VERSION 1.0

NOVEMBER 19, 2017

EXERCISE 2

PROJECT TWEET WORD COUNT

PRESENTED BY: K R V, PRAGNESH

SUMMER 2017 GRAD STUDENT
UC BERKELEY

EXERCISE 2

PROJECT COMMUNICATION DOCUMENTS

PROJECT COMMUNICATION TABLE

Document	Recipients	Responsibilities	Update frequency
Executive status report	Presenter	Reviewer	1

TEAM STRUCTURE

Pragnesh KRV - Developer

TEAM GOALS

- Develop a tweet word count application using Apache Storm
- Store the tweet word count results in a POSTGRESQL Database called "TCOUNT" and in a table called "TWEETWORDCOUNT"
- Make a report of words and their counts using Python program "FINALRESULTS.PY". The program should print the word and its count provided the word is passed as an input to the program; otherwise, it should print the entire list of tweet words and their respective counts
- Generate histograms using a Python program to list words and their counts between a certain range
- Generate bar chart of 20 most frequently used words

TEAM ASSIGNMENTS

[Use the following table to outline the project's marketing teams, team goals, team leads, and team roles.]

Project Tweet Word Count

Name of team	Team goals	Team leads	Team roles
Lake and Brown insights	Submit exercise 2	Edward Fine	

APPLICATION ARCHITECTURE

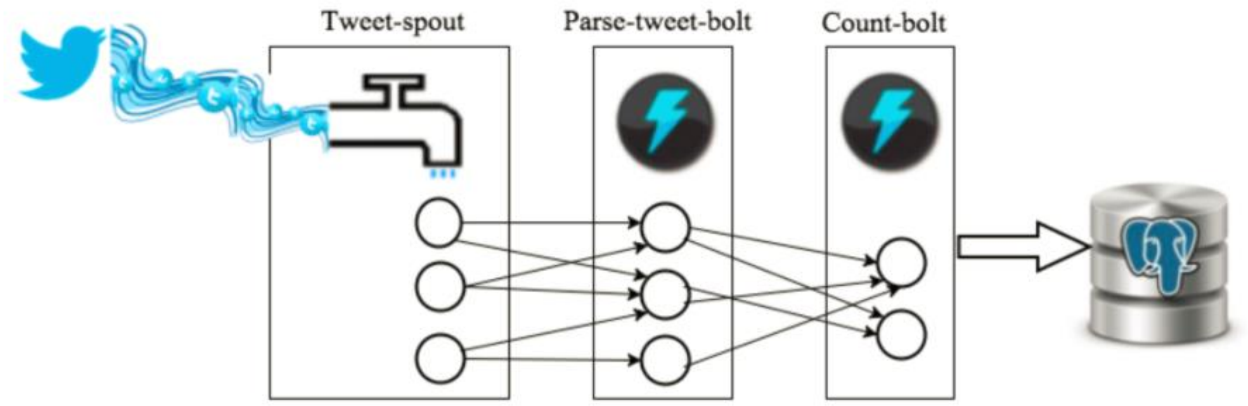


Figure 1: Application Topology

- The spout (3 in number) captures the Tweet and sends it to 3 bolts. These bolts check for valid English words and sends them to the next set of 2 bolts that count these words and store in a POSTGRESQL Database. The name of the Database is TCOUNT with the table name as TWEETWORDCOUNT

STEPS TO RUN THE MAIN APPLICATION

- Copy the code from github
- Sparse run
- After 1 minute, you may logon to POSTGRESQL using the command
- $ psql -U postgres -d tcount
- SELECT * FROM tweetwordcount

STEPS TO RUN FINALRESULTS

- The finalresults.py can be executed with 2 options
- Option 1: $ python finalresults.py
  - Displays all words with the associated wordcount
- Option 2: python finalresults.py <word>
  - Displays the word count

STEPS TO RUN HISTOGRAM.PY

- The histogram.py can be executed with 2 arguments
- Option 1: $ python histogram.py 3 8
  - Displays all words with the associated wordcount between 3 and 8

