

BIG DATA PROGRAMING

ASSIGNMENT-5

1. As a Primary step of this assignment spark context and session is created After this The two json files provided (tweets and city maps) are read as Data frames using spark session and this data frames are shown as below.

Once the data frames are created these data frames contains same columns with different column names these columns are used to join based on the city names and after they had joined there will be an extra column of the city that is city name's column which is dropped off to remove the extra column The joined output data frame is as shown in section two figures

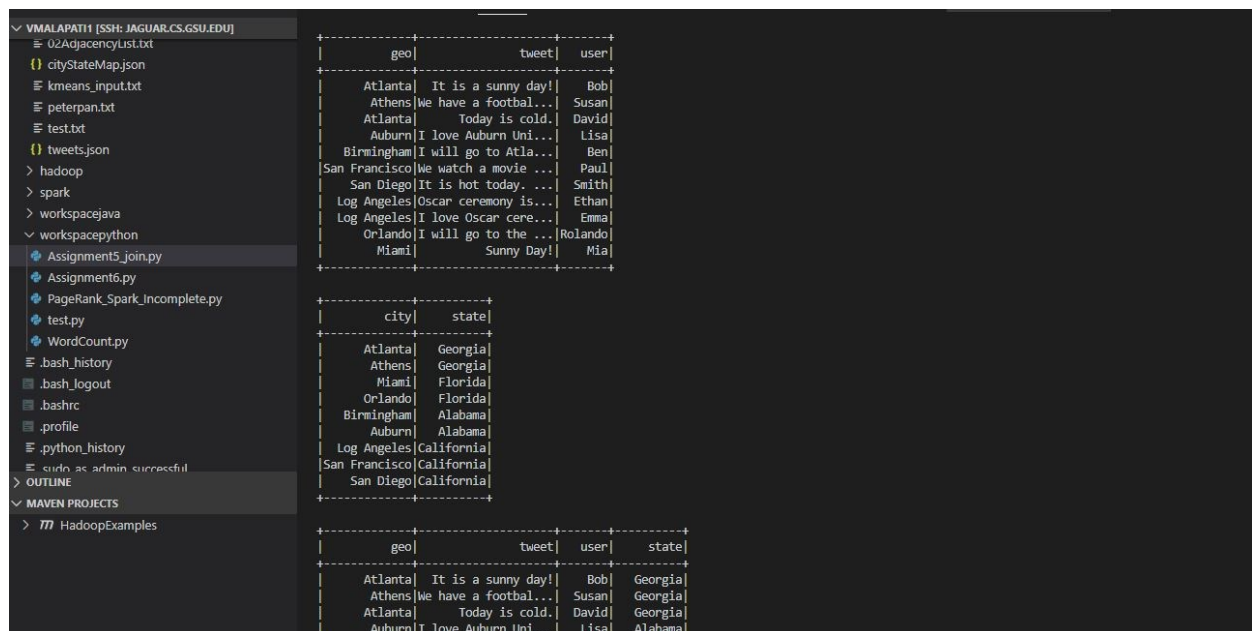
Once the columns are joined, the data frame is grouped by states such that the cities in each state fall in one place and the count of these cities are calculated and displayed as shown below in the data frame.

```
joined = tweets.join(citymap, tweets["geo"] == citymap["city"]).drop('city')
```

this is the code that plays a main role in joining data frames

2. The results are same as the results shown in the output image contains states and number of cities in each state as data frame.

The below are the screenshot showing data frames output in kvm.



geo	tweet	user
Atlanta	It is a sunny day!	Bob
Athens	We have a footbal...	Susan
Atlanta	Today is cold.	David
Auburn	I love Auburn Uni...	Lisa
Birmingham	I will go to Atla...	Ben
San Francisco	We watch a movie ...	Paul
San Diego	It is hot today. ...	Smith
Log Angeles	Oscar ceremony is...	Ethan
Log Angeles	I love Oscar care...	Emma
Orlando	I will go to the ...	Rolando
Miami	Sunny Day!	Mia

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

geo	tweet	user	state
Atlanta	It is a sunny day!	Bob	Georgia
Athens	We have a footbal...	Susan	Georgia
Atlanta	Today is cold.	David	Georgia
Auburn	I love Auburn Uni...	Lisa	Alabama

