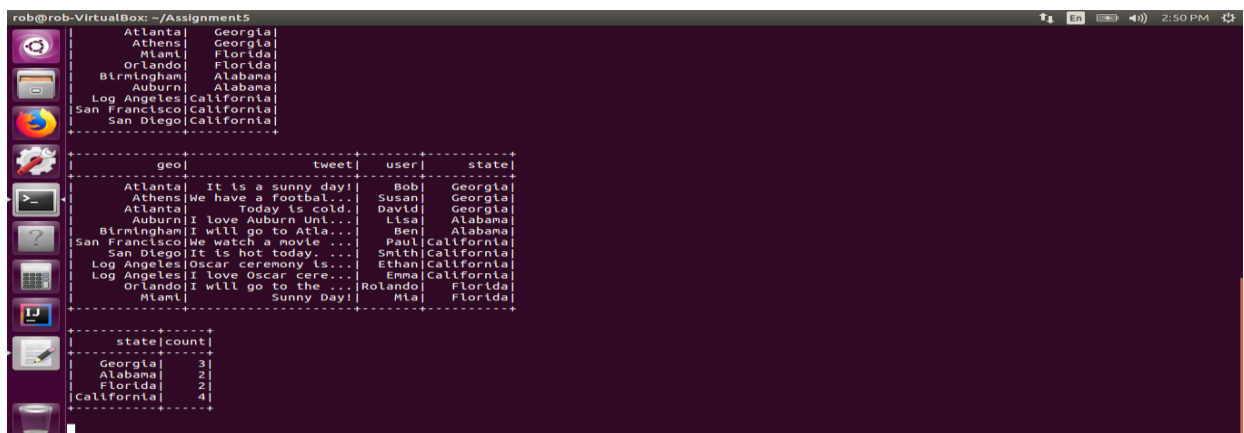John W. Martin
April 4, 2018
CSc 6760 – Big Data Programming
Dr. Yubao Wu
Homework #5

For this homework assignment, I used PySpark to count the number of tweets published in each state. Two initial tables were given, one showing a list of tweets by user and city, and the other showing a list of cities and states. The two tables were joined to display tweets by state, and then the number of tweets published in each state. The "city" field in the second table was joined to the "geo" field in the first table, and this gave the desired results. Furthermore, the Spark DataFrame was implemented to show the data in each table

Source Code Explanation:

The source code implements the use of the Spark DataFrame to read data. Two dataframes were created to show the data in the JSON tables supplied for this assignment: tweets.json and CityStateMap.json. A third dataframe was then created to join the initial dataframes, joining the city fields in the two tables and dropping the city field in the second table in the joined table. This determines which state each city is in, given the data. A fourth dataframe is then created to count the number of tweets in each state, by using the groupBy command. The output of the third and fourth tables is shown in the screenshot below: