Universitat d'Alacant
Universidad de Alicante

maESL
Máster Universitario en Estudios Literarios

Tema 6

Text Mining con *Topic Modeling*

Recursos informáticos para la investigación literaria

Máster en Estudios Literarios
Universidad de Alicante
Curso 2014-2015

Borja Navarro Colorado
borja@dlsi.ua.es
@bncolorado

Contenidos

+ Introducción.
+ ¿Qué es un *topic*?
+ Cómo funciona *topic modeling*.
+ Análisis de resultados. Problemas.
+ Aspectos prácticos: MALLET.

Text Mining

- Identificación de patrones recurrentes en amplias colecciones de texto.

- *Topic Modeling* es una forma concreta de hacer *Text Mining*.

Topic Modeling

Conjunto de algoritmos capaces de detectar y extraer **relaciones semánticas latentes** de amplios corpus.

- *Latent Dirichlet Allocation*

Topic!

¿Qué es un *topic*?

- *Topic*: conjunto de palabras que tienden a aparecer juntas en los mismos contextos.
  - Teoría semántica distribucional & Wittgenstein.
  - Si dos palabras co-ocurren (tienden a aparecer en los mismos contextos), se asume que hacen referencia a los mismos temas.
    - Esta tendencia debe ser muy marcada: función de probabilidad.

¿Qué es un *topic*?

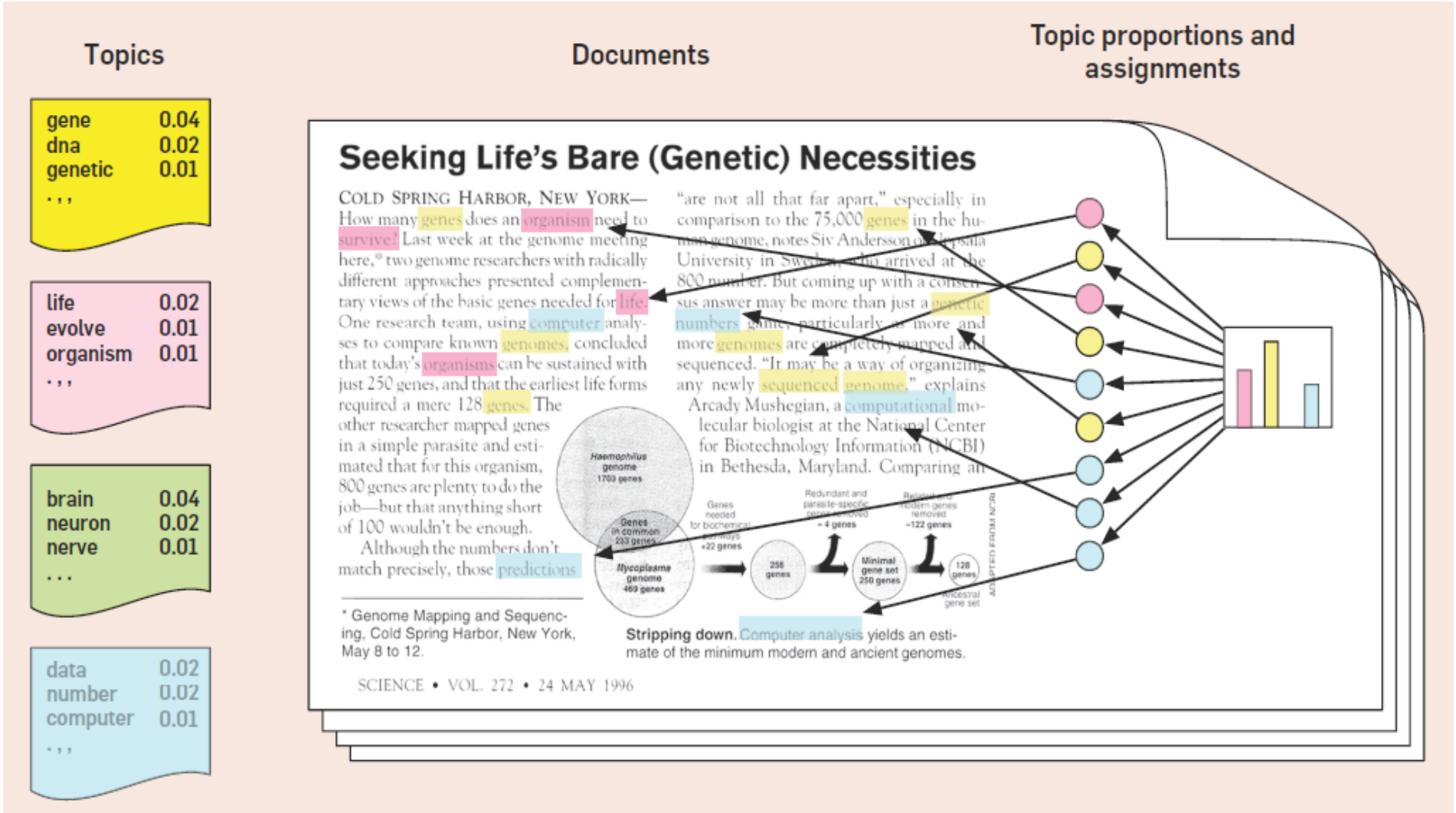
Blei (2012). 17000 artículos Science. 100 topics.

