DA5402: Assignment 9

We learned about Apache Spark for parallelization and distributed computing. In the classroom training, we used an Amazon reviews dataset to compute the average star-rating for a Gourmet product. The dataset is already shared with you. Let's spice up the use case and solve it.

Task 1 [30 points]

Let's use pretrained sentiment analysis pipeline to reprocess the review texts from the data file. Getting Started with Sentiment Analysis using Python provides a good introduction to sentiment analysis. Let's use the solution presented in Section 2 (sentiment-analysis pipeline) of that article.

Follow the map-reduce paradigm to script the distributed processing of the datafile for sentiment analysis. You will process every record for sentiment classification using the `pipeline` from the article into POSITIVE or NEGATIVE label. The records should be processed in a parallel processing style across the available CPUs in your machine.

Task 2 [20 points]

We have a rating for each item in the dataset, which needs to discretized in to POSITIVE and NEGATIVE label. Let's use rating >= 3.0 as the threshold for becoming POSITIVE. Define a map-reduce logic to compute the Precision and Recall of the sentiment classifier model, assuming that the labels from the dataset are the ground truth. Display the confusion matrix.

Please follow the usual routines in maintaining your code's neatness.