que en el contexto en el que se lanzó no tenía intención de que se perpetrara un magnicidio La Fiscalía entiende que de las explicaciones del propio Tardà sobre dicha proclama y de los actos anteriores y posteriores del diputado se deduce que dicha frase hace referencia a la crítica a la monarquía como institución constituyendo una aclamación a la abolición del régimen de monarquía parlamentaria y advenimiento de la república El escrito de archivo firmado por la Fiscal Superior de Cataluña Teresa Compte añade que las palabras de Tardà se encuadran en un discurso general de autoafirmación a través de la descalificación de varias de las instituciones de este país ciertamente con un grado de visceralidad por fortuna inusual en un diputado pero no por ello criminalizable sin más

```
 La el DA0FS0 0.972146
Fiscalía fiscalía NP00O00 1
archiva archivar VMIP3S0 0.993015
las el DA0FP0 0.97051
diligencias diligencia NCFP000 0.962932
por por SPS00 1
la el DA0FS0 0.972146
proclama proclama NCFS000 0.782094
muera morir VMSP1S0 0.416667
el el DA0MS0 1
Borbón_de_Joan_Tardà_Joan_Tardà_Joan_Tardà
borbón_de_joan_tardà_joan_tardà_joan_tardà NP00O00 1
en en SPS00 1
la el DA0FS0 0.972146
redacción redacción NCFS000 1
de de SPS00 0.999919
20minutos 20minutos Z 1
Miquel_Taverna_Ampliar miquel_taverna_ampliar NP00V00 1
La el DA0FS0 0.972146
Fiscalía_de_el_Tribunal_Superior_de_Cataluña
fiscalía_de_el_tribunal_superior_de_cataluña NP00O00 1
ha haber VAIP3S0 0.998141
archivado archivar VMP00SM 1
las el DA0FP0 0.97051
diligencias diligencia NCFP000 0.962932
```

```
abiertas abrir VMP00PF 1
a a SPS00 1
el el DA0MS0 1
diputado diputado NCMS000 0.833333
de de SPS00 0.999919
ERC_Joan_Tardà erc_joan_tardà NP00O00 1
por por SPS00 1
lanzar lanzar VMN0000 1
la el DA0FS0 0.972146
proclama proclama NCFS000 0.782094
muera morir VMSP1S0 0.416667
el el DA0MS0 1
Borbón borbón NP00O00 1
tras tras SPS00 1
un uno DI0MS0 0.986987
acto acto NCMS000 1
de de SPS00 0.999919
las el DA0FP0 0.97051
juventudes juventud NCFP000 1
de de SPS00 0.999919
Esquerra esquerra NP00O00 1
a a SPS00 1
el el DA0MS0 1
estimar estimar VMN0000 1
que que CS 0.4375
en en SPS00 1
el el DA0MS0 1
contexto contexto NCMS000 1
en en SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
se se P0000000 0.465602
lanzó lanzar VMIS3S0 1
no no RN 0.99778
tenía tener VMII1S0 0.5
intención intención NCFS000 1
de de SPS00 0.999919
que que CS 0.4375
se se P0000000 0.465602
perpetrara perpetrar VMSI1S0 0.5
un uno DI0MS0 0.986987
magnicidio magnicidio NCMS000 1
La_Fiscalía la_fiscalía NP00O00 1
entiende entender VMIP3S0 0.9375
```

```
que que CS 0.4375
de de SPS00 0.999919
las el DA0FP0 0.97051
explicaciones explicación NCFP000 1
de de SPS00 1
el el DA0MS0 1
propio propio AQ0MS0 0.98
Tardà tardà NP00SP0 1
sobre sobre SPS00 0.994203
dicha decir VMP00SF 0.357143
proclama proclama NCFS000 0.782094
y y CC 0.999812
de de SPS00 0.999919
los el DA0MP0 0.97623
actos acto NCMP000 1
anteriores anterior AQ0CP0 1
y y CC 0.999812
posteriores posterior AQ0CP0 1
de de SPS00 1
el el DA0MS0 1
diputado diputado NCMS000 0.833333
se se P0000000 0.465602
deduce deducir VMIP3S0 0.993015
que que CS 0.4375
dicha decir VMP00SF 0.357143
frase frase NCFS000 1
hace_referencia hacer_referencia VMIP3S0 1
a a SPS00 0.99585
la el DA0FS0 0.972146
crítica crítica NCFS000 0.47619
a a SPS00 0.99585
la el DA0FS0 0.972146
monarquía monarquía NCFS000 1
como como CS 0.998668
institución institución NCFS000 1
constituyendo constituir VMG0000 1
una uno DI0FS0 0.951241
aclamación aclamación NCFS000 1
a a SPS00 0.99585
la el DA0FS0 0.972146
abolición abolición NCFS000 1
de de SPS00 1
el el DA0MS0 1
régimen régimen NCMS000 1
```

```
de de SPS00 0.999919
monarquía monarquía NCFS000 1
parlamentaria parlamentario AQ0FS0 0.75
y y CC 0.999812
advenimiento advenimiento NCMS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
república república NCFS000 1
El el DA0MS0 1
escrito escribir VMP00SM 0.958333
de de SPS00 0.999919
archivo archivo NCMS000 0.25
firmado firmar VMP00SM 1
por por SPS00 1
la el DA0FS0 0.972146
Fiscal_Superior_de_Cataluña_Teresa_Compte
fiscal_superior_de_cataluña_teresa_compte NP00O00 1
añade añadir VMIP3S0 0.833333
que que CS 0.4375
las el DA0FP0 0.97051
palabras palabra NCFP000 1
de de SPS00 0.999919
Tardà tardà NP00SP0 1
se se P0000000 0.465602
encuadran encuadrar VMIP3P0 1
en en SPS00 1
un uno DI0MS0 0.986987
discurso discurso NCMS000 0.75
general general AQ0CS0 0.90625
de de SPS00 0.999919
autoafirmación autoafirmación NCFS000 1
a_través_de a_través_de SPS00 1
la el DA0FS0 0.972146
descalificación descalificación NCFS000 1
de de SPS00 0.999919
varias varios DI0FP0 0.939394
de de SPS00 0.999919
las el DA0FP0 0.97051
instituciones institución NCFP000 1
de de SPS00 0.999919
este este DD0MS0 0.956743
país país NCMS000 1
ciertamente ciertamente RG 1
con con SPS00 1
```

```
un uno DI0MS0 0.986987
grado grado NCMS000 0.964286
de de SPS00 0.999919
visceralidad visceralidad NCFS000 1
por por SPS00 1
fortuna fortuna NCFS000 1
inusual inusual AQ0CS0 1
en en SPS00 1
un uno DI0MS0 0.986987
diputado diputado NCMS000 0.833333
pero pero CC 0.998821
no no RN 0.99778
por por SPS00 1
ello él PP3NS000 1
criminalizable criminalizable AQ0CS0 0.946584
sin sin SPS00 1
más más RG 1
```

Ceuta

El Gobierno de Ceuta coordinará la atención a los pasajeros que puedan quedar bloqueados en el Estrecho El Consejo de Gobierno de Ceuta acordó hoy encomendar a su consejero de Economía Empleo y Turismo Guillermo Martínez coordinar la atención que se prestará a los pasajeros de la línea marítima que une la ciudad autónoma con Algeciras Cádiz durante los próximos días en previsión de que el temporal de viento que se registra en el Estrecho pueda obligar a alguno a quedarse bloqueado en cualquiera de las dos estaciones marítimas o incluso a pernoctar en ellas Según explicó la portavoz del Ejecutivo Yolanda Bel en declaraciones a los medios de comunicación Martínez se puso esta misma mañana en contacto con la Autoridad Portuaria de la Bahía de Algeciras APBA para conocer sus preparativos para hacer frente a tal eventualidad En la ciudad autónoma en coordinación directa con la Autoridad Portuaria será la propia Administración local la que se encargue de auxiliar a los pasajeros que pudieran ver alterados sus planes de viaje suministrándoles alimento o prendas de abrigo en caso de que lo necesiten Bel explicó que las previsiones meteorológicas que maneja el Ejecutivo local prevén que durante la jornada de mañana aumente la fuerza del viento y llueva con intensidad moderada Ceuta permanece en alerta naranja por alerta de lluvia viento y fenómenos costeros desde ayer aunque mañana el nivel se rebaja a amarillo

```
 El el DA0MS0 1
Gobierno_de_Ceuta gobierno_de_ceuta NP00O00 1
coordinará coordinar VMIF3S0 1
la el DA0FS0 0.972146
atención atención NCFS000 0.964286
a a SPS00 0.99585
los el DA0MP0 0.97623
pasajeros pasajero NCMP000 0.490442
que que PR0CN000 0.5625
puedan poder VMSP3P0 0.928571
quedar quedar VMN0000 1
bloqueados bloqueado NCMP000 0.442539
en en SPS00 1
el el DA0MS0 1
Estrecho_El_Consejo_de_Gobierno_de_Ceuta
estrecho_el_consejo_de_gobierno_de_ceuta NP00O00 1
acordó acordar VMIS3S0 1
hoy hoy RG 0.877358
encomendar encomendar VMN0000 1
a a SPS00 0.99585
su su DP3CS0 1
consejero consejero NCMS000 0.490442
de de SPS00 0.999919
Economía_Empleo economía_empleo NP00V00 1
y y CC 0.999812
Turismo_Guillermo_Martínez turismo_guillermo_martínez NP00SP0 1
coordinar coordinar VMN0000 1
la el DA0FS0 0.972146
atención atención NCFS000 0.964286
que que PR0CN000 0.5625
se se P0000000 0.465602
prestará prestar VMIF3S0 1
a a SPS00 0.99585
los el DA0MP0 0.97623
pasajeros pasajero NCMP000 0.490442
de de SPS00 0.999919
la el DA0FS0 0.972146
línea línea NCFS000 0.875
marítima marítimo AQ0FS0 1
que que PR0CN000 0.5625
une unir VMIP3S0 0.875
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
autónoma autónomo AQ0FS0 1
```

```
con con SPS00 1
Algeciras_Cádiz algeciras_cádiz NP00V00 1
durante durante SPS00 1
los el DA0MP0 0.97623
próximos próximo AQ0MP0 1
días día NCMP000 1
en en SPS00 1
previsión previsión NCFS000 1
de de SPS00 0.999919
que que CS 0.4375
el el DA0MS0 1
temporal temporal NCMS000 0.625
de de SPS00 0.999919
viento viento NCMS000 1
que que PR0CN000 0.5625
se se P0000000 0.465602
registra registrar VMIP3S0 0.75
en en SPS00 1
el el DA0MS0 1
Estrecho estrecho NP00G00 1
pueda poder VMSP1S0 0.490196
obligar obligar VMN0000 1
a a SPS00 0.99585
alguno alguno PI0MS000 0.5
a a SPS00 0.99585
quedar quedar VMN0000 1
se se PP3CN000 1
bloqueado bloqueado NCMS000 0.442539
en en SPS00 1
cualquiera cualquiera PI0CS000 0.863636
de de SPS00 0.999919
las el DA0FP0 0.97051
dos 2 Z 0.988722
estaciones estación NCFP000 0.875
marítimas marítimo AQ0FP0 1
o o CC 0.998845
incluso incluso RG 0.988095
a a SPS00 0.99585
pernoctar pernoctar VMN0000 1
en en SPS00 1
ellas ellos PP3FP000 1
Según según SPS00 0.986667
explicó explicar VMIS3S0 1
la el DA0FS0 0.972146
```

```
portavoz portavoz NCCS000 1
de de SPS00 1
el el DA0MS0 1
Ejecutivo_Yolanda_Bel ejecutivo_yolanda_bel NP00O00 1
en en SPS00 1
declaraciones declaración NCFP000 1
a a SPS00 0.99585
los el DA0MP0 0.97623
medios medio NCMP000 0.95
de de SPS00 0.999919
comunicación comunicación NCFS000 1
Martínez martínez NP00SP0 1
se se P0000000 0.465602
puso poner VMIS3S0 1
esta este DD0FS0 0.982143
misma mismo AQ0FS0 0.916667
mañana mañana NCFS000 0.397436
en en SPS00 1
contacto contacto NCMS000 0.916667
con con SPS00 1
la el DA0FS0 0.972146
Autoridad_Portuaria_de_la_Bahía_de_Algeciras_APBA
autoridad_portuaria_de_la_bahía_de_algeciras_apba NP00V00 1
para para SPS00 0.998507
conocer conocer VMN0000 1
sus su DP3CP0 0.998462
preparativos preparativo NCMP000 0.490442
para para SPS00 0.998507
hacer_frente hacer_frente VMN0000 1
a a SPS00 0.99585
tal tal DD0CS0 0.741379
eventualidad eventualidad NCFS000 1
En en SPS00 1
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
autónoma autónomo AQ0FS0 1
en en SPS00 1
coordinación coordinación NCFS000 1
directa directo AQ0FS0 1
con con SPS00 1
la el DA0FS0 0.972146
Autoridad_Portuaria autoridad_portuaria NP00O00 1
será ser VSIF3S0 1
la el DA0FS0 0.972146
```

```
propia propio AQ0FS0 1
Administración administración NP00O00 1
local local AQ0CS0 0.5
la el DA0FS0 0.972146
que que PR0CN000 0.5625
se se P0000000 0.465602
encargue encargar VMSP1S0 0.462493
de de SPS00 0.999919
auxiliar auxiliar VMN0000 0.280702
a a SPS00 0.99585
los el DA0MP0 0.97623
pasajeros pasajero NCMP000 0.490442
que que PR0CN000 0.5625
pudieran poder VMSI3P0 1
ver ver VMN0000 0.987179
alterados alterar VMP00PM 1
sus su DP3CP0 0.998462
planes plan NCMP000 1
de de SPS00 0.999919
viaje viaje NCMS000 0.946429
suministrando suministrar VMG0000 1
les les PP3CPD00 1
alimento alimento NCMS000 0.928571
o o CC 0.998845
prendas prenda NCFP000 0.666667
de de SPS00 0.999919
abrigo abrigo NCMS000 0.833333
en_caso_de en_caso_de SPS00 1
que que PR0CN000 0.5625
lo lo PP3CNA00 0.271163
necesiten necesitar VMSP3P0 0.860063
Bel bel NP00V00 1
explicó explicar VMIS3S0 1
que que CS 0.4375
las el DA0FP0 0.97051
previsiones previsión NCFP000 1
meteorológicas meteorológico AQ0FP0 1
que que PR0CN000 0.5625
maneja manejar VMIP3S0 0.75
el el DA0MS0 1
Ejecutivo ejecutivo NP00O00 1
local local AQ0CS0 0.5
prevén prever VMIP3P0 1
que que CS 0.4375
```

```
durante durante SPS00 1
la el DA0FS0 0.972146
jornada jornada NCFS000 1
de de SPS00 0.999919
mañana mañana NCFS000 0.397436
aumente aumentar VMSP1S0 0.444444
la el DA0FS0 0.972146
fuerza fuerza NCFS000 0.968254
de de SPS00 1
el el DA0MS0 1
viento viento NCMS000 1
y y CC 0.999812
llueva llover VMSP3S0 1
con con SPS00 1
intensidad intensidad NCFS000 1
moderada moderar VMP00SF 0.557461
Ceuta ceuta NP00000 1
permanece permanecer VMIP3S0 0.875
en en SPS00 1
alerta alerta RG 0.925
naranja naranja NCCS000 0.833333
por por SPS00 1
alerta alerta RG 0.925
de de SPS00 0.999919
lluvia lluvia NCFS000 1
viento viento NCMS000 1
y y CC 0.999812
fenómenos fenómeno NCMP000 1
costeros costero AQ0MP0 0.509558
desde desde SPS00 1
ayer ayer NCMS000 0.433333
aunque aunque CC 1
mañana mañana RG 0.205128
el el DA0MS0 1
nivel nivel NCMS000 1
se se P0000000 0.465602
rebaja rebajar VMIP3S0 0.777778
a a SPS00 0.99585
amarillo amarillo NCMS000 0.553571
```

