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

Índice

Introducción

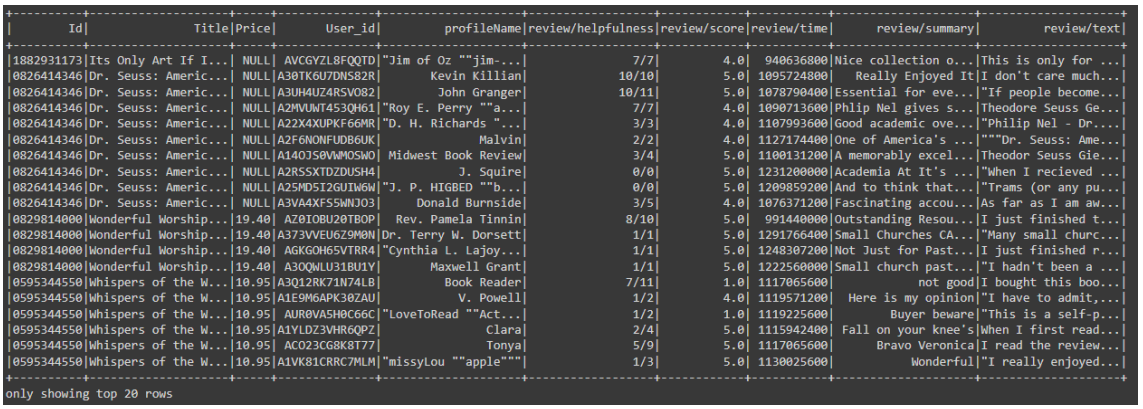
El proyecto se centra en realizar un análisis de sentimientos en un DataSet de 2.86 GB, en este caso, un DataSet de libros de Amazon. El análisis de sentimientos es sobre los reviews de los libros, todo esto utilizando PySpark. Además, se ha desarrollado un dashboard interactivo en Tableau para visualizar 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.

Conjunto de Datos Inicial

El conjunto de datos inicial, con un tamaño de 2.86 GB, consiste de información detallada sobre libros de Amazon, incluyendo revisiones proporcionadas por los usuarios. Antes de realizar cualquier análisis, es fundamental comprender la estructura y composición de los datos.



Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
1882931173	Its Only Art If I...	NULL	AVCGYZL8FQQTD	"Jim of Oz ""jim-...	7/7	4.0	940636800	Nice collection o...	This is only for ...
0826414346	Dr. Seuss: Americ...	NULL	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much...
0826414346	Dr. Seuss: Americ...	NULL	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for eve...	"If people become...
0826414346	Dr. Seuss: Americ...	NULL	A2MVUWT453QH61	"Roy E. Perry ""a...	7/7	4.0	1090713600	Phlip Nel gives s...	Theodore Seuss Ge...
0826414346	Dr. Seuss: Americ...	NULL	A22X4XUPKF66MR	"D. H. Richards "...	3/3	4.0	1107993600	Good academic ove...	"Philip Nel - Dr....
0826414346	Dr. Seuss: Americ...	NULL	A2F6NONFUDB6UK	Malvin	2/2	4.0	1127174400	One of America's ...	"""Dr. Seuss: Ame...
0826414346	Dr. Seuss: Americ...	NULL	A14OJS0VWMOSWO	Midwest Book Review	3/4	5.0	1100131200	A memorably excel...	Theodor Seuss Gie...
0826414346	Dr. Seuss: Americ...	NULL	A2RSSXTDZDUSH4	J. Squire	0/0	5.0	1231200000	Academia At It's ...	"When I recieved ...
0826414346	Dr. Seuss: Americ...	NULL	A25MD5I2GUIW6W	"J. P. HIGBED ""b...	0/0	5.0	1209859200	And to think that...	"Trams (or any pu...
0826414346	Dr. Seuss: Americ...	NULL	A3VA4XFS5WNJO3	Donald Burnside	3/5	4.0	1076371200	Fascinating accou...	As far as I am aw...
0829814000	Wonderful Worship...	19.40	AZ0IOBU20TBOP	Rev. Pamela Tinnin	8/10	5.0	991440000	Outstanding Resou...	I just finished t...
0829814000	Wonderful Worship...	19.40	A373VVEU6Z9M0N	Dr. Terry W. Dorsett	1/1	5.0	1291766400	Small Churches CA...	"Many small churc...
0829814000	Wonderful Worship...	19.40	AGKGOH65VTRR4	"Cynthia L. Lajoy...	1/1	5.0	1248307200	Not Just for Past...	I just finished r...
0829814000	Wonderful Worship...	19.40	A3OQWLU31BU1Y	Maxwell Grant	1/1	5.0	1222560000	Small church past...	"I hadn't been a ...
0595344550	Whispers of the W...	10.95	A3Q12RK71N74LB	Book Reader	7/11	1.0	1117065600	not good	I bought this boo...
0595344550	Whispers of the W...	10.95	A1E9M6APK30ZAU	V. Powell	1/2	4.0	1119571200	Here is my opinion	"I have to admit,...
0595344550	Whispers of the W...	10.95	AUR0VA5H0C66C	"LoveToRead ""Act...	1/2	1.0	1119225600	Buyer beware	"This is a self-p...
0595344550	Whispers of the W...	10.95	A1YLDZ3VHR6QPZ	Clara	2/4	5.0	1115942400	Fall on your knee's	When I first read...
0595344550	Whispers of the W...	10.95	AC023CG8K8T77	Tonya	5/9	5.0	1117065600	Bravo Veronica	I read the review...
0595344550	Whispers of the W...	10.95	A1VK81CRRC7MLM	"missyLou ""apple"""	1/3	5.0	1130025600	Wonderful	"I really enjoyed...

