CS226 Assignment 2: SparkRDD

Description

The goal of this assignment is to write some Spark RDD jobs and observe their running times.

Details

The input file is a log file that is stored in a tab-separated-value (TSV) format with the following attributes: ⟨host, logname, time, method, URL, response, bytes, referrer, useragent⟩. Each line contains a record with the fields separated by the tab character '\t'. You need to implement the following two queries in Spark RDD using Java, Scala, or Python.

A. (Grouped aggregation) Find the **average** number of bytes for lines of each response code. The answer should be in the following format:

Code 200, average number of bytes = 17230.603516

Code 304, average number of bytes = 0.000000

Code 404, average number of bytes = 0.000000

Code 302, average number of bytes = 73.253525

B. (Self Join) Find pairs of requests that ask for the same URL, same host, and happened within an hour of each other (i.e., difference in time $\leq$ 3600). In other words, it produces all the pairs of the tuples (t1, t2) that satisfy the following conditions.
   a. t1.host = t2.host
   b. t1.url = t2.url
   c. |t1.timestamp – t2.timestamp| <= 3600
   d. t1 != t2

For example, the following two records should appear in the answer because they satisfy the join conditions.

("n1123083.ksc.nasa.gov", "-" , 807294692, "GET", "/ksc.html", 200, 7280)
("n1123083.ksc.nasa.gov", "-", 807295150, "GET", "/ksc.html", 200, 7280)

The output is produced as a file that has one pair of records per line which lists all the attributes of the first record (separated by tab) then another tab character, and finally the attributes of the second record separated by tabs.

Submission instructions

- The assignment is due on Wednesday, 11/25/2020, at 11:59 PM Pacific Time.
- Late submissions are allowed with a 20% penalty for each calendar day up-to four days late.
- You can use either Java or Scala for this assignment.
- A sample file can be accessed on the following link: http://www.cs.ucr.edu/~eldawy/20FCS226/demos/nasa.tsv

- Please upload your answer in a single ZIP file named ‘cs226-asg2-<UCRnetID>.zip’ where <UCRnetID> is replaced with your ID.
- The ZIP file should contain the source code and an executable script that takes one parameter, the input file path, and does all of the following in order:
    - Compile the code into a binary file, e.g., JAR
    - Run the program on the given file and produces two output files named ‘task1.txt’ and ‘task2.txt’. While running the program, assume that ‘spark-submit’ is in the executable path.
- Failing to follow the instructions above might result in losing some points.