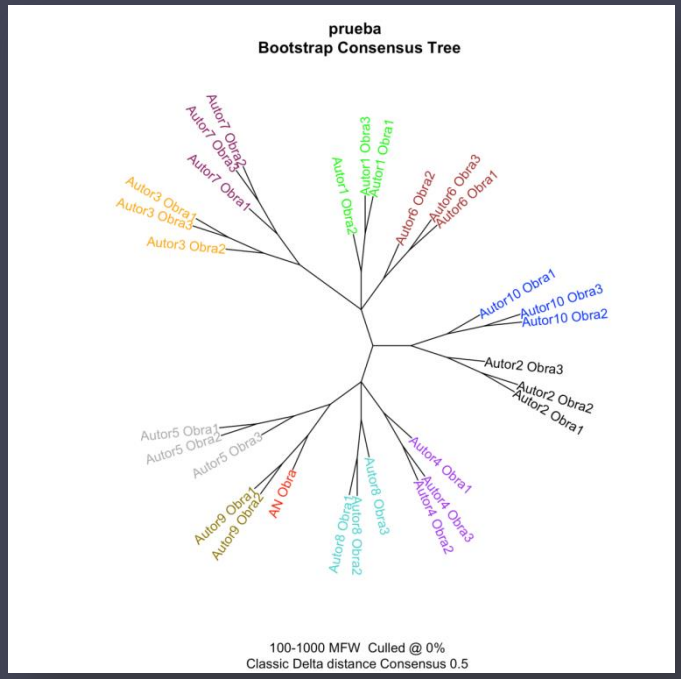
Introducción a la Estilometría



Laura Hernández Lorenzo

Humanidades Digitales. Del corpus a la interpretación: Estilometría con R (Universidad de Burgos)

Contenidos

- ¿Qué es la Estilometría?
- La huella autorial y atribución de autoría no tradicional
- ¿Qué necesito para hacer Estilometría?
- Medidas estilométricas
- Cuestiones importantes: tamaño, limpieza, corpus
- Métodos: análisis de grupos y árbol de consenso
- Práctica 1

¿Qué es la Estilometría?

- **Stylometry**
  - Style + metrics = estilo + medidas (estadísticas)

¿Qué es la Estilometría?

- Estilometría / Estilística computacional
- SIG ADHO Digital Literary Stylistics ("Estilística literaria digital")
- Otros usos de la palabra Estilometría en la bibliografía en español
  - Estudios sobre la frecuencia de hápax, n-gramas, cuantificación de determinadas palabras, de la longitud de estas y hasta en un estudio de variación dialectal

¿Qué es la Estilometría?

- Se incluye dentro de las disciplinas de tecnologías de análisis textual (Text Analysis), estudios cuantitativos (Moretti 2013, Jockers 2013) y minería de textos (Jockers & Underwood, 2016).

- Atribución de autoría, análisis de otras cuestiones estilísticas (géneros textuales, movimientos literarios, cuestiones influidas por el género masculino o femenino del autor, evolución del estilo en la obra de un autor…)

¿Qué es la Estilometría?

- En un sentido más amplio, también se incluyen dentro de la Estilometría:
  - *Topic Modelling*
  - Análisis de sentimiento (*Sentiment Analysis*)
  - Aplicación de herramientas de PLN a textos literarios

- En este curso, nos centramos fundamentalmente en atribución de autoría y, hasta cierto punto, en el análisis de otras cuestiones estilísticas, como géneros textuales (Estilometría más allá de la autoría).

Laura Hernández
Lorenzo

La huella autorial

La huella autorial

- Constituida por aquellos rasgos lingüísticos y de estilo que son inconscientes, difíciles de detectar con el ojo humano → más difíciles de imitar o plagiar

- Estos se pueden medir de forma estadística → uso de las palabras-función

Realizamos un pequeño experimento:

- ¿Cuántas *ſ* hay en el siguiente texto?
- ¡Tienes 7 segundos!

Tomado del documental de Mike Kestemont sobre Hildegarda de Bingen desde la Estilometría: https://vimeo.com/70881172

Finished files are the result of years of scientific study combined with the experience of many years.