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

¿Cómo funciona?

- Para interpretar el resultado, necesario saber bien cómo funciona.
- Dos metáforas:
  - Texto subrayado.
  - Compra del mercado.

TM y el texto subrayado



Blei (2012)

TM y el mercado

- Rhody 2012

El algoritmo

- Entrada:
  - Corpus plano separado en *m* textos.
  - *n* tópicos, *x* iteraciones.
- Proceso:
  - Por cada iteración, por cada documento, por cada palabra, asigna un topic a la palabra.
    - Topics de la palabra en otros documentos.
    - Topics del documento.
  - Inicio aleatorio

Ejemplo (Mimno 2012)

etrusco	comercio	precio	templo	mercado

Ejemplo (Mimno 2012)

Documento:

3	2	1	3	1
etrusco	comercio	precio	templo	mercado

Ejemplo (Mimno 2012)

Documento:

3	2	1	3	1
etrusco	comercio	precio	templo	mercado

Corpus:

	1	2	3
etrusco	1	0	35
mercado	50	0	1
precio	42	1	0
comercio	10	8	1
...			

Ejemplo (Mimno 2012)

Documento:

3	?	1	3	1
etrusco	comercio	precio	templo	mercado

Corpus:

	1	2	3
etrusco	1	0	35
mercado	50	0	1
precio	42	1	0
comercio	10	7	1
...			

Ejemplo (Mimno 2012)

Documento:

3	?	1	3	1
etrusco	comercio	precio	templo	mercado

Seleccionar nuevo t*opic* a "comercio" en este documento:

1. Resto de *topics* en el documento
   T1 = 2 veces; T3= 2 veces.
2. Resto de *topics* para "comercio"
   T1 = 10 veces; T2 = 7 veces; T3 = 1 vez

Corpus:

	1	2	3
comercio	10	7	1

Ejemplo (Mimno 2012)

Documento:

3	?	1	3	1
etrusco	comercio	precio	templo	mercado

Asigna *topic*:

T1 = 2 (documento) * 10 (palabra) = 20
T2 = 0 (documento) * 7 (palabra) = 0
T3 = 2 (documento) * 1 (palabra) = 2

Corpus:

	1	2	3
comercio	10	7	1

Ejemplo (Mimno 2012)

Documento:

3	1	1	3	1
etrusco	comercio	precio	templo	mercado

Corpus:

	1	2	3
etrusco	1	0	35
mercado	50	0	1
precio	42	1	0
comercio	**11**	**7**	1
...			

Resultado

- La probabilidad de pertenencia de cada palabra de cada documento a cada *topic*.
- Las palabras con más peso en cada *topic:*

Los resultados no son ni fáciles ni evidentes de interpretar.

Resultado

A. Machado. Poesía completa

0	2,5	¡oh hay siempre agua canta visto he suena dijo ¡qué cantar ¡y tienen buena hora ve otro flores cuatro
1	2,5	alma caminos españa buen eres guerra paz melancolía fuerte ¿quién fue castilla ríos quiere caballero guarda rincón manchego mujeres
2	2,5	sol viejo sueño jardín agua sed pasa ceño brilla sombras guiomar noble (a primavera tristeza memoria arco arena boca
3	2,5	sombra luna noche sueños yo voz vino dulce humilde estrellas fiesta espada calle quimera señora florida fiebre soy fantasma
4	2,5	ser sino mundo poeta sí pensamiento pensar conciencia otro puede pretende cada metafísica propia sea lógica real decir lírica
5	2,5	verde río flor alto sierra azul campo vi monte piedra nadie nube santa vii tren encina roca romero encinar
6	2,5	poeta mairena tiempo ha arte recuerdo verso barroco rima imágenes conceptos función espíritu tanto estrofa español queda artista versos
7	2,5	tierra campo hacia campos duero montes cielo verdes oro nieve soria luz ramas fría río castilla álamos sol grises
8	2,5	fuente agua tarde clara vieja piedra yo triste aire verano pena cristal alegría historia fue amores labios silencio copla
9	2,5	día vida tiempo don plaza olivares maestro mal lluvia toda mil infantil paso mancha negros cuerpo tic-tic divino quisiera
10	2,5	corazón mar dios ha señor dice niño hizo aguarda hace espera fe esperanza ilusión cabeza mares vida ¡ay gota

