Trabajo Práctico n 2

Minería de la Web 2017

Alumno: Santiago R Barboza

Para realizar la búsqueda basada en contexto el programa bbc.py (Busqueda Basada en Contexto) realiza 3 tareas principales, cada una implementada en una función de python. lee la consulta original o dato original del archivo ‘./data.txt’.

La primera función se encarga de buscar los 20 términos más relevantes del archivo ‘data.txt’. Para esto lee el archivo, les elimina las tildes, los stopwords y los links. Luego calcula la frecuencia de aparición de cada termino del archivo. Para poder destacar los sustantivos propios (Son los que le dan identidad a la noticia), se le multiplica por 2 la cantidad de apariciones. Por último selecciona los 5 términos más usados en el documento, junto con otros 10 términos seleccionados al azar de una ruleta sesgada , donde la probabilidad de que un elemento salga seleccionado depende de la frecuencia de aparición del mismo, y 5 frases formadas por 3 palabras elegidas al azar de los primeros 10 términos con 2 palabras elegidas al azar del resto de los términos. Estas últimas palabras no tienen ningún tipo de sesgo o preferencia.

La segunda parte la realiza una función llamada buscarTweets, que se encarga de realizar una busqueda de 5 tweets por cada uno de los 20 términos extraídos previamente y devolver la lista con todos los tweets no repetidos. Este metodo se encarga de realizar las peticiones a la REST API de Twitter.

Por último, la última función se llama filtrarTweets, que toma los tweets recolectados, los limpia y calcula para cada uno la similitud por coseno con el archivo original, los ordena de mayor a menor y muestra en pantalla los tweets cuya similitud sea mayor que 0.45.

Los archivos de prueba tienen la información utilizada y los resultados obtenidos.