The program checks whether the arguments are complex numbers, and if yes takes the real part and converts it into an integer and then executes the SQL query

FOLDER STRUCTURE

Exercise_2/src/spouts: contains the code for the spout

Exercise_2/src/bolts: contains the programs for the bolts

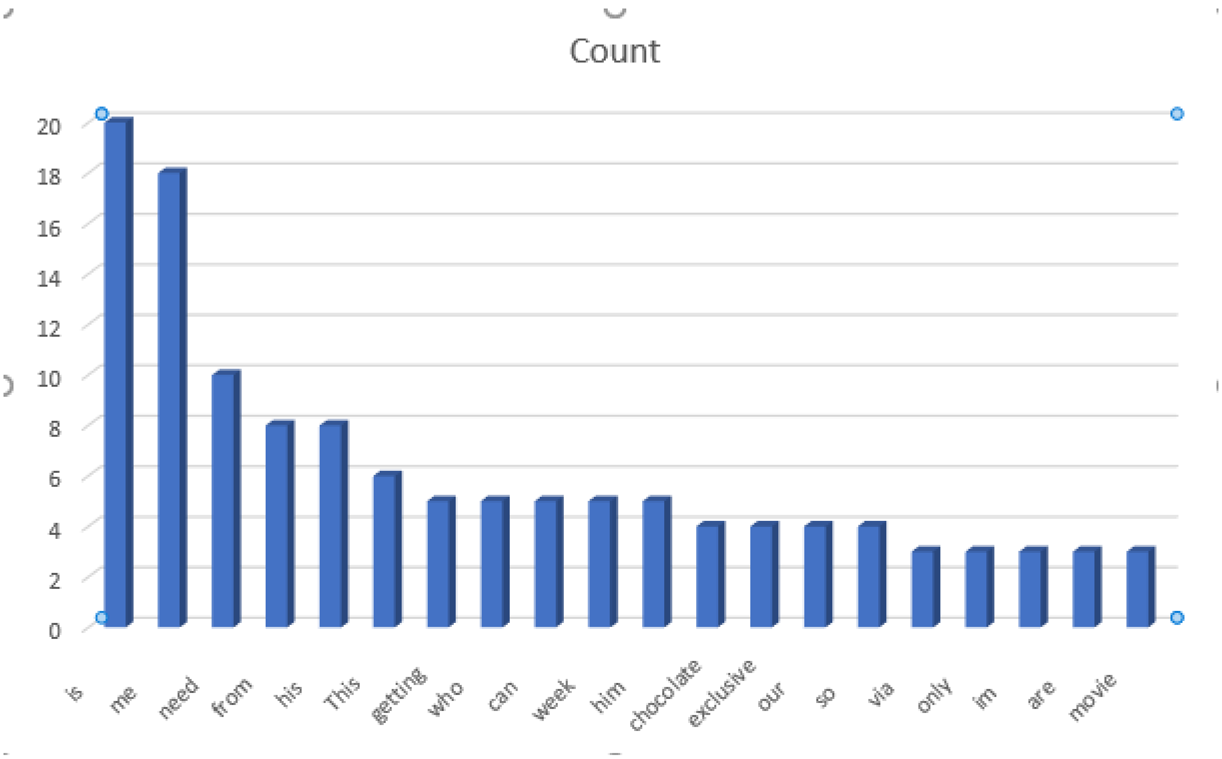
Exercise_2/hello-stream-twitter.py: Using Twitter stream API, print all the tweets in the stream containing the term "Hello" in a 1 min period

Exercise_2/Twittercredentials.py – contains access tokens for the hello-stream-twitter program

Exercise_2/topologies/extweetwordcount.clj has the topology

TOP 20 WORDS

Word	Count
is	20
me	18
need	10
from	8
his	8
This	6
getting	5
who	5
can	5
week	5
him	5
chocolate	4
exclusive	4
our	4
so	4
via	3
only	3
im	3
are	3
movie	3


Count
20
18
16
14
12
10
8
6
4
2
0
is
me
need
from
his
This
getting
who
can
week
him
chocolate
exclusive
our
so
via
only
im
are
movie