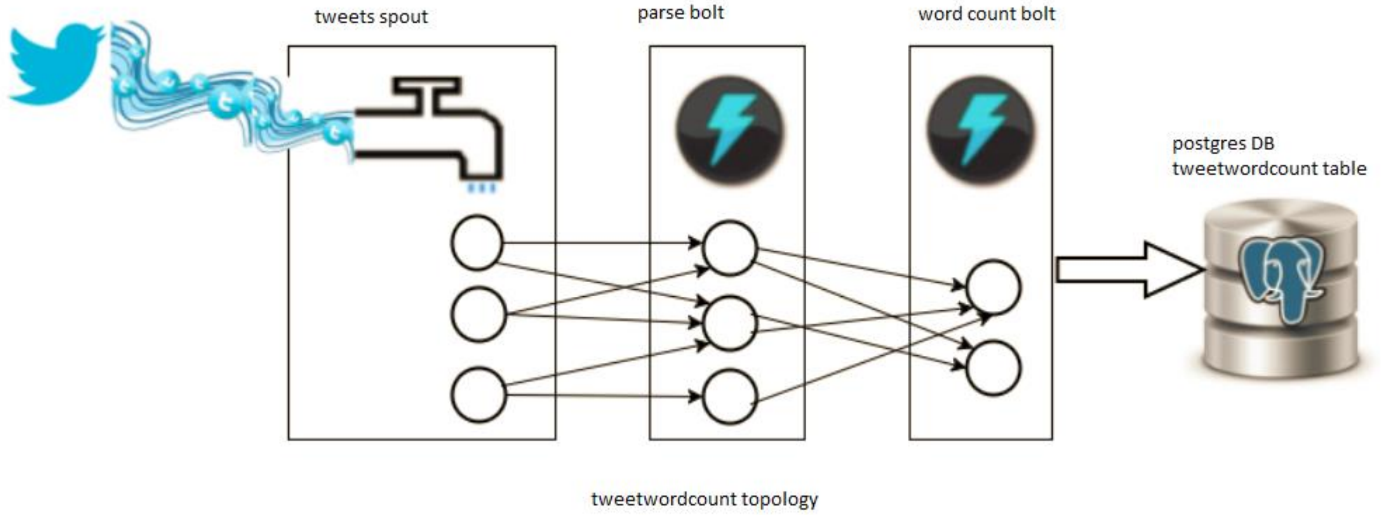
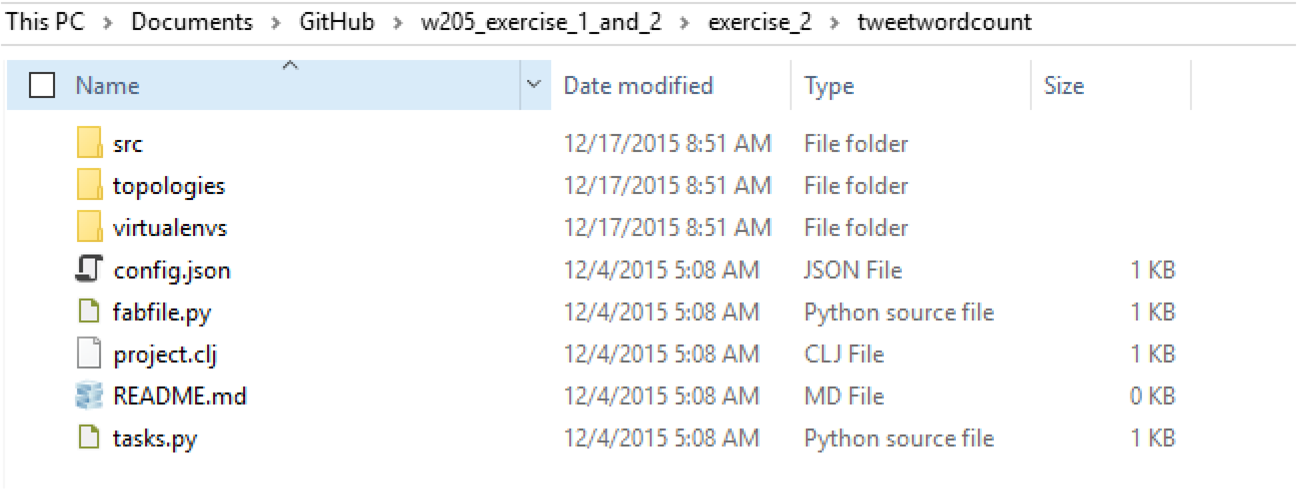
Architecture

