



Universitat d'Alacant
Universidad de Alicante

maESL
Máster Universitario en Estudios Literarios

Tema 2

Un nuevo enfoque: la literatura “desde lejos”

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

- + Nuevo enfoque investigador: cambio de escala.
- + Nuevo marco metodológico en los estudios literarios:
 - La “periferia contextual” de García Berrio
 - La “literatura vista desde lejos” de F. Moretti.
 - El “macroanálisis” de M. Jockers.
- + Visión general de los métodos computacionales.
- + Ejemplo: *culturomics*.
- + Conclusiones.

Aproximación histórica

- 1ª Etapa: 1940-1990.
 - Aparición y desarrollo del ordenador, en especial del ordenador personal.
 - Digitalización de textos.
- 2ª Etapa: 1990-2014.
 - Aparición y desarrollo de la web y consolidación de internet.
 - Análisis computacional del texto digital.

Situación actual (2004-2014)

- Novedades en los estudios literarios:
 - Nuevo marco de análisis literario: la literatura “desde lejos” o macroanálisis.
 - Aplicación de métodos computacionales avanzados al análisis de amplias colecciones de texto literario:
 - *Text Mining, Data Mining, Topic Modeling, Machine Learning, ...*
 - Ejemplo: Michel et al (2011).
 - Lenguaje de marcado (TEI) y anotación masiva de corpus.
 - Lingüística Computacional y Procesamiento del Lenguaje Natural.

Cambio de escala

Fenómeno común en otras áreas de conocimiento:
Big Data.

- Física, economía, matemáticas, politología, sociología, lingüística, etc.
- Acceso a cantidad masiva de datos producidos por y sobre personas y cosas:
 - Secuencias genéticas, interacción redes sociales, archivos médicos, informes gubernamentales, registros telefónicos, etc.
 - Soporte digital: bases de datos y textos.

Boyd & Crawford (2012)

Aplicación humanidades

Cambio de escala en estudios literarios:

- Análisis de la periferia contextual (García Berrio 2009)
- *Distant Reading* (Moretti 2007)
- Macroanálisis (M. Jockers 2013, 2014)

Periferia contextual

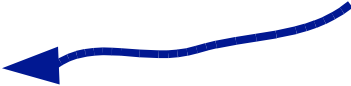
“creo que el único camino útil y razonable para construir la **periferia contextual de la cultura literaria** de Garcilaso, o de cualquier otro de nuestros autores clásicos, hay que esperarlo a partir del **tratamiento informático masivo** sobre las canteras de cultura clásica y clasicista, para configurar con todo ello **parámetros de interpretación global** sobre el comportamiento individual o social en la cultura.”

García Berrio 2009, p. 230

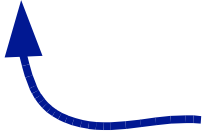
Distant reading (Moretti 2007)

- Buscar lo común en la historia de la literatura, no lo excepcional.
- Analizar las obras de grandes periodos como un todo.
- Métodos cuantitativos:
 - Obtención de datos.
 - Análisis e interpretación de datos.

Los datos plantean los problemas



*Responded a la pregunta:
¿Por qué?*



Distant reading

- Estudio de la novela inglesa ss. XVIII-XIX.
- Datos: fechas de publicación, género novelístico.
- Patrón: cambio de género cada 20 años.
 - Un género no se desarrolla hasta que se agota el anterior (ciclos).
 - La obra canónica de cada género se publica en la etapa anterior.
- Interpretación de los datos: cambios en el gusto lector: ciclos generacionales.

Distant reading

Conclusiones:

- "Los datos cuantitativos son útiles porque no dependen de las interpretaciones de los investigadores individuales"
- (Los datos cuantitativos) "son interesantes porque a veces exigen una interpretación que se sitúa fuera del universo cuantitativo"
- "de modo más radical, observamos que" (los datos cuantitativos) "falsan las teorías existentes y sugieren la necesidad de una teoría, no tanto de la novela, cuando de toda una familia de formas novelescas"

Distant reading

“Problemas carentes de solución es exactamente lo que necesita la historia literaria, donde planteamos sólo las preguntas de respuestas ya conocidas y, por lo tanto, nunca nos enfrentamos con los límites y las lagunas de nuestro conocimiento.”

Moretti 2007, p. 45

Macroanálisis

- Jocker & Mimno 2013:
 - Aplicación de métodos computacionales basados en *Text Mining: Topic Modeling*.
 - Corpus: 3346 novelas del siglo XIX (inglesa, irlandesa y norteamericana).
 - Extracción de tópicos recurrentes.
 - Distribución de tópicos por sexo del autor: femenino, masculino, desconocido.

Table 1

Twenty-five useful features in distinguishing between two classes.

