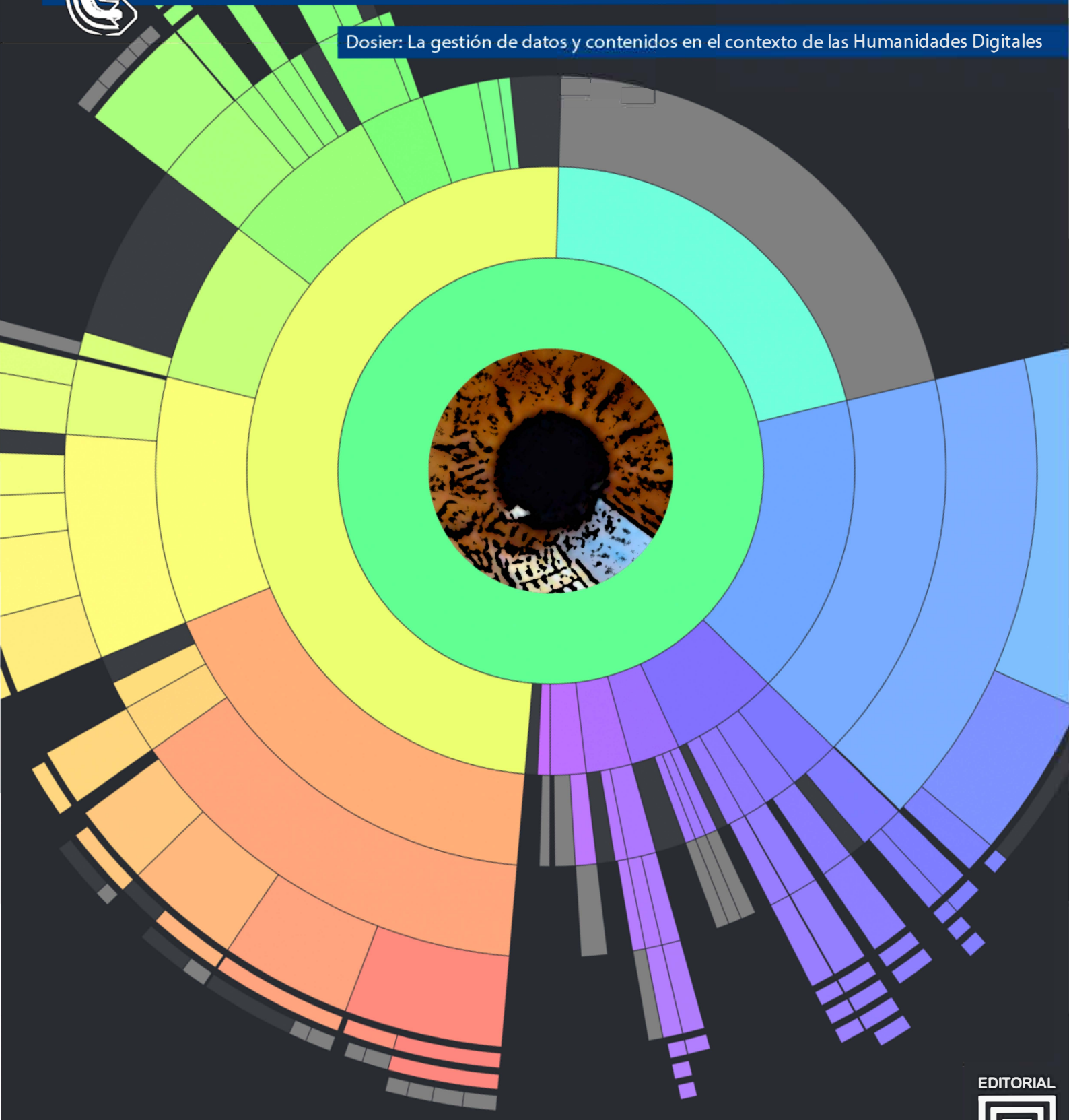
ISSN: 2254-4496 / Vol. 7 Nº 2 / Noviembre de 2018

CARACTERES

Estudios culturales y críticos de la esfera digital

En este número participan María Álvarez, Pablo Brescia, Rafael Cano Tenorio, Álvaro Cuéllar González, María Isabel Escalas Ruiz, Felipe Fernández García, Myriam Ferreira-Fernández, Candelas Gala, Carmen Gaona-Pisonero, Daniel Herrera Arenas, Anais Holgado Lage, Elena Martínez Carro, Eva Martínez Díaz, David Olay Varillas, Lucas Ramada Prieto, Alexandra María Sandulescu Budea, Alex Saum-Pascual, Ana Sedeño-Valdellós, Olga Taravilla Baquero, Andrea Mediana Téllez Girón, Alberto Venegas Ramos

Dosier: La gestión de datos y contenidos en el contexto de las Humanidades Digitales



EDITORIAL DELIRIO

Caracteres. Estudios culturales y críticos de la esfera digital

Caracteres es una revista académica interdisciplinar y plurilingüe orientada al análisis crítico de la cultura, el pensamiento y la sociedad de la esfera digital. Esta publicación prestará especial atención a las colaboraciones que aporten nuevas perspectivas sobre los ámbitos de estudio que cubre, dentro del espacio de las Humanidades Digitales. Puede consultar las normas de publicación en la web (http://revistacaracteres.net/normativa/).