only showing top 20 rows

El conjunto de datos consiste de las siguientes columnas:

- **Id**: Identificador único del libro.
- **Title**: Título del libro.
- **Price:** Precio del libro.
- **User_id:** Identificador del usuario que realizó la revisión.
- **profileName:** Nombre del usuario que realizó la revisión.
- **review/helpfulness**: Puntuación de utilidad de la revisión (ejemplo: 7/7).
- **review/score:** Calificación de la revisión en una escala de 0 a 5.
- **review/time:** Timestamp de la revisión.
- **review/summary:** Resumen del texto de la revisión.
- **review/text**: Texto completo de la revisión.

Posteriormente, se evaluó la presencia de valores nulos.

```
Column 'Id': 0 null values
Column 'Title': 208 null values
Column 'Price': 2517579 null values
Column 'User_id': 562250 null values
Column 'profileName': 562200 null values
Column 'review/helpfulness': 367 null values
Column 'review/score': 130 null values
Column 'review/time': 27 null values
Column 'review/summary': 65 null values
Column 'review/text': 43 null values
```

Al ver que columnas esenciales para el objetivo de nuestro proyecto contenían valores nulos, se hizo una limpieza para abordar estos valores nulos. Estos valores fueron eliminados y este proceso de eliminación sienta las bases para el proyecto.

Creación de la Columna “sentiment”

Con el objetivo de facilitar el análisis de sentimientos en los reviews de los libros, se procedió a crear una nueva columna denominada “sentiment”. Esta columna se diseñó para representar la polaridad de los reviews de acuerdo con la calificación proporcionada en la columna “review/score”.

La metodología adoptada para la creación de esta columna es la siguiente:

- **Asignación de valor 1:** Se asigna el valor 1 a la columna “sentiment” si el “review/score” es mayor a 3.0. Este valor indica un review positivo, ya que una calificación superior a 3.0 sugiere una experiencia satisfactoria por parte del usuario.
- **Asignación de valor 0:** En caso contrario, es decir, cuando el “review/score” es igual o inferior a 3.0, se asigna el valor 0 a la columna “sentiment”. Este valor representa una revisión menos positiva o negativa.

```
# Se agrega una columna 'sentiment' basada en la 'review/score'
# donde se asigna 1 si 'review/score' es mayor a 3.0, y 0 en caso contrario
data = data.withColumn("sentiment", when(col("review/score") > 3.0, 1).otherwise(0))
```

Esta estrategia de asignación binaria simplifica el análisis de sentimientos al convertir las calificaciones numéricas en una categoría más manejable y fácilmente interpretable.

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).

Hacer predicciones con un Input del usuario

```
# Pedir un review al usuario
user_review = input("Enter your review: ")

# Meter el input del usuario en un spark dataframe
user_data = spark.createDataFrame([(user_review,)], ["review/text"])

# Hacer la prediccion
prediction = model.transform(user_data)

# Extraer la prediccion del sentimiento
predicted_sentiment = prediction.select("prediction").collect()[0][0]

# Mostrar la predicccion de sentimiento
if predicted_sentiment == 1.0:
    print("Sentiment: Positive")
else:
    print("Sentiment: Negative")
```

```
Enter your review: I like this book, it is very fun!
Sentiment: Positive
```

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