MODELADO DE TEXTO USANDO EL MODELO BAG OF WORDS

INTRODUCCIÓN AL MODELO BAG OF WORDS

QUÉ ES EL MODELO BAG OF WORDS

El modelo Bag of Words es una técnica común en el procesamiento del lenguaje natural (NLP) donde se representa un documento como un conjunto de palabras, ignorando el orden y la estructura, centrándose solo en la presencia de palabras.

IMPORTANCIA EN EL MODELADO DE TEXTO

Es fundamental en el análisis de texto, ya que simplifica la complejidad al tratar el texto como una colección de palabras únicas, permitiendo la comparación y clasificación de documentos.

VENTAJAS DEL MODELO BAG OF WORDS

Permite una fácil implementación y comprensión. Es eficiente para grandes conjuntos de datos textuales. Es útil en tareas como clasificación de textos y análisis de sentimientos.

PROCESO DE MODELADO

DETALLES DEL PROCESO

El proceso de modelado de texto con Bag of Words implica tokenizar el texto, construir un vocabulario, representar el texto como un vector de frecuencias de palabras y aplicar técnicas de NLP y aprendizaje automático.

VENTAJAS DEL MODELO BAG OF WORDS

EXPLORANDO LAS VENTAJAS

Facilita la comparación de documentos. Es eficaz para documentos cortos. Permite identificar términos clave en un texto de manera sencilla.

EJEMPLO PRÁCTICO

RESULTADOS ESPERADOS

Se obtiene un vector numérico que representa la presencia de términos en el documento, permitiendo analizar la similitud entre distintos textos.

APLICACIÓN EN LA INDUSTRIA

Un ejemplo real sería la clasificación de correos electrónicos como spam o no spam mediante el conteo de palabras clave utilizando el modelo Bag of Words.

CONSIDERACIONES Y LIMITACIONES

ASPECTOS IMPORTANTES A TENER EN CUENTA

El modelo Bag of Words no considera el significado contextual de las palabras. Puede haber problemas con palabras homónimas. La dimensionalidad de los vectores puede afectar la eficacia en textos largos.

Pasos para construir el modelo BoW

1. Recopilar y preparar los datos:
    1. Reúne el texto que deseas analizar.
    2. Limpia el texto eliminando caracteres especiales, convirtiendo a minúsculas, etc.
2. Tokenización:
    1. Divide el texto en palabras individuales (tokens).
3. Eliminar palabras vacías
4. Construir el vocabulario:
    1. Crea una lista de todas las palabras únicas en el corpus.
5. Contar las frecuencias:
    1. Cuenta cuántas veces aparece cada palabra en cada documento.
6. Crear el vector BoW:
    1. Representa cada documento como un vector de frecuencias de palabras.