Problemas

- No siempre es posible asignar un nombre único a un *topic*.
- Mala calidad de los *topics*:
  - Palabras intrusas (no relacionadas)
  - Palabras aburridas (genéricas)
  - ...

Soluciones

- Realizar diversos experimentos, con diferentes configuraciones, hasta hallar los mejores *topics* posibles:
  - Número de *topics*.
  - Tamaño de los documentos (contexto).
  - Número de iteraciones
  - Tamaño del corpus
  - Preprocesos: filtro stopwords, lematización, etc.
  - ...

Aspectos prácticos

- Para hacer Topic Modeling necesitamos:
  - Un corpus amplio
  - Software:
    - MALLET
    - Standford Topic Modeling Toolbox

Instalar MALLET

- Descargar: http://mallet.cs.umass.edu/download.php
- Descomprimir en c:\
  - c:\mallet-2.0.7\bin
- Si fuera necesario, cambiar las *Environment Variables* (ver enlace):
  - Detalles:

http://programminghistorian.org/lessons/topic-modeling-and-mallet

Ejecutar MALLET

- Dos pasos:
  - Cargar corpus (import-file)
  - Entrenar tópicos (train-topics)

Ejecutar MALLET

- MALLET no dispone de interfaz gráfica.
  - Línea de comandos.
- Inicio > cmd
  - Para cambiar de directorio: `cd`
    - `cd ..` → sube un nivel
  - Para ver los ficheros de un directorio: `dir`

Paso 1 – Cargas corpus

- Ponemos el corpus en una carpeta dentro de la carpeta \mallet-2.0.7.
- Mediante línea de comandos nos situamos dentro de la carpeta \mallet-2.0.7
  - Ahí ejecutamos el programa -->

Paso 1 – Cargar corpus

Llama al programa

```
bin\mallet import-file
 --input misdatos\micorpus
 --output misdatos\resultados\resultados.mallet
--keep-sequence
--remove-stopwords
   --stoplist-file stoplists\es.txt
--encoding UTF-8
   --token-regex '[\p{L}\p{M}]+'
```

Orden para cargar el corpus

Ruta al corpus

Ruta donde guardar el resultado

Para eliminar stopwords (opcional)

Expresión regular para que separe bien las palabras acentuadas del español

Paso 1 – Cargar corpus

- Se genera un fichero nuevo llamado

  xxx.mallet.
- Este fichero es la matriz del corpus. Necesario para el siguientes paso...

Paso 2 - Extraer *topics*

Llama al programa

```
bin\mallet train-topics
--input  misdatos\resultados\resultados.mallet
--num-topics 20
--output-topic-keys
misdatos\resultados\mistopics_keys.txt
```

Orden para extraer Topics

Ruta al fichero mallet

Cantidad de Topics

Agrupación de palabras por Topics (fin).
Orden + ruta donde guarda el fichero con los topics.

Resultado

- El fichero xxx_keys.txt contiene las palabras más frecuentes de cada *topic*.

Más opciones

- bin/mallet import-file –help
- bin/mallet train-topics –help

- Explicación sencilla de todo el proceso:

http://programminghistorian.org/lessons/topic-modeling-and-mallet

Actividad

- Analizar con MALLET los sonetos de Garcilaso de la Vega...

Bibliografía citada

- Blei, D.M., 2012. Probabilistic Topic Models. Communications of the ACM, 55(4), pp.77–84.
- Brett, M.R., 2012. Topic Modeling: A Basic Introduction. Journal of Digital Humanities, 2(1).
- Jockers, M.L. & Mimno, D., 2014. Significant Themes in 19th-Century Literature. Poetics.
- Mimno, D., 2012. The details: Training and Validating Big Models on Big Data. Journal of Digital Humanities, 2(1).
- Rhody, L.M., 2012. Topic Modeling and Figurative Language. Journal of Digital Humanities, 2(1).