Realizamos un pequeño experimento:

- ¿Cuántas has contado?
  - La mayoría de las personas cuentan 3 o 4
  - Se tiende a obviar la f en la palabra “of”

- ¡¡Hay un total de 6!!

Las palabras-función (*function words*)

En un lugar **de la** Mancha, **de cuyo** nombre no quiero acordarme, no ha mucho tiempo **que** vivía **un** hidalgo **de los de** lanza **en** astillero, adarga antigua, rocín flaco **y** galgo corredor. [...] Frisaba **la** edad **de nuestro** hidalgo **con los** cincuenta años; era **de** complexión recia, seco **de** carnes, enjuto **de** rostro, gran madrugador **y** amigo **de la** caza. Quieren decir **que** tenía **el** sobrenombre **de** Quijada, **o** Quesada, **que en** esto hay alguna diferencia **en los** autores **que deste** caso escriben; **aunque por** conjeturas verosímiles **se** deja entender **que se** llamaba Quijana. **Pero** esto importa poco **a nuestro** cuento: basta **que en la** narración **dél** no **se** salga **un** punto **de la** verdad.

Laura Hernández Lorenzo

Quijote / Pride and prejudice

1	que	20.369	1	the	4.322
2	de	17.914	2	to	4.122
3	y	17.884	3	of	3.619
4	la	10.204	4	and	3.519
5	a	9.667	5	her	2.196
6	el	8.043	6	i	2.002
7	en	8.036	7	a	1.926
8	no	5.988	8	in	1.868
9	los	4.692	9	was	1.843
10	se	4.644	10	she	1.674
11	con	4.122	11	that	1.534
12	por	3.811	12	it	1.512
13	las	3.422	13	not	1.417
14	lo	3.404	14	he	1.312
15	le	3.367	15	you	1.303

La huella autorial

- La huella autorial se mide a través de la frecuencia de las palabras más frecuentes (en gran medida, palabras-función)

Atribución de autoría no tradicional

- Texto de autoría dudosa o anónimo
- Textos de otros autores conocidos (lo más cercanos posibles al texto anónimo e incluyen los principales candidatos)
- Objetivo: encontrar el autor más cercano al texto anónimo

¿Qué necesito para hacer Estilometría?

- **Textos en un formato electrónico adecuado**
- Medidas estilométricas
- Métodos para aplicar esas medidas o visualizar los resultados
- Parámetros
  - Elemento a usar: palabras, letras, anotación (POS, métrica…)
  - Número de elementos. Ej.: 1.000 MFW (palabras más frecuentes)

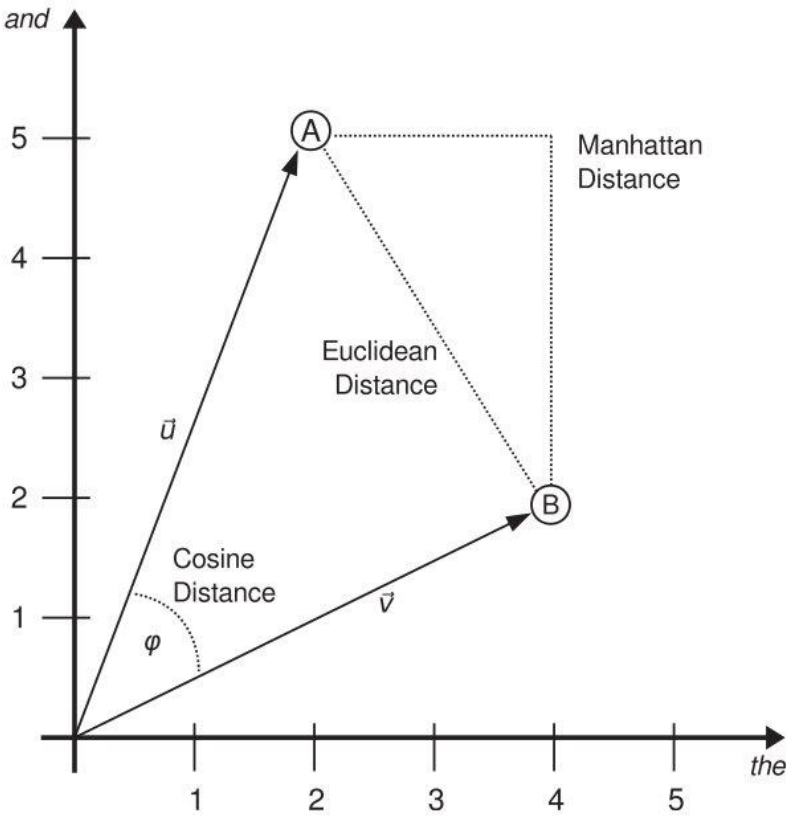
¿Qué necesito para hacer Estilometría?

