Minería de medios sociales: Práctica Bloque II: Minería de Texto y de la Web

Máster en ciencia de datos e ingeniería de computadores - UGR

1-5-2018

Mª Cristina Heredia Gómez

crstnheredia@correo.ugr.es

Índice

Datos seleccionados

Para la práctica, he decidido analizar los tweets relacionados con el caso Cifuentes para extraer las 60 palabras más usadas por la gente en relación con el caso. Para ello, usando una app asociada a mi cuenta de twitter extraje de la red social los 300 tweets más recientes que mencionaran la palabra *Cifuentes*, creando así un corpus de texto.

Extracción de tweets

Aunque primero intenté extraer los tweets instalando el plugin disponible en KNIME para twitter usando los nodos **Twitter API connector** y **Twitter Search**, me encontré con el problema de que la implementación disponible en KNIME no está actualizada a la última modificación de twitter: la de leer tweets con 280 caracteres, de tal forma que los tweets de más de 140 caracteres no los leía enteros.
Probé entonces a extraerlos con la herramienta **TwitteR**, la interfaz para twitter desde R más usada hasta hace poco. Sin embargo, tampoco han actualizado su código a los nuevos cambios, así que finalmente, investigando encontré un nuevo paquete para R llamado **rtweet** publicado este año, que permite extraer tweets de haciendo más sencilla y segura la autenticación además de que está adaptado para leer el máximo de 280 caracteres por tweet y permite especificar con un parámetro si se desean o no leer los tweets que son retweets (comienzan por RT). Los tweets extraídos se guardan en un csv.

Conjunto de tweets

A continuación se muestran las primeras líneas del csv de tweets estraídos:

```
Cifuentes_tweets$text
"""Cuando visioné el vídeo de Cristina Cifuentes en las mazmorras espectrales de un supermercado
perdido en el tiempo me acordé de que en lo que fue Galerías Preciados de Palma"", por
@juan_planas https://t.co/1JpHTAaRH0"
Pide otro escaño para Cifuentes que está en paro! https://t.co/tLYIOYzWos
"@protestona1 @salvadorsanvil1 Del policía que decidió dejar libre y sin cargos a Cifuentes cuando
la cogieron robando en Eroski, ¿ha dicho algo?"
La juez imputa al director del máster de Cifuentes  @lavanguardia https://t.co/QBxS4rFOBZ
El 'caso Cifuentes' hunde al PP y dispara a Ciudadanos en Madrid https://t.co/g1uLG8nXNU via
@elpais_espana
Cifuentes y las conspiraciones políticas a lo 'House of Cards' https://t.co/8wvhJuYArr https://
t.co/WbsiUi6wyE
Hace un año escribía este artículo en @elconfidencial para valorar la dimisión de Esperanza
Aguirre que también se producía finalizando el mes de abril del 2017. Sigue siendo plenamente
válido para valorar la dimisión de Cifuentes. https://t.co/MyrR0fiasb
Olay se convierte en una marca viral tras los memes del hurto de Cifuentes https://t.co/ptjtnV8sWO
https://t.co/APue2NKbkF
VÍDEO | La Cifuentes de Polònia desvalija a Rajoy en su despacho https://t.co/iuyxXvSKdC
"Una juez imputa por falsedad al director del máster de Cristina Cifuentes https://t.co/bOS7LNnCNp
Cifuentes no sabía que no había ""robado"" él Master y por eso va asegur siendo diputada
autonómica con el consentimiento  de C'S"
Ahí va Cifuentes!!!!! https://t.co/InN0XVx0rh
El vídeo que demuestra que Cristina Cifuentes robó 2 cremas Olay en Eroski https://t.co/JORM4aSpiF
Cuatro presidentes madrileños y más de 20 años salpicados por la corrupción https://t.co/mI0Fg3Flsi
Reacciones del Santa Fe 1 - Tolima 1. Juan Navidad y Rafael Cifuentes https://t.co/DrVmy7QXTn
"▶ Imputado Enrique Álvarez Conde, director del máster de @ccifuentes #ComunidadDeMadrid https://
t.co/AMKJLnObYd https://t.co/Ma0fmEE7BU"
Cifuentes y las conspiraciones políticas a lo 'House of Cards' https://t.co/bhFFLDmRqC
"@kikonovoa @aleperseco a propósito del escandelete de la Sra. Cifuentes, descrito sabiamente por
el Sr. De Prada. Moi boa reflexión académica. https://t.co/INMbeMakXv"
"@LydiaLozanoOf Tú eres más de CIFUENTES ,esa sí que es perfecta,menuda zorraca"
"Garrido no descarta que Cifuentes deje el escaño: ""Tomará una decisión en los próximos días
sobre si continuar como diputada"" https://t.co/WO2WRFAt9O vía @eldiario_Madrid "
@Rafa_Hernando @andresiniesta8 Y de cifuentes no dices nada??? Jajajaja
Cifuentes se robó la gracia de este Tweet q
El 'caso Cifuentes' hunde al PP y dispara a Ciudadanos en Madrid https://t.co/LWL11z1iAd Gabilondo
el único líder que aprueba!!!! El Presidente que Madrid necesita!!!!
Tiene que ser mayor de edad ? Porque no lo encontraréis https://t.co/QBhwb5aZpw
"Granados, sobre la dimisión de Cifuentes: ""Si buscas venganza, cava dos fosas"" https://t.co/
ifazVklWc2 vía @eldiarioes"
Cifuentes y las conspiraciones políticas a lo 'House of Cards' https://t.co/39G9JqJOF4
""No pararemos hasta matarla, hasta acabar con ella", la amenaza de un empresario a Cifuentes
https://t.co/Hhpr7OYUBF"
"Esta escena me recuerda q antiguamente mandaban los + valientes y sacrificados y no esta castuza
d eunucos como Marota, Cifuentes o Errejona. Decadencia total y millones d subnormales
siguiéndoles Letizio VI es 1 reyezuelo muy representativo d esta mierda https://t.co/Gm4damM9gE"
@Xavinoriguis Iniesta es un mito. Cifuentes una cita.
```