Granada

Expertos y representantes institucionales participarán hoy en el Foro Hispano Marroquí de Juristas Expertos y representantes institucionales como Mohamed Lididi secretario general del Ministerio de Justicia del Reino de Marruecos participarán hoy en el Foro Hispano-Marroquí de Juristas que se inaugurará a las 1700 horas en la sede de la Fundación Euroárabe un día después de que concluya la Cumbre UE-Marruecos que ha tenido lugar en Granada El secretario general del Foro Fernando Olivan manifestó en un comunicado la importancia de este encuentro porque es preciso apostar por una definitiva consolidación de la Unión por el Mediterráneo UPM Nos reconocemos en una convivencia formulada siempre desde el Derecho entendiendo esto no en la necesidad de un derecho homogéneo sino a través de un proceso de acercamiento y armonización de las distintas estructuras jurídicas que reconozca y proteja la autonomía de voluntad de los distintos pueblos dijo La jornada del Foro Hispano-Marroquí de Juristas que se celebra mañana lunes 8 de marzo en Granada cuenta con la colaboración de la Fundación Euroárabe y el Legado Andalusí

```
 Expertos experto NCMP000 0.9
y y CC 0.999812
representantes representante NCCP000 0.875
institucionales institucional AQ0CP0 1
participarán participar VMIF3P0 1
hoy hoy RG 0.877358
en en SPS00 1
el el DA0MS0 1
Foro_Hispano_Marroquí_de_Juristas_Expertos
foro_hispano_marroquí_de_juristas_expertos NP00V00 1
y y CC 0.999812
representantes representante NCCP000 0.875
institucionales institucional AQ0CP0 1
como como CS 0.998668
Mohamed_Lididi mohamed_lididi NP00SP0 1
secretario secretario NCMS000 1
general general AQ0CS0 0.90625
de de SPS00 1
el el DA0MS0 1
Ministerio_de_Justicia_de_el_Reino_de_Marruecos
ministerio_de_justicia_de_el_reino_de_marruecos NP00O00 1
participarán participar VMIF3P0 1
hoy hoy RG 0.877358
en en SPS00 1
el el DA0MS0 1
```

```
Foro_Hispano-Marroquí_de_Juristas
foro_hispano-marroquí_de_juristas NP00V00 1
que que PR0CN000 0.5625
se se P0000000 0.465602
inaugurará inaugurar VMIF3S0 1
a a SPS00 0.99585
las el DA0FP0 0.97051
1700 1700 Z 1
horas hora NCFP000 1
en en SPS00 1
la el DA0FS0 0.972146
sede sede NCFS000 0.8125
de de SPS00 0.999919
la el DA0FS0 0.972146
Fundación_Euroárabe fundación_euroárabe NP00O00 1
un uno DI0MS0 0.986987
día día NCMS000 1
después_de después_de SPS00 1
que que CS 0.4375
concluya concluir VMSP1S0 0.462493
la el DA0FS0 0.972146
Cumbre cumbre NCFS000 1
UE-Marruecos ue-marruecos AQ0MP0 0.599983
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
tenido_lugar tener_lugar VMP00SM 1
en en SPS00 1
Granada granada NP00G00 1
El el DA0MS0 1
secretario secretario NCMS000 1
general general AQ0CS0 0.90625
de de SPS00 1
el el DA0MS0 1
Foro_Fernando_Olivan foro_fernando_olivan NP00O00 1
manifestó manifestar VMIS3S0 1
en en SPS00 1
un uno DI0MS0 0.986987
comunicado comunicado NCMS000 0.442539
la el DA0FS0 0.972146
importancia importancia NCFS000 1
de de SPS00 0.999919
este este DD0MS0 0.956743
encuentro encuentro NCMS000 0.5625
porque porque CS 1
```

```
es ser VSIP3S0 1
preciso preciso AQ0MS0 0.9
apostar apostar VMN0000 1
por por SPS00 1
una uno DI0FS0 0.951241
definitiva definitivo AQ0FS0 1
consolidación consolidación NCFS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
Unión unión NP00O00 1
por por SPS00 1
el el DA0MS0 1
Mediterráneo_UPM_Nos mediterráneo_upm_nos NP00O00 1
reconocemos reconocer VMIP1P0 1
en en SPS00 1
una uno DI0FS0 0.951241
convivencia convivencia NCFS000 1
formulada formular VMP00SF 1
siempre siempre RG 1
desde desde SPS00 1
el el DA0MS0 1
Derecho derecho NP00V00 1
entendiendo entender VMG0000 1
esto este PD0NS000 1
no no RN 0.99778
en_la_necesidad_de en_la_necesidad_de SPS00 1
un uno DI0MS0 0.986987
derecho derecho NCMS000 0.884615
homogéneo homogéneo AQ0MS0 1
sino sino CC 0.981707
a_través_de a_través_de SPS00 1
un uno DI0MS0 0.986987
proceso proceso NCMS000 0.973684
de de SPS00 0.999919
acercamiento acercamiento NCMS000 1
y y CC 0.999812
armonización armonización NCFS000 1
de de SPS00 0.999919
las el DA0FP0 0.97051
distintas distinto AQ0FP0 1
estructuras estructura NCFP000 0.9
jurídicas jurídico AQ0FP0 1
que que PR0CN000 0.5625
reconozca reconocer VMSP1S0 0.462493
```

```
y y CC 0.999812
proteja proteger VMSP1S0 0.462493
la el DA0FS0 0.972146
autonomía autonomía NCFS000 1
de de SPS00 0.999919
voluntad voluntad NCFS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
distintos distinto AQ0MP0 1
pueblos pueblo NCMP000 1
dijo decir VMIS3S0 1
La el DA0FS0 0.972146
jornada jornada NCFS000 1
de de SPS00 1
el el DA0MS0 1
Foro_Hispano-Marroquí_de_Juristas
foro_hispano-marroquí_de_juristas NP00V00 1
que que PR0CN000 0.5625
se se P0000000 0.465602
celebra celebrar VMIP3S0 0.993015
mañana mañana RG 0.205128
lunes_8_de_marzo [L:8/3/??:??.??:??] W 1
en en SPS00 1
Granada granada NP00G00 1
cuenta contar VMIP3S0 0.219697
con con SPS00 1
la el DA0FS0 0.972146
colaboración colaboración NCFS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
Fundación_Euroárabe fundación_euroárabe NP00O00 1
y y CC 0.999812
el el DA0MS0 1
Legado legado NCMS000 0.75
Andalusí andalusí RG 1
```

Jaén

Griñán defiende el cambio de la capital con Peñalver y pide que no se detenga tras este salto de gigante El secretario general del PSOE de Andalucía José Antonio Griñán ha defendido la transformación experimentada por

Jaén capital durante el mandato dirigido por la actual alcaldesa y candidata socialista a revalidar el cargo Carmen Peñalver por lo que ha insistido en que no debe detenerse ahora que ha dado un salto de gigante El secretario general del PSOE de Andalucía José Antonio Griñán ha defendido la transformación experimentada por Jaén capital durante el mandato dirigido por la actual alcaldesa y candidata socialista a revalidar el cargo Carmen Peñalver por lo que ha insistido en que no debe detenerse ahora que ha dado un salto de gigante Griñán defiende el cambio de la capital con Peñalver y pide que no se Así lo asegura en una entrevista concedida a Diario Jaén recogida por Europa Press en la que se ha mostrado optimista ante el próximo 22 de mayo considerando que los ciudadanos tienen como primer aval la gestión No hay más que darse una vuelta por Jaén por el centro o por los barrios para comprobarlo no sólo porque la ciudad está moderna y hermosa sino por la cantidad de servicios sociales que ha impulsado ha destacado Igualmente ha resaltado que Peñalver tienen una gran credibilidad porque es luchadora y trabajadora por lo que ha dicho respaldarla por completo A juicio de Griñán no hay un candidato que la iguale de modo que se ha mostrado convencido por las encuestas y sobre todo por la gente de que el PSOE va a seguir gobernando Jaén Frente a su trabajo asimismo Griñán ha aludido a la labor de oposición del PP que se ha visto desbordada por el tremendo ímpetu de la alcaldesa y de la cantidad de proyectos e inversiones que se han puesto en marcha en la ciudad En su opinión en cuatro años con Peñalver se ha hecho más que en doce años de gobiernos populares algo en lo ha influido la predisposición de los responsables municipales En este sentido ha precisado que con los populares costó mucho trabajo sacar proyectos adelante porque preferían el discurso del agravio al de las actuaciones mientras que la alcaldesa socialista ha sabido dar facilidades poniendo a Jaén por encima de todo Por otro lado el líder de los socialistas andaluces se ha referido al PSOE de Jaén como un partido siempre preocupado por la provincia al tiempo que ha señalado que mantendrán su gran peso político en Andalucía porque es el que le da la ciudadanía Además preguntado por si la incidencia del paro puede influir en los resultados del partido ha dicho que la gente sabe distinguir el sentido de cada elección En cualquier caso ha añadido que la gente sabe que el paro no tiene color político de modo que en todas las provincias regiones y países la crisis ha pasado factura frente a lo cual la Junta se ha volcado para que la recuperación económica se traduzca cuanto antes en creación de empleo

```
 Griñán griñán NP00V00 1
defiende defender VMIP3S0 0.875
el el DA0MS0 1
cambio cambio NCMS000 0.97619
de de SPS00 0.999919
la el DA0FS0 0.972146
```

```
capital capital NCFS000 0.372549
con con SPS00 1
Peñalver peñalver NP00V00 1
y y CC 0.999812
pide pedir VMIP3S0 0.875
que que CS 0.4375
no no RN 0.99778
se se P0000000 0.465602
detenga detener VMSP1S0 0.462493
tras tras SPS00 1
este este DD0MS0 0.956743
salto salto NCMS000 0.833333
de de SPS00 0.999919
gigante gigante NCCS000 0.375
El el DA0MS0 1
secretario secretario NCMS000 1
general general AQ0CS0 0.90625
de de SPS00 1
el el DA0MS0 1
PSOE psoe NP00O00 1
de de SPS00 0.999919
Andalucía_José_Antonio_Griñán
andalucía_josé_antonio_griñán NP00O00 1
ha haber VAIP3S0 0.998141
defendido defender VMP00SM 0.75
la el DA0FS0 0.972146
transformación transformación NCFS000 1
experimentada experimentar VMP00SF 1
por por SPS00 1
Jaén jaén NP00G00 1
capital capital NCFS000 0.372549
durante durante SPS00 1
el el DA0MS0 1
mandato mandato NCMS000 1
dirigido dirigir VMP00SM 1
por por SPS00 1
la el DA0FS0 0.972146
actual actual AQ0CS0 1
alcaldesa alcaldesa NCFS000 1
y y CC 0.999812
candidata candidato NCFS000 1
socialista socialista AQ0CS0 0.875
a a SPS00 0.99585
revalidar revalidar VMN0000 1
```

```
el el DA0MS0 1
cargo cargo NCMS000 0.916667
Carmen_Peñalver carmen_peñalver NP00SP0 1
por por SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
insistido insistir VMP00SM 1
en en SPS00 1
que que CS 0.4375
no no RN 0.99778
debe deber VMIP3S0 0.974359
detener detener VMN0000 1
se se PP3CN000 1
ahora ahora RG 1
que que CS 0.4375
ha haber VAIP3S0 0.998141
dado dar VMP00SM 0.972222
un uno DI0MS0 0.986987
salto salto NCMS000 0.833333
de de SPS00 0.999919
gigante gigante NCCS000 0.375
El el DA0MS0 1
secretario secretario NCMS000 1
general general AQ0CS0 0.90625
de de SPS00 1
el el DA0MS0 1
PSOE psoe NP00000 1
de de SPS00 0.999919
Andalucía_José_Antonio_Griñán
andalucía_josé_antonio_griñán NP00000 1
ha haber VAIP3S0 0.998141
defendido defender VMP00SM 0.75
la el DA0FS0 0.972146
transformación transformación NCFS000 1
experimentada experimentar VMP00SF 1
por por SPS00 1
Jaén jaén NP00G00 1
capital capital NCFS000 0.372549
durante durante SPS00 1
el el DA0MS0 1
mandato mandato NCMS000 1
dirigido dirigir VMP00SM 1
por por SPS00 1
```

```
la el DA0FS0 0.972146
actual actual AQ0CS0 1
alcaldesa alcaldesa NCFS000 1
y y CC 0.999812
candidata candidato NCFS000 1
socialista socialista AQ0CS0 0.875
a a SPS00 0.99585
revalidar revalidar VMN0000 1
el el DA0MS0 1
cargo cargo NCMS000 0.916667
Carmen_Peñalver carmen_peñalver NP00SP0 1
por por SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
insistido insistir VMP00SM 1
en en SPS00 1
que que CS 0.4375
no no RN 0.99778
debe deber VMIP3S0 0.974359
detener detener VMN0000 1
se se PP3CN000 1
ahora ahora RG 1
que que CS 0.4375
ha haber VAIP3S0 0.998141
dado dar VMP00SM 0.972222
un uno DI0MS0 0.986987
salto salto NCMS000 0.833333
de de SPS00 0.999919
gigante gigante NCCS000 0.375
Griñán griñán NP00V00 1
defiende defender VMIP3S0 0.875
el el DA0MS0 1
cambio cambio NCMS000 0.97619
de de SPS00 0.999919
la el DA0FS0 0.972146
capital capital NCFS000 0.372549
con con SPS00 1
Peñalver peñalver NP00V00 1
y y CC 0.999812
pide pedir VMIP3S0 0.875
que que CS 0.4375
no no RN 0.99778
se se P0000000 0.465602
```

```
Así así RG 0.994118
lo lo PP3CNA00 0.271163
asegura asegurar VMIP3S0 0.875
en en SPS00 1
una uno DI0FS0 0.951241
entrevista entrevista NCFS000 0.75
concedida conceder VMP00SF 1
a_Diario a_diario RG 1
Jaén jaén NP00V00 1
recogida recogida NCFS000 0.75
por por SPS00 1
Europa_Press europa_press NP00000 1
en en SPS00 1
la el DA0FS0 0.972146
que que PR0CN000 0.5625
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
mostrado mostrar VMP00SM 1
optimista optimista AQ0CS0 1
ante ante SPS00 0.990566
el el DA0MS0 1
próximo_22_de_mayo [??:22/5/??:??.??:??] W 1
considerando considerar VMG0000 0.375
que que CS 0.4375
los el DA0MP0 0.97623
ciudadanos ciudadano NCMP000 0.875
tienen tener VMIP3P0 1
como como CS 0.998668
primer 1 A00MS0 1
aval aval NCMS000 1
la el DA0FS0 0.972146
gestión gestión NCFS000 1
No no NP00SP0 1
hay haber VMIP3S0 1
más más RG 1
que que CS 0.4375
dar dar VMN0000 1
se se PP3CN000 1
una uno DI0FS0 0.951241
vuelta vuelta NCFS000 0.954545
por por SPS00 1
Jaén jaén NP00G00 1
por por SPS00 1
el el DA0MS0 1
```

```
centro centro NCMS000 0.966667
o o CC 0.998845
por por SPS00 1
los el DA0MP0 0.97623
barrios barrio NCMP000 1
para para SPS00 0.998507
comprobar comprobar VMN0000 1
lo lo PP3CNA00 0.5
no_sólo no_sólo CC 1
porque porque CS 1
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
está estar VAIP3S0 0.996032
moderna moderno AQ0FS0 0.916667
y y CC 0.999812
hermosa hermoso AQ0FS0 1
sino sino CC 0.981707
por por SPS00 1
la el DA0FS0 0.972146
cantidad cantidad NCFS000 1
de de SPS00 0.999919
servicios servicio NCMP000 1
sociales social AQ0CP0 0.9375
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
impulsado impulsar VMP00SM 1
ha haber VAIP3S0 0.998141
destacado destacar VMP00SM 1
Igualmente igualmente NP00SP0 1
ha haber VAIP3S0 0.998141
resaltado resaltar VMP00SM 1
que que CS 0.4375
Peñalver peñalver NP00SP0 1
tienen tener VMIP3P0 1
una uno DI0FS0 0.951241
gran gran AQ0CS0 1
credibilidad credibilidad NCFS000 1
porque porque CS 1
es ser VSIP3S0 1
luchadora luchador AQ0FS0 0.509558
y y CC 0.999812
trabajadora trabajador AQ0FS0 0.509558
por por SPS00 1
lo el DA0NS0 0.457393
```

```
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
dicho decir VMP00SM 0.958333
respaldar respaldar VMN0000 1
la lo PP3FSA00 1
por por SPS00 1
completo completo AQ0MS0 0.75
A a SPS00 0.99585
juicio juicio NCMS000 1
de de SPS00 0.999919
Griñán griñán NP00SP0 1
no no RN 0.99778
hay haber VMIP3S0 1
un uno DI0MS0 0.986987
candidato candidato NCMS000 1
que que PR0CN000 0.5625
la lo PP3FSA00 0.0277626
iguale igualar VMSP1S0 0.462493
de_modo_que de_modo_que CS 1
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
mostrado mostrar VMP00SM 1
convencido convencer VMP00SM 1
por por SPS00 1
las el DA0FP0 0.97051
encuestas encuesta NCFP000 0.962932
y y CC 0.999812
sobre sobre SPS00 0.994203
todo todo PI0MS000 0.430982
por por SPS00 1
la el DA0FS0 0.972146
gente gente NCFS000 1
de de SPS00 0.999919
que que CS 0.4375
el el DA0MS0 1
PSOE psoe NP00O00 1
va ir VMIP3S0 1
a a SPS00 0.99585
seguir seguir VMN0000 1
gobernando gobernar VMG0000 1
Jaén jaén NP00G00 1
Frente_a frente_a SPS00 1
su su DP3CS0 1
trabajo trabajo NCMS000 0.940476
```

```
asimismo asimismo RG 1
Griñán griñán NP00SP0 1
ha haber VAIP3S0 0.998141
aludido aludir VMP00SM 1
a a SPS00 0.99585
la el DA0FS0 0.972146
labor labor NCFS000 1
de de SPS00 0.999919
oposición oposición NCFS000 1
de de SPS00 1
el el DA0MS0 1
PP pp NP00O00 1
que que PR0CN000 0.5625
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
visto ver VMP00SM 0.982759
desbordada desbordar VMP00SF 1
por por SPS00 1
el el DA0MS0 1
tremendo tremendo AQ0MS0 1
ímpetu ímpetu NCMS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
alcaldesa alcaldesa NCFS000 1
y y CC 0.999812
de de SPS00 0.999919
la el DA0FS0 0.972146
cantidad cantidad NCFS000 1
de de SPS00 0.999919
proyectos proyecto NCMP000 1
e e CC 0.973684
inversiones inversión NCFP000 1
que que PR0CN000 0.5625
se se P0000000 0.465602
han haber VAIP3P0 1
puesto poner VMP00SM 0.547619
en en SPS00 1
marcha marcha NCFS000 0.948718
en en SPS00 1
la el DA0FS0 0.972146
ciudad ciudad NCFS000 1
En en SPS00 1
su su DP3CS0 1
opinión opinión NCFS000 1
```

```
en en SPS00 1
cuatro 4 Z 1
años año NCMP000 1
con con SPS00 1
Peñalver peñalver NP00V00 1
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
hecho hacer VMP00SM 0.618421
más más RG 1
que que CS 0.4375
en en SPS00 1
doce 12 Z 1
años año NCMP000 1
de de SPS00 0.999919
gobiernos gobierno NCMP000 1
populares popular AQ0CP0 1
algo algo PI0CS000 0.896341
en en SPS00 1
lo lo PP3CNA00 0.271163
ha haber VAIP3S0 0.998141
influido influir VMP00SM 1
la el DA0FS0 0.972146
predisposición predisposición NCFS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
responsables responsable NCCP000 0.5
municipales municipal AQ0CP0 1
En en SPS00 1
este este DD0MS0 0.956743
sentido sentido NCMS000 0.948276
ha haber VAIP3S0 0.998141
precisado precisar VMP00SM 1
que que CS 0.4375
con con SPS00 1
los el DA0MP0 0.97623
populares popular AQ0CP0 1
costó costar VMIS3S0 1
mucho mucho DI0MS0 0.275362
trabajo trabajo NCMS000 0.940476
sacar sacar VMN0000 1
proyectos proyecto NCMP000 1
adelante adelante RG 0.911111
porque porque CS 1
preferían preferir VMII3P0 1
```

```
el el DA0MS0 1
discurso discurso NCMS000 0.75
de de SPS00 1
el el DA0MS0 1
agravio agravio NCMS000 0.962932
a a SPS00 1
el el DA0MS0 1
de de SPS00 0.999919
las el DA0FP0 0.97051
actuaciones actuación NCFP000 1
mientras_que mientras_que CS 1
la el DA0FS0 0.972146
alcaldesa alcaldesa NCFS000 1
socialista socialista AQ0CS0 0.875
ha haber VAIP3S0 0.998141
sabido saber VMP00SM 1
dar dar VMN0000 1
facilidades facilidad NCFP000 1
poniendo poner VMG0000 1
a a SPS00 0.99585
Jaén jaén NP00G00 1
por por SPS00 1
encima_de encima_de SPS00 1
todo todo PI0MS000 0.430982
Por_otro_lado por_otro_lado RG 1
el el DA0MS0 1
líder líder NCCS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
socialistas socialista NCCP000 0.5
andaluces andaluz AQ0MP0 0.833333
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
referido referir VMP00SM 1
a a SPS00 1
el el DA0MS0 1
PSOE psoe NP00O00 1
de de SPS00 0.999919
Jaén jaén NP00G00 1
como como CS 0.998668
un uno DI0MS0 0.986987
partido partido NCMS000 0.886364
siempre siempre RG 1
preocupado preocupar VMP00SM 1
```

```
por por SPS00 1
la el DA0FS0 0.972146
provincia provincia NCFS000 1
a_el_tiempo_que al_tiempo_que CS 1
ha haber VAIP3S0 0.998141
señalado señalar VMP00SM 1
que que PR0CN000 0.5625
mantendrán mantener VMIF3P0 1
su su DP3CS0 1
gran gran AQ0CS0 1
peso peso NCMS000 0.958333
político político AQ0MS0 0.785714
en en SPS00 1
Andalucía andalucía NP00G00 1
porque porque CS 1
es ser VSIP3S0 1
el el DA0MS0 1
que que PR0CN000 0.5625
le le PP3CSD00 1
da dar VMIP3S0 0.983871
la el DA0FS0 0.972146
ciudadanía ciudadanía NCFS000 1
Además además RG 1
preguntado preguntar VMP00SM 1
por por SPS00 1
si si CS 0.997706
la el DA0FS0 0.972146
incidencia incidencia NCFS000 1
de de SPS00 1
el el DA0MS0 1
paro paro NCMS000 0.904762
puede poder VMIP3S0 0.995614
influir influir VMN0000 1
en en SPS00 1
los el DA0MP0 0.97623
resultados resultado NCMP000 0.958333
de de SPS00 1
el el DA0MS0 1
partido partido NCMS000 0.886364
ha haber VAIP3S0 0.998141
dicho decir VMP00SM 0.958333
que que CS 0.4375
la el DA0FS0 0.972146
gente gente NCFS000 1
```

```
sabe saber VMIP3S0 0.983333
distinguir distinguir VMN0000 1
el el DA0MS0 1
sentido sentido NCMS000 0.948276
de de SPS00 0.999919
cada cada DI0CS0 1
elección elección NCFS000 1
En en SPS00 1
cualquier cualquiera DI0CS0 1
caso caso NCMS000 0.990741
ha haber VAIP3S0 0.998141
añadido añadir VMP00SM 0.9
que que CS 0.4375
la el DA0FS0 0.972146
gente gente NCFS000 1
sabe saber VMIP3S0 0.983333
que que CS 0.4375
el el DA0MS0 1
paro paro NCMS000 0.904762
no no RN 0.99778
tiene tener VMIP3S0 1
color color NCMS000 1
político político AQ0MS0 0.785714
de_modo_que de_modo_que CS 1
en en SPS00 1
todas todo DI0FP0 0.945312
las el DA0FP0 0.97051
provincias provincia NCFP000 1
regiones región NCFP000 1
y y CC 0.999812
países país NCMP000 1
la el DA0FS0 0.972146
crisis crisis NCFN000 1
ha haber VAIP3S0 0.998141
pasado pasar VMP00SM 0.647727
factura factura NCFS000 0.782094
frente_a frente_a SPS00 1
lo el DA0NS0 0.457393
cual cual PR0CS000 0.9
la el DA0FS0 0.972146
Junta junta NP00O000 1
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
volcado volcar VMP00SM 0.557461
```

```
para para SPS00 0.998507
que que CS 0.4375
la el DA0FS0 0.972146
recuperación recuperación NCFS000 1
económica económico AQ0FS0 1
se se P0000000 0.465602
traduzca traducir VMSP1S0 0.462493
cuanto_antes cuanto_antes RG 1
en en SPS00 1
creación creación NCFS000 1
de de SPS00 0.999919
empleo empleo NCMS000 0.785714
```

