¿QUE ES TF-IDF?

TF-IDF es la abreviatura de Term Frequency - Inverse Document Frecuency. Esta técnica tiene como objetivo asignar un peso a cada palabra en un documento, basado en la cantidad de veces que aparece en el documento (frecuencia del término) y en la cantidad de documentos en el corpus completo en los que aparece (frecuencia inversa del documento)

FORMULA DE LA FRECUENCIA DE TÉRMINOS (TF) —
Se calcula dividiendo el número de repeticiones de un término en un documento por el núm total de términos en ese documento. Esto nos da una medida de la importancia relativa de un término en un documento específico.

FORMULA DE LA FRECUENCIA INVERSA DE DOCUMENTOS (IDF)
La frecuencia inversa de documentos (IDF) se calcula tomando el logaritmo del número total del documento en el corpus dividido por el núm de documentos que contienen un término específico. Esta medida nos da una idea de lo común o raro que es un término en el corpus completo.

CÁLCULO DE TF-IDF
Multiplicamos la frecuencia de términos (TF) de un término por su frecuencia inversa de documentos (IDF). Esto nos da un valor que indica la importancia de un término en un documento específico en relación con el corpus completo.

¿En qué situaciones es más efectivo usar TF-IDF para tareas de clasificación de texto?
- Documentos con términos distintos: cuando se desea resaltar términos que son más representativos de ciertos documentos y reducir el peso de términos comunes que aparecen en muchos documentos.
- Clasificación de texto en dominios específicos: En conjunto de datos donde ciertos términos técnicos o específicos de una industria tiene un peso informativo importante, el TF-IDF ayuda a distinguir estos términos.
- Conjunto de datos de texto de tamaño moderado
- Problemas donde el contexto no es primordial: Funciona bien cuando las relaciones semánticas entre palabras no son fundamentales para la tarea, y la frecuencia de las palabras individuales es mas relevante que el significado contextual
- Clasificacion de categorías bien definidas: como clasificación de correos electronicos o categorización de artículos de noticias.

Ventajas
- considera tanto la frecuencia de un término en un documento como su importancia en el corpus completo
- Asigna un peso mayor a los ~~términos más frecuentes pero menos comunes en corpus~~
- Eficiente computacionalmente y fácil implementar.

¿Con que bibliotecas puedes implementar?
- viene incluido en sklearn como una feature extraction
  - from sklearn.feature_extraction.text import TfidTransformer
- Gensim: ofrece herramientas. Tiene implementación propias de TF-IDF
  - from gensim import corpora, models
- NLTK y spacy no tienen una función propia pero se puede usar en conjunto con otras bibliotecas.
  - import nltk
  - import spacy

Problema de los N-gram resuelto por LAPLACE SMOOTHING

El "laplace smoothing" (o suavizado laplace) resuelve el problema de probabilidades cero en los modelos N-gram. Este problema ocurre cuando una secuencia de palabras no aparecen en el corpus de entrenamiento. Sin suavizado, las secuencias con probabilidad cero afectan la capacidad del modelo para generar o predecir correctamente

¿Cómo trabaja el LAPLACE SMOOTHING?

Laplace smoothing trabaja añadiendo una pequeña cantidad (generalmente 1) al conteo de cada palabra o secuencia de palabras en el N-gram. La fórmula general para un modelo de N-gram con Laplace smoothing es:

$$P(w_n|w_i^{n-1}) = \frac{count(w_i^n)+1}{count(w_i^{n-1})+V}$$

Donde

- $count(w_i^n)$ es el conteo de la secuencia en el corpus
- V es el tamaño del vocabulario
- El 1 añadido en el numerador y V en el denominador permiten evitar probabilidades cero, ajustando todas las probabilidades para que sumen 1.

Impacto en el modelo de NLP:

Al emplear laplace smoothing, el modelo:

- Evita probabilidades cero, permitiendo que asigne una pequeña probabilidad a eventos que no se han observado en los datos de entrenamiento.
- Generaliza mejor en presencia de secuencias nuevas, pero puede subestimar las secuencias frecuentes y sobreestimar las raras, ya que el suavizado afecta todos los eventos por igual.

¿Qué pasa cuando una palaba en el test set no está en el vocabulario del modelo de N-gram?

Cuando una palabra del conjunto de prueba no se encuentra en el vocabulario del modelo, se denomina Out-of-Vocabulary (OOV). En este caso, la probabilidad asignada a dicha palabra sería cero si no se implementa alguna técnica para manejarlo, lo cual es problemático.

Modelar la probabilidad de palabras OOV.

Se pueden usar varias estrategias:

- Asignación de un token especial (e.g., <UNK>): Se introduce un token para representar todas las palabras que no están en el vocabulario. Se calculan la probabilidad de <UNK> durante el entrenamiento, sumando los conteos de las palabras OOV.
- Back-off models. Si una palabra o secuencia no está en el corpus, se retrocede a modelos de N-gram de menor orden (eg. de trigramas a bigramas o unigrams) para calcular la probabilidad basada en menos contexto.
- Suavizado avanzado: Métodos como el suavizado de Kneser-Ney son más avanzadas que el suavizado de Laplace y manejan las probabilidades de palabras OOV ajustando las frecuencias de palabras observadas y no observadas.

Referencias

KeepCoding. (s.f.). ¿Qué es el algoritmo TF-IDF Vectorizer? KeepCoding. Recuperado de https://keepcoding.io/blog/que-es-el-algoritmo-tf-idf-vectorizer/

Aggarwal, C. C., & Zhai, C. (2012). **Mining Text Data**. *Springer Science & Business Media*. https://doi.org/10.1007/978-1-4614-3223-4

Naik, K. (2024, febrero 14). *TF-IDF: Una forma poderosa de vectorizar texto*. Toolify. Recuperado de https://www.toolify.ai/es/ai-news-es/tfidf-una-forma-poderosa-de-vectorizar-texto-1195359