30/09/2024

Tarea 2: Investigación TF-IDF

1: Estrategia de vectorización TF-IDF:

TF-IDF (Term Frequency - Inverse Document Frequency) se calcula multiplicando la frecuencia de un término en un documento por el algoritmo inverso de la frecuencia de documentos en los que aparece.

Fórmula:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{1 + DF(t)}\right)$$

t = término d = documento N = núm. total de documentos
DF(t) = la cantidad de documentos q+ contienen t

Situaciones de uso:

Reducir el impacto de palabras comunes (stopwords) y resaltar términos distintivos del corpus. Básicamente, clasificación de documentos o análisis de sentimientos.

Bibliotecas:

- Scikit-learn
- NLTK

2- Problema N-gram

Problema: Resuelve el problema de asignar probabilidad cero a N-grams que no se observan en los datos de entrenamiento.

Funcionamiento: Laplace Smoothing añade valor pequeño (normalmente 1) a las cuentas de todos los N-grams, lo que garantiza que cada combinación tenga una probabilidad mínima no nula.

¿Qué pasa con el modelo? Evita el problema de probabilidad cero, lo que ayuda a mejorar la robustez del modelo, especialmente para corpus pequeños o limitados.

3- Palabras fuera del vocabulario

Problema: Cuando una palabra no está en el vocabulario el modelo no le puede asignar una probabilidad.

Solución:

- Smoothing: Además de Laplace, existen técnias como add-k que retroceden a modelos de menor orden.
- Reservar tokens OOV: Incluir un token especial para capturar todas las palabras fuera del vocabulario. con probabilidad pequeña.