Madrid

En la región se registra un accidente laboral cada cinco minutos Siniestralidad Laboral El coche fúnebre retira el cadáver del operario que murió ayer en una explosión química en Aranjuez ESPINOSA EFE ESPINOSA EFE Ampliar Los trabajadores madrileños han sufrido 26689 siniestros en lo que va de año 21 de ellos mortales Aun así Madrid es de las comunidades más seguras Cada día tienen algún percance 264 empleados El riesgo de sufrir un accidente en el trabajo es constante De hecho cada cinco minutos un empleado es víctima de un siniestro en la región En lo que va de año 26689 trabajadores han sufrido accidentes laborales 26518 leves 150 graves y 21 mortales según UGT Madrid Cada día tienen algún percance 264 empleados La situación ha mejorado respecto al mismo periodo de 2007 cuando hubo 37337 siniestros 285 menos con 27 mortales 37menos como muestran las estadísticas del Ministerio de Trabajo Sin embargo en los últimos años ha ido en aumento con un crecimiento del 106 de los accidentes desde 2000 Madrid es la tercera región con más muertes en el tajo 98 en 2007 pero es la cuarta mejor situada en términos relativos 51 accidentes por cada 1000 asalariados Los sectores más peligrosos son construcción 117 accidentados por 1000 e industria 80 por 1000

```
 En en SPS00 1
la el DA0FS0 0.972146
región región NCFS000 1
se se P0000000 0.465602
registra registrar VMIP3S0 0.75
un uno DI0MS0 0.986987
accidente accidente NCMS000 0.8125
```

```
laboral laboral AQ0CS0 1
cada cada DI0CS0 1
cinco 5 Z 0.98
minutos minuto NCMP000 0.958333
Siniestralidad_Laboral_El siniestralidad_laboral_el NP00SP0 1
coche coche NCMS000 0.964286
fúnebre fúnebre AQ0CS0 1
retira retirar VMIP3S0 0.75
el el DA0MS0 1
cadáver cadáver NCMS000 1
de de SPS00 1
el el DA0MS0 1
operario operario NCMS000 1
que que PR0CN000 0.5625
murió morir VMIS3S0 1
ayer ayer RG 0.566667
en en SPS00 1
una uno DI0FS0 0.951241
explosión explosión NCFS000 1
química químico AQ0FS0 0.833333
en en SPS00 1
Aranjuez aranjuez NP00G00 1
ESPINOSA espinoso AQ0FS0 1
EFE efe NCFS000 1
ESPINOSA espinoso AQ0FS0 1
EFE efe NCFS000 1
Ampliar ampliar VMN0000 1
Los el DA0MP0 0.97623
trabajadores trabajador NCMP000 0.928571
madrileños madrileño AQ0MP0 0.5
han haber VAIP3P0 1
sufrido sufrir VMP00SM 0.875
26689 26689 Z 1
siniestros siniestro NCMP000 0.25
en en SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
va ir VMIP3S0 1
de de SPS00 0.999919
año_21 [??:??/??/21:??.??:??] W 1
de de SPS00 0.999919
ellos ellos PP3MP000 1
mortales mortal AQ0CP0 0.3
Aun aun CS 0.15625
```

```
así así RG 0.994118
Madrid madrid NP00G00 1
es ser VSIP3S0 1
de de SPS00 0.999919
las el DA0FP0 0.97051
comunidades comunidad NCFP000 1
más más RG 1
seguras seguro AQ0FP0 1
Cada cada DI0CS0 1
día día NCMS000 1
tienen tener VMIP3P0 1
algún alguno DI0MS0 1
percance percance NCMS000 1
264 264 Z 1
empleados empleado NCMP000 0.7
El el DA0MS0 1
riesgo riesgo NCMS000 1
de de SPS00 0.999919
sufrir sufrir VMN0000 1
un uno DI0MS0 0.986987
accidente accidente NCMS000 0.8125
en en SPS00 1
el el DA0MS0 1
trabajo trabajo NCMS000 0.940476
es ser VSIP3S0 1
constante constante AQ0CS0 0.958333
De de SPS00 0.999919
hecho hecho NCMS000 0.381579
cada cada DI0CS0 1
cinco 5 Z 0.98
minutos minuto NCMP000 0.958333
un uno DI0MS0 0.986987
empleado empleado NCMS000 0.25
es ser VSIP3S0 1
víctima víctima NCFS000 0.961538
de de SPS00 0.999919
un uno DI0MS0 0.986987
siniestro siniestro NCMS000 0.166667
en en SPS00 1
la el DA0FS0 0.972146
región región NCFS000 1
En en SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
```

```
va ir VMIP3S0 1
de de SPS00 0.999919
año_26689 [??:??/??/26689:??.??:??] W 1
trabajadores trabajador NCMP000 0.928571
han haber VAIP3P0 1
sufrido sufrir VMP00SM 0.875
accidentes accidente NCMP000 0.875
laborales laboral AQ0CP0 0.833333
26518 26518 Z 1
leves leve AQ0CP0 0.980769
150 150 Z 1
graves grave AQ0CP0 0.833333
y y CC 0.999812
21 21 Z 1
mortales mortal NCCP000 0.7
según según SPS00 0.986667
UGT_Madrid_Cada ugt_madrid_cada NP00V00 1
día día NCMS000 1
tienen tener VMIP3P0 1
algún alguno DI0MS0 1
percance percance NCMS000 1
264 264 Z 1
empleados empleado NCMP000 0.7
La el DA0FS0 0.972146
situación situación NCFS000 1
ha haber VAIP3S0 0.998141
mejorado mejorar VMP00SM 1
respecto_a respecto_a SPS00 1
el el DA0MS0 1
mismo mismo AQ0MS0 0.961905
periodo periodo NCMS000 1
de de SPS00 0.999919
2007 2007 Z 1
cuando cuando CS 0.983796
hubo haber VMIS3S0 0.95
37337 37337 Z 1
siniestros siniestro NCMP000 0.25
285 285 Z 1
menos menos RG 1
con con SPS00 1
27 27 Z 1
mortales mortal NCCP000 0.7
37menos 37menos Z 1
como como CS 0.998668
```

```
muestran mostrar VMIP3P0 1
las el DA0FP0 0.97051
estadísticas estadística NCFP000 0.75
de de SPS00 1
el el DA0MS0 1
Ministerio_de_Trabajo_Sin ministerio_de_trabajo_sin NP00000 1
embargo embargo NCMS000 0.962932
en en SPS00 1
los el DA0MP0 0.97623
últimos último AO0MP0 1
años año NCMP000 1
ha haber VAIP3S0 0.998141
ido ir VMP00SM 1
en en SPS00 1
aumento aumento NCMS000 0.916667
con con SPS00 1
un uno DI0MS0 0.986987
crecimiento crecimiento NCMS000 1
de de SPS00 1
el el DA0MS0 1
106 106 Z 1
de de SPS00 0.999919
los el DA0MP0 0.97623
accidentes accidente NCMP000 0.875
desde desde SPS00 1
2000 2000 Z 1
Madrid madrid NP00G00 1
es ser VSIP3S0 1
la el DA0FS0 0.972146
tercera 3 AO0FS0 0.928571
región región NCFS000 1
con con SPS00 1
más más RG 1
muertes muerte NCFP000 1
en en SPS00 1
el el DA0MS0 1
tajo tajo NCMS000 0.75
98 98 Z 1
en en SPS00 1
2007 2007 Z 1
pero pero CC 0.998821
es ser VSIP3S0 1
la el DA0FS0 0.972146
cuarta cuarta NCFS000 0.141304
```

```
mejor mejor AQ0CS0 0.868421
situada situar VMP00SF 1
en en SPS00 1
términos término NCMP000 1
relativos relativo AQ0MP0 0.509558
51 51 Z 1
accidentes accidente NCMP000 0.875
por por SPS00 1
cada cada DI0CS0 1
1000 1000 Z 1
asalariados asalariado NCMP000 0.442539
Los el DA0MP0 0.97623
sectores sector NCMP000 1
más más RG 1
peligrosos peligroso AQ0MP0 1
son ser VSIP3P0 0.986772
construcción construcción NCFS000 1
117 117 Z 1
accidentados accidentado NCMP000 0.442539
por por SPS00 1
1000 1000 Z 1
e e CC 0.973684
industria industria NCFS000 0.833333
80_por_1000 80/1000 Zp 1
```


