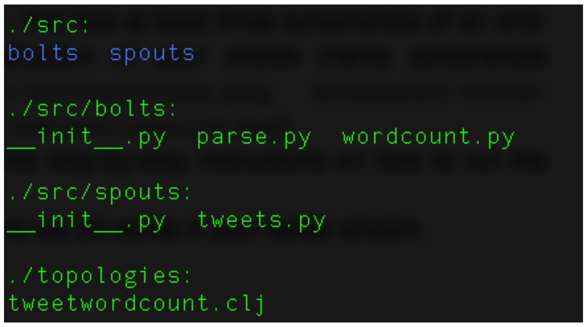
EXERCISE 2 ARCHITECTURE DOCUMENT

Kent Owen
8/9/16

OVERVIEW: FILE STRUCTURE, GOALS OF EACH FILE

Key files are set in the following folders to for the storm setup:



```
./src:
bolts  spouts

./src/bolts:
__init__.py  parse.py  wordcount.py

./src/spouts:
__init__.py  tweets.py

./topologies:
tweetwordcount.clj
```

src/bolts/parse.py: Parses tweets into individual words. Cleans out non-ASCII characters. Makes words lowercase, for consistency.

src/bolts/wordcount.py: Takes results of parse.py words and feeds into postgres sql database, incrementing wordcounts from words that come in.

src/spouts/tweets.py: Connects to twitter API, provides credentials necessary, pulls in twitter feed.

topologies/tweetwordcount.clj: Defines topology to take tweets (tweets.py) > parse (parse.py) > refine counts and upload to postgres database (wordcount.clj).

In main folder:
finalresults.py: Emits wordcount tuples for all or certain words. Explained in more detail later.
histogram.py: Emits wordcount tuples within a certain wordcount range. Explained in more detail later.
plot.png: Simple bar chart plot of top-20 words from when tweetstream was open.
README.txt: Directions on how to run files. These are replicated in this file in the subsequent section.
Architecture.pdf: This document.

Other files (not listed above):
In the "screenshots" folder, there are screenshots showing end-to-end run from sparse run to getting histogram data as well as top-20 words.

ADDITIONAL INFORMATION AND INSTRUCTIONS TO RUN FILE

After you connect to an instance, here are the commands that you need to create the postgres database, then begin querying data through the finalresults.py and histogram.py files:

#NAVIGATE TO Tweetwordcount project folder
cd Tweetwordcount

#FOR INITIAL RUN OF TWITTER STREAM, MUST FIRST CREASE tcount IN POSTGRES before sparse run: *psql -U postgres*
CREATE DATABASE tcount;
\q
sparse run

#RE-RUNNING TWITTER STREAM:
## must delete out table tweetwordcount first then re-run sparse run
psql -U postgres
\c tcount
DROP TABLE tweetwordcount;
\q
sparse run

#### RUNNING PYTHON PROGRAMS ####

#RUNNING finalresults.py
## TO GET ALL WORDCOUNTS:
python finalresults.py

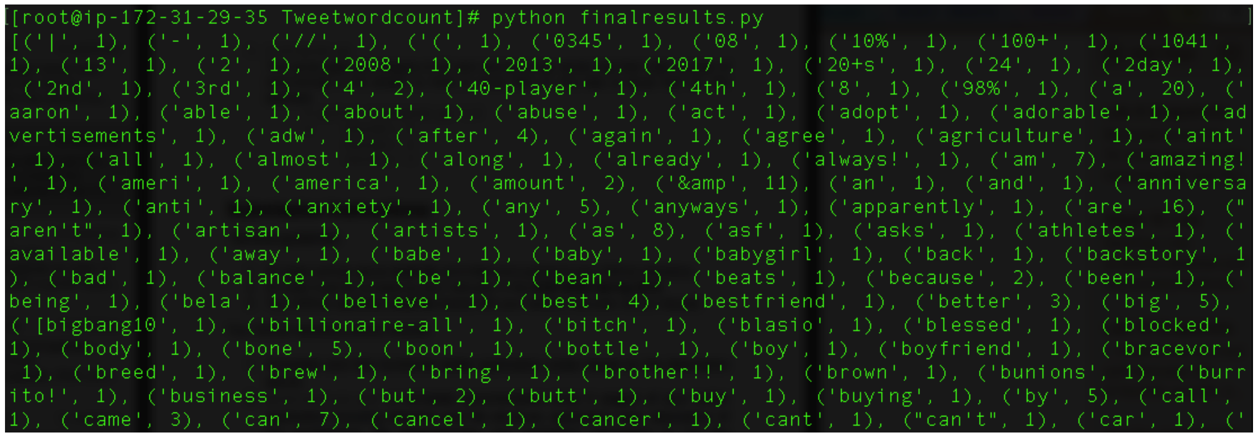
## TO GET CUSTOM WORDCOUNT OF 'day'
python finalresults.py day

