Taller de Preprocesamiento de Texto y Análisis de Sentimientos

Objetivo

En este taller trabajaremos con un dataset de reseñas de películas de IMDb. El objetivo es aplicar técnicas de preprocesamiento de texto y luego representar el texto utilizando dos métodos: Bag of Words (BoW) y TF-IDF. Finalmente, exploraremos cómo estas representaciones pueden ser útiles para entrenar un modelo de análisis de sentimientos.

Dataset

El dataset que utilizaremos contiene 50,000 reseñas de películas con etiquetas de sentimientos (positivas o negativas). Está disponible en Kaggle. Deben descargarlo para usarlo en las siguientes actividades.

Actividades

1. Exploración del Dataset

El primer paso es familiarizarse con el dataset. Asegúrense de que entienden cómo está organizado y qué tipo de datos contiene.

Instrucciones:

- Cargar el dataset en un entorno de trabajo (Jupyter Notebook, Google Colab, etc.).
- Realizar un análisis exploratorio inicial que incluya:
  - Número total de reseñas.
  - Proporción de reseñas etiquetadas como positivas y negativas.

Entregables:

- Un resumen breve del dataset: tamaño, distribución de etiquetas, tipos de variables.

2. Preprocesamiento de Texto

Antes de entrenar un modelo, es importante limpiar y transformar el texto. Los pasos a seguir son:

a. Conversión a minúsculas

Todas las palabras deben estar en minúsculas para evitar que el modelo distinga entre "Película" y "película".

b. Eliminación de puntuación

Se deben eliminar los signos de puntuación para simplificar el análisis.

c. Eliminación de stopwords

Las *stopwords* son palabras comunes que no aportan información útil (como "de", "la", "el"). Utilicen una lista de stopwords en inglés para eliminarlas.

d. Lematización y stemming

- **Lematización:** Transforma las palabras a su forma base (ejemplo: "corriendo" a "correr").
- **Stemming:** Recorta las palabras a su raíz (ejemplo: "jugando" a "jug").

Herramientas sugeridas: Utilicen las librerías `nltk` y `re` para este proceso.

Entregables:

- Código donde se apliquen los pasos mencionados al texto de las reseñas.
- Ejemplos del texto antes y después del preprocesamiento (al menos 5 reseñas).

3. Representación del Texto

Una vez que el texto está preprocesado, necesitamos convertir las palabras en números para que puedan ser usadas por un modelo de machine learning. Utilizaremos dos métodos diferentes para representar el texto.

a. Bag of Words (BoW)

Generen una matriz donde cada fila representa una reseña y cada columna una palabra. Las celdas de la matriz contendrán la cantidad de veces que una palabra aparece en cada reseña.

b. TF-IDF (Term Frequency-Inverse Document Frequency)

El método TF-IDF ajusta la frecuencia de las palabras basándose en cuántos documentos contienen esa palabra, destacando las más importantes para el análisis.

Herramientas sugeridas: `CountVectorizer` y `TfidfVectorizer` de la librería `scikit-learn`.

Entregables:

- Matrices de texto representadas mediante BoW y TF-IDF.
- Comparen ambas representaciones: ¿Cuáles son las diferencias más notables entre los dos métodos? ¿Cómo afecta esto al análisis?