Como se puede ver, cada tweet es un string y están separados por un salto de línea. Algunos tweets incluyen substrings, que son mayoritariamente citas de los usuarios.

Procedimiento

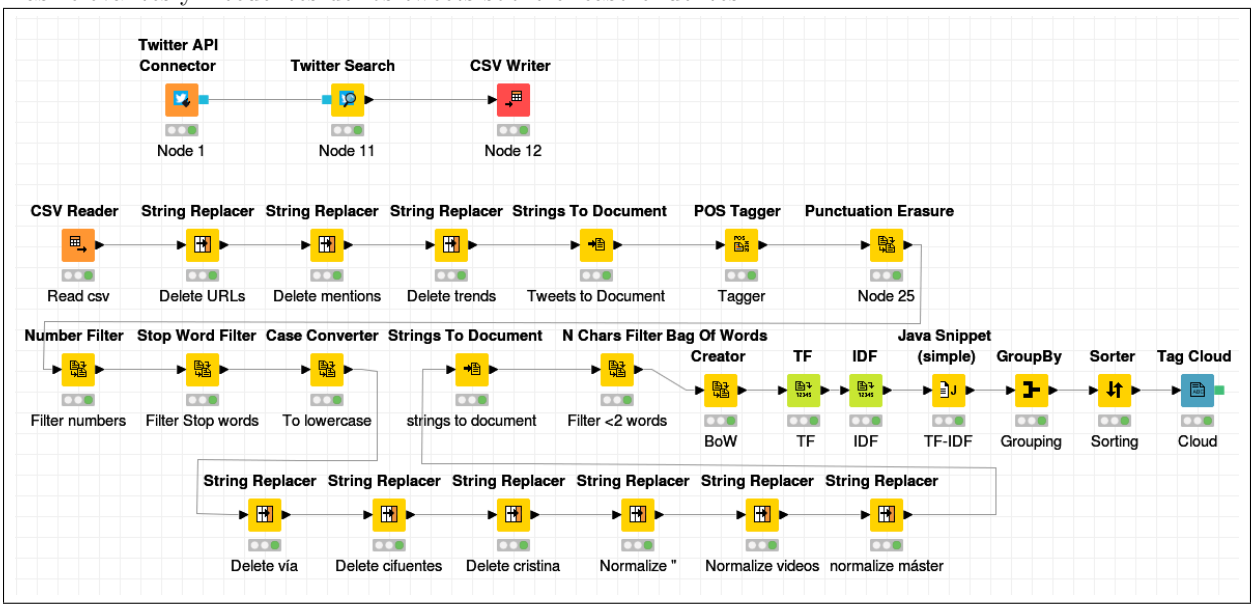
Handmade cleaning

Trabajar con tweets tiene un problema añadido, que es la forma de escribir de los usuarios. Para que el CSV se pudiera leer correctamente, hubo que tratar algunos tweets a mano que presentaban alguno de estos problemas:

- Líneas en blanco en el tweet: en la imágen anterior se muestra el conjunto de tweets ya procesado, pero en el conjunto de tweets original había muchos tweets que incluían saltos de línea dentro del mismo tweet, lo que el lector de CSV lo interpretaba como distintos tweets. Por tanto, tuve que eliminar esos saltos de línea dentro de cada tweet a mano.
- Comillas sin cerrar: muchos tweets incluyen citaciones o referencias entre comillas que muchos usuarios olvidan de cerrar o abrir. Esto ocasiona que el lector de CSV no pueda parsear bien los datos, por lo que tuve que revisar los tweets en las líneas que daba fallo para cerrar las comillas que faltaban.

