Minería de Medios Sociales.
Bloque 2. Minería de textos.

Marta Verona Almeida

May 20, 2019

1 Introducción

En esta práctica se utilizará el software `KNIME` para analizar un conjunto de tweets, extraídos de Twitter, con el fin de aplicar técnicas de procesamiento de textos, aplicar técnicas de minería sobre ellos y visualizar su tag cloud.

2 Datos seleccionados

Los datos utilizados en esta práctica han sido extraídos de Twitter mediante la api proporcinada por `KNIME`. Para ello, se han usado los nodos `Twitter API Connector` y `Twitter Search`, en la cual se han buscado 400 tweets a partir de las palabras "John nieve". Los nodos utilizados para realizar la búsqueda pueden verse en la imagen 2.

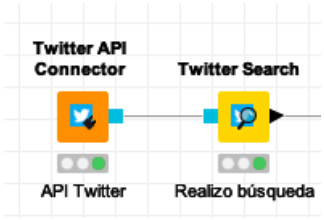

Figure 1: Workflow: Twitter

En la siguiente imagen 2 puede verse la tabla de entrada de datos. Como puede observarse, el número de columnas obtenidas es 26, sin embargo para esta práctica sólo se necesita el tweet, por lo que se filtran las columnas utilizando el nodo `Column Filter`.

Figure 2: Entrada de datos

3 Procesamiento de textos

Para el procesamiento de los tweets extraídos se han realizado diferentes acciones.

En primer lugar, se eliminan los retweets mediante el nodo `Row Filter`. A continuación, mediante el nodo `String Replacer`, utilizando expresiones regulares, se eliminan los enlaces, las menciones a otros usuarios y los números contenidos en los tweets.

En segundo lugar, se obtienen las etiquetas POS, se eliminan los signos de puntuación y se eliminan las palabras con más de tres letras. Para ello, se utilizan los nodos `POS tagger`, texttPunctuation Erasure y `N Chars Filter` respectivamente.

Por último, se utiliza el nodo `Stop word Filter` para eliminar de los tweets aquellas palabras vacías de significado.

El flujo realizado para llevar a cabo estas tareas puede verse en la imagen 3.

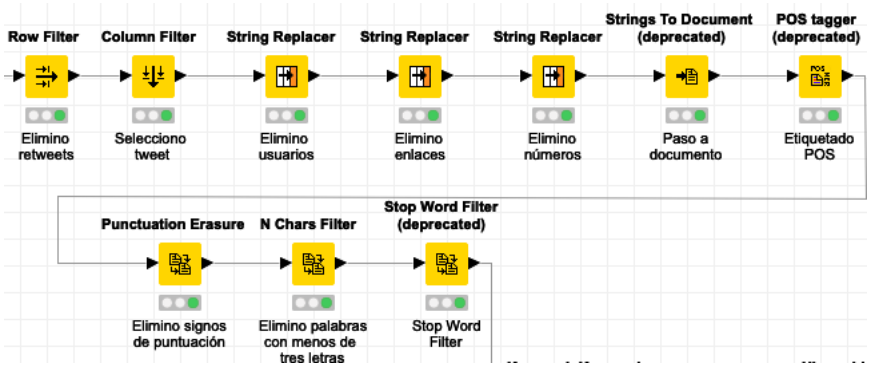

Figure 3: Workflow: preprocesado

4 Ténicas de minería

La técnica de minería aplicada sobre el conjunto de tweets obtenido en la sección anterior ha sido clústering jerárquico, a través del nodo `Hierarchical Clusterng`.

Para ello, se utiliza el nodo Keygraph keywork extractor para extraer las palabras relevantes de cada tweet mediante el uso de grafos. Se trata de asigar una puntuación a cada palabra dentro del tweet. Por último, se genera un vector a partir del conjunto de palabras del tweet, que será la entrada para el nodo de clústering jerárquico.

En la imagen 4 puede verse el flujo de nodos realizado para llevar a cabo estas tareas.