ISSN: 2254-4496

Editorial Delirio (www.delirio.es)

Editorial, PÁG. 6

Artículos de investigación

- Una aproximación a la fisionomía del mundo textual: texto, referente y lector. DE ANDREA MEDINA TÉLLEZ GIRÓN, PÁG. 13
- La problemática de la imagen como forma de transmisión histórica en la cultura digital. DE ALBERTO VENEGAS RAMOS, PÁG. 36
- Análisis del nivel de emisión de contenido en las cuentas oficiales de Twitter de los artistas flamencos. DE RAFAEL CANO TENORIO, PÁG. 57
- ¿Es YouTube una nueva esfera digital para leyendas urbanas? La representación de la infancia perturbadora a través del fenómeno cultural de los *Black-Eyed-Kids* (BEKS). DE MARÍA ISABEL ESCALAS RUIZ, PÁG. 88
- Paisaje-cuerpo-música en el videoclip musical y el videoarte actuales. DE CARMEN GAONA-PISONERO Y ANA SEDEÑO-VALDELLÓS, PÁG. 110
- El *souvenir* digital, la memoria en la nube. DE OLGA TARAVILLA BAQUERO, PÁG. 139

Reseñas

- *El español en la red*, de Mabel Giammateo, Patricia Gubitosi y Alejandro Parini (eds.). POR ANAIS HOLGADO LAGE PÁG. 150
- *Narrativas mutantes: anomalía viral en los genes de la ficción*, de Mihai Iacob y Adolfo R. Posada (eds.). POR ALEX SAUM-PASCUAL, PÁG. 163
- *Video Games As Culture. Considering the Role and Importance of Video Games in Contemporary Society,* de Daniel Muriel y Garry Grawford. POR LUCAS RAMADA PRIETO, PÁG. 169
- *Elogio de lo mínimo. Estudios sobre microrrelato y minificción en el siglo XXI,* de Ana Calvo Revilla (ed.). POR PABLO BRESCIA, PÁG. 177
- *Idea súbita. Ensayos sobre epifanía creativa,* de Amelia Gamoneda y Francisco González (eds.). POR CANDELAS GALA, PÁG. 185

Dosier: La gestión de datos y contenidos en el contexto de las Humanidades Digitales

- Los desafíos digitales en el mercado de la traducción. DE MARÍA ÁLVAREZ, PÁG. 193
- La nueva modalidad discursiva en Twitter: el discurso político como ejemplo. DE EVA MARTÍNEZ DÍAZ, PÁG. 216
- Una interpretación digital de dos tragedias lorquianas: *Yerma* y *Doña Rosita la soltera*. DE ELENA MARTÍNEZ CARRO, PÁG. 240
- Cartografía de todo y para todo. La información geográfica en internet. DE DANIEL HERRERA ARENAS, DAVID OLAY VARILLAS Y FELIPE FERNÁNDEZ GARCÍA, PÁG. 268
- La necesidad de la validación cruzada en Stylo y cómo programarla en R. DE ÁLVARO CUÉLLAR GONZÁLEZ, PÁG. 301
- Reputación digital en la gestión de las redes sociales de artes escénicas. DE ALEXANDRA MARÍA SANDULESCU BUDEA, PÁG. 321
- El uso de las tecnologías digitales en los museos españoles: estado de la cuestión. DE MYRIAM FERREIRA-FERNÁNDEZ, PÁG. 343

Petición de contribuciones, PÁG. 368

DOSIER: LA GESTIÓN DE DATOS Y CONTENIDOS EN EL CONTEXTO DE LAS HUMANIDADES DIGITALES
DOSSIER: DATA AND CONTENT MANAGEMENT WITHIN THE DIGITAL HUMANITIES

Coord. María Pilar Celma Valero

LA NECESIDAD DE LA VALIDACIÓN CRUZADA EN STYLO Y CÓMO PROGRAMARLA EN R

THE NECESSITY OF CROSS-VALIDATION IN STYLO AND HOW TO PROGRAM IT IN R

ÁLVARO CUÉLLAR GONZÁLEZ
UNIVERSITY OF KENTUCKY

ARTÍCULO RECIBIDO: 09-09-2018 | ARTÍCULO ACEPTADO: 19-11-2018

RESUMEN:
La validación cruzada es una herramienta estadística utilizada para evaluar nuestro grupo de control antes de aplicarlo a nuestro grupo problemático. Este artículo reflexiona sobre la necesidad de incorporar la validación cruzada en nuestros estudios de atribución de autoría con Stylo y muestra cómo programarla en R. Así, en primer lugar, se recogen y presentan las principales carencias que se observan de forma asidua en las más recientes investigaciones. Se explica, en segundo lugar, una posibilidad de programación de la validación cruzada en el lenguaje de programación R, que aúna el código ya creado por los desarrolladores de Stylo con unas nuevas líneas que realizarán nuestro propósito.

