Big Data Analytics Techniques and Applications

Homework III

Due Date: 2020/12/01 23:59:59

Goal

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

1. 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?
2. Implement a program to calculate the average amount in credit card trip for different number of passengers which are from one to four passengers in **2017.09** 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.
3. For each of the above task 1 and 2, compare the execution time on **local worker** and **yarn cluster**. Also, give some discussions on your observation.

Dataset

Taxi data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Requirements

1. Submit a zip file named “Hw3_StudentID}.zip” that includes the following items:
   - Source codes (including comment)
   - A report of PDF or Word file
     - Program workflow
     - Execution commands
     - Answers to Questions 1~3
     - Anything else worth mentioning (e.g. other valuable observations)