Aum Sri Sai Ram

MDSC 304(P): Practical Hadoop Programming

Assignment III

Opening date:1st Oct

Due date: 10th Oct 2021

Follow Academic Integrity and Honour Code.

Note: Dataset files are named as per the exercise number.
The link for the respective datasets: https://drive.google.com/drive/folders/1sJ2wXPC-wekVp2OtO_mDfC2iu2DS6GF7?usp=sharing

Exercises

1) Given an Input data set with name, age and city
   if age > 18 add a new column that's populated with 'Y' else 'N'

2) Given the input file where columns are stationId, timeOfTheReading, readingType, temperatureRecorded, and few other columns...

   We need to find the minimum temperature for each station id.

3) Given the input data, find the top 10 words searched.

4) Find the top 10 customers who spend the maximum amount on shopping.

   File Details
   CustomerId, ProductId, Amount
   Map the above column name in the data respectively.

5) Given a data set, write a spark code to answer below questions.
   - How many times do movies rated 5 stars?
   - How many times do movies rated 4 stars?
   - How many times do movies rated 3 stars?
   - How many times do movies rated 2 start?
   - How many times do movies rated 1 start?

Data Set Columns Respectively: userid, movieid, rating, timestamp