CSC 4760/6760 Big Data Programming

Assignment 5

Due Date: 11:59 pm, April 7, 2021

1. (100 points) (Counting Tweets)

Input Datasets:

Tweets (tweets.json):

user	geo	tweet
Bob	Atlanta	It is a sunny day!
Susan	Athens	We have a football game today :)
David	Atlanta	Today is cold.
Lisa	Auburn	I love Auburn University
Ben	Birmingham	I will go to Atlanta today!
Paul	San Francisco	We watch a movie today!
Smith	San Diego	It is hot today. Summer comes.
Ethan	Log Angeles	Oscar ceremony is wonderful!
Emma	Log Angeles	I love Oscar ceremony!
Rolando	Orlando	I will go to the beach!
Mia	Miami	Sunny Day!

City and State lookup table (cityStateMap.json):

city	state
Atlanta	Georgia
Athens	Georgia
Miami	Florida
Orlando	Florida
Birmingham	Alabama
Auburn	Alabama
Log Angeles	California
San Francisco	California
San Diego	California

Problem and Output Data:

We want to count the number of tweets published in each state. The following table shows the desired results.

state	count
Georgia	3
Florida	2
Alabama	2

California	4

Implementation:

Design and implement a PySpark program to solve the problem. We did not provide any template python file this time. You may want to create one python file from scratch.

You are required to use Spark Dataframe to implement this function.

Report:

Please write a report illustrating your experiments. You need to explain your basic idea about how to count tweets in each state. You may add comments to the source code such that the source code can be read and understood by the graders.

In the report, you should include the answers to the following questions.

1) Explanation of the source code

2) Experimental Results

2.1) Screenshots of the output. Since we plan to use Dataframe in Spark, it is easy to type in "DF.show()" to visualize the table in the terminal. Please do so and take a screenshot of the output in the terminal. The screenshot "output.PNG" of the output in my VM is given. You can use it to verify your outputs.

2.2) Explain your results. Does your implementation give the right answer?

Submission Materials:

a) Your report

b) Source code (.py file)

c) The screenshot of the outputs in the terminal