ABSTRACT:
Cross-validation is a tool that is used to analyze our control group before applying it to our problematic group. This article reflects on the need to incorporate cross-validation into our author attribution studies with Stylo and shows how to program it in R. Thus, in the first place, we recover and present the main deficiencies that are observed assiduously in the majority of most recent studies. It is explained, secondly, the possibility of programming the cross-validation in the programming language R, which combines the code already designed by Stylo developers with new lines that do our purpose.

Caracteres. Estudios culturales y críticos de la esfera digital
Vol. 7 (2), 2018: 301-320
ISSN: 2254-4496
http://revistacaracteres.net

PALABRAS CLAVE:
Estilometría, autoría, validación cruzada, plagio, estilo

KEYWORDS:
Stylometry, authorship, cross-validation, plagiarism, style

Álvaro Cuéllar González. finalizó sus estudios en Filología Hispánica en la Universidad de Valladolid (España) en 2018 y, desde entonces, sigue un programa de doctorado en la Universidad de Kentucky (EE. UU). Su principal área de interés es la aplicación de las nuevas tecnologías a la literatura y, concretamente, al teatro del Siglo de Oro y sus numerosos problemas de autoría.

1. Introducción

Cuando abordamos el estudio de una posible atribución con la herramienta Stylo[1], nos enfrentamos, con toda probabilidad, a una serie de dudas. No sabemos que método estadístico (Delta, NSC, SVM…) es más conveniente para nuestro corpus; ignoramos que número de palabras más frecuentes (MFW) utilizar y que *culling* (porcentaje de aparición en nuestros textos) aplicar; desconocemos si para nuestro corpus concreto es mejor utilizar palabras independientes o conjuntos de ellas (bigramas, trigramas…) y a partir de qué tamaño nuestros textos empiezan a ser correctamente atribuidos.

Una posible solución a este problema es aplicar una validación cruzada, técnica del análisis estadístico que trabaja con el corpus indubitado con anterioridad al estudio del dubitado, y que permite averiguar qué parámetros son los más adecuados para nuestro tipo de textos. Por desgracia, este procedimiento no es accesible de manera sencilla con Stylo, sino que debe ser programado y sus resultados dirigidos para que puedan ser interpretados de forma correcta.

Así, este artículo reflexiona, en primer lugar, sobre la necesidad de incorporar la validación cruzada en nuestros estudios de atribución de autoría con Stylo. Para ello se enuncian las principales carencias que se observan de forma asidua en las más recientes investigaciones. Se explica, en segundo lugar, una

[1] Todo el artículo hace referencia a Stylo, paquete para R desarrollado por el *Computational Stylistics Group*, equipo formado por miembros de las universidades de Cracovia y Amberes (Maciej et al., 2016).

posibilidad de programación en R, que aúna el código ya creado por los desarrolladores de Stylo con unas nuevas líneas que realizarán nuestro propósito. Aunque esta no sea nuestra herramienta de trabajo para el estudio de atribuciones, conviene tener los preceptos que aquí se exponen siempre presentes para no incurrir en errores que arruinen nuestra credibilidad científica.

2. Principales carencias en los estudios de atribución con Stylo

En la estilometría hay mejores intenciones que procedimientos. Cada cierto tiempo aparecen nuevos artículos que utilizan la estilometría y, en especial, Stylo para intentar dar respuesta a alguno de los muchos interrogantes que plantea la historia de la literatura.

Muchos de estos estudios, aunque bienintencionados e incluso casualmente certeros, carecen del debido rigor que debe acompañar cualquier investigación científica. Así, es habitual observar una serie de carencias que merman la calidad de los resultados obtenidos.

2.1. Arbitrariedad en las opciones escogidas

Es frecuente observar presentaciones en las que se ofrece un dendrograma o un *consensus tree* sobre un problema autorial con un determinado modelo matemático (como Delta, Eder, Argamon…), unas determinadas MFW (palabras más usuales a tener en cuenta) y un determinado *culling* (la aparición de estas en

un porcentaje de los textos). ¿Por qué se han elegido, entre los miles de posibilidades a nuestra elección, esas que se nos presentan?

Pareciera, en algunas ocasiones, que el autor de la investigación ha ensayado distintas opciones hasta dar con la que mejor se ajusta a aquello que quiere demostrar. Peor incluso es quien solo utiliza unos determinados parámetros, sin justificación ninguna, y acepta los resultados. La validación cruzada, como explicaremos más adelante, puede ayudarnos a descubrir qué opciones son las más aceptables para nuestro corpus y, gracias a ello, paliar este problema recurrente.