- Herramienta que lleve a cabo el análisis
  - Programación: R, Python
  - JGAAP (Juola 2005) → software comercial
  - Paquete *stylo* en R (Eder, Rybicki, Kestemont, 2016)

Medidas

- Manhattan
- Euclídea
- Cosine

Representación de la distancia entre los puntos 'and' y 'the' en dos textos A y B en un espacio bidimensional. La distancia entre dos puntos puede definirse de distintas formas. Imagen tomada de "Understanding and explaining Delta measures for authorship attribution" (Evert et al., 2017).



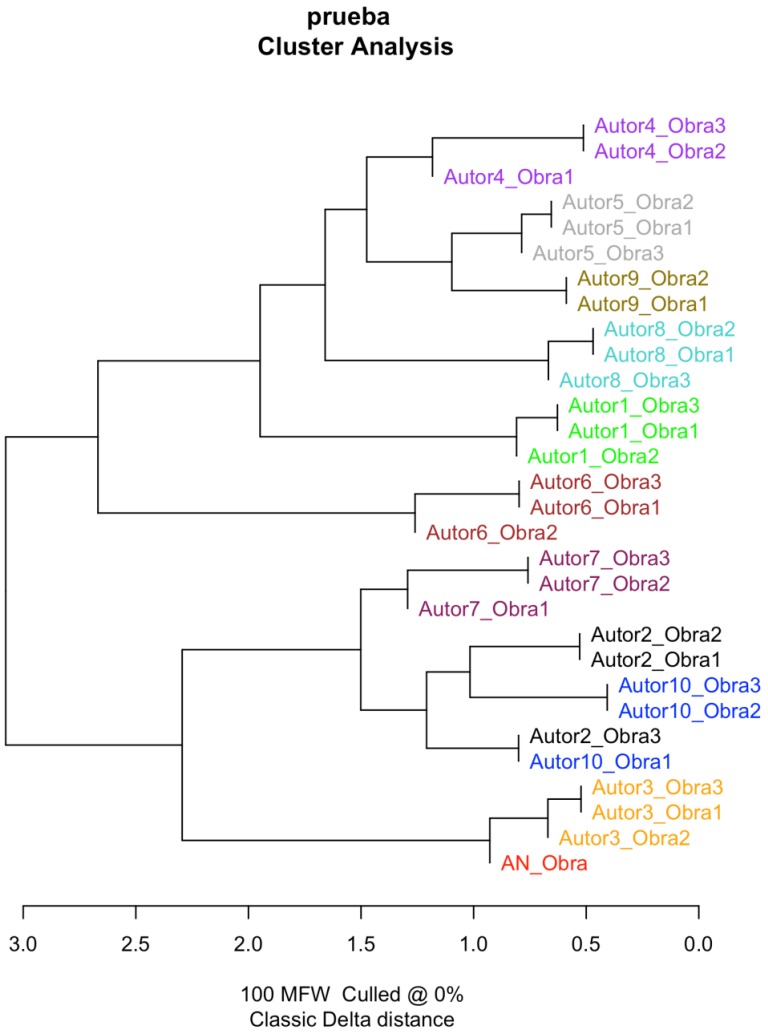
Medidas estilométricas

- Delta de Burrows (Burrows, 2002)
- Cosine Delta (Evert et al., 2017)
- Delta de Eder (Eder, 2013)
- Minmax (Kestemont et al., 2016)