Architecture diagram



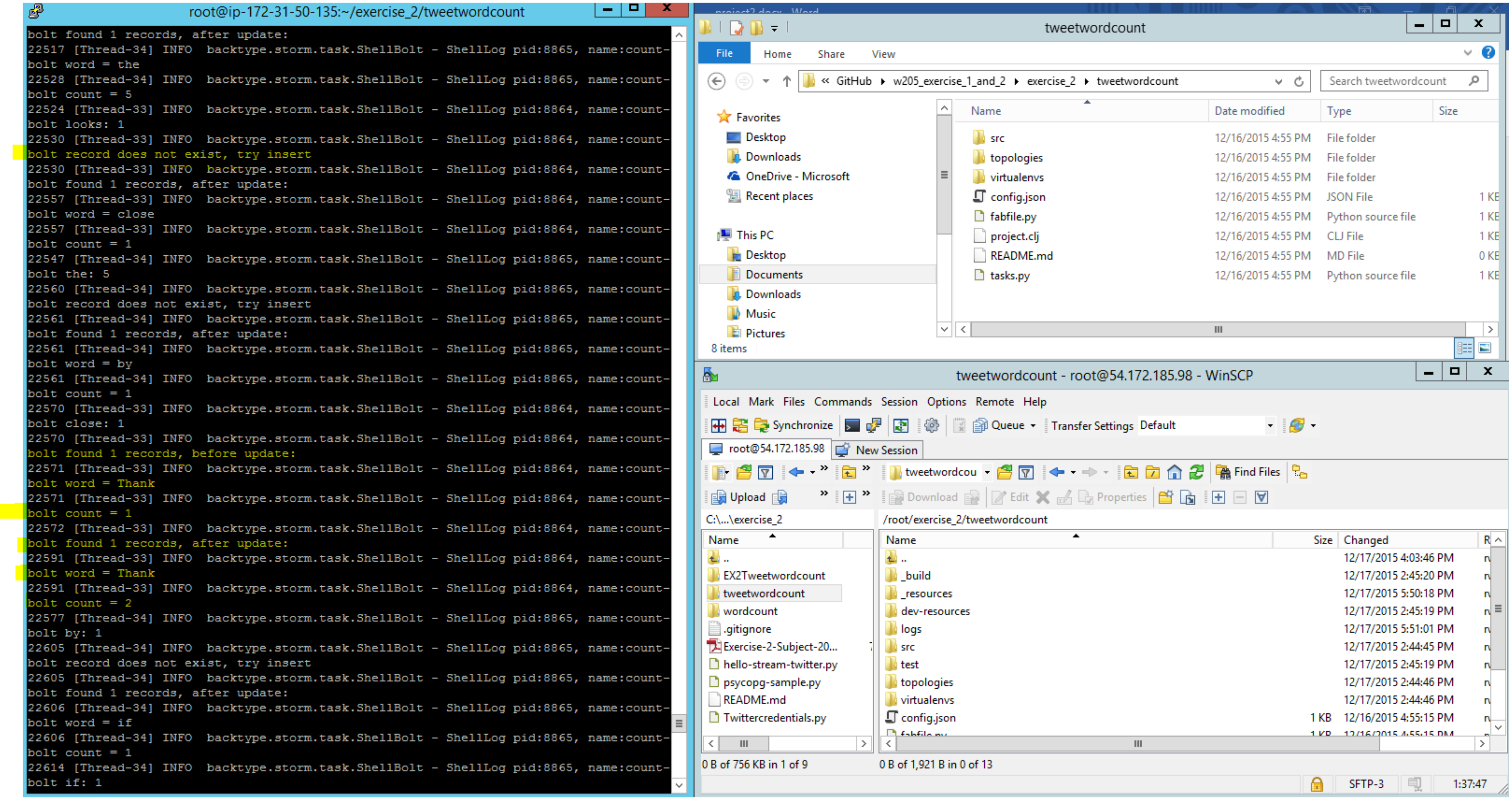
Folder structure


This PC › Documents › GitHub › w205_exercise_1_and_2 › exercise_2 › tweetwordcount