2.2. Aplicación directa a la obra problemática

Es habitual leer artículos que abordan un problema autorial concreto. Para ello clasifican numérica o gráficamente la obra cuestionada entre varios títulos de algunos autores. El problema reside en la falta de cotejo de estos textos de autoría fiable. Reflexionemos sobre qué sentido tiene aplicar la estilometría para intentar encontrar el autor de una obra dubitada cuando ni siquiera sabemos si nuestro método funciona con obras seguras.

Resulta imprescindible asegurar que el método y los parámetros de la estilometría responden de forma adecuada con obras fiables y, solo en ese momento, nos plantearemos aplicarlo, todavía con mil precauciones, a nuestra obra dubitada.

2.3. Textos demasiado cortos

Suele ser recurrente el intento de descubrir la autoría de textos ciertamente escuetos, como poemas o cartas. ¿A partir de qué

número de palabras la estilometría funciona en unos porcentajes aceptables para nuestro tipo de corpus?

Más adelante en este artículo se propone una forma de comprobar, mediante la validación cruzada, a partir de qué tamaño nuestros textos son clasificados de forma satisfactoria.

3. La validación cruzada

Una buena solución para paliar muchos de estos problemas es recurrir a una *cross-validation* o validación cruzada. Esta es utilizada en investigaciones estadísticas de todo tipo para comprobar la fiabilidad de un grupo de prueba antes de aplicar el análisis a un grupo problemático. Se suele concebir como una forma de regular los parámetros, de ajustarlos, para luego emplearlos con el conjunto desconocido. No estaremos haciendo trampas, por tanto, cuando ajustemos las opciones para que nuestros textos indubitados se clasifiquen de forma certera; estamos averiguando qué opciones son mejores para luego poder aplicarlas, sin cambiarlas, por supuesto, a nuestro grupo dubitado.

Stylo contiene una función de cross-validation: crossv( ), que es una extensión de la función classify( ). Esta, con el argumento *leaveoneout*: 'deja uno fuera', extrae uno por uno los textos del corpus y los enfrenta al resto aplicando las opciones que elijamos. Si se acierta en la clasificación del autor (para ello es muy importante anotar con corrección el corpus con el nombre del autor antes del guion bajo), arroja un resultado de 100; si falla, un resultado de 0. Es, por tanto, la manera ideal de comprobar qué opciones (MFW, *culling*, método probabilístico…) son mejores para nuestro corpus concreto. A priori, por ejemplo, no sabemos si

es mejor utilizar las 100 palabras más frecuentes o las 5000, o si utilizar palabras que puedan no estar en todos los textos o que obligatoriamente deban aparecer en todos ellos.

El escollo con el que nos enfrentamos al intentar aplicar esta función es que no es fácilmente accesible para los neófitos en programación, no está dotada de una interfaz gráfica como stylo( ), classify( ) o rolling.delta( ), sino que es imprescindible programar varias líneas de código para conseguir los resultados, y algunas más si queremos repetir el proceso numerosas veces y que todo se almacene convenientemente.

3.1. Programación de la validación cruzada

Veamos una posibilidad de programación de la validación cruzada en R. Por supuesto, todos los códigos y procesos pertenecen a la librería Stylo y sus creadores, pero son muy poco accesibles sin una explicación adecuada. Además, es necesario programar una serie de bucles que repitan numerosas veces la validación cruzada, que es en lo que aquí se va a hacer hincapié.

Imaginemos que queremos averiguar la autoría de un texto determinado. Para ello contamos con un corpus de candidatos formado por 5 autores con 10 textos cada uno. A través de la validación cruzada comprobaremos qué opciones clasifican mejor nuestro tipo de textos, para, una vez descubiertas, aplicarlas a nuestro texto dudoso. Nos olvidamos, por tanto, de nuestra obra problemática para realizar la validación cruzada. Conviene recordar que, si los resultados no son satisfactorios para ninguna de nuestras opciones, debemos renunciar a intentar aplicar la estilometría a nuestro problema autorial hasta no contar con un corpus indubitado que funcione correctamente.

En primer lugar, como es habitual, debemos crear una carpeta *corpus* en la que deben estar estos 50 textos anotados de forma correcta, esto es, con el nombre del autor, guion bajo, y el título del texto. Es imprescindible que haya más de un texto de cada autor, puesto que el proceso que va a realizar la validación cruzada es comprobar si cada texto se clasifica con el resto de obras del autor con los parámetros que le indiquemos.