#RUNNING HISTOGRAM
python histogram.py 4,8

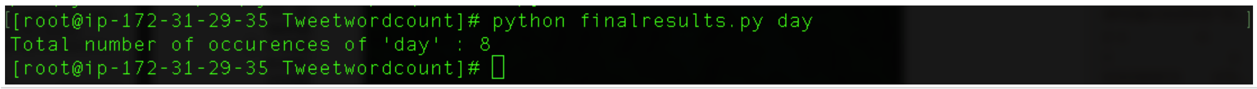
GOALS OF PYTHON QUERY FILES

Finalresults.py: Two main functions.

1) If the user enters no input, return the tuples of all words and counts.
2) If the user enters a specific word, see how many times that word was mentioned.

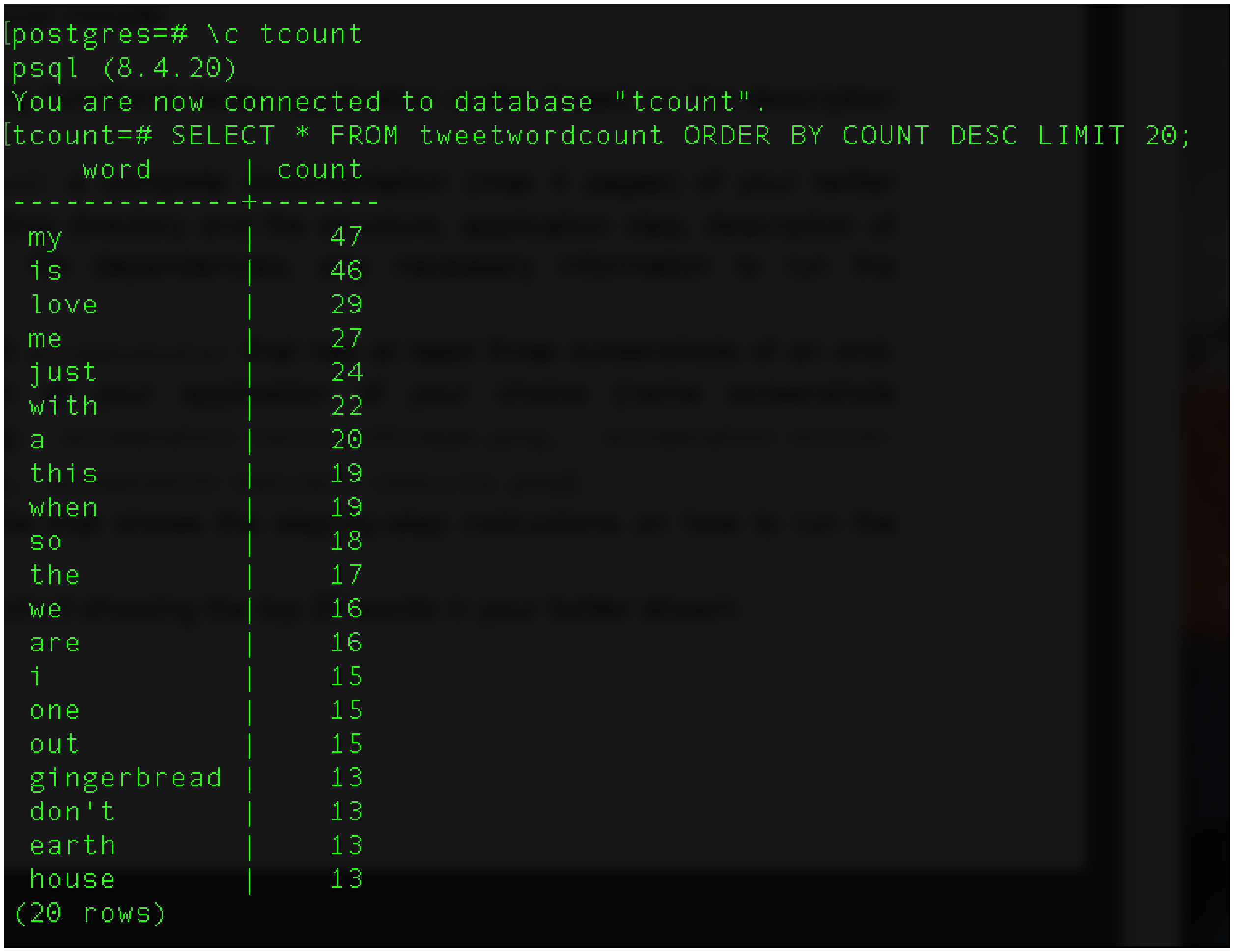


```
[root@ip-172-31-29-35 Tweetwordcount]# python finalresults.py
[('|', 1), ('-', 1), ('//', 1), ('(', 1), ('0345', 1), ('08', 1), ('10%', 1), ('100+', 1), ('1041',
1), ('13', 1), ('2', 1), ('2008', 1), ('2013', 1), ('2017', 1), ('20+s', 1), ('24', 1), ('2day', 1),
 ('2nd', 1), ('3rd', 1), ('4', 2), ('40-player', 1), ('4th', 1), ('8', 1), ('98%', 1), ('a', 20), ('
aaron', 1), ('able', 1), ('about', 1), ('abuse', 1), ('act', 1), ('adopt', 1), ('adorable', 1), ('ad
vertisements', 1), ('adw', 1), ('after', 4), ('again', 1), ('agree', 1), ('agriculture', 1), ('aint'
, 1), ('all', 1), ('almost', 1), ('along', 1), ('already', 1), ('always!', 1), ('am', 7), ('amazing!
', 1), ('ameri', 1), ('america', 1), ('amount', 2), ('&amp', 11), ('an', 1), ('and', 1), ('anniversa
ry', 1), ('anti', 1), ('anxiety', 1), ('any', 5), ('anyways', 1), ('apparently', 1), ('are', 16), ("
aren't", 1), ('artisan', 1), ('artists', 1), ('as', 8), ('asf', 1), ('asks', 1), ('athletes', 1), ('
available', 1), ('away', 1), ('babe', 1), ('baby', 1), ('babygirl', 1), ('back', 1), ('backstory', 1
), ('bad', 1), ('balance', 1), ('be', 1), ('bean', 1), ('beats', 1), ('because', 2), ('been', 1), ('
being', 1), ('bela', 1), ('believe', 1), ('best', 4), ('bestfriend', 1), ('better', 3), ('big', 5),
('[bigbang10', 1), ('billionaire-all', 1), ('bitch', 1), ('blasio', 1), ('blessed', 1), ('blocked',
1), ('body', 1), ('bone', 5), ('boon', 1), ('bottle', 1), ('boy', 1), ('boyfriend', 1), ('bracevor',
 1), ('breed', 1), ('brew', 1), ('bring', 1), ('brother!!', 1), ('brown', 1), ('bunions', 1), ('burr
ito!', 1), ('business', 1), ('but', 2), ('butt', 1), ('buy', 1), ('buying', 1), ('by', 5), ('call',
1), ('came', 3), ('can', 7), ('cancel', 1), ('cancer', 1), ('cant', 1), ("can't", 1), ('car', 1), ('
```



```
[root@ip-172-31-29-35 Tweetwordcount]# python finalresults.py day
Total number of occurences of 'day' : 8
[root@ip-172-31-29-35 Tweetwordcount]# 
```

