Tarea 2 Parte 1)

- Investiga la estrategia de Vectorización TF-IDF ¿Como se calcula?
- Se calcula en dos pasos, el primero el TF, se refiere a "frecuencia del termino, mientras mayor sea la frecuencia del término en el documento mayor sera su importancia (numero de veces / total) y IDF significa "frecuencia inversa de documento": mayor frecuencia, menor importancia del termino

- ¿En qué situaciones es más efectivo usar TF-IDF para tareas de clasificación de texto?
- Cuando el documento tiene muchas palabras reutilizadas, cuando se necesita obtener palabras claves y para situaciones que requieran filtrar texto.

- ¿Con qué bibliotecas se puede implementar?
- En muchas bibliotecas, pero tres de las mas populares son
  - Scikit-learn
  - Nltk (Natural language toolkit)
  - Gensim

Tarea 2 Parte 2

1) ¿Qué problemas de los N-gram resuelve el "Laplace Smoothing"? ¿Cómo trabaja? y qué pasa con un modelo de NLP cuando se emplea esta técnica?

- Resuelve el problema de cero probabilidad en n-grams, trabaja añadiendo uno mas a todos los Ngramas posibles, Cuando se aplica esta tecnica se evitan las probabilidades de cero, por lo que evita que el modelo sea arruinado.

$$\text{Formula: } P(w_i|class) = \frac{freq(w_i, class) + 1}{N_{class} + V}$$

2) ¿Qué pasa cuando una palabra en el test set no se encuentra en el vocabulario del modelo de los N-gram? ¿Cómo se puede modelar la probabilidad de palabras out-of-vocabulary? (OOV)

- Lo que pasa es que se genera un problema conocido como palabras out-of-Vocabulary (OOV), en donde el modelo no tiene información sobre la palabra y esta no puede asignar una probabilidad a esa palabra, para poder modelar estas palabras se puede utilizar la tokenización de subpalabras, para que el modelo pueda tomar en cuenta para poder obtener la probabilidad de las palabras.