Assignment #4

Date: 31st May, 2022

Due date: 6th June 2022 Total Marks: 60

Late Submission and Plagiarism Policy

- Late submissions are not encouraged. Late submission with in 24 hours of due date will get 30% Marks deduction. After 24 hours all submissions will be marked zero
- You need to submit the code files named after Question#, a pdf report file and a demo video of running code.
- Viva may be taken for any assignment(s) with suspected plagiarism

Problem # 1 [Marks: 20]

You are provided with file containing information regarding FYPs from Fall-2019 to Spring-2021 for FAST NUCES Islamabad. In this question you need to answer following questions:

a) Design a MapReduce Algorithm (Pseudo-code and Java Code) to find what are the most common pair of technologies that are used together (You need to find the Bigrams). Tools and Technologies used in one project is given in Comma separated format in the CSV file.

b) Describe the algorithm designed. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the output (key, value) pairs of the map stage are processed by the reduce stage to get the final answer(s). (Word count: 200 words)

Problem # 2 [Marks : 40]

You are provided with file containing information for NUCES admission stats. File contains data of applicants of all over Pakistan for all the disciplines. There are many empty fields you should devise a mechanism to handle them. University administration is interested in knowing answers to the questions mentioned below to define better future policies. Design a mechanism to answer the questions below using map-reduce. You need to provide pseudo code and Java Code for all the questions below.

a) Which BS program is a popular choice for admission (campus wise analysis required)? You need to find the count of applicants for each campus.

b) Find which region's Intermediate board performed best in the admission test by finding the average marks scored by students? You can categorize the boards into five regions (4 provinces + Federal). There are two types of Tests in the file. NU Admission Test or SAT Test Score. You need to find the average scores of all boards are find the board having highest average scores in the Admission Test.

Describe the algorithm designed for both the cases. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the output (key, value) pairs of the map stage are processed by the reduce stage to get the final answer(s). (Word count: 300 words)

What to submit:

- You need to submit a pdf file containing the report. Attach Screenshot of the results in the report.
- Java file of each program. Follow the naming conventions I19-xxx_QNumber_Part
- A demo video of the code