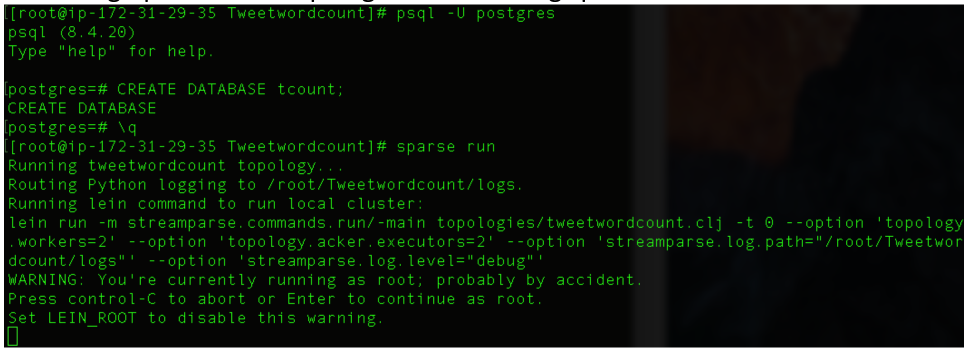
Histogram.py: Return all words and counts for a specific range (i.e. between 4 and 8) in terms of number of times the word was written while the twitter stream was open.



```
[postgres=# \c tcount
psql (8.4.20)
You are now connected to database "tcount".
[tcount=# SELECT * FROM tweetwordcount ORDER BY COUNT DESC LIMIT 20;
    word     | count
-------------+-------
 my          |    47
 is          |    46
 love        |    29
 me          |    27
 just        |    24
 with        |    22
 a           |    20
 this        |    19
 when        |    19
 so          |    18
 the         |    17
 we          |    16
 are         |    16
 i           |    15
 one         |    15
 out         |    15
 gingerbread |    13
 don't       |    13
 earth       |    13
 house       |    13
(20 rows)
```