Descripción del workflow

A continuación se muestra una imágen del workflow seguido en KNIME para extraer las palabras mencionadas más relevantes y frecuentes de los tweets sobre el caso cifuentes:



Los tres primeros nodos que no están conectados al wokflow original es como se intentó originalmente hacer la extracción y lectura de tweets que finalmente se hizo desde R con **rtweet**. A partir de ahí, el workflow seguido es el siguiente, donde se puede observar que en torno al 85 % de los pasos están dedicados a preprocesamiento y limpieza de tweets:

1. Se leen los datos del csv
2. Se borran URLs, hastags (#) y menciones (@) de los tweets, ya que son elementos inválidos que pueden meter ruido al hacer el análisis morfosintáctico
3. Pasamos los tweets a documento, ya que vienen como una serie de strings y el etiquetador requiere un documento a etiquetar
4. Etiquetamos los tweets con su etiqueta morfosintáctica

5. Una vez etiquetadas las palabras, eliminamos todos los signos de puntuación (.,:?¿!¡...) ya que no representan información importante para nuestro problema

6. Filtramos los números de los tweets. Muchos tweets contienen números escritos por los usuarios que no aportan información a nuestro problema

7. Eliminamos stopwords, es decir, palabras que no aportan nada. Para ello configuramos el nodo indicando que queremos eliminar Stopwords típicas del español, como son (de, para, y, o, en ...etc)

8. Normalizamos el texto a minúsculas, ya que hay tweets escritos sólo con mayúsculas, minúsculas o entremezcla de ambas

9. Una vez que todos los tweets están en minúsculas, eliminamos otras palabras que sabemos que no son interesantes para resolver el problema, como **vía, Cifuentes** o **Cristina**, ya que son palabras que aparecen mucho y que no nos dicen nada, pues ya sabemos que estamos buscando en tweets donde se habla de Cristina Cifuentes, y muchos tweets incluyen referencias a noticias u otras cuentas indicandolo con la palabra **vía**

10. Arreglamos las comillas, pues hay muchas palabras en las que los usuarios no han puesto espacios las ellas, por ejemplo **“esto**, en lugar de **“ esto**, por lo que se tokeniza como una sola palabra cuando en realidad no lo es

11. Sustituimos todas las palabras que se refieren a lo mismo, por ejemplo *vídeo, video* que algunos usuarios escriben con tilde y otros sin ella

12. Hacemos lo mismo con la palabra *máster*, ya que algunos usuarios escriben la palabra con tilde y otros sin ella

13. Pasamos los tweets, que de nuevo son Strings, a documento

14. Eliminamos aquellas palabras que tienen longitud menor que dos, ya que no aportan mucho o nada, y son, sobre todo, caracteres que representan emojis

15. Creamos una bolsa de términos donde se reflejan los términos que aparecen en el documento sin estar agrupados

16. Extraemos los términos más frecuentes, primero utilizando sólo el algoritmo TF absoluto, y luego usando TF absoluto+IDF. Para este último usamos un snippet de java para hacer la multiplicación de TFabs por IDF

17. Ya tenemos cuanto aparece cada palabra en cada tweet de forma ponderada con su relevancia. Ahora queda agrupar todas las palabras que se repitan en todos los tweets en una sola palabra, de forma que no queden palabras repetidas

18. Tras esto, las ordenamos de más a valor a menor valor de TF-IDF, dejando así las más relevantes al principio

19. Representamos en un gráfico de nube de palabras las 60 primeras palabras con tamaño proporcional a su relevancia

Resultados

La tagcloud obtenida usando tf-idf es la siguiente:

Donde se observa que las palabras más relevantes en el caso Cifuentes según la comunidad de twitter son **máster, madrid** y **pp**, en relación a las irregularidades de la expresidenta de Madrid con su supuesto máster de la URJC. También aparece en grande la palabra **vídeo** y **eroski**, incluso en un tono más ligero de azul, aparece la palabra **cremas**, lo que indica que también se ha twiteado bastante sobre el vídeo difundido del hurto de cremas Olay en el eroski en 2011. Aparece también la palabra **venganza** en grande porque hay muchos tweets que hablaban de la filtración del vídeo por una supuesta venganza.
Podemos ver que, en general, en la wordcloud de 60 palabras se muestran palabras realmente relevantes relativas al escándalo por el caso Cifuentes que permiten describir la historia conociendo un poco el contexto.