Karl Roth
GEOG 479
Lab 11

In this lab we explored the use of PIG grunt shell for processing twitter data and learned how to run several pig scripts on HDFS. Figure 1 demonstrates the files created in our HDFS for this laboratory. The lab 11 folder was imported in from ROGER and the result.txt and result_keywords.txt file were created. The result.txt file was created by loading the data into the grunt shell "cleaning" the data and filtering out only the tweets originating in New York. The result_keywords.txt file was created by limiting to tweets that contain the word keywords "happy" or "sad". Fig 2 demonstrates the contents of results.txt and result_keywords.txt respectively.

```
drwx------   - karoth4 hdfs          0 2017-11-15 00:00 .Trash
drwx------   - karoth4 hdfs          0 2017-11-24 11:17 .staging
drwxr-xr-x   - karoth4 hdfs          0 2017-11-24 10:50 lab11
-rw-r--r--   3 karoth4 hdfs 2781761974 2017-11-14 15:49 ny_taxi_1.csv
drwxr-xr-x   - karoth4 hdfs          0 2017-11-24 11:08 result.txt
drwxr-xr-x   - karoth4 hdfs          0 2017-11-24 11:17 result_keywords.txt
```

Fig 1: The contents of the HDFS after lab 11 was completed.

The contents o

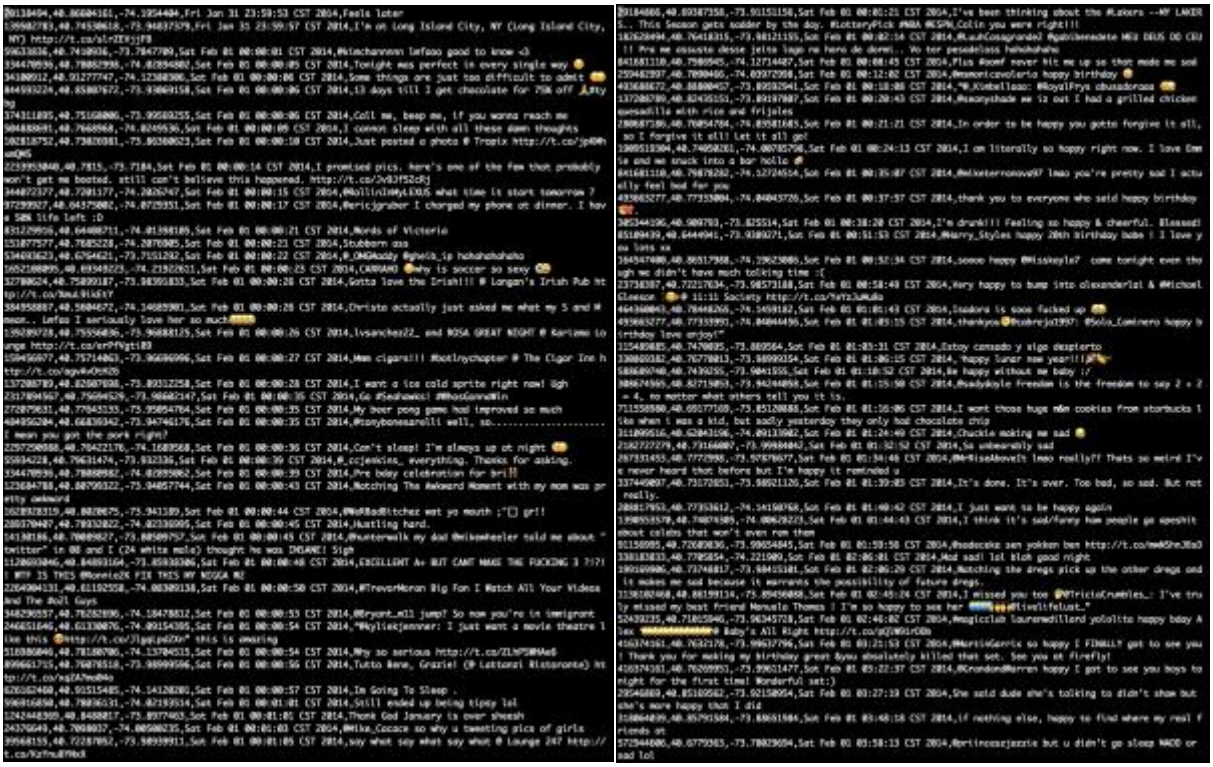

Fig 2: the contents of result.txt (left) and result_keywords.txt (right)

Then we ran the heatmap.pig script. This file loads radiation data from HDFS, cleans the data, and filters it for radiation and locations data. The area is separated into a grid with each block having a latitude and longitude “key”. The radiation value is then averaged over each block on the grid using the user-defined function “.getAverage()” and the result is outputted into the heatmap_output folder in HDFS.

Finally, the kmeans.pig script was run. This file loads the data from HDFS and cleans the data. Then the “request” field is filtered for entries regarding “Graffiti Removal”, but other requests also exist. The data is then filtered by time and streamed through the kmeans.py script. The kmeans.py script uses the kmeans clustering method to “group” the instances of Graffiti Removal geographically. Each entry is assigned a group number from 0-4 inclusive. The output is then returned to the pig script where it is saved in HDFS. Streaming data through a python script allows us to use simple python code and industry standard packages to accomplish tasks not achievable by PIG alone.