Nallely Serna
A00833111

① Investigue la estrategia de vectorización TF-IDF

- ¿Cómo se calcula?

TF (Frecuencia de término): mide la frecuencia con la que un término aparece en un documento en particular.

IDF (Frecuencia inversa del documento): mide qué tan común o raro es un término en todos los documentos.

$$TF\text{-}IDF = TF \times \log\left(\frac{\text{Número total de documentos}}{\text{Núm. de doc. que contienen el término}}\right)$$

- Efectividad

Es más efectivo en tareas de clasificación de texto cuando es necesario identificar la relevancia de las palabras y filtrar aquellas que son muy comunes (como "el", "la"). Funciona bien en sistemas de búsqueda de información, análisis de sentimientos, y clasificación de emails (spam/no spam).

- Implementación

Se pueden usar bibliotecas como
- Scikit-learn (TfidfVectorizer)
- NLTK

② Laplace Smoothing en N-gramas

- Problema que resuelve: resuelve el problema de asignar una probabilidad de Ø a las secuencias de N-gramas que no aparecen en el conjunto de entrenamiento, lo que impide hacer predicciones en esos casos.

- Cómo trabaja: Añade una pequeña cantidad (Normalmente 1) a todas las frecuencias de los N-gramas, incluso a los que no aparecen, para evitar que su probabilidad sea cero.

Nallely Serna
A00833111

- Impacto en NLP: Con esta técnica, se evitan resultados nulos en el modelo de lenguaje, lo que permite que el modelo maneje mejor las palabras o combinaciones raras o no vistas.

3 Palabras fuera del vocabulario (OOV) en N-gramas

- Problema: Cuando una palabra en el set de prueba no está en el vocabulario del modelo, el modelo no puede asignar una probabilidad correcta a esa palabra o secuencia de palabras.

- Modelar OOV
  - Asignar una probabilidad pequeña a palabras OOV, como suavizado. (Laplace).
  - Usar una categoría especial "UNK" para representar palabras desconocidas en el entrenamiento. Con esto, el modelo puede estimar la probabilidad de palabras no vistas sin detenerse.