CSC 4760/6760 DSCI 4760 Big Data Programming

Assignment 3

Due Date: 10.30 pm ET, 11/13/2022

1. (100 points) (Counting Tweets)

Input Datasets (Two are provided tweets.json and cityStateMap.json**):**

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:

As the final output - 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:

Implement a PySpark program to solve the problem. We have provided an *almost complete* python code. You must implement the code in a JuPyter Notebook.

"*almost complete*" phrasing is important here – some Hints:

1. **The python code provided does not include the required statements on the findspark package needed on your JuPyter Notebook cell.**
2. **The python code, although correct, may or may not be ready to use as is (Python is strict with indentation issues, for example). It is strictly part of your assignment to verify.**

You must use Spark Dataframe to implement this function, as in the code provided.

Report:

Please write a report illustrating your steps of execution. **Uploading screenshots to iCollege is NOT writing a report**. From the implementation, you need to explain the basic ideas to count tweets in each state. To the existing comments, you may add your own comments to the source code.

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

1) Your Explanation of the source code in addition to the comments provided -with screenshots

2) Your Results

Submission Materials:

a) Your report in PDF containing the screenshots of the outputs

b) Source code (JuPyter Notebook)