W205 Exercise 2

Name: Manish Shah

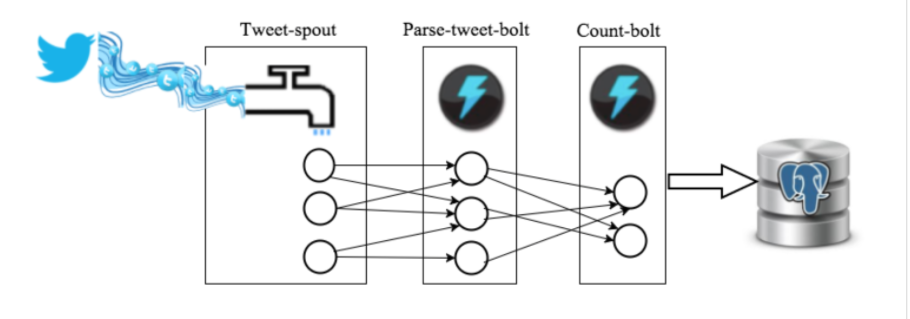
Date: 12/3/2017

# The Application

This application subscribes to a tweeter feed and writes the number of occurrences of these words in the database. The tweets containing the words, ["a", "the", "i", "you", "u"] are tracked and only the tweets in English language are selected.

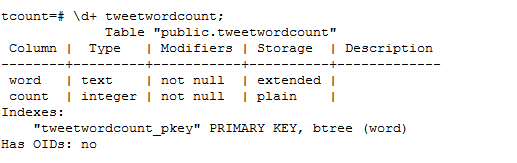
The application uses Storm technology to subscribe to the tweets and process them. The Tweet spout program subscribes to the tweet, queues them and emits it to the Parse bolt. The parse bolt processes the tweet, excludes retweets and tweets with non-ascii characters and creates a list of valid words and emits it to the wordcount bolt. The wordcount bolt updates the database. The application is implemented in python 2.7.

The application topology is illustrated in the diagram below.



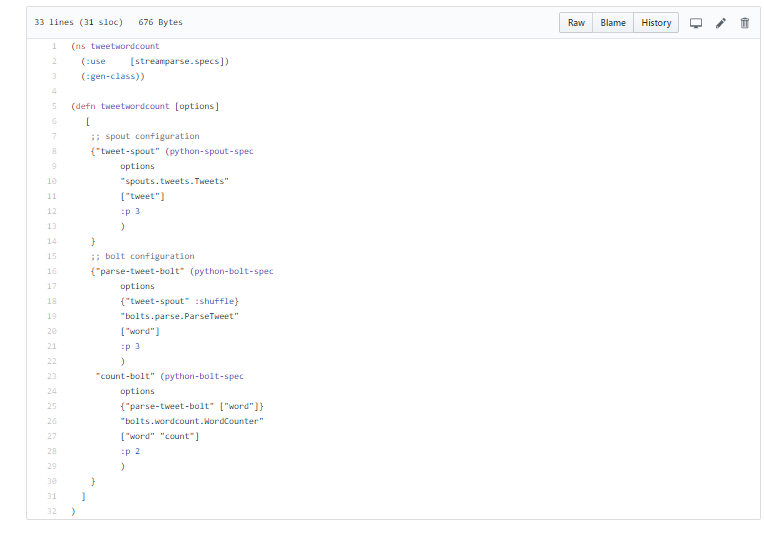
# Database

Postgres database is used to store the words and the counts of the words. The name of the database is tcount and it has one table, tweetwordcount. The tweetwordcount table has the following structure.



# Topology

The application has one spout, tweet, with 3 instances, one parse bolt with 3 instances and one count bolt with 2 instances. The topology is defined as.



# Directory Structure

The directory structure is as follows.

|  |  |
| --- | --- |
| Files/Folders | Description |
| Create\_database.py | Creates the database tcount and table tweetwordcount |
| Extweetwordcount | The streamparse project folder used by streamparse run |
| Final\_results.py | Queries the database for words |
| Histogram.py | Queries the database for words with counts >= k and <= l |
| Main.sh | Main program to run the script |
| Top20.py | Python program to query top 20 words based on count |
| Top20\_plot.py | Python program to plot the bar chart (tested on windows) |
| Plot.png | Top 20 bar graph. |
| Output | Folder for the output files |
| Screen shots | Screen shots of the project |

The topology, spout and bolts are defined in the Extweetwordcount folder.

# Run Instructions:

The analysis has been tested on AWS hosted EC2 instance with 100 GB of diskspace, 8 GB of Ram and 2 processors and using Linux operating system.

The AMI used is "UCB W205 Spring 2016 (ami-be0d5fd4)" .

The application needs tweepy and matlibplot libraries to be installed for python.

To run the program, on the command prompt, type,

./main.sh

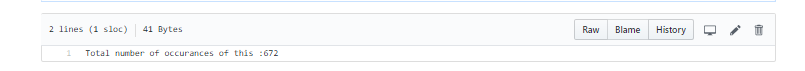
This script will create the database, run the storm application for 5 mins and then run the result scripts.

The main.sh script is shown below.

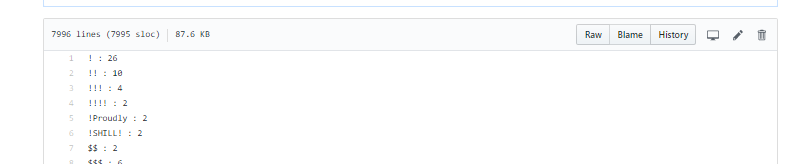


# Output:

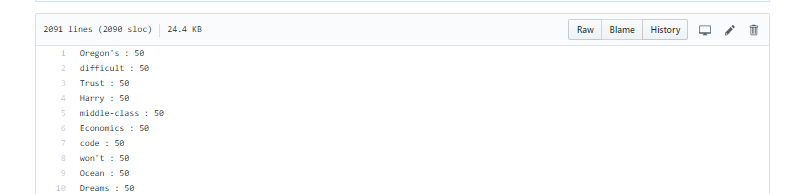
## Output of number of occurrences of word “this”.



## Output of all words sorted alphabetically with counts.



## Histogram of all words >= 20 and <= 50



## Histogram bar graph.