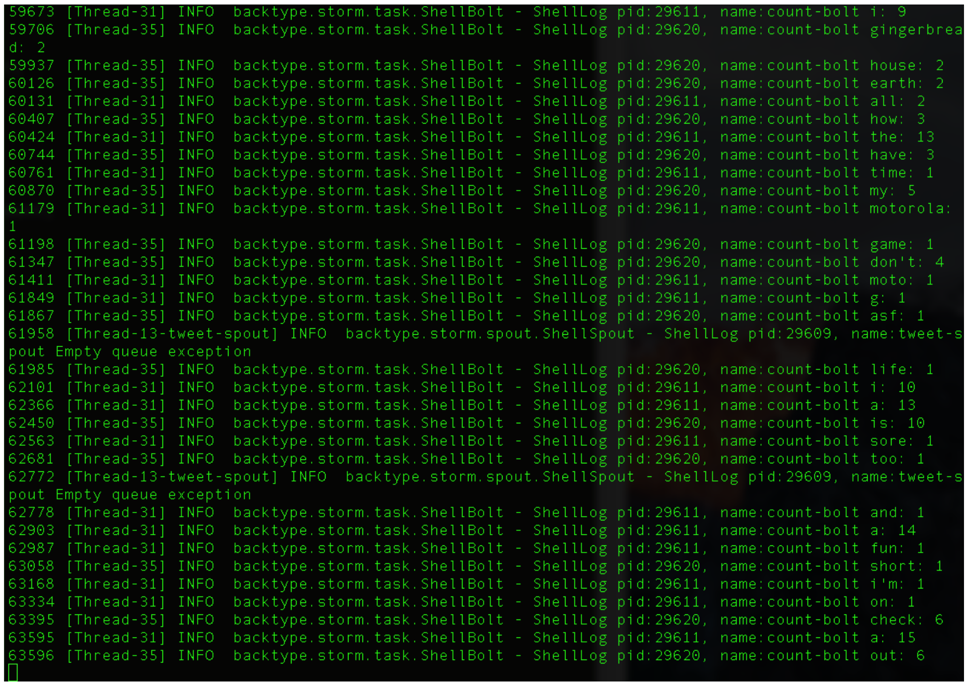
ADDITIONAL SCREENSHOTS

Setting up database in postgres and running sparse run



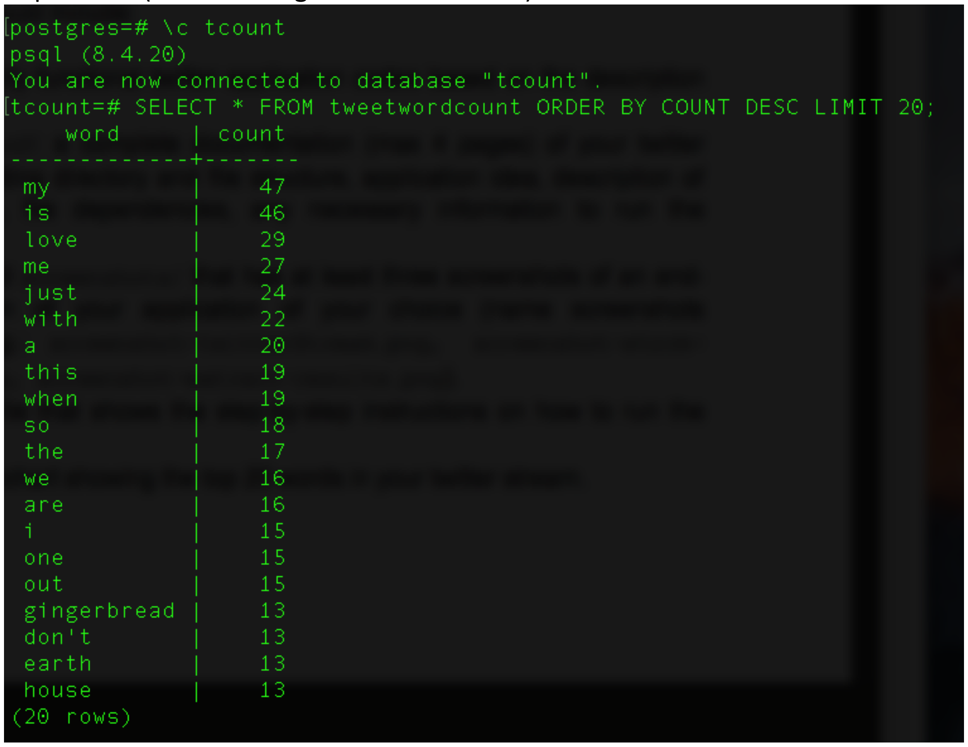
```
[[root@ip-172-31-29-35 Tweetwordcount]# psql -U postgres
psql (8.4.20)
Type "help" for help.

[postgres=# CREATE DATABASE tcount;
CREATE DATABASE
[postgres=# \q
[[root@ip-172-31-29-35 Tweetwordcount]# sparse run
Running tweetwordcount topology...
Routing Python logging to /root/Tweetwordcount/logs.
Running lein command to run local cluster:
lein run -m streamparse.commands.run/-main topologies/tweetwordcount.clj -t 0 --option 'topology
.workers=2' --option 'topology.acker.executors=2' --option 'streamparse.log.path="/root/Tweetwor
dcount/logs"' --option 'streamparse.log.level="debug"'
WARNING: You're currently running as root; probably by accident.
Press control-C to abort or Enter to continue as root.
Set LEIN_ROOT to disable this warning.
```

Streaming tweets and parsing to wordcounts:



```
59673 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt i: 9
59706 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt gingerbrea
d: 2
59937 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt house: 2
60126 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt earth: 2
60131 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt all: 2
60407 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt how: 3
60424 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt the: 13
60744 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt have: 3
60761 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt time: 1
60870 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt my: 5
61179 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt motorola:
1
61198 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt game: 1
61347 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt don't: 4
61411 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt moto: 1
61849 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt g: 1
61867 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt asf: 1