Santander

Circo Balagan inicia el 20 de diciembre en Santander su 2º gira por España en la que invita a soñar que se puede volar El show representado por 25 artistas de distintos países combina música coreografía acrobacia y comedia con la commedia dellárte El Circo Balagan iniciará el próximo 20 de diciembre en Santander su segunda gira por España con un espectáculo en el que 25 artistas de diferentes países invitarán al público a soñar que pueden volar Circo Balagan inicia el 20 de diciembre en Santander su 2º gira por España en la que invita a soñar que se puede volar La gira que recalará en el Palacio de Festivales de la capital cántabra hasta el día 23 para ofrecer un total de seis funciones continuará después por otros siete escenarios españoles el Teatro Principal de Vitoria del 27 al 30 de diciembre el Teatro Principal de Alicante del 4 al 8 de enero el Teatro Principal de Mallorca del 11 al 15 de enero el Teatro Cervantes de Málaga del 17 al 19 de enero el Nuevo Teatro Circo de Cartagena del 21 al 22 de enero el Teatro

Romea de Murcia del 25 al 29 de enero y el Teatre Principal de Alcoi del 3 al 5 de febrero Tanto la gira como el espectáculo han sido presentados este miércoles en rueda de prensa por el productor Misha Matorin creador del Circo Balagan y también integrante durante muchos años del Circo del Sol y Juan Calzada director del Palacio de Festivales de Santander que han coincidido en señalar que esta producción aúna la grandeza del Circo del Sol en la intimidad de un teatro Así artistas de Estados Unidos Rusia Ucrania China e Inglaterra volarán y danzarán en el aire para representar en un espectáculo inolvidable el sueño que muchas personas tienen desde pequeños que pueden volar De este modo gimnastas malabaristas acróbatas trapecistas clowns y bailarines tratarán de provocar la imaginación y sentimientos del público para que crea que puede volar El show apto para personas desde cinco a noventa y cinco años combina música coreografía acrobacia y comedia con la clásica commedia dellárte y destaca también por el vestuario la luz o la música que han sido creados específicamente para esta producción Un pequeño mundo que viaja Matorin que también fue director del circo de Moscú ha subrayado que el Circo Balagan que en ruso significa circo de mercado constituye un pequeño mundo compuesto por habitantes de distintos países que viajan por todo el planeta y que tratan de coger las mejores ideas en cada sitio para lograr un espectáculo atractivo para el público Así tras indicar que los personajes son como elementos de otros planetas y que por tanto no hacen falta animales el productor ha señalado que uno de los mayores retos de la compañía es adaptar un espectáculo circense al escenario de un teatro que es bastante duro como ha admitido Es un reto ha apostillado para destacar que el fin último es traer el circo al teatro y que la gente se divierta y salga contenta

```
 Circo_Balagan circo_balagan NP00000 1
inicia iniciar VMIP3S0 0.993015
el el DA0MS0 1
20_de_diciembre [??:20/12/??:??.??:??] W 1
en en SPS00 1
Santander santander NP00G00 1
su su DP3CS0 1
2º 2º Z 1
gira gira NCFS000 0.782094
por por SPS00 1
España españa NP00G00 1
en en SPS00 1
la el DA0FS0 0.972146
que que PR0CN000 0.5625
invita invitar VMIP3S0 0.993015
a a SPS00 0.99585
soñar soñar VMN0000 1
```

```
que que CS 0.4375
se se P0000000 0.465602
puede poder VMIP3S0 0.995614
volar volar VMN0000 1
El el DA0MS0 1
show show NCMS000 1
representado representar VMP00SM 1
por por SPS00 1
25 25 Z 1
artistas artista NCCP000 0.75
de de SPS00 0.999919
distintos distinto AQ0MP0 1
países país NCMP000 1
combina combinar VMIP3S0 0.75
música música NCFS000 0.928571
coreografía coreografía NCFS000 1
acrobacia acrobacia NCFS000 1
y y CC 0.999812
comedia comediar VMM02S0 0.166667
con con SPS00 1
la el DA0FS0 0.972146
commedia commedia AQ0FS0 0.61524
dellárte dellárte NCFS000 0.712647
El_Circo_Balagan el_circo_balagan NP00O00 1
iniciará iniciar VMIF3S0 1
el el DA0MS0 1
próximo_20_de_diciembre [??:20/12/??:??.??:??] W 1
en en SPS00 1
Santander santander NP00G00 1
su su DP3CS0 1
segunda 2 A00FS0 0.958333
gira gira NCFS000 0.782094
por por SPS00 1
España españa NP00O00 1
con con SPS00 1
un uno DI0MS0 0.986987
espectáculo espectáculo NCMS000 1
en en SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
25 25 Z 1
artistas artista NCCP000 0.75
de de SPS00 0.999919
diferentes diferente AQ0CP0 1
```

```
países país NCMP000 1
invitarán invitar VMIF3P0 1
a a SPS00 1
el el DA0MS0 1
público público NCMS000 0.59375
a a SPS00 0.99585
soñar soñar VMN0000 1
que que CS 0.4375
pueden poder VMIP3P0 1
volar volar VMN0000 1
Circo_Balagan circo_balagan NP00000 1
inicia iniciar VMIP3S0 0.993015
el el DA0MS0 1
20_de_diciembre [??:20/12/??:??.??:??] W 1
en en SPS00 1
Santander santander NP00G00 1
su su DP3CS0 1
2º 2º Z 1
gira gira NCFS000 0.782094
por por SPS00 1
España españa NP00G00 1
en en SPS00 1
la el DA0FS0 0.972146
que que PR0CN000 0.5625
invita invitar VMIP3S0 0.993015
a a SPS00 0.99585
soñar soñar VMN0000 1
que que CS 0.4375
se se P0000000 0.465602
puede poder VMIP3S0 0.995614
volar volar VMN0000 1
La el DA0FS0 0.972146
gira gira NCFS000 0.782094
que que PR0CN000 0.5625
recalará recalar VMIF3S0 1
en en SPS00 1
el el DA0MS0 1
Palacio_de_Festivales palacio_de_festivales NP00G00 1
de de SPS00 0.999919
la el DA0FS0 0.972146
capital capital NCFS000 0.372549
cántabra cántabro AQ0FS0 0.509558
hasta hasta SPS00 0.955172
el el DA0MS0 1
```

```
día_23 [??:23/??/??:??.??:??] W 1
para para SPS00 0.998507
ofrecer ofrecer VMN0000 1
un uno DI0MS0 0.986987
total total AQ0CS0 0.944444
de de SPS00 0.999919
seis 6 Z 1
funciones función NCFP000 0.916667
continuará continuar VMIF3S0 1
después después RG 1
por por SPS00 1
otros otro DI0MP0 0.56875
siete 7 Z 1
escenarios escenario NCMP000 1
españoles español AQ0MP0 0.488889
el el DA0MS0 1
Teatro_Principal_de_Vitoria teatro_principal_de_vitoria NP00O00 1
de de SPS00 1
el el DA0MS0 1
27 27 Z 1
a a SPS00 1
el el DA0MS0 1
30_de_diciembre [??:30/12/??:??.??:??] W 1
el el DA0MS0 1
Teatro_Principal_de_Alicante
teatro_principal_de_alicante NP00V00 1
de de SPS00 1
el el DA0MS0 1
4 4 Z 1
a a SPS00 1
el el DA0MS0 1
8_de_enero [??:8/1/??:??.??:??] W 1
el el DA0MS0 1
Teatro_Principal_de_Mallorca
teatro_principal_de_mallorca NP00O00 1
de de SPS00 1
el el DA0MS0 1
11 11 Z 1
a a SPS00 1
el el DA0MS0 1
15_de_enero [??:15/1/??:??.??:??] W 1
el el DA0MS0 1
Teatro_Cervantes_de_Málaga teatro_cervantes_de_málaga NP00V00 1
de de SPS00 1
```

```
el el DA0MS0 1
17 17 Z 1
a a SPS00 1
el el DA0MS0 1
19_de_enero [??:19/1/??:??.??:??] W 1
el el DA0MS0 1
Nuevo_Teatro_Circo_de_Cartagena
nuevo_teatro_circo_de_cartagena NP00O00 1
de de SPS00 1
el el DA0MS0 1
21 21 Z 1
a a SPS00 1
el el DA0MS0 1
22_de_enero [??:22/1/??:??.??:??] W 1
el el DA0MS0 1
Teatro_Romea_de_Murcia teatro_romea_de_murcia NP00O00 1
de de SPS00 1
el el DA0MS0 1
25 25 Z 1
a a SPS00 1
el el DA0MS0 1
29_de_enero [??:29/1/??:??.??:??] W 1
y y CC 0.999812
el el DA0MS0 1
Teatre_Principal_de_Alcoi teatre_principal_de_alcoi NP00V00 1
de de SPS00 1
el el DA0MS0 1
3 3 Z 1
a a SPS00 1
el el DA0MS0 1
5_de_febrero [??:5/2/??:??.??:??] W 1
Tanto tanto RG 0.857895
la el DA0FS0 0.972146
gira gira NCFS000 0.782094
como como CS 0.998668
el el DA0MS0 1
espectáculo espectáculo NCMS000 1
han haber VAIP3P0 1
sido ser VSP00SM 1
presentados presentar VMP00PM 1
este este DD0MS0 0.956743
miércoles [X:??/??/??:??.??:??] W 1
en en SPS00 1
rueda rueda NCFS000 0.666667
```

```
de de SPS00 0.999919
prensa prensa NCFS000 0.925926
por por SPS00 1
el el DA0MS0 1
productor productor NCMS000 0.490442
Misha_Matorin misha_matorin NP00SP0 1
creador creador NCMS000 0.833333
de de SPS00 1
el el DA0MS0 1
Circo_Balagan circo_balagan NP00V00 1
y y CC 0.999812
también también RG 1
integrante integrante AQ0CS0 0.509558
durante durante SPS00 1
muchos mucho DI0MP0 0.541667
años año NCMP000 1
de de SPS00 1
el el DA0MS0 1
Circo_de_el_Sol circo_de_el_sol NP00V00 1
y y CC 0.999812
Juan_Calzada juan_calzada NP00SP0 1
director director NCMS000 0.970588
de de SPS00 1
el el DA0MS0 1
Palacio_de_Festivales_de_Santander
palacio_de_festivales_de_santander NP00V00 1
que que PR0CN000 0.5625
han haber VAIP3P0 1
coincidido coincidir VMP00SM 1
en en SPS00 1
señalar señalar VMN0000 1
que que CS 0.4375
esta este DD0FS0 0.982143
producción producción NCFS000 1
aúna aunar VMIP3S0 0.993015
la el DA0FS0 0.972146
grandeza grandeza NCFS000 1
de de SPS00 1
el el DA0MS0 1
Circo_de_el_Sol circo_de_el_sol NP00O00 1
en en SPS00 1
la el DA0FS0 0.972146
intimidad intimidad NCFS000 0.833333
de de SPS00 0.999919
```

```
un uno DI0MS0 0.986987
teatro teatro NCMS000 1
Así así RG 0.994118
artistas artista AQ0CP0 0.25
de de SPS00 0.999919
Estados_Unidos_Rusia_Ucrania_China
estados_unidos_rusia_ucrania_china NP00000 1
e e CC 0.973684
Inglaterra inglaterra NP00G00 1
volarán volar VMIF3P0 1
y y CC 0.999812
danzarán danzar VMIF3P0 1
en en SPS00 1
el el DA0MS0 1
aire aire NCMS000 0.986111
para para SPS00 0.998507
representar representar VMN0000 1
en en SPS00 1
un uno DI0MS0 0.986987
espectáculo espectáculo NCMS000 1
inolvidable inolvidable AQ0CS0 1
el el DA0MS0 1
sueño sueño NCMS000 0.966667
que que CS 0.4375
muchas mucho DI0FP0 0.819444
personas persona NCFP000 0.983871
tienen tener VMIP3P0 1
desde desde SPS00 1
pequeños pequeño AQ0MP0 0.958333
que que PR0CN000 0.5625
pueden poder VMIP3P0 1
volar volar VMN0000 1
De de SPS00 0.999919
este este DD0MS0 0.956743
modo modo NCMS000 1
gimnastas gimnasta NCCP000 1
malabaristas malabarista NCCP000 1
acróbatas acróbata NCCP000 1
trapecistas trapecista NCCP000 1
clowns clown NCMP000 1
y y CC 0.999812
bailarines bailarín NCMP000 0.490442
tratarán tratar VMIF3P0 1
de de SPS00 0.999919
```

```
provocar provocar VMN0000 1
la el DA0FS0 0.972146
imaginación imaginación NCFS000 1
y y CC 0.999812
sentimientos sentimiento NCMP000 1
de de SPS00 1
el el DA0MS0 1
público público NCMS000 0.59375
para para SPS00 0.998507
que que CS 0.4375
crea creer VMSP1S0 0.366667
que que CS 0.4375
puede poder VMIP3S0 0.995614
volar volar VMN0000 1
El el DA0MS0 1
show show NCMS000 1
apto apto AQ0MS0 0.666667
para para SPS00 0.998507
personas persona NCFP000 0.983871
desde desde SPS00 1
cinco 5 Z 0.98
a a SPS00 0.99585
noventa_y_cinco 95 Z 1
años año NCMP000 1
combina combinar VMIP3S0 0.75
música música NCFS000 0.928571
coreografía coreografía NCFS000 1
acrobacia acrobacia NCFS000 1
y y CC 0.999812
comedia comediar VMM02S0 0.166667
con con SPS00 1
la el DA0FS0 0.972146
clásica clásico AQ0FS0 1
commedia commedia NCFS000 0.307855
dellárte dellárte NCFS000 0.712647
y y CC 0.999812
destaca destacar VMIP3S0 0.993015
también también RG 1
por por SPS00 1
el el DA0MS0 1
vestuario vestuario NCMS000 1
la el DA0FS0 0.972146
luz luz NCFS000 1
o o CC 0.998845
```

```
la el DA0FS0 0.972146
música música NCFS000 0.928571
que que PR0CN000 0.5625
han haber VAIP3P0 1
sido ser VSP00SM 1
creados crear VMP00PM 1
específicamente específicamente RG 1
para para SPS00 0.998507
esta este DD0FS0 0.982143
producción producción NCFS000 1
Un uno DI0MS0 0.986987
pequeño pequeño AQ0MS0 0.868421
mundo mundo NCMS000 1
que que PR0CN000 0.5625
viaja viajar VMIP3S0 0.75
Matorin matorin NP00SP0 1
que que CS 0.4375
también también RG 1
fue ser VSIS3S0 0.932292
director director NCMS000 0.970588
de de SPS00 1
el el DA0MS0 1
circo circo NCMS000 1
de de SPS00 0.999919
Moscú moscú NP00G00 1
ha haber VAIP3S0 0.998141
subrayado subrayar VMP00SM 0.557461
que que CS 0.4375
el el DA0MS0 1
Circo_Balagan circo_balagan NP00O00 1
que que CS 0.4375
en en SPS00 1
ruso ruso NCMS000 0.25
significa significar VMIP3S0 0.958333
circo circo NCMS000 1
de de SPS00 0.999919
mercado mercado NCMS000 0.9
constituye constituir VMIP3S0 0.875
un uno DI0MS0 0.986987
pequeño pequeño AQ0MS0 0.868421
mundo mundo NCMS000 1
compuesto componer VMP00SM 0.5
por por SPS00 1
habitantes habitante NCCP000 1
```

```
de de SPS00 0.999919
distintos distinto AQ0MP0 1
países país NCMP000 1
que que PR0CN000 0.5625
viajan viajar VMIP3P0 1
por por SPS00 1
todo todo DI0MS0 0.559816
el el DA0MS0 1
planeta planeta NCMS000 1
y y CC 0.999812
que que PR0CN000 0.5625
tratan tratar VMIP3P0 1
de de SPS00 0.999919
coger coger VMN0000 1
las el DA0FP0 0.97051
mejores mejor AQ0CP0 0.964286
ideas idea NCFP000 0.939394
en en SPS00 1
cada cada DI0CS0 1
sitio sitio NCMS000 0.916667
para para SPS00 0.998507
lograr lograr VMN0000 1
un uno DI0MS0 0.986987
espectáculo espectáculo NCMS000 1
atractivo atractivo AQ0MS0 0.5
para para SPS00 0.998507
el el DA0MS0 1
público público NCMS000 0.59375
Así así RG 0.994118
tras tras SPS00 1
indicar indicar VMN0000 1
que que CS 0.4375
los el DA0MP0 0.97623
personajes personaje NCMP000 1
son ser VSIP3P0 0.986772
como como CS 0.998668
elementos elemento NCMP000 1
de de SPS00 0.999919
otros otro DI0MP0 0.56875
planetas planeta NCMP000 1
y y CC 0.999812
que que CS 0.4375
por por SPS00 1
tanto tanto RG 0.857895
```

```
no no RN 0.99778
hacen hacer VMIP3P0 1
falta falta NCFS000 0.693182
animales animal NCMP000 0.974359
el el DA0MS0 1
productor productor NCMS000 0.490442
ha haber VAIP3S0 0.998141
señalado señalar VMP00SM 1
que que CS 0.4375
uno uno PI0MS000 0.964744
de de SPS00 0.999919
los el DA0MP0 0.97623
mayores mayor AQ0CP0 0.8125
retos reto NCMP000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
compañía compañía NCFS000 1
es ser VSIP3S0 1
adaptar adaptar VMN0000 1
un uno DI0MS0 0.986987
espectáculo espectáculo NCMS000 1
circense circense AQ0CS0 1
a a SPS00 1
el el DA0MS0 1
escenario escenario NCMS000 1
de de SPS00 0.999919
un uno DI0MS0 0.986987
teatro teatro NCMS000 1
que que PR0CN000 0.5625
es ser VSIP3S0 1
bastante bastante RG 0.763889
duro duro AQ0MS0 0.925926
como como CS 0.998668
ha haber VAIP3S0 0.998141
admitido admitir VMP00SM 1
Es es NP00SP0 1
un uno DI0MS0 0.986987
reto reto NCMS000 0.833333
ha haber VAIP3S0 0.998141
apostillado apostillar VMP00SM 1
para para SPS00 0.998507
destacar destacar VMN0000 1
que que CS 0.4375
el el DA0MS0 1
```

```
fin fin NCMS000 1
último último AO0MS0 1
es ser VSIP3S0 1
traer traer VMN0000 1
el el DA0MS0 1
circo circo NCMS000 1
a a SPS00 1
el el DA0MS0 1
teatro teatro NCMS000 1
y y CC 0.999812
que que CS 0.4375
la el DA0FS0 0.972146
gente gente NCFS000 1
se se P0000000 0.465602
divierta divertir VMSP1S0 0.462493
y y CC 0.999812
salga salir VMSP1S0 0.416667
contenta contento NCFS000 0.125
```

