Tarea 2.

Adrian Pineda - A00834710

① Vectorización TF-IDF

- ¿Cómo se calcula?

- Frecuencia de Término (TF): Mide cuántas veces aparece una palabra en un documento relacionado con el número total de palabras del documento.

  - $TF = \frac{\#\ \text{Término del documento}}{\#\ \text{Total palabras del documento}}$

- Frecuencia Inversa del doc (IDF): Evalua la importancia de una palabra en la colección total de documentos.

  - $IDF = \log\left(\frac{N}{df(t)}\right) = \log\left(\frac{\#\ \text{Total de documentos}}{\#\ \text{Numero de documentos que tienen al término "t"}}\right)$

[1][3]

Al multiplicar ambos términos obtenemos TF-IDF:

$TF\text{-}IDF = TF(IDF)$

- ¿En qué situaciones es más efectivo usar TF-IDF?

Realmente las situaciones que más destacan son: "clasificación de texto", como:

- Filtrado y Clasificación de Spam.
- Análisis de Sentimientos o clasificación de documentos.

Da mayor importancia a palabras que tienen mayor valor informativo y considera irrelevantes textos o palabras de uso común. [1][3]

- ¿Con qué bibliotecas se puede implementar?

  - SciKit-Learn en Python, usando TfidVectorizer [3]
  - NLTK en Python para procesar textos y generar los vectores TF-IDF [3]

② ¿Qué problemas de los N-grams resuelve el "Laplace Smoothing"?

Resuelve el problema de las palabras con prob. 0% en un modelo de N-grams cuando esas palabras o combinaciones no aparecen en el conjunto de entrenamiento. De lo contrario podría arruinar el modelo con probs nulas.

- ¿Cómo trabaja?

El suavizado de Laplace añade una constante (1) al contador de ocurrencias de cada término.

$$P(\text{término}) = \frac{\#\text{Ocurrencias del término} + 1}{\#\text{Total términos} + \text{Vocabulario}}$$

Se evita la probabilidad 0 en el modelo. [1]

- Efecto en un modelo NLP

El modelo después del suavizado se vuelve más robusto para manejar palabras desconocidas en los datos de entrenamiento. Aunque es de observar que puede afectar reduciendo ligeramente las probs ya existentes debido ya que se ajustan de forma generalizada. [2]

③ ¿Qué pasa cuando una palabra en el test set no se encuentra en el vocabulario de N-gram?

Si una palabra en el conjunto de prueba no aparece en el vocabulario del modelo, su prob sería 0, podría ocasionar fuertes errores en las predicciones

- ¿Como se puede modelar la probabilidad de palabras OOV?

  * Suavizado de Laplace: Asegura que aún las palabras no vistas tengan una ligera probabilidad.

  * Modelos complejos basados en transformers o embeddings: Modelos modernos como: "Word2Vec" o "GloVe" pueden generalizar mejor para palabras OOV porque aprenden relaciones entre palabras basados en su contexto además de su frecuencia. [1] [2]

Referencias:

[1] Analytics Vidhya, 2021 (Step by Step Guide to Master NLP)

[2] Zilliz, 2024 (TF-IDF Understanding Term frecuency)

[3] CodeSignal, n.d (Implementing TF-IDF for feature Engineering in text Clasification)