PRÁCTICA: SIMILITUD DE DOCUMENTOS CON BAG-OF-WORDS Y SIMILITUD DEL COSENO

Pedro Sarmiento Yánez

1.- Explicación del código

En mi código definimos una función para eliminar los *stopwords* la cual usaremos en un bucle en el que cargaremos todos los textos y los procesaremos y limpiaremos un poco para hacer un mejor análisis, quitamos signos de puntuación y también eliminamos espacios que sobran, quitamos los números y luego nos quedamos solo con las palabras que no son *stopwords*, tras finalizar el bucle guardaremos cada texto y también los tokens de cada uno. Tras esto nos quedamos con la bolsa de palabra realizando una función que hemos llamado *unique*. Con la bolsa de palabras crearemos los vectores de cada uno de los textos los cuales luego pasaremos a la función de la similitud del coseno para así hallar qué documentos son más similares entre sí y cuáles son más diferentes y ver si tiene algo que ver con los temas de los que trata cada uno de ellos.

2.- Análisis de los resultados

Los documentos con una mayor similitud, dado que tienen una similitud del coseno mayor del 30% son el documento 11 y 12, junto al 9 y al 12, estos tres textos tratan distintos temas relacionados con los coches eléctricos por lo que tiene bastante sentido el resultado que hemos obtenido. Luego, en la otra cara de la moneda tenemos cerca de 15 resultados del 0% en la similaridad del coseno, esto se debe a que no todos los textos tratan sobre el mismo tema, aunque esto no es siempre así en el primer 0% que nos encontramos, por ejemplo es entre el texto 1 y el texto 3, el texto 1 trata sobre móviles y el segundo sobre aplicaciones para estos dispositivos, en teoría deberían de tener un cierto punto de similitud pero al usar términos diferentes para referirse a lo mismo, nuestro programa no lo detecta y por eso da una similitud nula. Aunque esto no es una tónica general entre los demás ejemplos.