Lab 11: Summarizing the number of Tweets in each state

Using Spark to work with Shapefile operations

1. Outline

In this lab, you will perform the task for summarizing the number of tweets authored in each state in the USA (excluding Alaska and Hawaii)

2. Materials

The data and scripts are stored in: /gpfs_scratch/geog479/lecture11

Before copying this folder to your home directory:

Delete the same folder in the your home directory from previous lecture:

```
>> rm -r ~/lecture11
```

And then copy

```
>> cp -r /gpfs_scratch/geog479/lab11 ~/
```

3. Tasks

Task 1:

- Getting to know Shapefile.py (https://pypi.python.org/pypi/pyshp)
- It is located in the “demo” folder
- Read the script and understand what it does
  >> nano createShapefile.py
- Execute the script
  >> python createShapefile.py
- Visualize in ArcGIS
  scp -r netID@roger-login.ncsa.illinois.edu:~/lecture11/demo/shapefiles Desktop/

Task 2:

- Summarizing the number of tweets in each state during January 1st, 2014
- Navigate to the folder at ~/lecture11
- 1. You need to prepare the data
- 2. You need to read the script to figure out the flow and fill in the blanks to complete the code
- 3. I will explain it before execution.
- Run the script and visualize the results in ArcGIS
- spark-submit --master yarn-client --executor-memory 5g spark_pinP.py
- What about visualize the results with D3.js?