Sevilla

Arcángel presenta Quijote de los sueños con letras de Juan Cobos Wilkins y la colaboración de Antonio Orozco El cuarto trabajo del cantaor onubense es viaje entre la ortodoxia y la heterodoxia e incluye temas para todos los gustos La directora del Instituto Andaluz del Flamenco IAF María de los Ángeles Carrasco ha presentado este martes en Sevilla Quijote de los sueños el último trabajo discográfico del cantaor onubense Arcángel que incluye letras del escritor Juan Cobos Wilkins algunos extraídos de su última obra Biografía impura y otros escritos especialmente para Arcángel Además cuenta con la colaboración de Antonio Orozco Arcángel presenta Quijote de los sueños con letras de Juan Cobos Wilkins y la colaboración de Antonio Orozco En rueda del prensa en la sede del IAF Carrasco ha destacado la calidad del trabajo porque cuando dos genios se unen el resultado es un pequeño tesoro en el que nos encontramos un creador maduro en relación a Arcángel Además la director ha definido Quijote de los sueños como un equilibrio entre la celebración de la vida y el desgarre de la muerte Asimismo Cobos Wilkins ha explicado que fue una sorpresa para él que poemas suyos estuvieran en la voz de Arcángel desde hacía dos años porque siempre pensé que los cantaría alguien como Amancio Prada pero no alguien como Arcángel y me sorprendió más continúa cuando asumió cantar mi poema Un niño se confiesa que en el disco se titula No consigo por lo que me pareció revelador

su capacidad para innovar Además explica el escritor fui asumiendo los riesgos que me iba lanzando y aunque nunca había escrito atendiendo a la métrica y el ritmo del flamenco lo hice y me fui enganchando cada vez más a este proyecto que ha durado dos años y del que Cobos Wilkins destaca el absoluto rigor con el que ha trabajado el cantaor onubense que también lo produce quien no ha cambiado ni una palabra sin mi beneplácito Para mí ha sido un trabajo muy fértil y muy hermoso reconoce Por su parte Arcángel ha dicho que Quijote de los sueños es viaje entre la ortodoxia y la heterodoxia e incluye temas para todos los gustos Desde alegrías y soleá sin olvidar fandangos de Huelva y tres canciones una de ellas cuenta con la colaboración de Antonio Orozco al que me unen un vínculo artístico y personal y al que tenía ganas de acercar al flamenco tras una colaboración anterior continúa el cantaor Al hilo de anterior Arcángel ha señalado que él siempre llama a la concordia entre lo tradicional y la vanguardia y los clásicos tienen que respetar a la gente que tiene ganas de hacer otras cosas y la gente de vanguardia tiene que respetar a los clásicos del flamenco sino es así añade es difícil llegar a un equilibrio que tenemos que conseguir los aficionados a este arte El cantaor de Huelva se siente muy identificado con título de este trabajo el cuarto en su carrera artística me siento Quijote por su afán de luchador pero además es un homenaje a Paco Toronjo maestro del fandango onubense que siempre ha sido un poco quijotesco De igual modo en el libreto de este disco hay un mención especial a Enrique Morente al que Arcángel llama maestro Sobre Cobos Wilkins responsable del 80 por ciento de los textos de este disco Arcángel ha comentado que es un artista con el que me puedo tomar un café por eso le invité a participar en este disco que el propio artista define como familiar Los conceptos de estos textos van muy de acuerdo con lo que soy y lo que intento ser cada día resalta El resultado letras más complicadas que tienen cierto matiz social revuelven conciencias y hacen pensar Un trabajo arriesgado que Arcángel define afirmando que es bonito soñar aunque a veces los sueños te peguen un porrazo Trayectoria de arcángel Desde que la peña La Orden de Huelva le viera actuar por primera vez con apenas diez años la trayectoria artística de Arcángel se ha profesionalizado y consolidado convirtiéndose en uno de los cantaores imprescindibles del flamenco actual Niño de Pura Jesús Cayuela José Roca y Mario Maya fueron algunos de los artistas que reclamaron su colaboración al inicio de su carrera hasta que el ciclo Jueves Flamencos de la fundación Cajasol y la Bienal de Flamenco de Sevilla le cedieron un lugar protagonista y empezó a ser conocido por méritos propios No ha dejado de cantar para el baile a Israel Galván o Eva Yerbabuena y su espíritu buscador le lleva a participar en proyectos novedosos como Cus cus flamenco en 2001 con la Orquesta Chekara de Tetuán Hasta Quijote de los sueños ha editado tres trabajos discográficos Arcángel en 2001 que estuvo producido por Juan Carlos Romero y fue reconocido por los premios Andalucía Joven 2002 y el Nacional Flamenco Activo de Úbeda La calle perdía y Ropavieja son