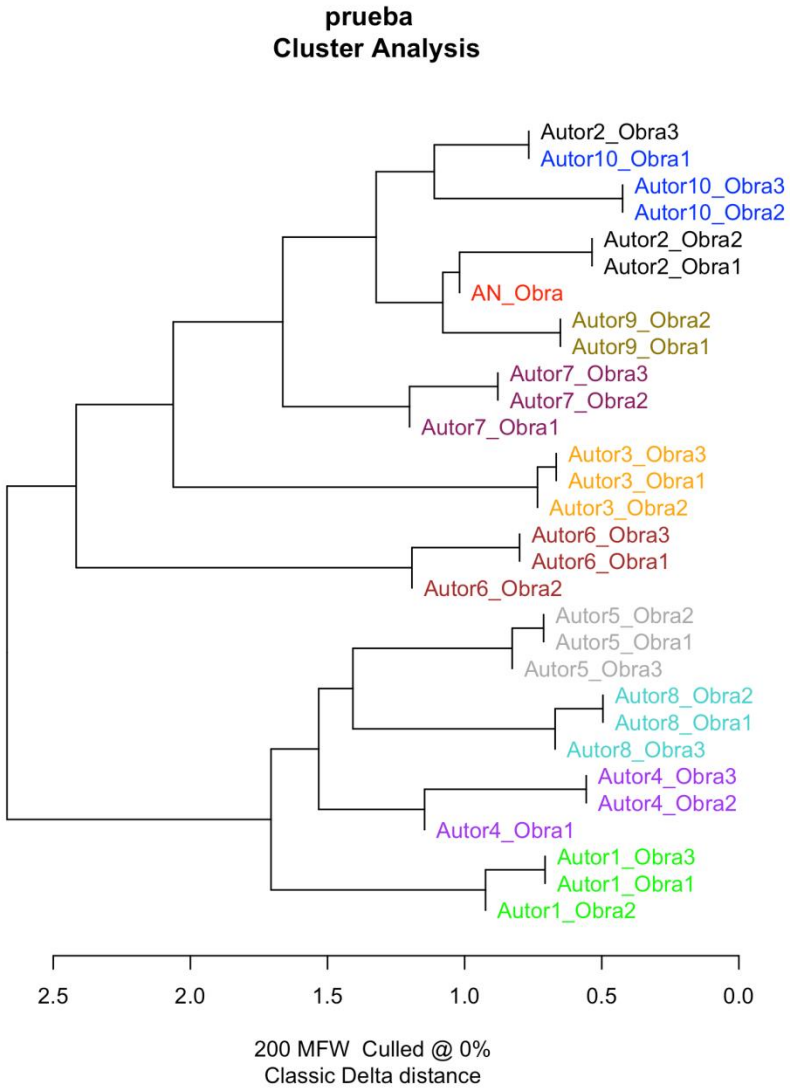
Métodos: Análisis de grupos o Cluster Analysis

Laura Hernández Lorenzo

- Los textos se agrupan en función de su similitud, proximidad o cercanía
- El dendograma se lee de derecha a izquierda → los textos más cercanos, o entre los que hay una distancia menor, se agruparán más a la derecha.
- Problema: ¿cómo elegir los parámetros?



