Ben Attix
MIDS w205 – Spring 2017
Exercise 2 Architecture

**Application Idea**

The idea of this project is to set up a Twitter streaming application to analyze Twitter data. The application will read and parse live Tweets, split out and count each word, and write the counts of each word to a Postgres table where is can be queried for further exploration.

**Architecture Description**

The application uses Apache Storm with a spout pulling data from the Twitter streaming API. This spout sends data to a "Parse-tweet" bolt where the the words are split out from each other and certain values are removed (hash tags, @, RT, URLs, and some punctuation) and then this bolt outputs the cleaned words. The cleaned words are picked up by the "count" bolt where each occurence of a word is counted. The count-bolt outputs the word counts to the screen as well as writes the word counts to a Postgres file.

**Notes for Running the Application**

Before running the application, the user must create a Postgres database called tcount and in that database, have a table called tweetwordcount. These can both be created by running the create-tcount.py program in the exercise_2 folder.

Additionally, the user must input their Twitter credentials in the "Twitter credentials" section of the tweet.py spout before running the application.

Please check the README.txt file for a full description of how to run the application.

Directory and file structures are shown on the next page.

**Directory and File Structure**

```
exercise_2
|-- Architecture.pdf
|-- README.txt
|-- create-tcount.py
|-- Plot.png
|-- finalresults.py
|-- histogram.py
|-- screenshots
|     |-- screenshot-StormRunning.png
|     |-- screenshot-finalresults-Working.png
|     |-- screenshot-postgresPopulation.png
|-- extweetwordcount
|     -- topologies
|        |-- tweetwordcount.clj
|     -- src
|        -- bolts
|           |-- __init__.py
|           |-- parse.py
|           |-- wordcount.py
|        -- spouts
|           |-- __init__.py
|           |-- tweets.py
|     -- _build
|     -- _resources
|     -- logs
|     -- virtualenvs
|     -- config.json
|     -- project.clj
```