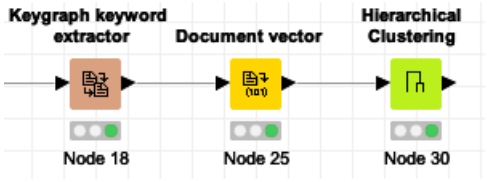

Figure 4: Workflow: Clustering

Por último, el resultado obtenido puede verse en la imagen 4. Se puede observar que existe un único clúster que agrupa todos los tweets.

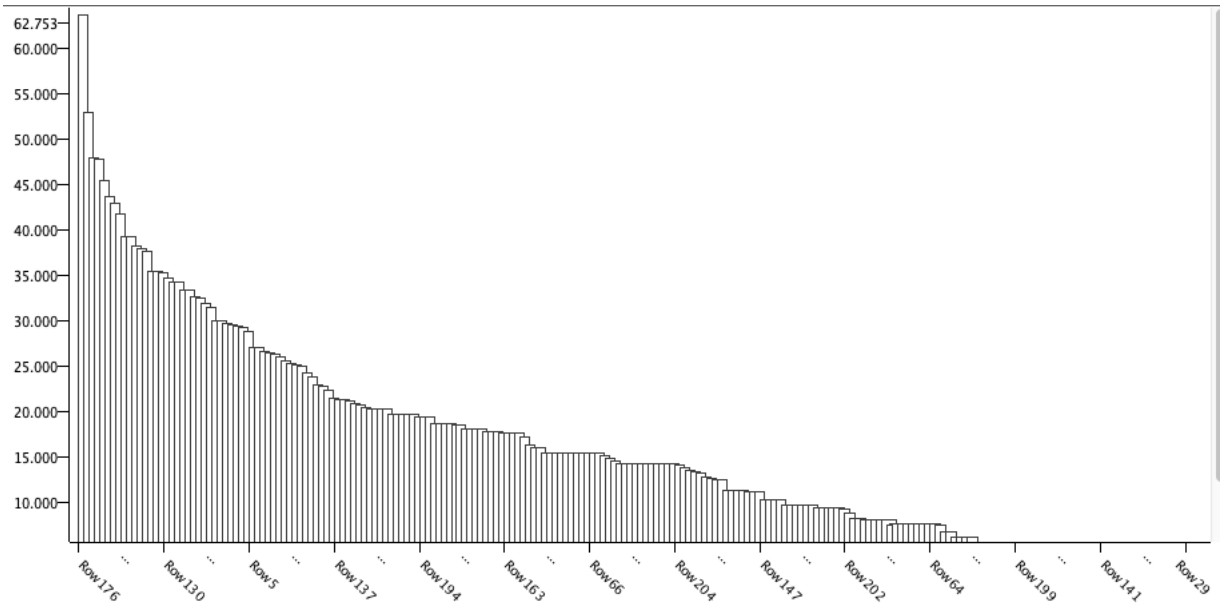

Figure 5: Clustering

5 Tag Cloud

Para extraer el Tag Cloud se utiliza el nodo Bag of words creator para extraer todas las palabras aparecen en los tweets, en relación a éste. A partir de esto, se calculan la frecuencia absoluta y la IDF (*Inverse Document Frequency* de cada palabra. Posteriormente, mediante el nodo `Java Snippet`, se crea un nuevo campo representado por el producto de estas dos frecuencias. Por último, se filtra atendiendo a este nuevo campo y se muestra la Tag Cloud asociado a las 100 palabras con mayor valor. Este proceso puede verse en la imagen 5.

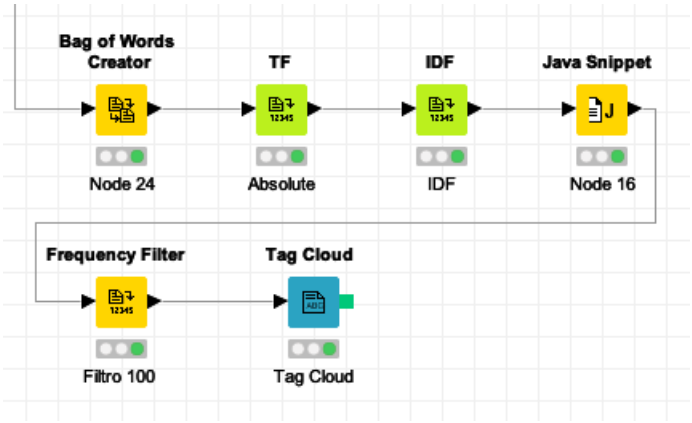

Figure 6: Workflow: Tag Cloud

En la imagen 5 pueden verse las palabras resultantes de este proceso. Aquellas de mayor tamaño indican una mayor frecuencia de aparición.