otros proyectos discográficos del cantaor que recibió en 2002 el Giraldillo de Cante reconocimiento que consagró su carrera que le llevará el próximo 4 de noviembre al Teatro Villamarta de Jerez dentro del ciclo Flamenco Viene del Sur

```
 Arcángel arcángel NCMS000 1
presenta presentar VMIP3S0 0.954545
Quijote quijote NCMS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
con con SPS00 1
letras letra NCFP000 1
de de SPS00 0.999919
Juan_Cobos_Wilkins juan_cobos_wilkins NP00SP0 1
y y CC 0.999812
la el DA0FS0 0.972146
colaboración colaboración NCFS000 1
de de SPS00 0.999919
Antonio_Orozco antonio_orozco NP00SP0 1
El el DA0MS0 1
cuarto 4 A00MS0 0.333333
trabajo trabajo NCMS000 0.940476
de de SPS00 1
el el DA0MS0 1
cantaor cantaor NCMS000 1
onubense onubense AQ0CS0 0.509558
es ser VSIP3S0 1
viaje viaje NCMS000 0.946429
entre entre SPS00 0.995223
la el DA0FS0 0.972146
ortodoxia ortodoxia NCFS000 1
y y CC 0.999812
la el DA0FS0 0.972146
heterodoxia heterodoxia NCFS000 1
e e CC 0.973684
incluye incluir VMIP3S0 0.75
temas tema NCMP000 0.625
para para SPS00 0.998507
todos todo DI0MP0 0.622419
los el DA0MP0 0.97623
gustos gusto NCMP000 1
La el DA0FS0 0.972146
directora director NCFS000 0.490442
```

```
de de SPS00 1
el el DA0MS0 1
Instituto_Andaluz_de_el_Flamenco_IAF_María_de_los_Ángeles_Carrasco
instituto_andaluz_de_el_flamenco_iaf_maría_de_los_ángeles_carrasco
NP00000 1
ha haber VAIP3S0 0.998141
presentado presentar VMP00SM 1
este este DD0MS0 0.956743
martes [M:??/??/??:??.??:??] W 1
en en SPS00 1
Sevilla_Quijote sevilla_quijote NP00V00 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
el el DA0MS0 1
último último AO0MS0 1
trabajo trabajo NCMS000 0.940476
discográfico discográfico AQ0MS0 1
de de SPS00 1
el el DA0MS0 1
cantaor cantaor NCMS000 1
onubense onubense AQ0CS0 0.509558
Arcángel arcángel NCMS000 1
que que PR0CN000 0.5625
incluye incluir VMIP3S0 0.75
letras letra NCFP000 1
de de SPS00 1
el el DA0MS0 1
escritor escritor NCMS000 1
Juan_Cobos_Wilkins juan_cobos_wilkins NP00SP0 1
algunos alguno PI0MP000 0.523438
extraídos extraer VMP00PM 1
de de SPS00 0.999919
su su DP3CS0 1
última último AO0FS0 1
obra obra NCFS000 0.972222
Biografía biografía NCFS000 0.666667
impura impuro AQ0FS0 1
y y CC 0.999812
otros otro PI0MP000 0.43125
escritos escribir VMP00PM 0.75
especialmente especialmente RG 1
para para SPS00 0.998507
Arcángel_Además arcángel_además NP00V00 1
```

```
cuenta contar VMIP3S0 0.219697
con con SPS00 1
la el DA0FS0 0.972146
colaboración colaboración NCFS000 1
de de SPS00 0.999919
Antonio_Orozco_Arcángel antonio_orozco_arcángel NP00SP0 1
presenta presentar VMIP3S0 0.954545
Quijote quijote NCMS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
con con SPS00 1
letras letra NCFP000 1
de de SPS00 0.999919
Juan_Cobos_Wilkins juan_cobos_wilkins NP00SP0 1
y y CC 0.999812
la el DA0FS0 0.972146
colaboración colaboración NCFS000 1
de de SPS00 0.999919
Antonio_Orozco_En antonio_orozco_en NP00SP0 1
rueda rodar VMIP3S0 0.266667
de de SPS00 1
el el DA0MS0 1
prensa prensa NCFS000 0.925926
en en SPS00 1
la el DA0FS0 0.972146
sede sede NCFS000 0.8125
de de SPS00 1
el el DA0MS0 1
IAF_Carrasco iaf_carrasco NP00O00 1
ha haber VAIP3S0 0.998141
destacado destacar VMP00SM 1
la el DA0FS0 0.972146
calidad calidad NCFS000 1
de de SPS00 1
el el DA0MS0 1
trabajo trabajo NCMS000 0.940476
porque porque CS 1
cuando cuando CS 0.983796
dos 2 Z 0.988722
genios genio NCMP000 1
se se P0000000 0.465602
unen unir VMIP3P0 1
el el DA0MS0 1
```

```
resultado resultado NCMS000 0.807692
es ser VSIP3S0 1
un uno DI0MS0 0.986987
pequeño pequeño AQ0MS0 0.868421
tesoro tesoro NCMS000 1
en en SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
nos nos PP1CP000 0.933333
encontramos encontrar VMIP1P0 0.5
un uno DI0MS0 0.986987
creador creador NCMS000 0.833333
maduro maduro AQ0MS0 0.981013
en en SPS00 1
relación relación NCFS000 1
a a SPS00 0.99585
Arcángel_Además arcángel_además NP00SP0 1
la el DA0FS0 0.972146
director director NCMS000 0.970588
ha haber VAIP3S0 0.998141
definido definir VMP00SM 0.75
Quijote quijote NCMS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
como como CS 0.998668
un uno DI0MS0 0.986987
equilibrio equilibrio NCMS000 1
entre entre SPS00 0.995223
la el DA0FS0 0.972146
celebración celebración NCFS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
vida vida NCFS000 1
y y CC 0.999812
el el DA0MS0 1
desgarre desgarrar VMSP1S0 0.462493
de de SPS00 0.999919
la el DA0FS0 0.972146
muerte muerte NCFS000 1
Asimismo_Cobos_Wilkins asimismo_cobos_wilkins NP00V00 1
ha haber VAIP3S0 0.998141
explicado explicar VMP00SM 1
que que CS 0.4375
```

```
fue ser VSIS3S0 0.932292
una uno DI0FS0 0.951241
sorpresa sorpresa NCFS000 1
para para SPS00 0.998507
él él PP3MS000 1
que que CS 0.4375
poemas poema NCMP000 1
suyos suyo PX3MP0C0 1
estuvieran estar VASI3P0 1
en en SPS00 1
la el DA0FS0 0.972146
voz voz NCFS000 1
de de SPS00 0.999919
Arcángel arcángel NP00V00 1
desde desde SPS00 1
hacía hacer VMII1S0 0.5
dos 2 Z 0.988722
años año NCMP000 1
porque porque CS 1
siempre siempre RG 1
pensé pensar VMIS1S0 1
que que CS 0.4375
los los PP3MPA00 0.023584
cantaría cantar VMIC1S0 0.5
alguien alguien PI0CS000 1
como como CS 0.998668
Amancio_Prada amancio_prada NP00SP0 1
pero pero CC 0.998821
no no RN 0.99778
alguien alguien PI0CS000 1
como como CS 0.998668
Arcángel arcángel NP00SP0 1
y y CC 0.999812
me me PP1CS000 0.889706
sorprendió sorprender VMIS3S0 1
más más RG 1
continúa continuar VMIP3S0 0.875
cuando cuando CS 0.983796
asumió asumir VMIS3S0 1
cantar cantar VMN0000 0.875
mi mi DP1CSS 0.995536
poema poema NCMS000 1
Un uno DI0MS0 0.986987
niño niño NCMS000 0.973684
```

```
se se P0000000 0.465602
confiesa confesar VMIP3S0 0.75
que que CS 0.4375
en en SPS00 1
el el DA0MS0 1
disco disco NCMS000 1
se se P0000000 0.465602
titula titular VMIP3S0 0.993015
No no NP00SP0 1
consigo conseguir VMIP1S0 0.776731
por por SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
me me PP1CS000 0.889706
pareció parecer VMIS3S0 1
revelador revelador AQ0MS0 0.509558
su su DP3CS0 1
capacidad capacidad NCFS000 1
para para SPS00 0.998507
innovar innovar VMN0000 1
Además además NP00V00 1
explica explicar VMIP3S0 0.95
el el DA0MS0 1
escritor escritor NCMS000 1
fui ser VSIS1S0 0.7
asumiendo asumir VMG0000 1
los el DA0MP0 0.97623
riesgos riesgo NCMP000 1
que que PR0CN000 0.5625
me me PP1CS000 0.889706
iba ir VMII1S0 0.5
lanzando lanzar VMG0000 1
y y CC 0.999812
aunque aunque CC 1
nunca nunca RG 1
había haber VAII1S0 0.414184
escrito escribir VMP00SM 0.958333
atendiendo atender VMG0000 1
a a SPS00 0.99585
la el DA0FS0 0.972146
métrica métrica NCFS000 0.490442
y y CC 0.999812
el el DA0MS0 1
ritmo ritmo NCMS000 0.9375
```

```
de de SPS00 1
el el DA0MS0 1
flamenco flamenco NCMS000 0.25
lo lo PP3CNA00 0.271163
hice hacer VMIS1S0 1
y y CC 0.999812
me me PP1CS000 0.889706
fui ir VMIS1S0 0.3
enganchando enganchar VMG0000 1
cada cada DI0CS0 1
vez vez NCFS000 1
más más RG 1
a a SPS00 0.99585
este este DD0MS0 0.956743
proyecto proyecto NCMS000 0.9375
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
durado durar VMP00SM 1
dos 2 Z 0.988722
años año NCMP000 1
y y CC 0.999812
de de SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
Cobos_Wilkins cobos_wilkins NP00SP0 1
destaca destacar VMIP3S0 0.993015
el el DA0MS0 1
absoluto absoluto AQ0MS0 0.875
rigor rigor NCMS000 1
con con SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
ha haber VAIP3S0 0.998141
trabajado trabajar VMP00SM 1
el el DA0MS0 1
cantaor cantaor NCMS000 1
onubense onubense AQ0CS0 0.509558
que que PR0CN000 0.5625
también también RG 1
lo lo PP3CNA00 0.271163
produce producir VMIP3S0 0.961538
quien quien PR0CS000 1
no no RN 0.99778
ha haber VAIP3S0 0.998141
```

```
cambiado cambiar VMP00SM 1
ni ni CC 1
una uno DI0FS0 0.951241
palabra palabra NCFS000 1
sin sin SPS00 1
mi mi DP1CSS 0.995536
beneplácito beneplácito NCMS000 1
Para para SPS00 0.998507
mí mí PP1CS000 1
ha haber VAIP3S0 0.998141
sido ser VSP00SM 1
un uno DI0MS0 0.986987
trabajo trabajo NCMS000 0.940476
muy muy RG 1
fértil fértil AQ0CS0 1
y y CC 0.999812
muy muy RG 1
hermoso hermoso AQ0MS0 1
reconoce reconocer VMIP3S0 0.75
Por por SPS00 1
su su DP3CS0 1
parte parte NCFS000 0.497967
Arcángel arcángel NP00SP0 1
ha haber VAIP3S0 0.998141
dicho decir VMP00SM 0.958333
que que CS 0.4375
Quijote quijote NP00SP0 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
es ser VSIP3S0 1
viaje viaje NCMS000 0.946429
entre entre SPS00 0.995223
la el DA0FS0 0.972146
ortodoxia ortodoxia NCFS000 1
y y CC 0.999812
la el DA0FS0 0.972146
heterodoxia heterodoxia NCFS000 1
e e CC 0.973684
incluye incluir VMIP3S0 0.75
temas tema NCMP000 0.625
para para SPS00 0.998507
todos todo DI0MP0 0.622419
los el DA0MP0 0.97623
```

```
gustos gusto NCMP000 1
Desde desde SPS00 1
alegrías alegría NCFP000 1
y y CC 0.999812
soleá soleá NCFS000 1
sin sin SPS00 1
olvidar olvidar VMN0000 1
fandangos fandango NCMP000 1
de de SPS00 0.999919
Huelva huelva NP00G00 1
y y CC 0.999812
tres 3 Z 1
canciones canción NCFP000 1
una uno DI0FS0 0.951241
de de SPS00 0.999919
ellas ellos PP3FP000 1
cuenta contar VMIP3S0 0.219697
con con SPS00 1
la el DA0FS0 0.972146
colaboración colaboración NCFS000 1
de de SPS00 0.999919
Antonio_Orozco antonio_orozco NP00SP0 1
a a SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
me me PP1CS000 0.889706
unen unir VMIP3P0 1
un uno DI0MS0 0.986987
vínculo vínculo NCMS000 1
artístico artístico AQ0MS0 1
y y CC 0.999812
personal personal AQ0CS0 0.710526
y y CC 0.999812
a a SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
tenía tener VMII1S0 0.5
ganas gana NCFP000 0.9375
de de SPS00 0.999919
acercar acercar VMN0000 1
a a SPS00 1
el el DA0MS0 1
flamenco flamenco NCMS000 0.25
tras tras SPS00 1
```

```
una uno DI0FS0 0.951241
colaboración colaboración NCFS000 1
anterior anterior AQ0CS0 1
continúa continuar VMIP3S0 0.875
el el DA0MS0 1
cantaor cantaor NCMS000 1
A_el_hilo al_hilo RG 1
de de SPS00 0.999919
anterior anterior AQ0CS0 1
Arcángel arcángel NP00V00 1
ha haber VAIP3S0 0.998141
señalado señalar VMP00SM 1
que que CS 0.4375
él él PP3MS000 1
siempre siempre RG 1
llama llamar VMIP3S0 0.948718
a a SPS00 0.99585
la el DA0FS0 0.972146
concordia concordia NCFS000 1
entre entre SPS00 0.995223
lo el DA0NS0 0.457393
tradicional tradicional AQ0CS0 1
y y CC 0.999812
la el DA0FS0 0.972146
vanguardia vanguardia NCFS000 1
y y CC 0.999812
los el DA0MP0 0.97623
clásicos clásico NCMP000 0.75
tienen tener VMIP3P0 1
que que CS 0.4375
respetar respetar VMN0000 1
a a SPS00 0.99585
la el DA0FS0 0.972146
gente gente NCFS000 1
que que PR0CN000 0.5625
tiene tener VMIP3S0 1
ganas gana NCFP000 0.9375
de de SPS00 0.999919
hacer hacer VMN0000 1
otras otro DI0FP0 0.731481
cosas cosa NCFP000 0.986842
y y CC 0.999812
la el DA0FS0 0.972146
gente gente NCFS000 1
```

```
de de SPS00 0.999919
vanguardia vanguardia NCFS000 1
tiene tener VMIP3S0 1
que que CS 0.4375
respetar respetar VMN0000 1
a a SPS00 0.99585
los el DA0MP0 0.97623
clásicos clásico NCMP000 0.75
de de SPS00 1
el el DA0MS0 1
flamenco flamenco NCMS000 0.25
sino sino CC 0.981707
es ser VSIP3S0 1
así así RG 0.994118
añade añadir VMIP3S0 0.833333
es ser VSIP3S0 1
difícil difícil AQ0CS0 1
llegar llegar VMN0000 1
a a SPS00 0.99585
un uno DI0MS0 0.986987
equilibrio equilibrio NCMS000 1
que que PR0CN000 0.5625
tenemos tener VMIP1P0 1
que que CS 0.4375
conseguir conseguir VMN0000 1
los el DA0MP0 0.97623
aficionados aficionado NCMP000 0.625
a a SPS00 0.99585
este este DD0MS0 0.956743
arte arte NCCS000 1
El el DA0MS0 1
cantaor cantaor NCMS000 1
de de SPS00 0.999919
Huelva huelva NP00G00 1
se se P0000000 0.465602
siente sentir VMIP3S0 0.911111
muy muy RG 1
identificado identificar VMP00SM 1
con con SPS00 1
título título NCMS000 1
de de SPS00 0.999919
este este DD0MS0 0.956743
trabajo trabajo NCMS000 0.940476
el el DA0MS0 1
```

```
cuarto cuarto NCMS000 0.641026
en en SPS00 1
su su DP3CS0 1
carrera carrera NCFS000 1
artística artístico AQ0FS0 1
me me PP1CS000 0.889706
siento sentar VMIP1S0 0.5
Quijote quijote NP00V00 1
por por SPS00 1
su su DP3CS0 1
afán afán NCMS000 1
de de SPS00 0.999919
luchador luchador NCMS000 0.490442
pero pero CC 0.998821
además además RG 1
es ser VSIP3S0 1
un uno DI0MS0 0.986987
homenaje homenaje NCMS000 1
a a SPS00 0.99585
Paco_Toronjo paco_toronjo NP00SP0 1
maestro maestro NCMS000 0.490442
de de SPS00 1
el el DA0MS0 1
fandango fandango NCMS000 1
onubense onubense AQ0CS0 0.509558
que que PR0CN000 0.5625
siempre siempre RG 1
ha haber VAIP3S0 0.998141
sido ser VSP00SM 1
un uno DI0MS0 0.986987
poco poco RG 0.54065
quijotesco quijotesco AQ0MS0 1
De de SPS00 0.999919
igual igual AQ0CS0 0.716667
modo modo NCMS000 1
en en SPS00 1
el el DA0MS0 1
libreto libreto NCMS000 1
de de SPS00 0.999919
este este DD0MS0 0.956743
disco disco NCMS000 1
hay haber VMIP3S0 1
un uno DI0MS0 0.986987
mención mención NCFS000 1
```

```
especial especial AQ0CS0 1
a a SPS00 0.99585
Enrique_Morente enrique_morente NP00SP0 1
a a SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
Arcángel arcángel NP00SP0 1
llama llamar VMIP3S0 0.948718
maestro maestro NCMS000 0.490442
Sobre_Cobos_Wilkins sobre_cobos_wilkins NP00SP0 1
responsable responsable NCCS000 0.416667
de de SPS00 1
el el DA0MS0 1
80_por_ciento 80/100 Zp 1
de de SPS00 0.999919
los el DA0MP0 0.97623
textos texto NCMP000 1
de de SPS00 0.999919
este este DD0MS0 0.956743
disco disco NCMS000 1
Arcángel arcángel NP00V00 1
ha haber VAIP3S0 0.998141
comentado comentar VMP00SM 1
que que CS 0.4375
es ser VSIP3S0 1
un uno DI0MS0 0.986987
artista artista NCCS000 0.75
con con SPS00 1
el el DA0MS0 1
que que PR0CN000 0.5625
me me PP1CS000 0.889706
puedo poder VMIP1S0 1
tomar tomar VMN0000 1
un uno DI0MS0 0.986987
café café NCMS000 1
por por SPS00 1
eso ese PD0NS000 1
le le PP3CSD00 1
invité invitar VMIS1S0 1
a a SPS00 0.99585
participar participar VMN0000 1
en en SPS00 1
este este DD0MS0 0.956743
disco disco NCMS000 1
```

```
que que CS 0.4375
el el DA0MS0 1
propio propio AQ0MS0 0.98
artista artista NCCS000 0.75
define definir VMIP3S0 0.75
como como CS 0.998668
familiar familiar AQ0CS0 0.9375
Los el DA0MP0 0.97623
conceptos concepto NCMP000 1
de de SPS00 0.999919
estos este DD0MP0 0.990566
textos texto NCMP000 1
van ir VMIP3P0 1
muy muy RG 1
de_acuerdo_con de_acuerdo_con SPS00 1
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
soy ser VSIP1S0 1
y y CC 0.999812
lo el DA0NS0 0.457393
que que PR0CN000 0.5625
intento intentar VMIP1S0 0.1
ser ser VSN0000 0.93883
cada cada DI0CS0 1
día día NCMS000 1
resalta resaltar VMIP3S0 0.993015
El el DA0MS0 1
resultado resultado NCMS000 0.807692
letras letra NCFP000 1
más más RG 1
complicadas complicar VMP00PF 1
que que PR0CN000 0.5625
tienen tener VMIP3P0 1
cierto cierto AQ0MS0 1
matiz matiz NCMS000 1
social social AQ0CS0 0.977273
revuelven revolver VMIP3P0 1
conciencias conciencia NCFP000 0.962932
y y CC 0.999812
hacen hacer VMIP3P0 1
pensar pensar VMN0000 1
Un uno DI0MS0 0.986987
trabajo trabajo NCMS000 0.940476
arriesgado arriesgar VMP00SM 1
```

```
que que PR0CN000 0.5625
Arcángel arcángel NP00SP0 1
define definir VMIP3S0 0.75
afirmando afirmar VMG0000 1
que que CS 0.4375
es ser VSIP3S0 1
bonito bonito AQ0MS0 0.928571
soñar soñar VMN0000 1
aunque aunque CC 1
a a SPS00 0.99585
veces vez NCFP000 1
los el DA0MP0 0.97623
sueños sueño NCMP000 1
te te PP2CS000 0.933962
peguen pegar VMSP3P0 0.860063
un uno DI0MS0 0.986987
porrazo porrazo NCMS000 1
Trayectoria trayectoria NP00V00 1
de de SPS00 0.999919
arcángel arcángel NCMS000 1
Desde desde SPS00 1
que que CS 0.4375
la el DA0FS0 0.972146
peña peña NCFS000 1
La_Orden_de_Huelva la_orden_de_huelva NP00O00 1
le le PP3CSD00 1
viera ver VMSI1S0 0.5
actuar actuar VMN0000 1
por por SPS00 1
primera 1 AO0FS0 0.987179
vez vez NCFS000 1
con con SPS00 1
apenas apenas RG 0.979167
diez 10 Z 1
años año NCMP000 1
la el DA0FS0 0.972146
trayectoria trayectoria NCFS000 1
artística artístico AQ0FS0 1
de de SPS00 0.999919
Arcángel arcángel NP00V00 1
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
profesionalizado profesionalizar VMP00SM 1
y y CC 0.999812
```

```
consolidado consolidar VMP00SM 1
convirtiendo convertir VMG0000 1
se se PP3CN000 1
en en SPS00 1
uno uno PI0MS000 0.964744
de de SPS00 0.999919
los el DA0MP0 0.97623
cantaores cantaor NCMP000 1
imprescindibles imprescindible AQ0CP0 1
de de SPS00 1
el el DA0MS0 1
flamenco flamenco NCMS000 0.25
actual actual AQ0CS0 1
Niño_de_Pura_Jesús_Cayuela_José_Roca
niño_de_pura_jesús_cayuela_josé_roca NP00V00 1
y y CC 0.999812
Mario_Maya mario_maya NP00SP0 1
fueron ser VSIS3P0 0.796296
algunos alguno PI0MP000 0.523438
de de SPS00 0.999919
los el DA0MP0 0.97623
artistas artista NCCP000 0.75
que que PR0CN000 0.5625
reclamaron reclamar VMIS3P0 1
su su DP3CS0 1
colaboración colaboración NCFS000 1
a a SPS00 1
el el DA0MS0 1
inicio inicio NCMS000 0.75
de de SPS00 0.999919
su su DP3CS0 1
carrera carrera NCFS000 1
hasta hasta SPS00 0.955172
que que CS 0.4375
el el DA0MS0 1
ciclo ciclo NCMS000 0.833333
Jueves_Flamencos jueves_flamencos NP00V00 1
de de SPS00 0.999919
la el DA0FS0 0.972146
fundación fundación NCFS000 1
Cajasol cajasol NP00O00 1
y y CC 0.999812
la el DA0FS0 0.972146
Bienal_de_Flamenco_de_Sevilla bienal_de_flamenco_de_sevilla NP00O00
```

```
1
le le PP3CSD00 1
cedieron ceder VMIS3P0 1
un uno DI0MS0 0.986987
lugar lugar NCMS000 1
protagonista protagonista AQ0CS0 0.15
y y CC 0.999812
empezó empezar VMIS3S0 1
a a SPS00 0.99585
ser ser VSN0000 0.93883
conocido conocer VMP00SM 0.954545
por por SPS00 1
méritos mérito NCMP000 1
propios propio AQ0MP0 0.884615
No no NP00SP0 1
ha haber VAIP3S0 0.998141
dejado dejar VMP00SM 1
de de SPS00 0.999919
cantar cantar VMN0000 0.875
para para SPS00 0.998507
el el DA0MS0 1
baile baile NCMS000 0.75
a a SPS00 0.99585
Israel_Galván israel_galván NP00V00 1
o o CC 0.998845
Eva_Yerbabuena eva_yerbabuena NP00SP0 1
y y CC 0.999812
su su DP3CS0 1
espíritu espíritu NCMS000 1
buscador buscador NCMS000 1
le le PP3CSD00 1
lleva llevar VMIP3S0 0.975
a a SPS00 0.99585
participar participar VMN0000 1
en en SPS00 1
proyectos proyecto NCMP000 1
novedosos novedoso AQ0MP0 1
como como CS 0.998668
Cus cus NP00SP0 1
cus cu NCFP000 1
flamenco flamenco AQ0MS0 0.75
en en SPS00 1
2001 2001 Z 1
con con SPS00 1
```

```
la el DA0FS0 0.972146
Orquesta_Chekara_de_Tetuán_Hasta_Quijote
orquesta_chekara_de_tetuán_hasta_quijote NP00000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
sueños sueño NCMP000 1
ha haber VAIP3S0 0.998141
editado editar VMP00SM 1
tres 3 Z 1
trabajos trabajo NCMP000 1
discográficos discográfico AQ0MP0 1
Arcángel arcángel NP00V00 1
en en SPS00 1
2001 2001 Z 1
que que PR0CN000 0.5625
estuvo estar VAIS3S0 1
producido producir VMP00SM 1
por por SPS00 1
Juan_Carlos_Romero juan_carlos_romero NP00SP0 1
y y CC 0.999812
fue ser VSIS3S0 0.932292
reconocido reconocer VMP00SM 1
por por SPS00 1
los el DA0MP0 0.97623
premios premio NCMP000 1
Andalucía_Joven andalucía_joven NP00000 1
2002 2002 Z 1
y y CC 0.999812
el el DA0MS0 1
Nacional_Flamenco_Activo_de_Úbeda
nacional_flamenco_activo_de_úbeda NP00V00 1
La el DA0FS0 0.972146
calle calle NCFS000 0.96875
perdía perder VMII1S0 0.5
y y CC 0.999812
Ropavieja ropavieja NP00SP0 1
son ser VSIP3P0 0.986772
otros otro DI0MP0 0.56875
proyectos proyecto NCMP000 1
discográficos discográfico AQ0MP0 1
de de SPS00 1
el el DA0MS0 1
cantaor cantaor NCMS000 1
que que PR0CN000 0.5625
```

```
recibió recibir VMIS3S0 1
en en SPS00 1
2002 2002 Z 1
el el DA0MS0 1
Giraldillo_de_Cante giraldillo_de_cante NP00V00 1
reconocimiento reconocimiento NCMS000 1
que que PR0CN000 0.5625
consagró consagrar VMIS3S0 1
su su DP3CS0 1
carrera carrera NCFS000 1
que que PR0CN000 0.5625
le le PP3CSD00 1
llevará llevar VMIF3S0 1
el el DA0MS0 1
próximo_4_de_noviembre [??:4/11/??:??.??:??] W 1
a a SPS00 1
el el DA0MS0 1
Teatro_Villamarta_de_Jerez teatro_villamarta_de_jerez NP00SP0 1
dentro_de dentro_de SPS00 1
el el DA0MS0 1
ciclo ciclo NCMS000 0.833333
Flamenco flamenco AQ0MS0 0.75
Viene venir VMIP3S0 1
de de SPS00 1
el el DA0MS0 1
Sur sur NCMS000 1
```