61958 [Thread-13-tweet-spout] INFO  backtype.storm.spout.ShellSpout - ShellLog pid:29609, name:tweet-s
pout Empty queue exception
61985 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt life: 1
62101 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt i: 10
62366 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt a: 13
62450 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt is: 10
62563 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt sore: 1
62681 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt too: 1
62772 [Thread-13-tweet-spout] INFO  backtype.storm.spout.ShellSpout - ShellLog pid:29609, name:tweet-s
pout Empty queue exception
62778 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt and: 1
62903 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt a: 14
62987 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt fun: 1
63058 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt short: 1
63168 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt i'm: 1
63334 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt on: 1
63395 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt check: 6
63595 [Thread-31] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29611, name:count-bolt a: 15
63596 [Thread-35] INFO  backtype.storm.task.ShellBolt - ShellLog pid:29620, name:count-bolt out: 6
```

Top counts (after running for a few minutes)



```
postgres=# \c tcount
psql (8.4.20)
You are now connected to database "tcount".
tcount=# SELECT * FROM tweetwordcount ORDER BY COUNT DESC LIMIT 20;
    word     | count
-------------+-------
 my          |    47
 is          |    46
 love        |    29
 me          |    27
 just        |    24
 with        |    22
 a           |    20
 this        |    19
 when        |    19
 so          |    18
 the         |    17
 we          |    16
 are         |    16
 i           |    15
 one         |    15
 out         |    15
 gingerbread |    13
 don't       |    13
 earth       |    13
 house       |    13
(20 rows)
```