1. Investigue la estrategia de vectorización TF-IDF.

- ¿Cómo se calcula?

Se calcula multiplicando el term frequency y el inverse document frequency

El TF es:

$$\frac{\#\text{ de veces que el término aparece en el documento}}{\#\text{ total de términos en el documento}}$$

El IDF es:

$$\log\left(\frac{\#\text{ de documentos en el corpus}}{\#\text{ de docs que contiene el término}}\right)$$

$$TF\text{-}IDF = TF \cdot IDF$$

El TF-IDF sirve para medir la importancia o relevancia de una palabra de acuerdo con su frecuencia y tomando en cuenta su presencia a través de los documentos

- Uso en clasificación. Se usa cuando se necesitan identificar palabras distintivas en documentos largos, es útil para clasificación como filtrado de spam, análisis de sentimiento y categorización de noticias

- Se puede implementar con scikit-learn, NLTK o spaCy

2. Laplace Smoothing en N-grams

- El problema que resuelve es que pueda haber N-grams con probabilidad 0 Este suavizado hace que la probabilidad sea $\neq 0$

- Funciona sumando 1 a los conteos de todos los N-grams tanto a los observados como no observados.

- Este suavizado mejora la generalización del modelo, sin embargo, puede introducir un sesgo hacia probabilidades más uniformes.

3. Palabras OOV

- Si una palabra no se encuentra en el vocabulario del modelo su probabilidad puede ser 0.

- Para solucionarlo se puede agregar una probabilidad pequeña a todas las palabras se pueden usar diferentes técnicas como

  - suavizado (Laplace, Good-Turing) que asigna probabilidades de forma uniforme a palabras no vistas.

  - Asignar una clase especial OOV. Asignar una probabilidad cuando se encuentra una palabra OOV, una probabilidad similar a otras palabras OOV