Tarea 2 Pt 1

1.- Investigue la estrategia de vetorizacion TF-IDF. ¿Como se calcula

Técnica que mide la importancia de una palabra en un documento dentro de un conjunto de documentos. Se calcula multiplicando la frecuencia de la palabra en el documento (TF) por la frecuencia inversa de esa palabra en todos los documentos (IDF)

2. Efectividad de TF-IDF

Es útil en textos con vocabularios diversos y documentos de diferentes longitudes, ~~etc~~ ayudando a ~~destacar~~ destacar términos relevantes y minimizar el impacto de palabras comunes.

3. Bibliotecas para implementarla

- scikit-learn
- NLTK
- Gensim
- spaCy

4.- Laplace Smoothing en N-grams

- Evita que N-gramas no observados tengan una probabilidad de cero, lo que podría causar fallos en el modelo
- Asigna una probabilidad a todos los N-gramas, incluso si aparecen en el corpus, añadiendo una constante a las frecuencias (generalmente 1).
- Evita probabilidades nulas, pero puede asignar probabilidades no realistas a sec. raras

5.- Palabras fuera de Voc (OOV)

- Si una palabra del test set no está en el voc. del modelo, no se puede calcular la probabilidad.
- Se puede usar suavizando (como Laplace), etiquetar especialer para OOV, o modelos que retroceden a N-gramas de menor orden (back-off models) para estimar probabilidades