Noviembre 2023

Tecnológico de Monterrey

Análisis de Sentimientos con PySpark y Dashboard en Tableau

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Querétaro

Inteligencia Artificial Avanzada para la Ciencia de Datos II

TC3007B.501

Presenta:
Ricardo Cáceres | A01706972

Introducción

El proyecto se centra en realizar un análisis de sentimientos en un DataSet de 2.86 GB, en este caso, un DataSet de libros de Amazon. El análisis de sentimientos es sobre los reviews de los libros, todo esto utilizando PySpark. Además, se ha desarrollado un dashboard interactivo en Tableau para visualizar datos del conjunto de datos y proporcionar funcionalidades específicas a los usuarios.

Objetivo del Proyecto

- Implementar un análisis de sentimientos en comentarios de usuarios utilizando PySpark.
- Desarrollar un dashboard en Tableau para explorar y analizar datos relacionados con libros.

Flujo del Proyecto

1. Configuración del Entorno:
    a. Se instaló Apache Spark.
2. Carga y Exploración de Datos:
    a. Se leyó el conjunto de datos de 2.86 GB (Book_rating.csv).
    b. El conjunto de datos fue sacado de Kaggle.com y la liga es la siguiente: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews/data
3. Preprocesamiento de Datos:
    a. Se hace la limpieza y la transformación de datos para prepararlos para el análisis.
    b. Selección de columnas relevantes y manejo de tipos de datos.
4. Tokenización y Filtrado de Stop Words:
    a. Se usa el Tokenizer y StopWordsRemover para procesar el texto de los reviews.
5. Creación del Modelo:
    a. Se define y se configura un modelo de regresión logística.
6. Entrenamiento del Modelo:
    a. Se dividen los datos en conjuntos de entrenamiento y prueba.
    b. Se entrena el modelo utilizando el conjunto de entrenamiento.
7. Predicciones:
    a. Se utiliza el modelo entrenado para realizar predicciones sobre nuevos reviews (Input del usuario).
8. Dashboard en Tableau:
    a. Se creó un dashboard interactivo en Tableau para visualizar datos del conjunto de datos.
    b. Dentro de las funcionalidades se incluyen la búsqueda de un libro específico y la visualización de su precio y el average review/score.
    c. Incluye la posibilidad de buscar los Top 10 libros por debajo de un precio ingresado, basándose en el review/score.

d. Link al Dashboard público de Tableau: https://public.tableau.com/views/AmazonBookDashboard/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link