Nos dirigimos a la carpeta donde esté situado nuestro corpus y activamos la librería de Stylo.

```
library(stylo)
```

Tomamos todos los textos de la carpeta *corpus* y seleccionamos ya los grupos de palabras que deseamos (una palabra, bigramas, trigramas…) variando el número en ngram.size =.

```
texts = load.corpus.and.parse(files = "all", corpus.dir = "corpus", ngram.size = 1)
```

Ahora creamos la lista de las palabras más frecuentes en el total de nuestros textos. Stylo usa por defecto 5000, pero nosotros podemos disponer cualquier otra cifra. Más adelante elegiremos qué palabras usar de entre estas y qué *culling* aplicar. Es posible que en nuestro corpus entero no haya 5000 palabras diferentes, lo cual producirá mensajes posteriores en forma de incontables *NA, NA, NA*… en nuestra consola. No interfiere en los resultados de la prueba, pero, si se quiere evitar, basta con reducirlo hasta un número adecuado.

```
freq.list = make.frequency.list(texts, head = 5000)
```

Una vez que tenemos la lista de palabras más frecuentes, es necesario ver en qué proporción están estas palabras en cada uno de nuestros textos, para ello aplicamos:

```
word.frequencies = make.table.of.frequencies(corpus = texts, features =
freq.list)
```

Ya tenemos una lista con los textos y la proporción de las palabras más frecuentes en ellos. Podemos querer aplicar *culling*, es decir, quedarnos solo con las palabras que se encuentren como mínimo en el 10%, 20%... de los textos de nuestro corpus. Es posible que esto evite problemas en el caso de que aparezcan los mismos nombres de personajes en varias obras o estas traten argumentos similares, pues con un *culling* alto solo resisten las palabras gramaticales y las más comunes, ajenas a cualquier particularidad de nuestro texto. Para seleccionar el porcentaje de *culling* cambiamos la primera cifra de la siguiente línea, entre el 0 y el 100.

```
tabla = perform.culling(word.frequencies, 0)
```

Ahora seleccionamos las palabras más frecuentes (MFW) que queremos usar para el análisis. Por tanto, donde ahora vemos 100, escribimos el número de palabras que deseamos que se tengan en cuenta.

```
tabla1 = t(head(t(tabla), 100))
```

Aplicamos ahora la validación cruzada con el comando *leaveoneout* y el método Delta. El programa irá extrayendo una por una las obras de nuestro corpus y enfrentándolas al resto con las opciones anteriormente elegidas. Podemos seleccionar otros métodos de clasificación, como SVM, NSC, KNN, etc.

```
crossv(tabla1, cv.mode = "leaveoneout", classification.method = "delta")
```

Si quisiéramos un modelo estadístico concreto dentro de Delta, como Eder, añadimos distance = y el método entre comillas.

```
crossv(tabla1, cv.mode = "leaveoneout", classification.method = "delta",
distance = "eder")
```

Ya está configurada la validación cruzada para palabras independientes, un *culling* de 0, unas MFW de 100 y un método estadístico Eder. Esta nos devuelve resultados como los siguientes, en los que 100 significa acierto y 0 error en el orden alfabético de nuestro corpus. El programa ha ido extrayendo los textos e intentando clasificarlos, cuando ha acertado lo marca con 100 y cuando falla con 0.

```
100 100 100 100 0 100 100 0 0 0 100 100 100 0 0 100 0 100 100 100
100 100 100 100 0 100 100 100 0 100 0 100 0 100 100 100 0 0 100 100
0 100 100 100 100 100 100 100 100 100
```

Si queremos ver con claridad qué obras se han adjudicado mal y cuáles bien podemos simplemente guardar los resultados de la validación en *resultados* y hacer una sencilla matriz.

```
resultados = crossv(tabla1, cv.mode = "leaveoneout",
classification.method = "delta", distance = "eder")

matrix(resultados)
```

Lo que nos devuelve una lista ordenada de los aciertos y errores siguiendo el orden alfabético de nuestro corpus. Así, vemos que la obra quinta y octava de nuestro corpus no se han adjudicado bien con los parámetros escogidos. Basta con acudir a nuestro corpus y observar qué obras son estas.

```
[1,]100
```

```
[2,]100

[3,]100

[4,]100

[5,]0

[6,]100

[7,]100

[8,]0

...
```

Para convertir esta lista de ceros y cienes en resultados porcentuales legibles, hacemos que la función crossv( ) se guarde en *resultados*, sumamos todas sus cifras y lo dividimos entre el número de ellas que hay. Así obtenemos el porcentaje de acierto; solo resta redondearlos a dos cifras decimales o a las que deseemos con la función round( ).

```
resultados = crossv(tabla1, cv.mode = "leaveoneout", classification.method = "delta", distance = "eder")

round(sum(resultados)/length(resultados), 2)
```

