Assignment 2

APACHE-SPARK

Due by Sunday, Feb 21, 2021

1. Note:

This assignment needs to be done by using Pyspark. Submit a compressed archive (zip, tar, etc.) of your code, along with the input and output files and CLI 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 file 'integer.txt' download it from the Quercus. Show your code and output.

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

3. [Marks: 10] Implement MapReduce using Pyspark on file 'shakespeare.txt' download it from the Quercus. Show how many times these particular words appear in the document: SHAKESPEARE, GUTENBERG, WILLIAM, LIBRARY, COLLEGE, WORLD and THIS.

4. [Marks: 10] Calculate top 20 and bottom 20 words using file 'shakespeare.txt' download it from the Quercus. Show 20 words with most count and 20 words with least count. You can limit by 20 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 collaborative filtering approach and answer following questions.

1. [Marks: 10] Describe your data. Calculate top 10 movies with highest ratings and top 10 users who provided highest ratings. Show your code and output.
2. [Marks: 10] Split dataset into train and test. Try 2 different combinations for e.g. (60/40, 70/30, 75/25 and 80/20). (Train your model and use 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] Compare and evaluate both of your models with evaluation metrics (RMSE or MSE) show your code and print your result. Describe which one works better and why?
4. [Marks: 30] Now tune parameters of the 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 best values and output.
5. [Marks: 20] **Bonus**: Calculate top 15 movies recommendations for user id 11 and user id 23. Show your code and output.