Toledo

CECAM considera necesaria la moderación salarial para el mantenimiento del empleo La Confederación Regional de Empresarios de Castilla-La Mancha CECAM afirmó hoy no entender tras los datos del Índice de Precios al Consumo IPC del mes de noviembre y ante las previsiones para los próximos meses la insistencia de los planteamientos realizados por las centrales sindicales respecto a unos incrementos salariales muy por encima de nuestra realidad que se demuestran de todo punto inviables y contraproducentes con el objetivo del mantenimiento del empleo En nota de prensa en la que valoró el Índice de Precios al Consumo del mes de noviembre la patronal regional consideró necesario realizar ahora más que nunca un ejercicio de responsabilidad que lleve a una moderación en torno a dichos incrementos con el objetivo final del mantenimiento

del empleo en nuestra región CECAM mantuvo que Castilla-La Mancha comenzó siendo una de las primeras regiones con tasas de inflación negativas y considerando los datos actuales será de las últimas en recuperar tasas positivas Previsiblemente será a principios del próximo ejercicio cuando se recuperen cifras positivas Por ello consideró que estos datos son significativos de la debilidad de la demanda interna fruto de la desconfianza existente motivada por la coyuntura económica que estamos atravesando y la ausencia de pruebas que puedan ayudar la idea de recuperación en 2010

```
 CECAM cecam NP00000 1
considera considerar VMIP3S0 0.9
necesaria necesario AQ0FS0 1
la el DA0FS0 0.972146
moderación moderación NCFS000 1
salarial salarial AQ0CS0 1
para para SPS00 0.998507
el el DA0MS0 1
mantenimiento mantenimiento NCMS000 1
de de SPS00 1
el el DA0MS0 1
empleo empleo NCMS000 0.785714
La_Confederación_Regional_de_Empresarios_de_Castilla-La_Mancha_CECAM
la_confederación_regional_de_empresarios_de_castilla-la_mancha_cecam
NP00000 1
afirmó afirmar VMIS3S0 1
hoy hoy RG 0.877358
no no RN 0.99778
entender entender VMN0000 0.9
tras tras SPS00 1
los el DA0MP0 0.97623
datos dato NCMP000 1
de de SPS00 1
el el DA0MS0 1
Índice_de_Precios_a_el_Consumo_IPC
índice_de_precios_a_el_consumo_ipc NP00V00 1
de de SPS00 1
el el DA0MS0 1
mes_de_noviembre [??:??/11/??:??.??:??] W 1
y y CC 0.999812
ante ante SPS00 0.990566
las el DA0FP0 0.97051
previsiones previsión NCFP000 1
para para SPS00 0.998507
los el DA0MP0 0.97623
```

```
próximos próximo AQ0MP0 1
meses mes NCMP000 0.964286
la el DA0FS0 0.972146
insistencia insistencia NCFS000 1
de de SPS00 0.999919
los el DA0MP0 0.97623
planteamientos planteamiento NCMP000 1
realizados realizar VMP00PM 1
por por SPS00 1
las el DA0FP0 0.97051
centrales central NCFP000 0.166667
sindicales sindical AQ0CP0 0.75
respecto_a respecto_a SPS00 1
unos uno DI0MP0 0.879032
incrementos incremento NCMP000 1
salariales salarial AQ0CP0 0.973684
muy muy RG 1
por por SPS00 1
encima_de encima_de SPS00 1
nuestra nuestro DP1FSP 0.892857
realidad realidad NCFS000 1
que que PR0CN000 0.5625
se se P0000000 0.465602
demuestran demostrar VMIP3P0 1
de_todo_punto de_todo_punto RG 1
inviables inviable AQ0CP0 1
y y CC 0.999812
contraproducentes contraproducente AQ0CP0 1
con con SPS00 1
el el DA0MS0 1
objetivo objetivo NCMS000 0.761905
de de SPS00 1
el el DA0MS0 1
mantenimiento mantenimiento NCMS000 1
de de SPS00 1
el el DA0MS0 1
empleo empleo NCMS000 0.785714
En en SPS00 1
nota nota NCFS000 0.3125
de de SPS00 0.999919
prensa prensa NCFS000 0.925926
en en SPS00 1
la el DA0FS0 0.972146
que que PR0CN000 0.5625
```

```
valoró valorar VMIS3S0 1
el el DA0MS0 1
Índice_de_Precios índice_de_precios NP00V00 1
a a SPS00 1
el el DA0MS0 1
Consumo consumo NP00O00 1
de de SPS00 1
el el DA0MS0 1
mes_de_noviembre [??:??/11/??:??.??:??] W 1
la el DA0FS0 0.972146
patronal patronal NCFS000 0.25
regional regional AQ0CS0 1
consideró considerar VMIS3S0 1
necesario necesario AQ0MS0 1
realizar realizar VMN0000 1
ahora ahora RG 1
más más RG 1
que que CS 0.4375
nunca nunca RG 1
un uno DI0MS0 0.986987
ejercicio ejercicio NCMS000 1
de de SPS00 0.999919
responsabilidad responsabilidad NCFS000 1
que que PR0CN000 0.5625
lleve llevar VMSP1S0 0.462493
a a SPS00 0.99585
una uno DI0FS0 0.951241
moderación moderación NCFS000 1
en_torno_a en_torno_a SPS00 1
dichos decir VMP00PM 0.75
incrementos incremento NCMP000 1
con con SPS00 1
el el DA0MS0 1
objetivo objetivo NCMS000 0.761905
final final AQ0CS0 0.148148
de de SPS00 1
el el DA0MS0 1
mantenimiento mantenimiento NCMS000 1
de de SPS00 1
el el DA0MS0 1
empleo empleo NCMS000 0.785714
en en SPS00 1
nuestra nuestro DP1FSP 0.892857
región región NCFS000 1
```

```
CECAM cecam NP00V00 1
mantuvo mantener VMIS3S0 1
que que CS 0.4375
Castilla-La_Mancha castilla-la_mancha NP00G00 1
comenzó comenzar VMIS3S0 1
siendo ser VSG0000 1
una uno DI0FS0 0.951241
de de SPS00 0.999919
las el DA0FP0 0.97051
primeras 1 AO0FP0 0.9
regiones región NCFP000 1
con con SPS00 1
tasas tasa NCFP000 0.962932
de de SPS00 0.999919
inflación inflación NCFS000 1
negativas negativo AQ0FP0 0.509558
y y CC 0.999812
considerando considerar VMG0000 0.375
los el DA0MP0 0.97623
datos dato NCMP000 1
actuales actual AQ0CP0 1
será ser VSIF3S0 1
de de SPS00 0.999919
las el DA0FP0 0.97051
últimas último AO0FP0 1
en en SPS00 1
recuperar recuperar VMN0000 1
tasas tasa NCFP000 0.962932
positivas positivo AQ0FP0 1
Previsiblemente previsiblemente NP00V00 1
será ser VSIF3S0 1
a_principios_de a_principios_de SPS00 1
el el DA0MS0 1
próximo próximo AQ0MS0 1
ejercicio ejercicio NCMS000 1
cuando cuando CS 0.983796
se se P0000000 0.465602
recuperen recuperar VMSP3P0 0.860063
cifras cifra NCFP000 0.75
positivas positivo AQ0FP0 1
Por por SPS00 1
ello él PP3NS000 1
consideró considerar VMIS3S0 1
que que CS 0.4375
```

```
estos este DD0MP0 0.990566
datos dato NCMP000 1
son ser VSIP3P0 0.986772
significativos significativo AQ0MP0 1
de de SPS00 0.999919
la el DA0FS0 0.972146
debilidad debilidad NCFS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
demanda demanda NCFS000 0.833333
interna interno AQ0FS0 0.75
fruto fruto NCMS000 1
de de SPS00 0.999919
la el DA0FS0 0.972146
desconfianza desconfianza NCFS000 1
existente existente AQ0CS0 1
motivada motivar VMP00SF 1
por por SPS00 1
la el DA0FS0 0.972146
coyuntura coyuntura NCFS000 1
económica económico AQ0FS0 1
que que PR0CN000 0.5625
estamos estar VAIP1P0 1
atravesando atravesar VMG0000 1
y y CC 0.999812
la el DA0FS0 0.972146
ausencia ausencia NCFS000 1
de de SPS00 0.999919
pruebas prueba NCFP000 0.9
que que PR0CN000 0.5625
puedan poder VMSP3P0 0.928571
ayudar ayudar VMN0000 1
la el DA0FS0 0.972146
idea idea NCFS000 0.974138
de de SPS00 0.999919
recuperación recuperación NCFS000 1
en en SPS00 1
2010 2010 Z 1
```

Valencia

La primera cría de jirafa nacida en Bioparc ya comparte con otros animales la Sabana verde Tumai la primera cría de jirafa nacida en Bioparc de Valencia ya comparte con otras especies animales la Sabana verde tras cinco meses de aclimatación según ha informado el parque en un comunicado Tumai la primera cría de jirafa nacida en Bioparc de Valencia ya comparte con otras especies animales la Sabana verde tras cinco meses de aclimatación según ha informado el parque en un comunicado La primera cría de jirafa nacida en Bioparc ya comparte con otros animales la Sabana verde Tumai fue rechazado por su madre al nacer y ha sido alimentado por sus cuidadores con biberones gigantes Tras cinco meses ha comenzado a salir con otros animales por la Sabana verde por la tarde cuando el sol ya ha bajado de intensidad En el tiempo que lleva saliendo se la ha visto integrado con el entorno Tumai fue el nombre elegido por los internautas para la primera cría de jirafa nacida en Bioparc con casi un 50 por ciento de votos Es un nombre africano que significa esperanza de vida Ahora se ha puesto en marcha la elección del nombre de la cría de hipopótamo que nació hace un mes el 19 de julio Para ello los cuidadores del animal eligen cuatro opciones de nombre entre las que se puede votar la preferida La votación se realiza a través de la página de Bioparc Valencia en Facebook en el siguiente link 927v app_48845543673 o entrando en la página wwwbioparcvalenciaes y accediendo al enlace de la elección Los cuatro nominados son Ceive significa libre en gallego Nanuk personaje de dibujos animados Eris nombre italiano que se ha elegido en memoria de un abuelo y Durban ciudad de Sudáfrica donde España ganó a Alemania en las semifinales del último mundial de fútbol


```
 La el DA0FS0 0.972146
primera 1 AO0FS0 0.987179
cría cría NCFS000 0.416667
de de SPS00 0.999919
jirafa jirafa NCFS000 1
nacida nacer VMP00SF 0.75
en en SPS00 1
Bioparc bioparc NP00G00 1
ya ya RG 0.996988
comparte compartir VMIP3S0 0.215152
con con SPS00 1
otros otro DI0MP0 0.56875
animales animal NCMP000 0.974359
la el DA0FS0 0.972146
Sabana sabana NP00V00 1
verde verde AQ0CS0 0.916667
```

```
Tumai tumai NP00SP0 1
la el DA0FS0 0.972146
primera 1 AO0FS0 0.987179
cría cría NCFS000 0.416667
de de SPS00 0.999919
jirafa jirafa NCFS000 1
nacida nacer VMP00SF 0.75
en en SPS00 1
Bioparc bioparc NP00G00 1
de de SPS00 0.999919
Valencia valencia NP00O00 1
ya ya RG 0.996988
comparte compartir VMIP3S0 0.215152
con con SPS00 1
otras otro DI0FP0 0.731481
especies especie NCFP000 1
animales animal NCMP000 0.974359
la el DA0FS0 0.972146
Sabana sabana NP00V00 1
verde verde AQ0CS0 0.916667
tras tras SPS00 1
cinco 5 Z 0.98
meses mes NCMP000 0.964286
de de SPS00 0.999919
aclimatación aclimatación NCFS000 1
según según SPS00 0.986667
ha haber VAIP3S0 0.998141
informado informar VMP00SM 1
el el DA0MS0 1
parque parque NCMS000 1
en en SPS00 1
un uno DI0MS0 0.986987
comunicado comunicado NCMS000 0.442539
Tumai tumai NP00V00 1
la el DA0FS0 0.972146
primera 1 AO0FS0 0.987179
cría cría NCFS000 0.416667
de de SPS00 0.999919
jirafa jirafa NCFS000 1
nacida nacer VMP00SF 0.75
en en SPS00 1
Bioparc bioparc NP00G00 1
de de SPS00 0.999919
Valencia valencia NP00O00 1
```

```
ya ya RG 0.996988
comparte compartir VMIP3S0 0.215152
con con SPS00 1
otras otro DI0FP0 0.731481
especies especie NCFP000 1
animales animal NCMP000 0.974359
la el DA0FS0 0.972146
Sabana sabana NP00V00 1
verde verde AQ0CS0 0.916667
tras tras SPS00 1
cinco 5 Z 0.98
meses mes NCMP000 0.964286
de de SPS00 0.999919
aclimatación aclimatación NCFS000 1
según según SPS00 0.986667
ha haber VAIP3S0 0.998141
informado informar VMP00SM 1
el el DA0MS0 1
parque parque NCMS000 1
en en SPS00 1
un uno DI0MS0 0.986987
comunicado comunicado NCMS000 0.442539
La el DA0FS0 0.972146
primera 1 AO0FS0 0.987179
cría cría NCFS000 0.416667
de de SPS00 0.999919
jirafa jirafa NCFS000 1
nacida nacer VMP00SF 0.75
en en SPS00 1
Bioparc bioparc NP00G00 1
ya ya RG 0.996988
comparte compartir VMIP3S0 0.215152
con con SPS00 1
otros otro DI0MP0 0.56875
animales animal NCMP000 0.974359
la el DA0FS0 0.972146
Sabana_verde_Tumai sabana_verde_tumai NP00V00 1
fue ser VSIS3S0 0.932292
rechazado rechazar VMP00SM 1
por por SPS00 1
su su DP3CS0 1
madre madre NCFS000 1
a a SPS00 1
el el DA0MS0 1
```

```
nacer nacer VMN0000 1
y y CC 0.999812
ha haber VAIP3S0 0.998141
sido ser VSP00SM 1
alimentado alimentar VMP00SM 1
por por SPS00 1
sus su DP3CP0 0.998462
cuidadores cuidador NCMP000 0.490442
con con SPS00 1
biberones biberón NCMP000 1
gigantes gigante AQ0CP0 0.25
Tras tras SPS00 1
cinco 5 Z 0.98
meses mes NCMP000 0.964286
ha haber VAIP3S0 0.998141
comenzado comenzar VMP00SM 1
a a SPS00 0.99585
salir salir VMN0000 1
con con SPS00 1
otros otro DI0MP0 0.56875
animales animal NCMP000 0.974359
por por SPS00 1
la el DA0FS0 0.972146
Sabana sabana NP00V00 1
verde verde AQ0CS0 0.916667
por por SPS00 1
la el DA0FS0 0.972146
tarde tarde NCFS000 0.269565
cuando cuando CS 0.983796
el el DA0MS0 1
sol sol NCMS000 1
ya ya RG 0.996988
ha haber VAIP3S0 0.998141
bajado bajar VMP00SM 1
de de SPS00 0.999919
intensidad intensidad NCFS000 1
En en SPS00 1
el el DA0MS0 1
tiempo tiempo NCMS000 1
que que PR0CN000 0.5625
lleva llevar VMIP3S0 0.975
saliendo salir VMG0000 1
se se P0000000 0.465602
la lo PP3FSA00 0.0277626
```

```
ha haber VAIP3S0 0.998141
visto ver VMP00SM 0.982759
integrado integrar VMP00SM 1
con con SPS00 1
el el DA0MS0 1
entorno entorno NCMS000 0.95
Tumai tumai NP00SP0 1
fue ser VSIS3S0 0.932292
el el DA0MS0 1
nombre nombre NCMS000 0.97973
elegido elegir VMP00SM 0.75
por por SPS00 1
los el DA0MP0 0.97623
internautas internauta NCCP000 1
para para SPS00 0.998507
la el DA0FS0 0.972146
primera 1 AO0FS0 0.987179
cría cría NCFS000 0.416667
de de SPS00 0.999919
jirafa jirafa NCFS000 1
nacida nacer VMP00SF 0.75
en en SPS00 1
Bioparc bioparc NP00G00 1
con con SPS00 1
casi casi RG 1
un uno DI0MS0 0.986987
50_por_ciento 50/100 Zp 1
de de SPS00 0.999919
votos voto NCMP000 1
Es es NP00V00 1
un uno DI0MS0 0.986987
nombre nombre NCMS000 0.97973
africano africano AQ0MS0 0.509558
que que PR0CN000 0.5625
significa significar VMIP3S0 0.958333
esperanza esperanza NCFS000 0.925926
de de SPS00 0.999919
vida vida NCFS000 1
Ahora ahora RG 1
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
puesto poner VMP00SM 0.547619
en en SPS00 1
marcha marcha NCFS000 0.948718
```

```
la el DA0FS0 0.972146
elección elección NCFS000 1
de de SPS00 1
el el DA0MS0 1
nombre nombre NCMS000 0.97973
de de SPS00 0.999919
la el DA0FS0 0.972146
cría cría NCFS000 0.416667
de de SPS00 0.999919
hipopótamo hipopótamo NCMS000 1
que que PR0CN000 0.5625
nació nacer VMIS3S0 1
hace hacer VMIP3S0 1
un uno DI0MS0 0.986987
mes mes NCMS000 1
el el DA0MS0 1
19_de_julio [??:19/7/??:??.??:??] W 1
Para para SPS00 0.998507
ello él PP3NS000 1
los el DA0MP0 0.97623
cuidadores cuidador NCMP000 0.490442
de de SPS00 1
el el DA0MS0 1
animal animal NCMS000 0.944444
eligen elegir VMIP3P0 1
cuatro 4 Z 1
opciones opción NCFP000 1
de de SPS00 0.999919
nombre nombre NCMS000 0.97973
entre entre SPS00 0.995223
las el DA0FP0 0.97051
que que PR0CN000 0.5625
se se P0000000 0.465602
puede poder VMIP3S0 0.995614
votar votar VMN0000 1
la el DA0FS0 0.972146
preferida preferir VMP00SF 1
La el DA0FS0 0.972146
votación votación NCFS000 1
se se P0000000 0.465602
realiza realizar VMIP3S0 0.75
a_través_de a_través_de SPS00 1
la el DA0FS0 0.972146
página página NCFS000 1
```

```
de de SPS00 0.999919
Bioparc_Valencia bioparc_valencia NP00O00 1
en en SPS00 1
Facebook facebook NP00G00 1
en en SPS00 1
el el DA0MS0 1
siguiente siguiente AQ0CS0 1
link link NCMS000 0.60039
927v 927v Z 1
app_48845543673 app_48845543673 Z 1
o o CC 0.998845
entrando entrar VMG0000 1
en en SPS00 1
la el DA0FS0 0.972146
página página NCFS000 1
wwwbioparcvalenciaes wwwbioparcvalenciaes AQ0CP0 0.233861
y y CC 0.999812
accediendo acceder VMG0000 1
a a SPS00 1
el el DA0MS0 1
enlace enlace NCMS000 0.625
de de SPS00 0.999919
la el DA0FS0 0.972146
elección elección NCFS000 1
Los el DA0MP0 0.97623
cuatro 4 Z 1
nominados nominar VMP00PM 1
son ser VSIP3P0 0.986772
Ceive ceive NP00SP0 1
significa significar VMIP3S0 0.958333
libre libre AQ0CS0 0.917647
en en SPS00 1
gallego gallego NCMS000 0.5
Nanuk nanuk NP00SP0 1
personaje personaje NCMS000 1
de de SPS00 0.999919
dibujos dibujo NCMP000 1
animados animar VMP00PM 1
Eris eris NP00SP0 1
nombre nombre NCMS000 0.97973
italiano italiano AQ0MS0 0.166667
que que PR0CN000 0.5625
se se P0000000 0.465602
ha haber VAIP3S0 0.998141
```

```
elegido elegir VMP00SM 0.75
en_memoria_de en_memoria_de SPS00 1
un uno DI0MS0 0.986987
abuelo abuelo NCMS000 1
y y CC 0.999812
Durban durban NP00V00 1
ciudad ciudad NCFS000 1
de de SPS00 0.999919
Sudáfrica sudáfrica NP00G00 1
donde donde PR000000 0.963542
España españa NP00G00 1
ganó ganar VMIS3S0 1
a a SPS00 0.99585
Alemania alemania NP00G00 1
en en SPS00 1
las el DA0FP0 0.97051
semifinales semifinal NCFP000 1
de de SPS00 1
el el DA0MS0 1
último último AO0MS0 1
mundial mundial NCMS000 0.0714286
de de SPS00 0.999919
fútbol fútbol NCMS000 1
```

