Big Data Analytics Techniques and Applications

Homework 3
Due Date: 2022/04/27 23:59:59

- Dataset
    1. “You’ve got to find what you love" text file. You can download the file on E3.
    2. NYC Taxi data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

  Practice Spark programming on the Hadoop platform. You may choose either one program language from Java, Scala, and Python to implement your program on Spark as follows.

- Questions:
    - Q1: Implement a program to calculate the average occurrences of each word in a sentence in the attached article (Youvegottofindwhatyoulove.txt).
        A. Show the top 30 most frequent occurring words and their average occurrences in a sentence.
        B. According to the result, what are the characteristics of these words?
    - Q2: In **YARN cluster mode**, implement a program to calculate the average amount (“Total_amount”) in credit card trip and cash trip for different numbers of passengers, which are from one to four passengers in 2018/10 NYC Yellow Taxi trip data. In NYC Taxi data, the "Passenger_count" is a driver-entered value. Explain also how you deal with the data loss issue.
    - Q3: Referring to Q2, monitor HDFS and YARN metrics through HTTP API; collect MapReduce counters-related information through the web UI. Please provide screenshots and observations regarding the metrics in your report. (Read through this guide to finish the question).

- Requirements
    - Submit a report named **“HW3_StudentID.pdf” and your source code** to E3 and describe clearly the following items:

        - The execution results by using Spark (Attach source code)
        - Descriptions of how you solve each question in detail.
        - Some figures or tables to illustrate your analyzed answers to each question.
        - Anything else worth mentioning (e.g., other valuable observations, or difficulties encountered in this work and how you resolve them).

- Penalty for late submission
    - If your work is submitted within one day after the deadline, a penalty of 20 percentage marks will be applied.
    - If your work is submitted within two days after the deadline, a penalty of 50 percentage marks will be applied.
    - If your work is submitted over two days after the deadline, you will get 0 in this homework.