Twitter Data Analyzing Package

This is a comprehensive package for analyzing tweets. It can do tasks such as block number label & count, drawing heat map, and import/export data to/from hive and do extra information. It can also perform several natural language processing analysis such as sentiment analysis and n-gram analysis.

How to compile

Prerequisites
Python [>=2.7.9] (https://www.python.org/downloads/)
pip [>=0.9.0] (https://pip.pypa.io/en/stable/installing/)
NLTK [>=3.0] (http://www.nltk.org/install.html)
Java Development Kit [>=SE 7] (http://www.oracle.com/technetwork/java/javase/downloads/index.html)

Install NLTK using pip. The package dependency should be resolved automatically by pip. If met problem, try to install in the local mode by adding --user parameter.

How to use

Block Number Label & Count
The block count code calculates the block number of the input tweets dataset, and append the number as the last field in each record. It also counts block number and outputs to stderr stream, which could be used for head map drawing. Before using, edit the java file and put the correct latitde and longitude range in rangeLati and rangeLongi arraies. Also, put the desired row/column split number in the splitNum array.

```
javac BlockCount.java
cat [tweets file] | java BlockCount >[output file] 2>[block number count file]
```

Drawing Heat Map
The code is actually an excel file :) Simply copy & paste put the block number count result into the excel, and a *heat map* is generated!

Sentiment Analysis
Put the tweets file into the same folder as the sentiment analysis package. Edit the “sentiment analysis.py” if necessary to custimize column name and column location of the tweet. The sentiment analysis module is provided by VADER Sentiment Analysis package (https://github.com/cjhutto/vaderSentiment). Four scores will be generated after the analysis: negative score, neutral score, positive score, and a compound score of the tweet.

```
python2 sentiment_analysis.py [tweets file]
```

N-Gram Analysis
The code is adapted from Jinglin Wang’s n-gram analysis package. This relies on Hadoop MapReduce framework and the input/output data file should be stored on HDFS. Two easy-to-use script files are provided for compiling and running the code on the Hadoop system. Make sure the two script files have enough permission. The input file should be stored in ‘work directory’/input/, and the result will be generated into ‘work directory’/output on the HDFS.

```
./compile.sh NGram
./run.sh NGram [work directory]
```

Hive Import, Query & Export
Detailed information can be found in the comments of the hive import/query/export codes in the hive folder. The Csv2Hive package (https://github.com/enahwe/Csv2Hive) is used to easily generate hive query from .csv file.