README

Para poder ejecutar correctamente nuestro código, considerando que es un fichero python, tenemos que abrirlo en un IDE que soporte ese tipo de lenguaje, preferiblemente Anaconda.

Primeramente, para empezar ejecutar nuestro código se debe ejecutar la serie de "imports" que podemos encontrar en el inicio de nuestro fichero python.
A continuación abrimos el fichero 'tw_hurricane_data.json' en 'lines'; que recoge información sobre diferentes tweets, para ello tenemos que poner el path donde guardamos dicho file en nuestro ordenador.
Hemos imprimido el doc 12 (lines[12]), para comprobar la información de dicho tweet. Convertimos el archivo 'lines' que es json, en un diccionario guardado en 'datos_diccionarios'; hemos hecho en print para comprobar cual es tweet id del doc 12 según: datos_diccionario[12]['id'].

Después de guardar el fichero, tenemos que ejecutar la sección de funciones que se llama 'Functions'. En este apartado, podemos encontrar las funciones de build_terms() y create_index().
Seguidamente, podemos encontrar un apartado llamado TF-IDF + cosine similarity, dónde tenemos que ejecutar por orden las funciones que encontramos. La primera corresponde a la función rank_documents_tf_idf() y la segunda se llama search_tf_idf(). Después, encontramos una sección llamada BM25, dónde encontramos las funciones rank_documents_bm25() y search_bm25(), se tienen que ejecutar por orden al que aparecen. Si seguimos bajando podemos encontrar la sección de Our score + cosine similarity, tenemos que ejecutar las funciones que nos aparecen por orden cronológico, es decir, primero rank_documents_our_score() y después, search_our_score().

Si continuamos viendo el código, tenemos que ejecutar la celda que contiene la llamada de la función create_index(), tened en cuenta que puede tardar un poco al completar la ejecución.

Seguidamente, tenemos que ejecutar la sección de las Queries. Aquí podemos encontrar diferentes celdas con diferentes queries avaladas por diferentes modelos y ránkings. Tenemos que ejecutar estas queries por orden cronológico donde aparecen las celdas.

Finalmente, la última parte de código que tenemos corresponde el Word2vec. En primer lugar, tenemos que ejecutar las celdas que contienen las funciones top20_2vec() y search(). Después de ejecutar estas dos funciones, tenemos que ejecutar las siguientes celdas de código por orden cronológico. Primero vamos a

llamar las funciones de arribar y después vamos a ejecutar las queries y ver su output.