Métodos: Análisis de grupos o *Cluster Analysis*



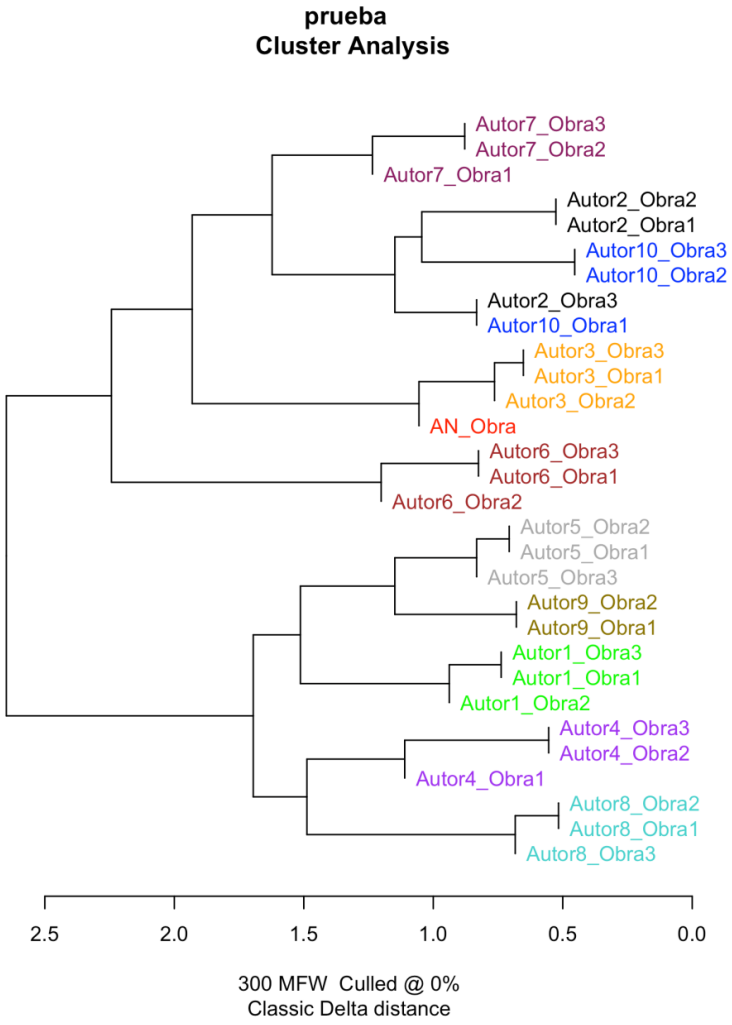
Laura Hernández Lorenzo

Métodos: Análisis de grupos o Cluster Analysis

Laura Hernández Lorenzo

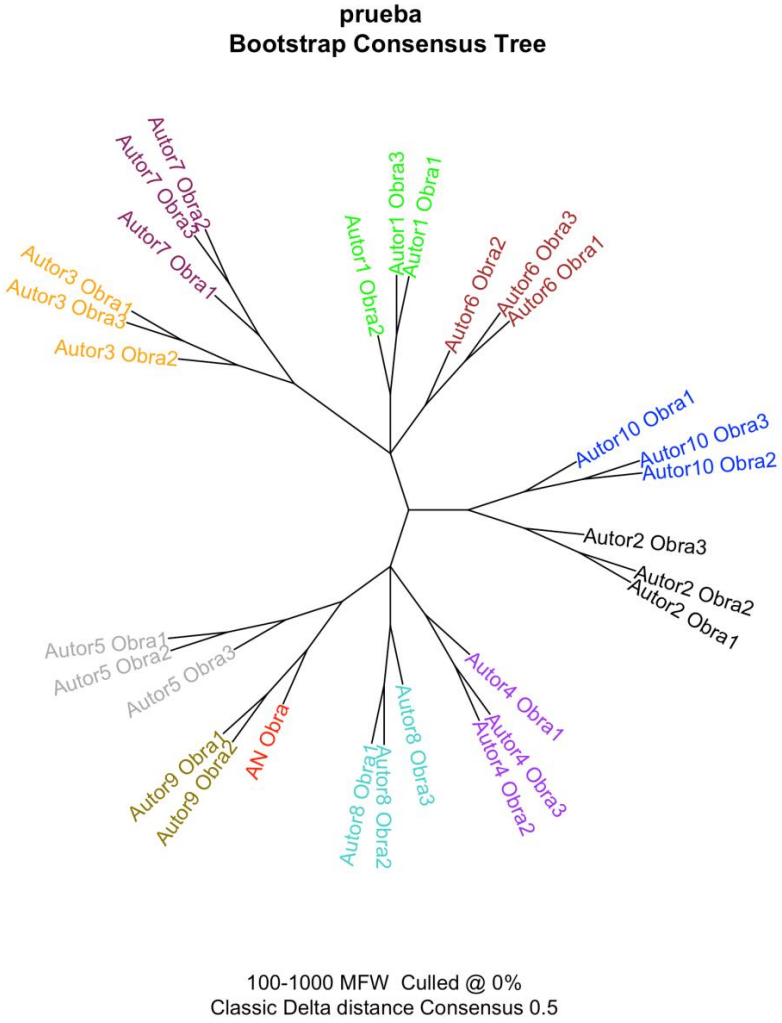
- Los dendogramas cambian según el número de palabras más frecuentes elegido

- ¿Cuál es el correcto?