Name	Date modified	Type	Size
src	12/17/2015 8:51 AM	File folder	
topologies	12/17/2015 8:51 AM	File folder	
virtualenvs	12/17/2015 8:51 AM	File folder	
config.json	12/4/2015 5:08 AM	JSON File	1 KB
fabfile.py	12/4/2015 5:08 AM	Python source file	1 KB
project.clj	12/4/2015 5:08 AM	CLJ File	1 KB
README.md	12/4/2015 5:08 AM	MD File	0 KB
tasks.py	12/4/2015 5:08 AM	Python source file	1 KB

Submission 1 and 2

I tried to create DB named Tcount, but for some reason I cannot access it from sparse, so I just used default DB named postgres, and also used default folder tweetwordcount.



root@ip-172-31-50-135:~/exercise_2/tweetwordcount

```
bolt found 1 records, after update:
22517 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt word = the
22528 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt count = 5
22524 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt looks: 1
22530 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt record does not exist, try insert
22530 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt found 1 records, after update:
22557 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt word = close
22557 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt count = 1
22547 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt the: 5
22560 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt record does not exist, try insert
22561 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt found 1 records, after update:
22561 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt word = by
22561 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt count = 1
22570 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt close: 1
22570 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt found 1 records, before update:
22571 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt word = Thank
22571 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt count = 1
22572 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt found 1 records, after update:
22591 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt word = Thank
22591 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8864, name:count-
bolt count = 2
22577 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt by: 1
22605 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt record does not exist, try insert
22605 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt found 1 records, after update:
22606 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt word = if
22606 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt count = 1
22614 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:8865, name:count-
bolt if: 1
```

tweetwordcount

« GitHub › w205_exercise_1_and_2 › exercise_2 › tweetwordcount

Name	Date modified	Type	Size
src	12/16/2015 4:55 PM	File folder	
topologies	12/16/2015 4:55 PM	File folder	
virtualenvs	12/16/2015 4:55 PM	File folder	
config.json	12/16/2015 4:55 PM	JSON File	1 KB
fabfile.py	12/16/2015 4:55 PM	Python source file	1 KB
project.clj	12/16/2015 4:55 PM	CLJ File	1 KB
README.md	12/16/2015 4:55 PM	MD File	0 KB
tasks.py	12/16/2015 4:55 PM	Python source file	1 KB

8 items

tweetwordcount - root@54.172.185.98 - WinSCP

C:\...\exercise_2

Name
..
EX2Tweetwordcount
tweetwordcount
wordcount
.gitignore
Exercise-2-Subject-20...
hello-stream-twitter.py
psycopg-sample.py
README.md
Twittercredentials.py