Bibliografía

Amitay, E., HarÉl, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference*, SIGIR '04, pages 273–280, New York, NY, USA. ACM. 1, 2.1.2, 3, 3.1, 3.2, 3.3, 3.4

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York. 1, 2.1.4

Beesley, K. R. and Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*. 2

Benzécri, J.-P. and Bellier, L. (1976). *Lánalyse des données*, volume 1. Dunod Paris. 6.1

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203. 6.4.1

Bigi, B. (2003). *Using Kullback-Leibler distance for text categorization*. Springer. 6.1, 6.1

Bilhaut, F., Charnois, T., Enjalbert, P., and Mathet, Y. (2003). Geographic reference analysis for geographic document querying. In *In Proceedings of Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT)*. 2

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM. 5.1.1

Burnham, K. P. and Anderson, D. (2003). Model selection and multi-model inference. *A Pratical informatio-theoric approch. Sringer*. 6.1

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19. 2.1.2

Buscaldi, D. and Magnini, B. (2010). Grounding toponyms in an italian local news corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, page 15. ACM. 2.1.1

Buscaldi, D. and Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22(3):301–313. 3.2

Cardie, C. (1996). Automating feature set selection for case-based learning of linguistic knowledge. 5.3.1

Cardoso, N. and Santos, D. (2008). To separate or not to separate: reflections about current gir practice. In *Workshop on Novel Methodologies for Evaluation in Information Retrieval*, page 19. Citeseer. 2.1.4

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM. 3.5, 3.5.2, 3.5.2

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 25–30. ACM. 2.1.3

Clough, P., Tang, J., Hall, M. M., and Warner, A. (2011). Linking archival data to location: a case study at the uk national archives. *ASLIB Proceedings*, 63(2/3):127–147. 2.1.6

Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 761–770, New York, NY, USA. ACM. 3.5.1

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *International AAAI Conference on Weblogs and Social Media*, page 58. 3.5.2

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *ACL*, pages 168–175. 3.1

Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics. 3

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., and Vilain, M. (1997). Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 348–355. Association for Computational Linguistics. 3.1

Devijver y Kittler (1982). *Pattern recognition: A statistical approach*, volume 761. Prentice-Hall London. 5

Ding, J., Shivakumar, N., and Gravano, L. (2000). Computing geographical scopes of web resources. pages 545–556. 3.1, 3.2, 3.3, 3.3, 3.4

Earle, P. S., Bowden, D. C., and Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6). 3.5

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874. 5.1.1

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics. 3

Frew, J., Freeston, M., Freitas, N., Hill, L., Janee, G., Lovette, K., Nideffer, R., Smith, T., and Zheng, Q. (1998). The alexandria digital library architecture. In *Research and Advanced Technology for Digital Libraries*, pages 61–73. Springer. 2.2.1

Fu, G., Jones, C. B., and Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the web. In *Databases and Applications*, pages 167–172. 1

Gan, Q., Attenberg, J., Markowetz, A., and Suel, T. (2008). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, pages 49–56. ACM. 1

Garbin, E. and Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 363–370. Association for Computational Linguistics. 2

Gey, F., Larson, R., Machado, J., and Yoshioka, M. (2011). Ntcir9-geotime overview - evaluating geographic and temporal search: Round 2, in this proceedings. 3.6

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221. 2.1.2

Harris, Z. S. (1954). Distributional structure. *Word*. 5.1.1

Hill, L. L. (2006). Georeferencing: The geographic associations of information (digital libraries and electronic publishing). 2.1.6, 3

Hopcroft, J. E. (2007). *Introduction to automata theory, languages, and computation*. Pearson Addison Wesley. 2

How, B. C. and Narayanan, K. (2004). An empirical study of feature selection for text categorization based on term weightage. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 599–602, Washington, DC, USA. IEEE Computer Society. 5.3.1

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S., and Sundararajan, S. (2008). A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on Machine learning*, pages 408–415. ACM. 5.1.1

Iyyer, M., Boyd-Graber, J. L., Claudino, L. M. B., Socher, R., and Daumé III, H. (2014). A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644. 7.4

Janowicz, K. and Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129–1157. 1

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446. 3

Jiménez, D. (1998). Dynamically weighted ensemble neural networks for classification. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pages 753–756. IEEE. 3.5.2

Jones, C. B. and Purves, R. S. (2009). *Geographical information retrieval*. Springer. 2.1

Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H. (2008). Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065. 2.1.3

Kinsella, S., Murdock, V., and O'Hare, N. (2011). ï'm eating a sandwich in glasgow": modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 61–68, New York, NY, USA. ACM. 3.5.2, 3.5.2, 5.1.2

Kittur, A., Chi, E. H., and Suh, B. (2009). What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1509–1512. ACM. 4.3

Kullback, S. (1987). Letter to the editor: the kullback-leibler distance. 6.1

Kullback, S. (1997). *Information theory and statistics*. Courier Dover Publications. 2, 6.1

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, pages 79–86. 6.1

Lancaster, H. O. and Seneta, E. (1969). *Chi-Square Distribution*. Wiley Online Library. 5.3.1

Larson, R. R. and Larson, R. R. (1996). *Geographic Information Retrieval and Spatial Browsing*, pages 81–124. University of Illinois. 3

Leidner, J. L. (2008). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Dissertation.Com. 2.1.2, 1, 3.2

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11. 2.1.1

Leveling, J. and Hartrumpf, S. (2006). On metonymy recognition for geographic ir. In *GIR*. 2.1.1

Lin, C.-J., Weng, R. C., and Keerthi, S. S. (2008). Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, 9:627–650. 5.1.1

Lloret, E., Gutiérrez, Y., Peregrino, F. S., Gómez, J. M., Guillén, A., and Llopis, F. (2015). Explotación y tratamiento de la información disponible en internet para la anotación y generación de textos adaptados al usuario. *Procesamiento del Lenguaje Natural*, 55:177–180. 7.5.3

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA. 6.4

Mahmud, J., Nichols, J., and Drews, C. (2012). *Where Is This Tweet From? Inferring Home Locations of Twitter Users*, pages 511–514. 3.5.2

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA. 3

Mitamura, T., Nyberg, E., Shima, H., Kato, T., Mori, T., Lin, C.-Y., Song, R., Lin, C.-J., Sakai, T., Ji, D., et al. (2008). Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access. In *Proceedings of the Seventh NTCIR Workshop Meeting*, pages 16–19. 2.1.7

Moh'd A Mesleh, A. (2007). Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, 3(6):430–435. 5.3.1

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204. 2.1.3

Mostern, R. and Johnson, I. (2008). From named place to naming event: creating gazetteers for history. *International Journal of Geographical Information Science*, 22(10):1091–1108. 1

Nadeau, D., Turney, P., and Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. 1

Osiński, S. and Weiss, D. (2005). Carrot2: Design of a flexible and efficient web information retrieval framework. In *Advances in Web Intelligence*, pages 439–444. Springer. 6.4

Overell, S. E. and Rüger, S. M. (2006). Identifying and grounding descriptions of places. In *GIR*. 4.3

Padró, L., Reese, S., Agirre, E., and Soroa, A. (2010). Semantic services in freeling 2.1: Wordnet and ukb. In Bhattacharyya, P., Fellbaum, C., and Vossen, P., editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India. Global Wordnet Conference 2010, Narosa Publishing House. 5.2.1

Peregrino, F. S., Díaz, D. T., and Pascual, F. L. (2012a). Estudio sobre el impacto de los componentes de un sistema de recuperación de información geográfica y temporal. *Linguamática*, 3(2):69–82. 7.5.1

Peregrino, F. S., Tomás, D., Clough, P., and Llopis, F. (2012b). Mapping routes of sentiments. In *Spanish Conference on Information Retrieval*. 7.5.3

Peregrino, F. S., Tomás, D., and Llopis, F. (2011a). Map-based filters for fuzzy entities in geographical information retrieval. In *Natural Language Processing and Information Systems*, pages 270–273. Springer. 2.1.3, 7.5.1

Peregrino, F. S., Tomás, D., and Llopis, F. (2011b). University of alicante at ntcir-9 geotime. 7.5.1

Peregrino, F. S., Tomás, D., and Llopis, F. (2012c). Una aproximación basada en corpus para la detección del foco geográfico en el texto. *Procesamiento del Lenguaje Natural*, 50(0). 1, 7.5.2

Peregrino, F. S., Tomás, D., and Llopis, F. (2013). Every move you make i'll be watching you: geographical focus detection on twitter. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 1–8. ACM. 7.5.2

Peregrino, F. S., Tomás, D., and Pascual, F. L. (2012d). Question answering and multi-search engines in geo-temporal information retrieval. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 342–352. Springer. 7.5.1

Peregrino Torregrosa, F., Tomás Díaz, D., Llopis Pascual, F., et al. (2014). Clasificación geográfica de textos informales. 7.5.2

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM. 3.5.2, 5.1.2

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., et al. (2007). The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, 21(7):717–745. 2.1, 2.1.6

Roberts, K., Bejan, C. A., and Harabagiu, S. M. (2010). Toponym disambiguation using events. In *FLAIRS Conference*. 3.6

Robertson, S. E., Walker, S., Beaulieu, M., and Willett, P. (1999). Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. *Nist Special Publication SP*, pages 253–264. 2.1.5

Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM. 3.5.2

Sakai, T. (2007). On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pages 32–43. 2

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM. 3.5

Schilder, F., Versley, Y., and Habel, C. (2008). Extracting spatial information : grounding, classifying and linking spatial expressions. 2

Serdyukov, P., Murdock, V., and van Zwol, R. (2009). Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 484–491, New York, NY, USA. ACM. 3.2, 3, 3.5.1

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399. 2.1.2, 2.1.4

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240. 3

Tomás, D., Peregrino, F. S., Llopis, F., Vázquez, S., Moreda, P., Saquete, E., Gómez, J. M., Izquierdo, R., and Ferrández, Ó. (2012). Tratamiento de la dimensión espacial en el texto y su aplicación a la recuperación de información. *Procesamiento del Lenguaje Natural*, 49:193–196. 7.5.2

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In *Advances in Spatial and Temporal Databases*, pages 218–235. Springer. 2.1.4

Van Kreveld, M., Reinbacher, I., Arampatzis, A., and Van Zwol, R. (2005). Multi-dimensional scattered ranking methods for geographic information retrieval*. *GeoInformatica*, 9(1):61–84. 2.1.5

Vázquez, S., Lloret, E., Peregrino, F., Gutiérrez, Y., Fernández, J., and Gómez, J. M. (2014). Tratamiento inteligente de la información para ayuda a la toma de decisiones. *Procesamiento del Lenguaje Natural*, 53:139–142. 7.5.2

Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005a). Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on Geographic information retrieval*, GIR '05, pages 17–24, New York, NY, USA. ACM. 2, 2.1, 2.1.2, 2.1.4

Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y., and Li, Y. (2005b). Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 424–431, New York, NY, USA. ACM. 2

Wang, X. and He, Q. (2004). Enhancing generalization capability of svm classifiers with feature weight adjustment. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 1037–1043. Springer. 5.3.1

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. 3.5.2, 5.3.1

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420. 5.3.1

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214. 5.1.2

Zhang, W. V., Rey, B., Stipp, E., and Jones, R. (2006). Geomodification in query rewriting. In *GIR*. 1

Zheng, Z., Wu, X., and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, 6(1):80–89. 5.3.1

Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. (2005). On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362. ACM. 3.1, 3.2, 3.3, 3.4, 3.6