DA5402 - Assignment 9

Anirudh Rao be21b004

Task 1

To process the Amazon 'Gourmet Foods' review data and apply sentiment analysis in a parallel fashion, a Python script called `sentiment_analysis.py` was created.

The script begins by setting up logging, Spark context, and Spark configuration to orchestrate the distributed processing. The dataset (`Gourmet_Foods.txt`) is structured in blocks, where each block represents a review in a key-value pair format. The `parse_review_block` function is responsible for parsing each block to extract the review score and the text. This ensures that only valid entries are retained for analysis, while any parsing errors are logged.

Sentiment analysis is performed via `apply_sentiment_partition`, a function designed to be used with Spark's `mapPartitions`. Each review text is passed to the Hugging Face sentiment analysis pipeline, which classifies it as either "`POSITIVE`" or "`NEGATIVE`". Any issues during prediction are caught and logged as "`ERROR`" labels.

Task 2

To evaluate model predictions, the script uses `partition_confusion_matrix` to compute components of the confusion matrix — True Positives, False Positives, False Negatives, and True Negatives — locally within each partition. These are then aggregated using reduce, reflecting the reduce step in the map-reduce paradigm.

Finally, the script computes and prints precision and recall, standard metrics for classification performance, along with a simple visualisation of the confusion matrix.

Files submitted

- `requirements.txt` − Text file that contains the Python libraries to be installed prior to running the scripts.
- `sentiment_analysis.py` − Performs distributed sentiment analysis on Amazon Gourmet Food reviews using a pre-trained pipeline and evaluates the model's performance by computing a confusion matrix, precision, and recall against discretised ground-truth labels derived from review star ratings, all within a map-reduce framework using Apache Spark. Also creates a log file called `script.log`.

Steps to run code

- Open a Linux environment.
- Run the following commands:

```
sudo apt update
sudo apt upgrade
sudo apt install python3
sudo apt install python3-pip
sudo apt install openjdk-11-jdk
pip install -r requirements.txt
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export PYSPARK_PYTHON=python3
```

- Download the `Gourmet_Foods.txt` data file from the course GitHub repository. The file can be found at `examples/data-engineering/Gourmet_Foods.txt`.
- Run the script using the command:

```
python3 sentiment_analysis.py
```

- Once the script has finished running, check the logfile `script.log`.