Assignment 2

APACHE-SPARK

Due by June 21, 2023

1. **Note:**

This assignment needs to be done by using Pyspark or SQL Spark. Submit a compressed archive (zip, tar, etc.) of your code, along with the input and output files and output screenshots (output/input commands with results). Please include a pdf document with answers to the questions below.

PART A:

1. [Marks: 10] Count the odd and even numbers using the file 'integer.txt' and download it from Quercus. Show your code and output.

2. [Marks: 10] Calculate the salary sum per department using the file 'salary.txt' and download it from Quercus. Show the department name and salary sum. Show your code and output.

3. [Marks: 10] Implement MapReduce using Pyspark on file 'shakespeare.txt' and download it from Quercus. Show how many times these particular words appear in the document: Shakespeare, why, Lord, Library, GUTENBERG, WILLIAM, COLLEGE and WORLD. (Count exact words only (marks will be deducted for incorrect lowercase/uppercase))

4. [Marks: 10] Calculate the top 15 and bottom 15 words using the file 'shakespeare.txt' and download it from Quercus. Show 15 words with the most count and 15 words with the least count. You can limit by 15 in ascending and descending order of count. Show your code and output.

PART B:

The purpose of this part is to work with a distributed recommender system. To do this, create a recommender system using Apache Spark. Things that were taken into consideration were the efficiency of the systems as well as Spark's complexity.

Data input

For part B implementation, the dataset is provided to you, download it from Quercus.

- movies.csv

Implementation

Load Dataset and import required libraries. Create a recommendation system using a collaborative filtering approach and answer the following questions.

1. [Marks: 10] Describe your data. Calculate the top 20 movies with the highest ratings and the top 15 users who provided the highest ratings. Show your code and output.

2. [Marks: 10] Split the dataset into train and test. Try 2 different combinations e.g. (60/40, 70/30, 75/25 and 80/20). (Train your model and use a collaborative filtering approach on 70 percent of your data and test with the other 30 percent and so on). Show your code and output.

3. [Marks: 10] Explain MSE, RMSE and MAE. Compare and evaluate both of your models with evaluation metrics (RMSE or MAE), show your code and print your results. Describe which one works better and why?

4. [Marks: 20] Now tune the parameters of your algorithm to get the best set of parameters. Explain different parameters of the algorithm which you have used for tuning your algorithm. Evaluate all your models again. Show your code with the best values and output.

5. [Marks: 10]: Calculate the top 15 movie recommendations for user id 10 and user id 14. Show your code and output.