Tarea 2
Erik Elizondo González - A01284899

1. Estrategia de vectorización TF-IDF
- Definición: Es una (~~estrategia~~) técnica de vectorización que combina la frecuencia de un término en un documento (TF) y su rareza en toda la colección (IDF).
- Cálculo: Se calcula como el producto de TF (Term Frequency) y IDF (Inverse Document Frequency)
  * $TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t,D)$
  * $TF(t,d) = \frac{\text{Número de veces que aparece el término } t \text{ en documento } d}{\text{Número total de términos en el documento } d}$
  * $IDF(t,D) = \log\left(\frac{\text{Número total de documentos}}{\text{Número de documentos que contienen } t}\right)$
- ¿Cuando es más efectivo?
  - Tareas de clasificación de texto y búsqueda de texto cuando:
    - Enfatizar términos que son importantes en un documento
    - Evitar que las palabras comunes tengan un peso en la clasificación
    - Procesar textos pequeños - medianos con vocabularios amplios
- Bibliotecas para implementar TF-IDF
  - Scikit-learn (TfidfVectoricer)
  - NLTK (Parcialmente)
  - spaCy (Parcialmente)

2. Laplace Smoothing
- ¿Qué problema de los N-grams resuelve Laplace smoothing?
  - Problema de probabilidad $\emptyset$ en los modelos N-gram al añadir un valor constante a los conteos, asignando probabilidades a secuencias no vistas. Esto mejora la generalización del modelo, pero puede dar (~~oportunidad~~) probabilidades más altas a secuencias poco comunes.

3. Modelado de palabras fuera del vocabulario (OOV)
- Cuando una palabra del test set no está en el vocabulario (OOV), el modelo no puede asignarle probabilidad. Para manejarlo, se usa un token OOV especial, técnicas de smoothing o vocabularios más amplios para evitar probabilidades nulas.