0 B of 756 KB in 1 of 9

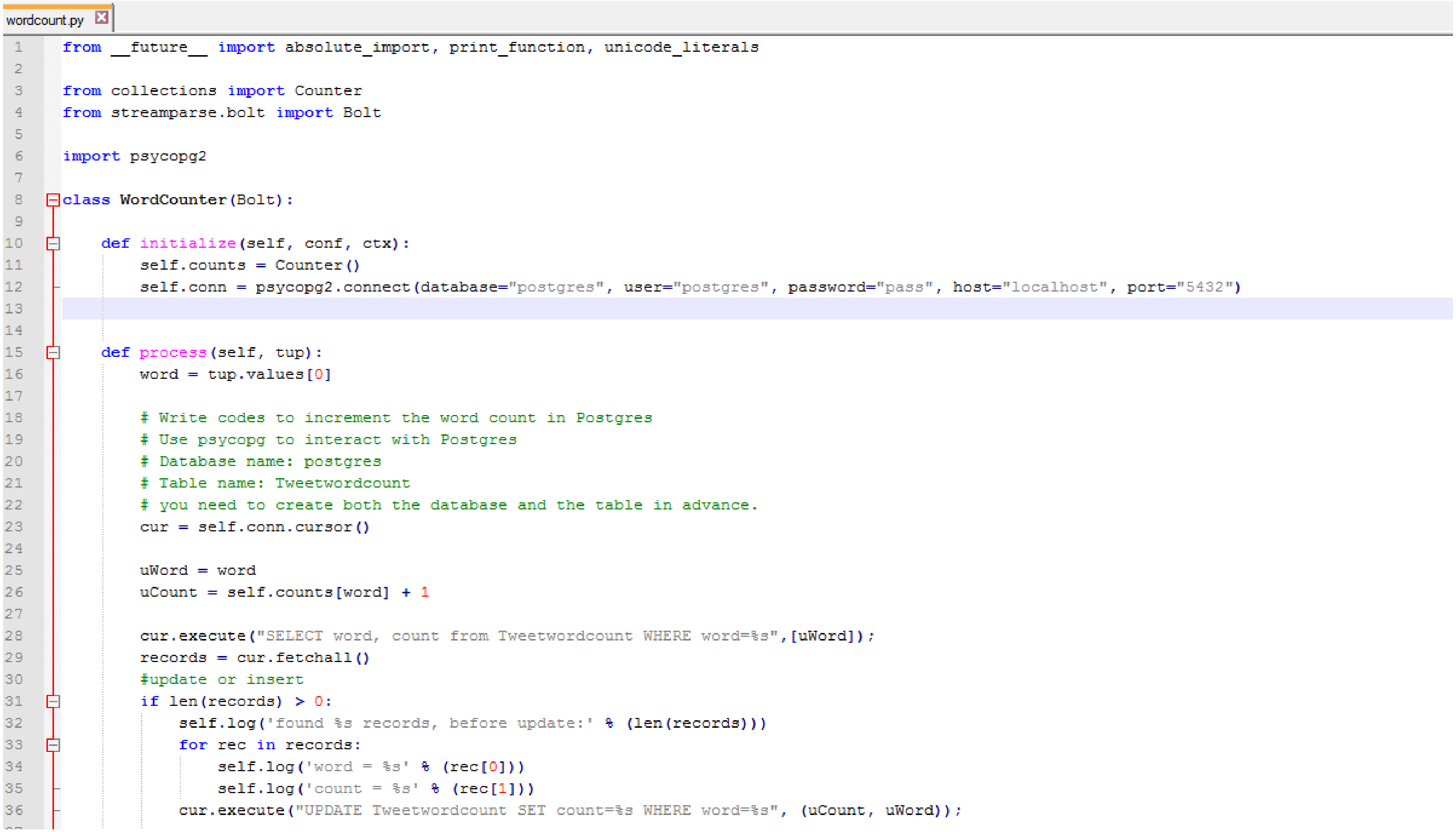
/root/exercise_2/tweetwordcount

Name	Size	Changed
..		12/17/2015 4:03:46 PM
_build		12/17/2015 2:45:20 PM
_resources		12/17/2015 5:50:18 PM
dev-resources		12/17/2015 2:45:19 PM
logs		12/17/2015 5:51:01 PM
src		12/17/2015 2:44:45 PM
test		12/17/2015 2:45:19 PM
topologies		12/17/2015 2:44:46 PM
virtualenvs		12/17/2015 2:44:46 PM
config.json	1 KB	12/16/2015 4:55:15 PM