La función que nos devuelve el porcentaje de acierto de nuestra validación cruzada es, por tanto:

```
library(stylo)

texts = load.corpus.and.parse(files = "all", corpus.dir = "corpus", ngram.size = 1)

freq.list = make.frequency.list(texts, head = 5000)
```

```
word.frequencies = make.table.of.frequencies(corpus = texts, features =
freq.list)

tabla = perform.culling(word.frequencies, 0)

tabla1 = t(head(t(tabla), 100))

resultados = crossv(tabla1, cv.mode = "leaveoneout",
classification.method = "delta", distance = "eder")

round(sum(resultados)/length(resultados), 2)
```

Hemos conseguido aplicar la validación cruzada con un método estadístico, un *culling* y un número de palabras más frecuentes concretos. Además, hemos utilizado palabras independientes, no bigramas, trigramas, etc. De poco nos sirve esta información si no la comparamos con los resultados que se obtienen a medida que vamos variando los parámetros. Podemos ir modificando estas opciones manualmente e ir anotando los resultados para comprobar cuáles son mejores para nuestro propósito. Otra opción, mucho más económica, es crear bucles que repetirán el proceso mientras varían los parámetros e irán registrando los resultados.

En R un bucle tiene la forma:

```
for(i in 1:5)

{

#Interior del bucle

}
```

En este ejemplo el programa realiza cinco vueltas, en la primera a *i* se le asignará el valor 1; en la segunda, 2… hasta 5. Podemos jugar con los bucles para conseguir que nuestro programa

produzca de una vez los resultados con *culling* incrementándose de 10 en 10 hasta 100, representado por la *e,* y las palabras más frecuentes utilizadas de 100 en 100 hasta 5000, representado por la *i*. Añadimos un pequeño *break* que detiene el bucle si intentamos tomar más palabras de las que tenemos disponibles, lo que produciría continuamente el mismo resultado.

```
for(e in 0:10)

{

for(i in 1:50)

{

tabla = perform.culling(word.frequencies, e*10)

tabla1 = t(head(t(tabla), i*100))

usadas = ncol(tabla)

if(i*100> disponibles)

{

break

}
```

Solo resta guardar los resultados en un archivo externo, para ello usamos el comando cat( ). La función completa que nos devuelve un archivo de texto con los resultados de la validación cruzada para palabras independientes, *culling* de 0 a 100 cada 10, MFW de 100 a 2000 cada 100 y método SVM es:

```
library(stylo)

#Leemos datos, indicamos n-grams
```

```
texts = load.corpus.and.parse(files = "all", corpus.dir = "corpus", ngram.size = 1)

#Indicamos palabras totales a tener en cuenta

freq.list = make.frequency.list(texts, head = 5000)

word.frequencies = make.table.of.frequencies(corpus = texts, features = freq.list)

#Preparamos unos enunciados para los archivos de resultados

culling = "Culling: "

mfw = "MFW: "

acierto = "Acierto: "

palabrasdisponibles = "Palabras disponibles: "

cambio = "Culling = "

#Bucle para culling

for(e in 0:10){

#Bucle para MFW

for(i in 1:20){

#Seleccionamos culling

tabla = perform.culling(word.frequencies, e*10)

#Seleccionamos MFW

tabla1 = t(head(t(tabla), i*100))

disponibles = ncol(tabla)

if(i*100> disponibles){
```

```
break}

#Se aplica la validación cruzada

resultados = crossv(tabla1, cv.mode = "leaveoneout", classification.method = "svm")

#Se imprimen los resultados en el archivo de texto

porcentaje = round(sum(resultados)/length(resultados), 2)

cat(culling, e*10, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(mfw,i*100, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(palabrasdisponibles, disponibles, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(acierto, porcentaje, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(file = "resultados.txt", append = TRUE, fill=TRUE)

}}
```

Como se puede comprobar, solo somos capaces de variar mediante bucles las MFW y el *culling*; debemos cambiar manualmente los n-grams y el método estadístico para poder también sumarlos a nuestra comparación.

Es muy posible que queramos hacer una validación cruzada en la que lo que varíe sea la longitud de los textos de nuestro corpus. Así podemos comprobar a partir de qué cantidad de palabras es efectiva la estilometría. Para ello incluiremos al recopilar los textos el comando *sample.size*, el cual tomará el número de palabras a nuestra elección de forma aleatoria de los textos. Aquí programamos que los textos sean de un tamaño de 2000 palabras.

