Minería de textos

Omar Díaz, Felipe Gerard, Andrés Villaseñor

4 de junio de 2015

0. Introducción y caso de uso original

Las colecciones grandes de documentos pueden ser difíciles de aprovechar. Por ejemplo, si queremos que una o varias palabras aparezcan en el documento, podemos hacer un filtro simple. Sin embargo, si las palabras son (razonablemente) comunes, el query nos podría regresar un porcentaje importante de la colección de documentos, sobre todo si son muy largos. Algo más refinado que podríamos hacer sería por ejemplo usar el número de veces que aparecen las palabras deseadas, aunque esto favorecería documentos extensos y las palabras comunes.

El presente trabajo consistió en minar una colección de *abstracts* de 129,000 artículos ganadores de premios de la National Science Foundation (NSF) entre 1990 y 2003. El conjunto de datos es interesante porque provee las versiones cortas de los artículos, de modo que podemos leerlos y darnos una idea rápida de lo que trata cada artículo. En la página estaba la información en crudo (.txt) y ya lista para utilizar el método de bolsa de palabras. Preferimos utilizar la información cruda original para hacer el proceso desde la limpieza de los datos. El objetivo es realizar búsquedas temáticas inteligentes dentro de la colección, es decir, dado un conjunto de palabras clave, queremos obtener los artículos más relevantes. En la siguiente sección describiremos la teoría del método que utilizamos para mitigar los problemas descritos anteriormente.

1. Teoría

Como decíamos arriba, no basta con filtrar los artículos según la aparición de las palabras clave, ni tampoco un conteo simple. Lo que hicimos fue tomar en cuenta los conteos, pero usando técnicas para disminuir la importancia de la longitud de los documentos y de la frecuencia de aparición de las palabras comunes.

Denotamos la frecuencia del término w (puede ser una palabra o la raíz de una palabra obtenida con *stemming*) en el documento d como $tf_{w,d}$. De manera natural, denotamos por tf_d al vector de todas las frecuencias de los términos conocidos del documento d. Supongamos que queremos comparar los documentos q y d. Un primer enfoque que podríamos utilizar sería usar la distancia coseno, que toma en cuenta únicamente la frecuencia relativa de las palabras dentro de los documentos:

$$d_{coseno}(q,d) = \frac{tf_q \cdot tf_d}{\|tf_q\|\|tf_d\|}$$

El problema con lo anterior es que las palabras comunes podrían tomar un papel protagónico y en nuestro caso compartir por ejemplo artículos o preposiciones no es muy interesante. Para mitigar esto utilizaremos la frecuencia inversa en documentos idf_w, definida para una palabra w. Si N es el número total de documentos en la colección y df_w es el número de documentos de la colección que contienen a w, entonces definimos la frecuencia inversa de documentos como sigue:

$$idf_w = log(\frac{N}{df_w})$$

Y entonces, en lugar de describir a d por medio de tf_d, lo describimos por medio de c_d, donde

$$c_{d,w} = idf_w \times tf_{d,w}$$

Y así, calculamos la distancia entre dos documentos como la distancia coseno entre sus vectores característicos:

$$d(q,d) = \frac{c_q \cdot c_d}{\|c_q\|\|c_d\|}$$

La idf_w es el logaritmo del inverso de la probabilidad de que el término w aparezca en un documento elegido al azar. El efecto que esto tiene es que las palabras comunes casi no tendrán ningún efecto en la distancia, puesto que su probabilidad es cercana a 1 y por lo tanto su *idf* es cercana a cero. Por el contrario, las palabras raras tendrán una probabilidad cercana a 0, por lo que su *idf* será grande y contribuirán mucho. La razón detrás de hacer esto es que queremos que discriminen las palabras especiales o específicas a un contexto y no las genéricas. El esquema expuesto está escrito con mayor detalle en el libro Mining of Massive Datasets, con la pequeña diferencia de que ahí además normalizan tf_w por la máxima frecuencia obtenida por el término en la colección de documentos.

2. Aplicación a colección de arte mexicano

Para aplicar lo antes descrito a la colección de arte mexicano, el primer paso es extraer el texto de los PDFs. Ya con esto podemos tomar la mínima unidad de texto acordada (¿página? ¿párrafo?) como "documento" y aplicar la técnica TF-IDF. La idea es entonces que proporcionando un *query* el programa nos regrese las unidades de texto más relevantes. Tal vez sea más conveniente utilizar páginas en lugar de párrafos (incluso si se pudieran separar) porque tenderán a tener una longitud estable, lo que limita el sesgo por esta razón.

En el trabajo anterior hicimos un refinamiento adicional: hicimos TF-IDF para los *abstracts* y para los títulos de los artículos por separado y los ponderamos, con el fin de dar mayor peso a uno u otro campo a voluntad. En la aplicación que nos concierne ahora podríamos hacer algo similar, por ejemplo usando el título de los libros, subtítulos dentro del texto o pies de página, por ejemplo. También podríamos hacer otro tipo de cosas como tener palabras especiales (como nombres de autores) que sean ponderadas más fuertemente que las palabras normales.

Entre las cosas que hay que tener cuidado es que este tipo de sistemas tienden a enfocarse demasiado. Es decir, si el *query* contiene "Diego Rivera", muy probablemente los primeros resultados de la búsqueda tenderán a ser siempre los mismos, independientemente de que se agregue otros términos. Aunando esto a un sistema de recomendación que tome en cuenta también a los usuarios podría aislar gran parte de la colección: si siempre recomendamos lo mismo, todos los usuarios verán las mismas páginas, lo que hará que el sistema se clave aún más en ese contenido. Tendremos que meter aleatoriedad o algún tipo de factor exploratorio para evitar estas problemáticas, como por ejemplo que uno de los primeros 5 resultados sea escogido al azar con probabilidad proporcional a su *score*. Habrá que investigar más a este respecto.

Referencias

Jure Leskovec, Anand Rajaraman and Jeff Ullman. (2014). *Mining of Massive Datasets.* Cambridge University Press, USA. URL: http://www.mmds.org

Michael J. Pazzani. (2003). *NSF Research Award Abstracts 1990-2003 Data Set.* University of California, Irvine, USA. URL: https://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003

Omar Díaz, Felipe Gerard, Andrés Villaseñor. (2015). *Minería de textos. Proyecto final de Métodos Analíticos.* Instituto Tecnológico Autónomo de México, México.