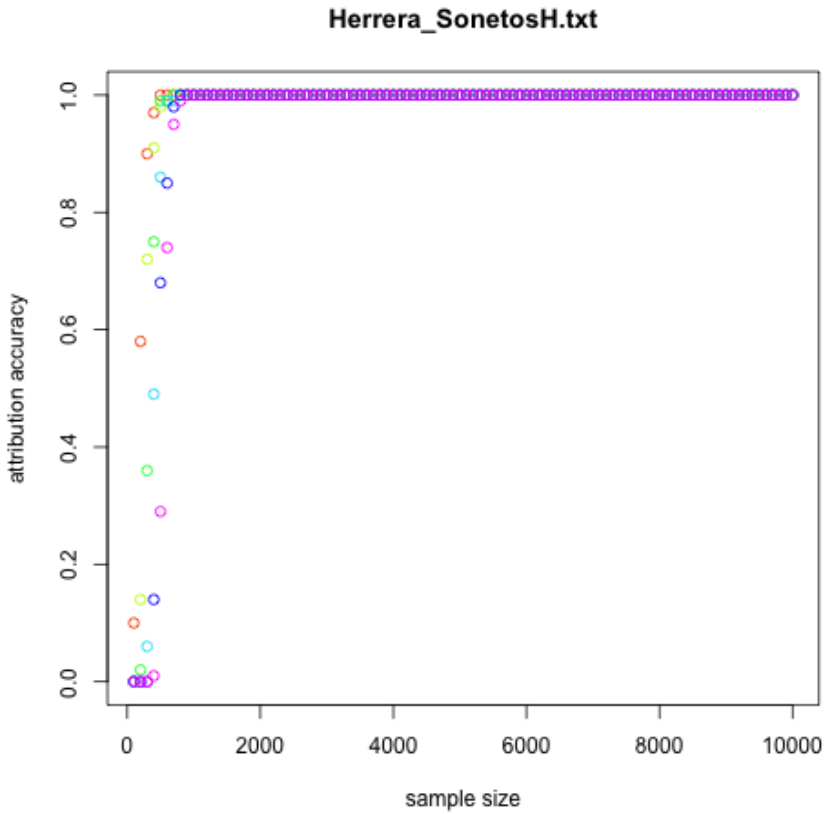
Métodos: Árbol de consenso o *Consensus Tree*

- Método desarrollado por Eder (2013)
- Superar el problema de *cherry-picking*
- Quedan recogidas únicamente las relaciones más estables entre los textos del corpus → las que permanecen a lo largo de un número mínimo de iteraciones o valor de consenso



Cuestiones importantes

- Tamaño de los textos
  - Estudios
    - Eder, 2015, 2017
    - Hernández-Lorenzo, 2019
  - Al menos, cerca de 2.000 palabras



Cuestiones importantes

- Limpieza de los textos
  - Eder, 2013

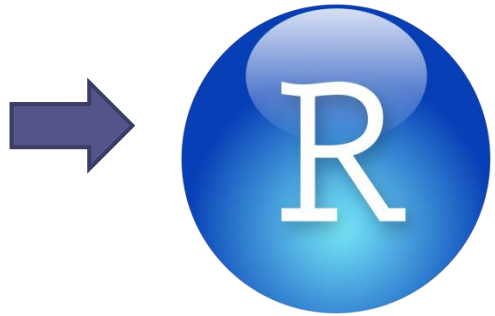
- Para atribución de autoría → textos más similares posibles, misma época, mismo género literario…

Laura Hernández
Lorenzo

Práctica 1: tu primer análisis estilométrico

(análisis de grupos y árbol de consenso)

¡Manos a la obra!

- Localiza en tu ordenador el corpus de novela española de CLiGS (Calvo Tello, 2017)
- Arranca R Studio 
- Si no lo tienes, instala *stylo*

```
install.packages("stylo")
```

Laura Hernández Lorenzo

Fíjate en que para stylo()…

- Tus textos deben estar dentro de una carpeta que se llame “corpus”
- Los nombres de los textos deben ser así:
  - Autor_obra.txt
- Los ficheros resultantes del análisis, se crearán en la carpeta matriz (tabla, imagen con figura, archivo con configuración del análisis)