El resto del código no varía, utilizamos el método SVM, 0 *culling* y 1000 MFW.

```
library(stylo)

texts = load.corpus.and.parse(files = "all", corpus.dir = "corpus", ngram.size = 1, sample.size = 2000, sampling = "random.sampling", number.of.samples = 1)

freq.list = make.frequency.list(texts, head = 5000)

word.frequencies = make.table.of.frequencies(corpus = texts, features = freq.list)

tabla = perform.culling(word.frequencies, 0)

tabla1 = t(head(t(tabla), 1000))

disponibles = ncol(tabla)

resultados = crossv(tabla1, cv.mode = "leaveoneout", classification.method = "svm")

round(sum(resultados)/length(resultados), 2)
```

Para poder saber a partir de cuántas palabras funciona la estilometría necesitamos un bucle que vaya variando el tamaño de nuestros textos. Aquí programamos desde 100 hasta 10000 palabras en intervalos de 100.

```
library(stylo)

for (a in 1:100) {

texts = load.corpus.and.parse(files = "all", corpus.dir = "corpus", ngram.size = 1, sample.size = a*100, sampling = "random.sampling", number.of.samples = 1)

freq.list = make.frequency.list(texts, head = 5000)
```

```
word.frequencies = make.table.of.frequencies(corpus = texts, features = freq.list)

acierto = "Acierto: "

palabrasdisponibles = "Palabras disponibles: "

tamaño = "Tamaño = "

tabla = perform.culling(word.frequencies, 0)

tabla1 = t(head(t(tabla), 1000))

disponibles = ncol(tabla)

resultados = crossv(tabla1, cv.mode = "leaveoneout", classification.method = "svm")

porcentaje = round(sum(resultados)/length(resultados), 2)

cat(tamaño, a*100, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(palabrasdisponibles, disponibles, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(acierto, porcentaje, file = "resultados.txt", append = TRUE, fill=TRUE)

cat(file = "resultados.txt", append = TRUE, fill=TRUE)

}
```

Estas líneas nos devolverán un archivo de resultados en el que comprobaremos el porcentaje de acierto a medida que los textos van aumentando de tamaño para unas MFW, un *culling* y un método estadístico concreto. Es común que aparezca en nuestra consola un mensaje de este tipo:

```
The following words/features could not be found in the corpus:
```

NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA...

No debe preocuparnos, no afecta al análisis ni a los resultados. Se debe a que hemos pedido a la máquina que tome las 5000 palabras más frecuentes del conjunto de nuestros textos para trabajar con ellas, pero es muy habitual no contar con tal cantidad de lemas diferentes, sobre todo cuando nuestros textos han reducido considerablemente su tamaño.

4. Conclusión

La validación cruzada se demuestra como una herramienta útil para ajustar los parámetros de nuestras investigaciones con *Stylo*. Con su puesta en práctica podemos determinar qué opciones son las más convenientes para nuestro problema autorial antes de abordar el texto conflictivo. Además, gracias a la validación cruzada comprobaremos si nuestro corpus indubitado, nuestro corpus de partida, funciona dentro de unos límites considerables para la estilometría, con el fin de no utilizarlo si el resultado es negativo.

Aunque no nos interese realizar la validación cruzada en los términos que aquí se exponen y nos decantemos por hacer un uso de Stylo más sencillo, mediante dendrogramas, *consensus trees* y demás esquemas gráficos, es muy recomendable tener siempre presentes los preceptos que se han comentado someramente. Así, si queremos comprobar gráficamente mediante un dendrograma dónde se sitúa un texto conflictivo, debemos realizar antes todo el proceso con nuestro corpus indubitado para probar si funciona de forma aceptable. Si este se desarrolla favorablemente entonces pasaremos a introducir nuestro texto dubitado, pero nunca antes de

haber regulado los parámetros. Solo cuando el dendrograma de nuestro corpus indubitado resulte aceptable pasaremos a introducir el texto conflictivo; si nunca llega a serlo, debemos renunciar a continuar nuestra investigación.

5. Bibliografía

Fradejas, José Manuel (2016). "El análisis estilométrico aplicado a la literatura española: las novelas policíacas e históricas". *Caracteres. Estudios culturales y críticos de la esfera digital* 5 (2). <http://revistacaracteres.net/revista/vol5n2noviembre2016/analisis-estilometrico/>. pp. 196-264 (3-9-2018).

Ilsemann, Hartmut (2017). "The two Oldcastles of London". *Digital Scholarship in the Humanities*, 32 (4). <https://academic.oup.com/dsh/article-abstract/32/4/788/2669798>. pp. 788–796. (3-9-2018).

Maciej, Eder (2010). "Does Size Matter? Authorship Attribution, Small Samples, Big Problem". *Proceedings of Digital Humanities* 30 (2). <https://academic.oup.com/dsh/article/30/2/167/390738>. pp. 132–135. (3-9-2018).

Maciej, Eder (2013). "Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing"*. *Literary and Linguistic Computing* 28 (4). <https://academic.oup.com/dsh/article-abstract/28/4/603/1077777>. pp. 603–614. (3-9-2018).

Maciej, Eder y Jan Rybicki (2013). "Do birds of a feather really flock together, or how to choose training samples for authorship attribution". *Literary and Linguistic Computing* 28 (2). <https://academic.oup.com/dsh/article-abstract/28/2/229/1034020>. pp. 229–236. (3-9-2018).

Maciej, Eder, Jan Rybicki y Mike Kestemont (2016). "Stylometry with R: a package for computational text analysis". *R Journal* 8 (1). <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>. pp. 107-121 (3-9-2018).

