Andrés Villarreal González A00833915

① Investigue la estrategia de vectorización TF-IDF. ¿Cómo se calcula? ¿En que situaciones es mas efectivo usarlo para tareas de clasificación de texto? ¿Con que bibliotecas se puede implementar?

- Es una técnica que evalua la importancia de una palabra dentro de un documento y en relación con un conjunto de documentos. Se calcula en dos pasos:

  • TF. La frecuencia con la que un termino aparece en un documento especifico en relación con el numero total de terminos en ese documento

$$TF(t,d) = \frac{\text{Numero de veces que aparece el termino } t \text{ en el doc } d.}{\text{Total de terminos en el documento } d.}$$

  • IDF: Importancia de un termino en todo el corpus. Reduce el peso de palabras comunes

$$IDF(t) = \log \frac{\text{num. total de docs}}{\text{num. total de docs que contienen el termino } t}$$

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

- Funciona bien en tareas de clasificación de texto como analisis de sentimiento, categorización de noticias, detección de spam, y recuperación de información, ya que filtra palabras comunes y entatiza palabras distintivas

- Bibliotecas:

  • scikit-learn

  • NLTK

  • Gensim

② ¿Qué problema de los N-gram resuelve el "laplace Smoothing"? ¿Cómo trabaja? ¿Que pasa con un modelo NLP cuando se emplea esta tecnica?

- En modelos de N-gram, a veces es posible que ciertos N-grams no aparezcan en el conjunto de entrenamiento, lo que puede llevar a una probabilidad de 0 para esas secuencias causando que el modelo falle en las predicciones cuando aparezcan nuevos N-grams en el test set

- El Laplace Smoothing añade un valor constante a todas las freccencias de N-grams evitando que las probabilidades sean cero.

$$P(\omega_i | \omega_{i-1}) = \frac{\text{Frecuencia de } (\omega_i - 1, \omega_i) + 1}{\text{Total de ocurrencias de } \omega_{i-1} + V}$$

donde V es el tamaño del vocabulario

- Este metodo puede hacer que las probabilidades esten mas sesgadas hacia los N-grams raros o nunca vistos, lo que puede afectar negativamente a la precisión

③ ¿Qué pasa cuando una palabra en el test no se encuentre en el vocabulario del modelo de los N-gram? ¿Cómo se puede modelar la probabilidad de palabras out-of-vocabulary?

- Cuando una palabra en el test no esta en el vocabulario del modelo entrenado, el modelo no tiene una probabilidad asociada a esa palabra, lo que puede llevar a fallos en la precisión

- Como modelar las OOV:
  - Una estrategia comun es añadir una 'token OOV' al vocabulario, lo que permite que cualquier palabra desconocida se trate como una palabra generica con baja probabilidad
  - Se pueden usar tecnicas de subword embeddings, como el modelo Byte Pair Encoding o WordPiece que permiten representar palabras desconocidas como combinaciones de subpalabras, evitando la necesidad de una unica representacion de cada palabra.