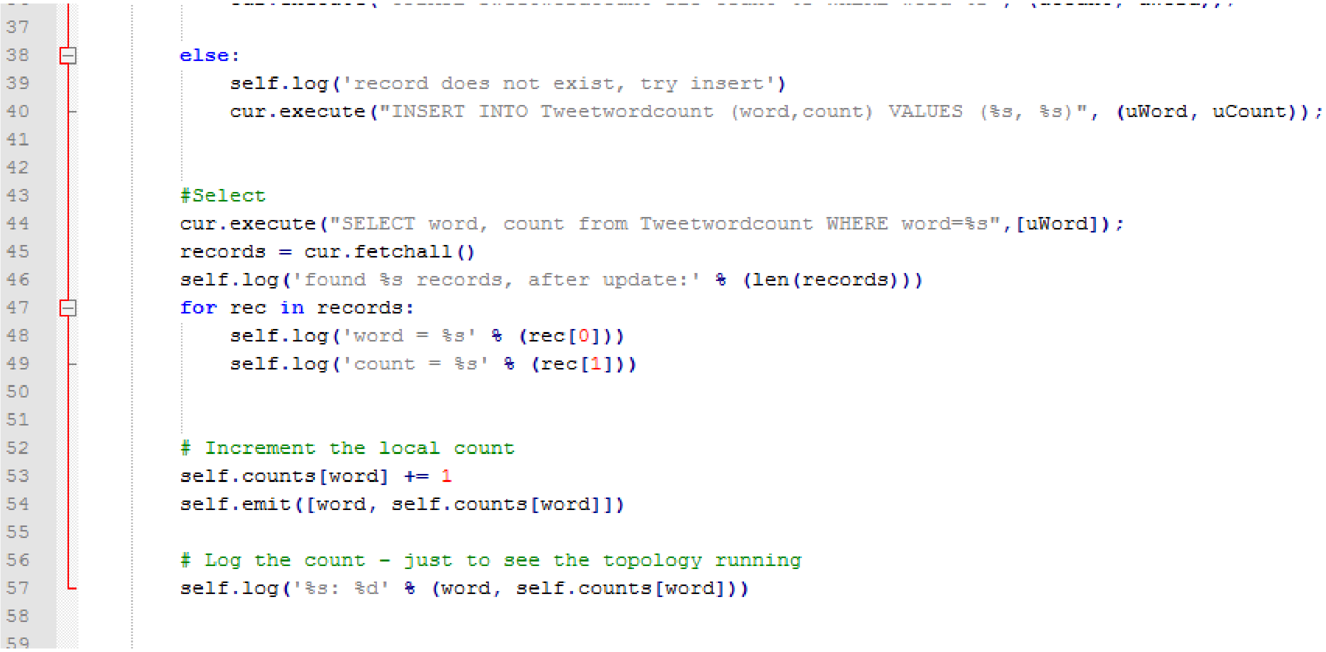
0 B of 1,921 B in 0 of 13

Submission 3

When word count bolt counts words from twitter feed, it also reads from postgres DB. If this word already existed, it will read the existing count and increment using the local word count value. If the word does not exist, it will insert a key using current word and count.

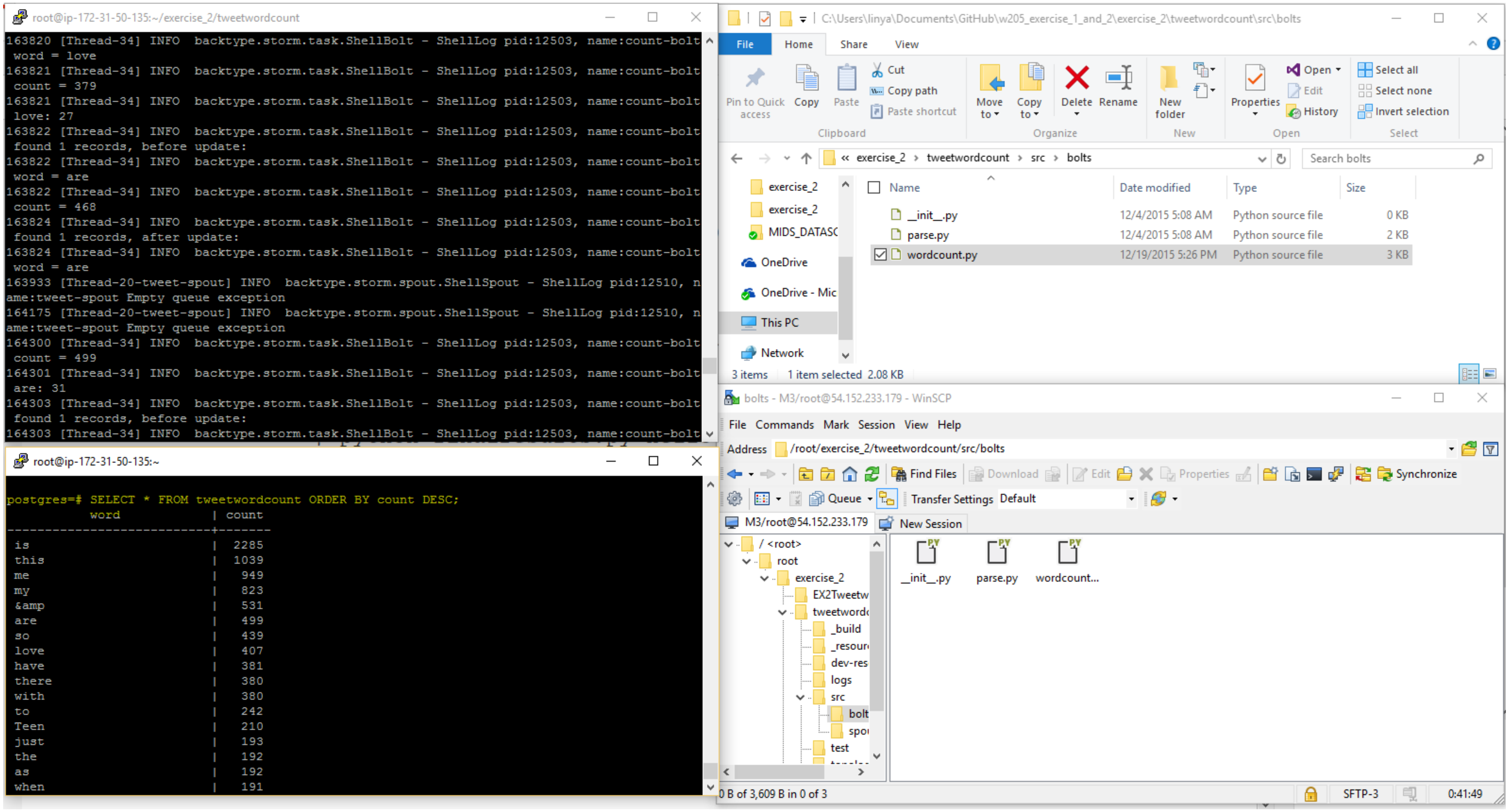


```python
from __future__ import absolute_import, print_function, unicode_literals

from collections import Counter
from streamparse.bolt import Bolt

import psycopg2

class WordCounter(Bolt):

    def initialize(self, conf, ctx):
        self.counts = Counter()
        self.conn = psycopg2.connect(database="postgres", user="postgres", password="pass", host="localhost", port="5432")


    def process(self, tup):
        word = tup.values[0]

        # Write codes to increment the word count in Postgres
        # Use psycopg to interact with Postgres
        # Database name: postgres
        # Table name: Tweetwordcount
        # you need to create both the database and the table in advance.
        cur = self.conn.cursor()

        uWord = word
        uCount = self.counts[word] + 1

        cur.execute("SELECT word, count from Tweetwordcount WHERE word=%s",[uWord]);
        records = cur.fetchall()
        #update or insert
        if len(records) > 0:
            self.log('found %s records, before update:' % (len(records)))
            for rec in records:
                self.log('word = %s' % (rec[0]))
                self.log('count = %s' % (rec[1]))
            cur.execute("UPDATE Tweetwordcount SET count=%s WHERE word=%s", (uCount, uWord));
```



```
else:
    self.log('record does not exist, try insert')
    cur.execute("INSERT INTO Tweetwordcount (word,count) VALUES (%s, %s)", (uWord, uCount));


#Select
cur.execute("SELECT word, count from Tweetwordcount WHERE word=%s",[uWord]);
records = cur.fetchall()
self.log('found %s records, after update:' % (len(records)))
for rec in records:
    self.log('word = %s' % (rec[0]))
    self.log('count = %s' % (rec[1]))


# Increment the local count
self.counts[word] += 1
self.emit([word, self.counts[word]])

# Log the count - just to see the topology running
self.log('%s: %d' % (word, self.counts[word]))
```