Como puede verse, ya que la palabra buscada ha sido “John Nieve”, las palabras más influyentes hacen referencia a la serie Juego de Tronos. Por ejemplo: “Drogon”, “muere”, “Danaeris”, “capítulo”, “trono”, “reinos” o “Snow”. Además es llamativo que aparezca la palabra “eurovisión”, ya que no tiene relación alguna con la búsqueda. Esto puede deberse a que en el momento en que se realizó la búsqueda coincidió

con la fecha en que este programa se emitía.

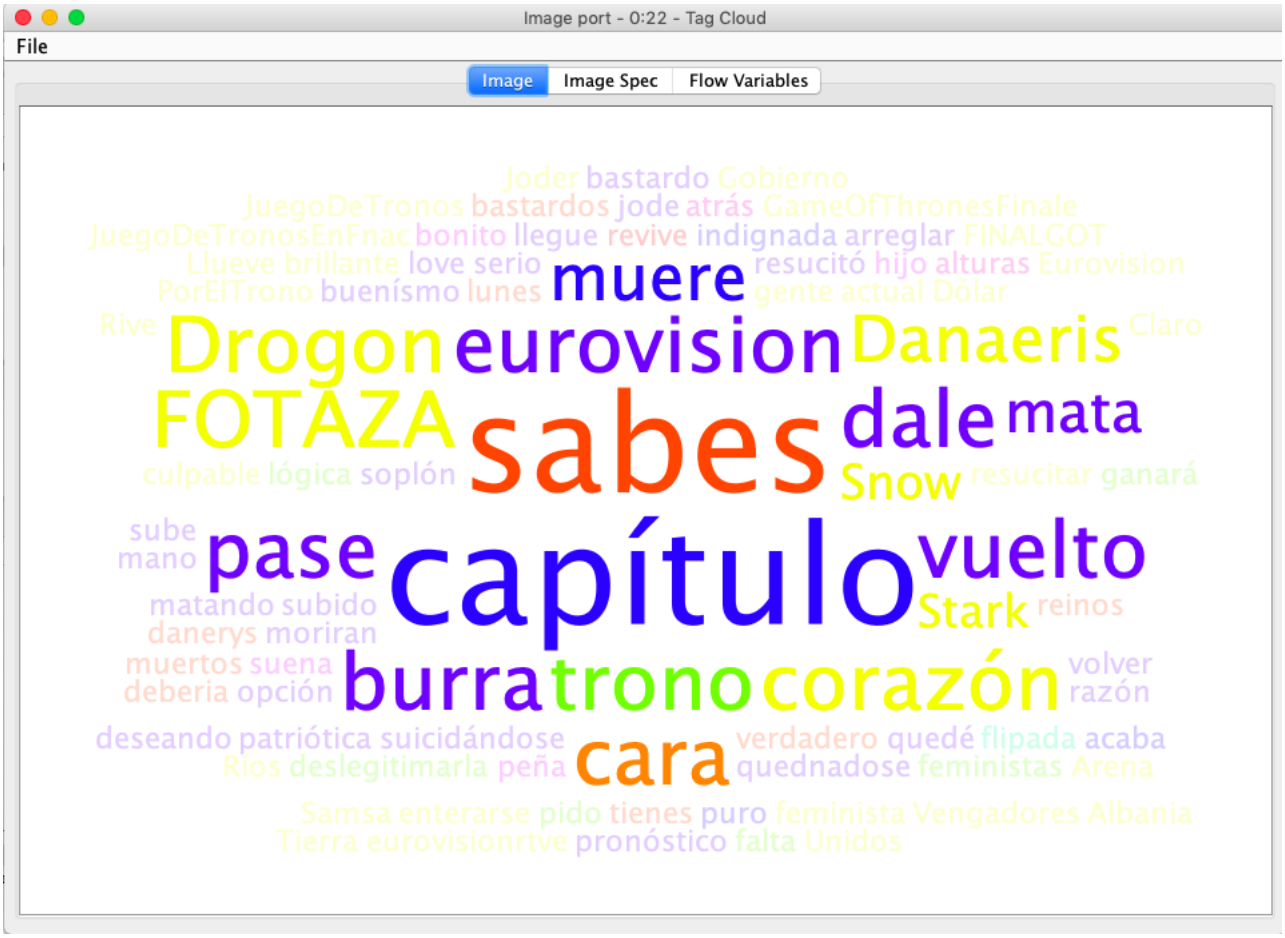

Figure 7: Tag Cloud

6 Workflow

A lo largo de la práctica se han mostrado secciones del flujo construido con la herramiento `KNIME`. En la imagen 8 puede observarse el flujo de trabajo completo llevado a cabo para realizar todas las tareas descritas a lo largo de la práctica.

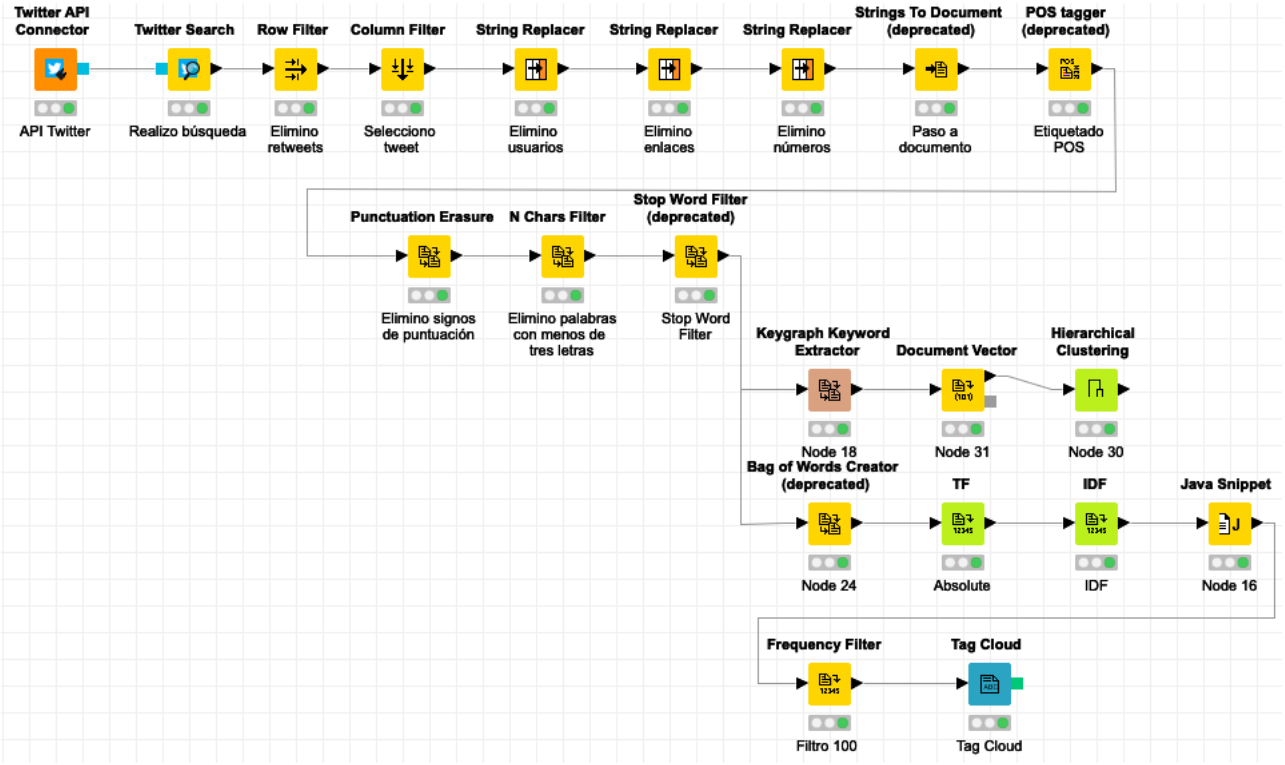

Figure 8: Workflow