INF 551 – Spring 2018

Homework #5: MapReduce & Spark

Due: April 22, Sunday (end of day)
100 points

In this homework, we consider a dataset "accounts.json" which contains the information of some bank accounts. For each account, it records its account number, name, age, gender, address, employer, email, city, and state of account holder, and account balance.

1. [50 points] Write a MapReduce program "Average.java" that computes, for each different state, the average balance of accounts owned by people of the age between 20 and 30 (inclusive). You should make use of combiner. Combiner will compute local sum and local count. Reducer will then compute global sum and count and then derive average.
   Execution format:

   hadoop jar avg.jar Average <input-dir> <output-dir>

   Note that the dataset may be split and stored as multiple files under the <input-dir> folder. Folder "accounts_task1" is provided for testing.

   Output format:

   WA 23243.0
   WI 34212. 666666666668
   …
2. [50 points]
   a. [25 points] Write a Spark program "avg.py" to compute the same average as Problem 1.
      Execution format:

      spark-submit avg.py <input-file> <output-file>

      Note that <input-file> is a single file that stores all the accounts that you are asked to compute the average. <output-file> stores the output data. File "accounts_task2.json" is provided for testing.

      Output format:

      WA,23243.0
      WI,34212. 6666667
      …
   b. [25 points] Write a Spark program "index.py" that creates an inverted index for the address field of given account data. It writes the index on disk.
      Execution format:

      spark-submit index.py <input-file> <output-file>

      Note <output-file> stores the index in the following format:

<word>:<list of account_numbers>

All words are in lower-case and do not contain numbers

For example,

madison:[13,22]

quentin:[32]

…

Deliverables:

1. Source code for all questions and the jar file for q1: Average.java, avg.py, index.py, avg.jar
2. Output files: q1.txt(renamed from Hadoop output part-r-00000), q2-a.txt, q2-b.txt

Important Notes:

1. Please prepend your name to all the submission files as before to facilitate the grading. e.g. firstname_lastname_q1.txt, firstname_lastname_avg.py … **DO NOT** place them in a folder or zip file.
2. For q1, **please do not use any library other than org.apache.hadoop.*, java.***
3. For q2, **please use python 2.7** and do not use any library other than Python Standard Library.
4. Please submit your **7 files** to Blackboard by the due time and note that only your last attempt will be viewed and graded.