Alumno:	García González Aarón Antonio
Grupo:	3CV9
Unidad de Aprendizaje:	Procesamiento de lenguaje natural
Profesor:	Kolesnikova Olga
Tarea #6:	Clustering por algoritmo K-means
Fecha:	Martes 12 de mayo de 2020

Explicación breve del desarrollo del algoritmo:

1. A partir del archivo de texto, obtener una lista de listas, donde cada lista es un mensaje que contiene los tokens ya normalizados, es decir se aplico minúscula, quitamos stopwords, lematización y partes de oración.
2. Obtenemos vocabulario.
3. Obtener un vector numérico, donde 1 representa que el mensaje en el índice X es spam y 0 que es ham.
4. Calculamos la matriz TF IDF con el objetivo de tener los datos de manera numérica.
5. Dentro del algoritmo de K-mens, inicializamos los K centroides, que serán dos vectores de los que ya tenemos del conjunto de datos, pero estos los elegimos de manera aleatoria por su índice, para calcular la distancia entre los puntos y los centroides utilice la formula de la norma de la resta de dos vectores, también lo hice con ángulo entre dos vectores, pero la primera entrego resultados mas acertados, posteriormente realizamos el movimiento de centroide o ajuste 100 veces, la mayoría de las veces en mi ejecución basta con menos de 30 veces para que el error disminuya practicante nada.

Nota: Dado que la selección aleatoria de los centroides iniciales es determinante para todo el algoritmo, pues no siempre obtendremos buenos resultados, así que muestro a continuación algunas de las pruebas que realice.