Tecnológico de Monterrey

Campus Querétaro

Nombre del trabajo:

Informe de Procesamiento y Análisis de Big Data

Curso:

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)

Estudiante:

Jaime López Hernández – A00571842

Fecha de entrega:

30 de octubre

Introducción

- El objetivo del proceso es crear un modelo de clasificación para predecir si una reseña de Amazon es positiva o negativa.
- El conjunto de datos utilizado es un archivo CSV de más de 1 GB de tamaño, que contiene 34.6 millones de reseñas.
- El proceso se lleva a cabo utilizando PySpark y MLlib.

Configuración del entorno de trabajo

En esta sección, puedes describir los pasos que seguiste para configurar tu entorno de trabajo en Colab para utilizar PySpark. Esto incluye la instalación de PySpark y la configuración de las variables de entorno.

Importación de los datos

En esta sección, puedes describir los pasos que seguiste para importar los datos desde Google Drive a PySpark. También puedes describir los pasos que seguiste para limpiar los datos, como eliminar las filas con valores faltantes.

El esquema tiene 3 columnas: polaridad, título y texto. La columna polaridad es un número entero que permite nulos, probablemente indicando la polaridad positiva/negativa de la reseña.

- Las columnas título y texto son cadenas que permiten nulos, por lo que pueden faltar título y/o texto en algunas reseñas.
- No hay columnas complejas, son simples valores únicos.
- No hay ID único ni otros metadatos.
- Polaridad como número entero sugiere es una clasificación codificada (ej. 1 = negativa, 2 = positiva).

Estadística Inicial

summary	polarity	title	text
count	3600000	3599952	3599987
mean	1.5	NaN	null
stddev	0.5000000694444585	NaN	null
min	1	The Worst Thing ...	this is the best...
max	2	♦ LOVE IT ♦	...were Marvin and ...

- El conjunto de datos contiene 3,6 millones de filas, lo cual indica que es extenso y adecuado para machine learning.
- La columna polaridad codifica el sentimiento como 1 (negativo) o 2 (positivo). Su media de 1.5 muestra una distribución balanceada de clases.
- La columna título tiene muchos valores nulos, por lo que habrá que gestionar los títulos faltantes.
- La columna texto también tiene nulos. Los valores mínimos y máximos muestran variedad en la longitud del texto.
- En resumen, se tratan de datos etiquetados para clasificación de sentimientos. El texto requerirá procesamiento y extracción de características.
- Los títulos desequilibrados podrían afectar ciertos análisis, por lo que quizá haya que aislar reseñas con/sin título.
- No hay valores atípicos evidentes en la polaridad según las estadísticas.

Creación del modelo de clasificación

En esta sección, puedes describir los pasos que seguiste para crear un modelo de clasificación utilizando MLlib. Esto incluye la selección del algoritmo de clasificación, el entrenamiento del modelo y la evaluación del modelo.

Visualización de los resultados

En esta sección, puedes describir los pasos que seguiste para crear un tablero de visualización con los resultados del modelo. Esto podría incluir la creación de gráficos y tablas para mostrar la precisión del modelo, la distribución de las clases y otros datos relevantes.

Prueba del modelo

El entrenamiento del modelo de regresión logística tardó más de 45 minutos en este conjunto de datos debido a su complejidad.

Para evaluar el funcionamiento de este modelo de clasificación binaria, se utilizó la métrica del Área Bajo la Curva (AUC). Esta mide qué tan bien el modelo puede distinguir entre la clase positiva y la negativa.

Luego de entrenar el modelo de regresión logística y probarlo con datos nuevos, se obtuvo un AUC de 0.919. Esto significa que el modelo puede separar las clases positiva y negativa con un alto grado de precisión. Un AUC de 1.0 representaría un modelo perfecto, mientras que uno de 0.5 sería equivalente a una suposición al azar.

Por lo tanto, con un AUC de 0.919 se puede afirmar que este modelo tiene una capacidad considerable para diferenciar los casos de las clases positiva y negativa. Si bien no es un desempeño perfecto, supera ampliamente el azar y permite hacer una distinción fiable entre ambas clases.

Conclusión