Rybicki Jan y Eder Maciej (2011). "Deeper Delta across genres and languages: Do we really need the most frequent words?" *Literary and Linguistic Computing* 26 (3). <https://academic.oup.com/dsh/article/26/3/315/1149353>. pp. 315–321. (3-9-2018).

Stefan Schöberlein (2017). "Poe or not Poe? A stylometric analysis of Edgar Allan Poe's disputed writings". *Digital Scholarship in the Humanities* 32 (3). <https://academic.oup.com/dsh/article-abstract/32/3/643/2669752>. pp. 643–659. (3-9-2018).

Este mismo texto en la web
http://revistacaracteres.net/revista/vol7n2noviembre2018/validacion


CARACTERES
Estudios culturales y críticos de la esfera digital

PETICIÓN DE CONTRIBUCIONES – CALL FOR CONTRIBUTIONS

Caracteres. Estudios culturales y críticos de la esfera digital es una publicación académica independiente **en torno a las Humanidades Digitales** con un reconocido consejo editorial, especialistas internacionales en múltiples disciplinas como consejo científico y un sistema de selección de artículos de doble ciego basado en informes de revisores externos de contrastada trayectoria académica y profesional. **El próximo número (vol. 8 n. 1, mayo 2019) está abierto a la recepción de colaboraciones**.

Los temas generales de la revista comprenden las disciplinas de Humanidades y Ciencias Sociales en su mediación con la tecnología y con las Humanidades Digitales. **La revista está abierta a recibir contribuciones misceláneas dentro de todos los temas de interés para la publicación.**

La revista está abierta a la recepción de artículos todo el año, pero hace especial hincapié en los tiempos máximos para garantizar la publicación en el número más próximo. Puede consultar las normas de publicación y la hoja de estilo a través de la sección específica de la web <http://revistacaracteres.net/normativa/>. Para saber más sobre nuestros objetivos, puede leer nuestra declaración de intenciones. **La recepción de artículos para el siguiente número se cerrará el 15 de marzo de 2019** (las colaboraciones recibidas con posterioridad a esa fecha podrían pasar a un número posterior). Los artículos deberán cumplir con las normas de publicación y la hoja de estilo. Se enviarán por correo electrónico a articulos@revistacaracteres.net.

Caracteres se edita en España bajo el ISSN 2254-4496 y está recogida en bases de datos, catálogos e índices nacionales e internacionales como **ESCI, ERIH Plus, Latindex, MLA**, Fuente Académica Premier o DOAJ. Puede consultar esta información en la sección correspondiente de la web <http://revistacaracteres.net/bases-de-datos/>.

Le agradecemos la posible difusión que pueda aportar a la revista informando sobre su disponibilidad y periodo de recepción de colaboraciones a quienes crea que les puede interesar.

PETICIÓN DE CONTRIBUCIONES – CALL FOR CONTRIBUTIONS

Caracteres. Estudios culturales y críticos de la esfera digital is an independent **journal on Digital Humanities** with a renowned editorial board, international specialists in a range of disciplines as scientific committee, and a double blind system of article selection based on reports by external reviewers of a reliable academic and professional career. **The next issue (vol. 8 n. 1, May 2019) is now open to the submission of contributions**.

The general topics of the journal include the disciplines of Humanities and Social Sciences in its mediation with the technology and the Digital Humanities. **The journal is now open to the submission of miscellaneous contributions** within all the relevant topics for this publication.

While the journal welcomes submissions throughout the year, it places special emphasis on the advertised deadlines in order to guarantee publication in the latest issue. Both the publication guidelines and the style sheet can be found in a specific section of our webpage <http://revistacaracteres.net/normativa/> . To know more about our objectives, the declaration of principles of the journal can be consulted. **The deadline for the reception of papers is March 15th, 2019** (contributions submitted at a later date may be published in the next issue). Articles should adhere to the publication guidelines and the style sheet, and should be sent by email to articulos@revistacaracteres.net.

Caracteres is published in Spain (ISSN: 2254-4496) and it appears in national and international catalogues, indexing organizations and databases, such as **ESCI, ERIH Plus, Latindex, MLA**, Fuente Académica Premier or DOAJ. More information is available in the website <http://revistacaracteres.net/bases-de-datos/>.

We appreciate the publicity you may give to the journal reporting the availability and the call for papers to those who may be interested.

Caracteres. Estudios culturales y críticos de la esfera digital

http://revistacaracteres.net

Noviembre de 2018. Volumen 7 número 2
http://revistacaracteres.net/revista/vol7n2noviembre2018

Contenidos adicionales

Campo conceptual de la revista Caracteres
http://revistacaracteres.net/campoconceptual/
Blogs
http://revistacaracteres.net/blogs/

Síguenos en

Twitter
http://twitter.com/caracteres_net

Facebook

http://www.facebook.com/RevistaCaracteres