Adam Reilly
W205 Exercise 2

Architecture Description

General Architecture Summary- The below section describes the various tools that are used for this assignment. The order of this section follows the flow of the data from the beginning to the end.

1. Twitter Apps- Allows the virtual computer to connect to twitter and receive tweet data
2. Tweet Spout- This spout connects to twitter and transmits the tweet data so that other parts of the virtual computer can use them
3. Parse Bolt- Removes extraneous information from the tweet (which is everything but individual words, which includes retweet notations, @'s, and the author's name) and separates the remainder into individual words
4. Wordcount Bolt- Connects to Postgres and emits both the words and their counts
5. Postgres- Allows for the storage of the words and counts in a retrievable table

Necessary Installed Packages or Dependencies

UCB AMI- UCB EX 2 Full AMI is created in Amazon Web Services. You want to attach a data volume to it (that can be moved around for various AMIs) and ensure that port 5432 (Postgres) is allowed.

Psycopg2- Allows python and postgres to interact. This is used in the Wordcount Bolt, as well as histogram.py and finalresults.py. This is to be installed on the AMI. The description of how to install is in the exercise 2 instructions in the github directory.

Tweepy- Allows for interaction between Twitter and the TweetSpout. This is to be installed on the AMI. The description of how to install is in the exercise 2 instructions in the github directory.

Postgres- An SQL based program allowing for separate databases. This is to be installed on the AMI. Instructions are in Lab 2 in the W205 Berkeley github.

Twitter Keys- Permission keys from Twitter that allows the overall architecture to connect to twitter and pull down the data. These keys are entered in the Tweet Spout and are obtained by signing up for access to developer tools in Twitter.

My Exercise 2 GITHUB Directory- This section outlines the location files in my github directory. Includes whether something is a directory or a file

Exercise_2 (directory in github)
→topologies (directory)
→→tweetwordcount.clj (file)
→src (directory)
→→spouts (directory)
→→→tweets.py (file)
→→bolts (directory)
→→→parse.py (file)
→→→wordcount.py (file)
→screenshots (directory)
→→screenshot-twitterStream.png
→→screenshot-extractResults.png
→→screenshot-histogram.py.png
→architecture.pdf (file)
→exercise_2.pdf(file; this is the exercise2 prompt with instructions)
→finalresults.py (file)
→histogram.py (file)
→plot.png (file)
→readme.txt (file)

Short descriptions of each file in directory (minus Architecture.pdf)
Tweetwordcount.clj- Topologies for Storm. This taken directly from the exercise 2 assignment with no changes.
Tweets.py- Spout that takes information from Twitter. The only change needed is the 4 Twitter Keys.
Parse.py- Directions to clean the data from twitter into words. This is taken directly from the exercise 2 assignment with no changes.
Wordcount.py- Instructions that tell Storm to connect to Postgres, store the words and their counts there, and emit the information in real time as well.
Screenshot-twitterStream.png- Image showing the words & their counts being emitted in real time
Screenshot-extractResults.png- Image showing the results in the tweetwordcount table
Screenshot-histogram.py.png- Image showing the result of running a histogram.py query
Architecture.pdf- File describing various facets of the directories and programs in this github
Exercise-2.pdf- Exercise2 prompt with instructions regarding some package installations as well as how to run both finalresults.py and histogram.py
Finalresults.py- Python file that allows you to get each the count for all words in the postgres table or the count for 1 specified word. Runs on the command line and can (but doesn't need to take 1 argument.

Histogram.py- Python file that allows you to get all the words (and their counts) with counts between two numbers. Runs on the command line and must take 2 arguments
Plot.png- A chart showing the 20 most common words in my dataset
Readme.txt- Instructions on how to run the complete program
Plot.png- A chart of my top 20 words