Modelado de texto usando el modelo TF-IDF

Problemas de BoW

Todas las palabras tienen la misma importancia.

No se preserva información semántica.

Modelo TF-IDF - Solución

- Cierta información semántica se conserva a medida que se da más importancia a las palabras poco comunes que a las palabras comunes.

Ejemplo: "Ella es hermosa". Aquí ”hermosa" tendrá más importancia que "Ella" o "es".

Modelo TF-IDF - Solución

Esto se debe a que TF-IDF no solo cuenta la frecuencia de las palabras en un documento (como lo hace el modelo Bag of Words), sino que también ajusta esta frecuencia según la cantidad de documentos en los que aparece la palabra.

En otras palabras, TF-IDF aumenta la relevancia de las palabras que son únicas o raras en el corpus, mientras que reduce la importancia de las palabras que son comunes en muchos documentos.

¿Qué es TF-IDF?

TF = Term Frequency

IDF = Inverse Document Frequency

TF-IDF = TF * IDF

Modelo TF-IDF – Term Frequency

Fórmula:

$$\frac{Número\ de\ apariciones\ de\ una\ palabra\ en\ un\ documento}{Número\ de\ palabras\ en\ ese\ documento}$$

Modelo TF-IDF – Inverse Document Frequency

Fórmula:

$$\log\left(\frac{(Número\ total\ de\ documentos)}{(Número\ de\ documentos\ que\ contienen\ la\ palabra)}\right)$$

Modelo TF-IDF

Fórmula: $TF - IDF(t, d) = TF(t, d) * IDF(t)$

Aplicaciones de TF-IDF

- Búsqueda de información
- Clasificación de texto
- Análisis de sentimiento
- Recomendación de contenido
- Análisis de temas

Descripción de cómo funciona TF-IDF

TF-IDF funciona al asignar un puntaje a cada palabra en un documento en función de su frecuencia en ese documento (TF) y su frecuencia en todos los documentos (IDF).

Cuanto más a menudo aparece una palabra en un solo documento, pero menos a menudo en todos los documentos, mayor es su puntaje TF-IDF.

Las palabras que aparecen con más frecuencia en un documento, pero raramente en otros documentos son más importantes.

Pasos para construir el modelo TF-IDF en Python

Paso 1: Cargar e importar librerías : Cargar NLTK y otras librerías necesarias.

Paso 2: Preprocesar los datos : Tokenizar, eliminar stopwords y convertir a minúsculas.

Paso 3: Calcular la frecuencia de términos (TF) : Calcular la frecuencia de cada palabra en cada documento.

Paso 4: Calcular la frecuencia inversa de documentos (IDF) : Calcular la frecuencia inversa de cada palabra en toda la colección.

Paso 5: Calcular TF-IDF : Combinar TF y IDF para obtener la importancia de cada palabra.

Paso 6: Normalizar TF-IDF (opcional) : Normalizar los valores de TF-IDF para mejorar la comparabilidad.

Paso 7: Almacenar y utilizar el modelo TF-IDF : Almacenar el modelo y utilizarlo para aplicaciones como búsqueda, clasificación y análisis de sentimiento.