BIG DATA ANALYTICS

Lab Manual 08

SPRING 2023

Input File

You are given a huge **logfiles** of NUCES students accessing google classroom. We want to analyze the file and extract some basic statistics and useful patterns from it.

The format of the input file is as follows:

RollNo Login time Logout time list of pages accessed during this time

Example Input file

L20-4305 Course: BigData Sem: Spring2020 Login:12-03-20-12:45 Logout:12-03-20-2:45 Accessed: stream, assignment
L20-1111 Course:DataMining Sem:Spring2020 Login:12-03-20-11:00 Logout:12-03-20-12:00 Accessed: quiz, material, assignment
L20-4305 Course:DataMining Sem:Spring2020 Login:12-03-20-12:00 Logout:12-03-20-2:00 Accessed: quiz, material, assignment
L20-1111 Course:DataMining Sem:Spring2020 Login:12-03-20-2:00 Logout:12-03-20-3:00 Accessed: quiz, material, assignment
L20-4305 Course:BigData Sem:Spring2020 Login:12-03-20-12:00 Logout:12-03-20-1:00 Accessed: quiz, material, assignment

Question 1: (10 marks)

Write a Map Reduce algorithm to find the number of times a student accessed his each class on google classroom during year 2020. You have to provide the pseudo-code for Mapper, Reducer and Combiner. You can use **associative memory (array) in Mapper** to make your program efficient.

Output for given input

L20-4305 Course: BigData Sem: Spring2020 2
L20-4305 Course: DataMining Sem:Spring2020 1
L20-1111 Course:DataMining Sem:Spring2020 1

Question 2: (10 marks) For each course output the number of distinct students who have accessed the course classroom Write an efficient MapReduce algorithm to perform above task.

Output:

BigData 1
DataMining 2

Question 3: (10 marks)

For each student find the percentage of time spent by the student in a course in Spring 2020. Write an efficient MapReduce algorithm to perform above task.

Percentage of the time spend by a student A in course B =

$$\frac{\text{Time spend by the student B in a course B classroom}}{\text{Total time spent by all students in course B classroom}}$$

Output format for above Question
(Student rollnumber, course -> Percentage of the time spend by a student A in course B (sem C)**)**

Output for given input
L20-4305, BigData -> 100%
L20-4305, DataMining -> 60%
L20-1111, DataMining -> 40%

Hint: This problem is similar to relative frequency word co-occurrence problem discussed in the class.