	Label	Male-authors	Female-authors
1	Female fashion	-0.2015	0.2614
2	Flowers and natural beauty	-0.1698	0.2203
3	Tears and sorrow	-0.1619	0.2101
4	Drawing rooms	-0.16	0.2076
5	Drink as in liquor and beer and tobacco	0.1489	-0.1932
6	Governesses and education of children	-0.1469	0.1906
7	Nurses for children	-0.1467	0.1904
8	Pistols and other guns	0.144	-0.1869
9	Children girls	-0.1374	0.1783
10	Pity	-0.1341	0.174
11	Children	-0.1333	0.173
12	Facial features	-0.1324	0.1719
13	Affection	-0.132	0.1712
14	Health and illness	-0.1314	0.1705
15	Landlords	-0.1301	0.1688
16	Men with guns	0.1298	-0.1684
17	Moments of confusion in battle	0.1292	-0.1677
18	Grief and sorrow	-0.1269	0.1646
19	Happiness	-0.1253	0.1627
20	Afternoon and tea time	-0.1243	0.1613
21	Swords and weapons	0.1241	-0.161
22	Male clothing	0.1234	-0.1601
23	Tea and coffee	-0.1232	0.1599
24	Soldiers	0.121	-0.157
25	Dear girls children creatures	-0.1198	0.1554

Métodos computacionales

Roe (2012) diferencia:

- Métodos clásicos
- Métodos modernos

Métodos computacionales

Roe (2012) diferencia:

- Métodos clásicos:
 - Frecuencia de palabras y n-gramas.
 - Concordancia y colocaciones.
- Métodos modernos

Métodos computacionales

Roe (2012) diferencia:

- Métodos clásicos
- Métodos modernos:
 - *Data Mining y Machine Learning.*
- Más:
 - Anotación de corpus con técnicas de Lingüística Computacional.

Métodos clásicos

- Ventajas:
 - Transparentes.
 - Robustos: diferentes idiomas.
- Problema:
 - Uso limitado en amplios corpus.
 - Complejo inferir generalizaciones.

Métodos modernos

- Ventaja:
 - Tratamiento de amplias colecciones de texto.
 - Extracción de generalizaciones.
- Desventaja:
 - Opacos.
 - Suposiciones previas no siempre correctas.

Métodos modernos

Tipos:

– Supervisados:

- Necesidad de un corpus anotado a mano.
- Ej.- Métodos bayesianos, árboles de decisión, etc.

– No supervisados:

- Actúan sobre texto “plano”.
- Ej.- modelos vectoriales, *topic modeling*.

Data Mining

“Data mining is the **extraction** of **implicit, previously unknown, and potentially useful information** from data.

The idea is to build computer programs that sift through databases **automatically, seeking regularities or patterns**. Strong patterns, if found, will likely generalize to make accurate predictions on future data.”

Witten & Frank (2005)

Text ~~Data~~ Mining

Text

“~~Data~~ mining is the **extraction** of **implicit, previously unknown, and potentially useful information** from ~~data~~. *texts*”

The idea is to build computer programs that sift through ~~databases~~ **automatically, seeking regularities or patterns**. Strong patterns, if found, will likely generalize to make accurate predictions on future data.”

text corpus

Witten & Frank (2005)

Data Mining

“Of course, there will be **problems. Many patterns will be banal and uninteresting** (...) spurious, contingent on accidental coincidences...

Real data is imperfect: Some parts will be garbled, and some will be missing. Anything discovered will be inexact: There will be exceptions to every rule and cases not covered by any rule.

Algorithms need to be robust enough **to cope with imperfect data and to extract regularities that are inexact but useful.**”

Witten & Frank (2005)

Machine learning

“Machine learning provides the **technical basis of data mining**. It is used to extract information from the raw data...

The process is one of abstraction: taking the data, warts and all, and inferring whatever structure underlies it.”

Witten & Frank (2005)

Data Mining

- Métodos computacionales capaces de procesar gran cantidad de textos y extraer de ellos información útil en forma de patrones y regularidades mediante un proceso de abstracción.
- Esta información puede ser predictiva o descriptiva.
- No son infalibles: siempre presentarán un porcentaje de error.
- Los datos finales deben ser interpretados por expertos.

Conclusiones

- Cambio de escala.
 - Nuevos métodos, nuevas cuestiones, nuevas descripciones.
 - Nuevos instrumentos: métodos computacionales.
- Complementario métodos tradicionales.
- Próximos temas:
 - Diseño y compilación de corpus
 - Herramientas de análisis computacional.

Ejemplo

Jean-Baptiste Michel, et al. (2011) "Quantitative Analysis of Culture Using Millions of Digitized Books" *Science* 331, 176

- Aspectos estructurales. Objetivos y uso de gráficos.
- Descripción del corpus y metadatos.
- Interpretación de los datos.
- Se pueden replicar sus datos en Google Ngrams:
<https://books.google.com/ngrams>
 - Obtención de datos | Análisis de datos (¿por qué?)

Bibliografía citada

- Boyd & Crawford (2012) "Critical Questions for Big Data", *Information, Communication & Society*, 15:5, 662-679.
- García Berrio, A, (2009) "Retórica figural. Esquemas argumentativos en los sonetos de Garcilaso". *El centro en lo múltiple (selección de ensayos) II. El contenido de las formas (1985-2005)*. Barcelona, Anthropos, pp. 228–240.
- Jockers, M.L. (2013) *Macroanalysis. Digital Media and Literary History*, Illinois, University of Illinois Press.
- Jockers, M.L., (2014) *Text Analysis with R for Students of Literature*,: Springer
<http://link.springer.com/10.1007/978-3-319-03164-4>
- Jockers and Mimno (2013) "Significant themes in 19th-century literature" *Poetics* 41.
- Michel, et al. (2011) "Quantitative Analysis of Culture Using Millions of Digitized Books" *Science* 331, 176.
- Moretti, F. (2007) *La literatura vista desde lejos*, Barcelona: Marbot ediciones.
- Roe (2012) "The Dangers and Delights of Data Mining". In *Digital Humanities Oxford Summer School*.
- Witten, I.H. & Frank, E., 2005. *Data Mining. Practical Machine Learning Tools and Techniques* 2nd ed., Amsterdam, etc.: Elsevier.