After running for a few minutes, the table tweetwordcount got updated



```
163820 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 word = love
163821 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 count = 379
163821 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 love: 27
163822 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 found 1 records, before update:
163822 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 word = are
163822 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 count = 468
163824 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 found 1 records, after update:
163824 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 word = are
163933 [Thread-20-tweet-spout] INFO  backtype.storm.spout.ShellSpout - ShellLog pid:12510, n
ame:tweet-spout Empty queue exception
164175 [Thread-20-tweet-spout] INFO  backtype.storm.spout.ShellSpout - ShellLog pid:12510, n
ame:tweet-spout Empty queue exception
164300 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 count = 499
164301 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 are: 31
164303 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
 found 1 records, before update:
164303 [Thread-34] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12503, name:count-bolt
```

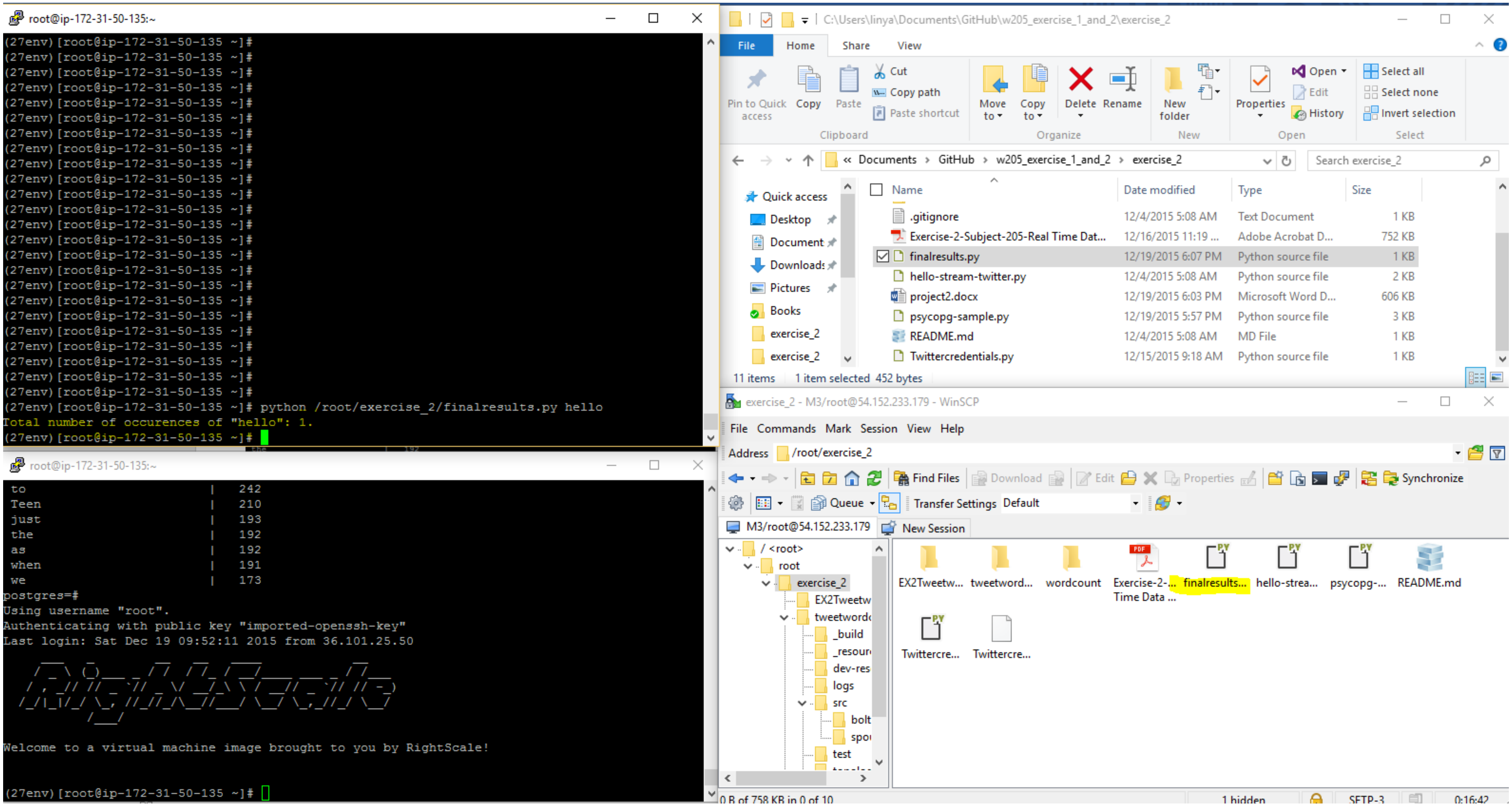
```
postgres=# SELECT * FROM tweetwordcount ORDER BY count DESC;
          word            | count
--------------------------+-------
 is                       |  2285
 this                     |  1039
 me                       |   949
 my                       |   823
 &amp                     |   531
 are                      |   499
 so                       |   439
 love                     |   407
 have                     |   381
 there                    |   380
 with                     |   380
 to                       |   242
 Teen                     |   210
 just                     |   193
 the                      |   192
 as                       |   192
 when                     |   191
```

Name	Date modified	Type	Size
init.py	12/4/2015 5:08 AM	Python source file	0 KB
parse.py	12/4/2015 5:08 AM	Python source file	2 KB
☑ wordcount.py	12/19/2015 5:26 PM	Python source file	3 KB

3 items   1 item selected 2.08 KB

Submission 4

Sample query result for word hello.



root@ip-172-31-50-135:~

```
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]#
(27env) [root@ip-172-31-50-135 ~]# python /root/exercise_2/finalresults.py hello
Total number of occurences of "hello": 1.
(27env) [root@ip-172-31-50-135 ~]#
```

root@ip-172-31-50-135:~

```
 to                         |   242
 Teen                       |   210
 just                       |   193
 the                        |   192
 as                         |   192
 when                       |   191
 we                         |   173
postgres=#
Using username "root".
Authenticating with public key "imported-openssh-key"
Last login: Sat Dec 19 09:52:11 2015 from 36.101.25.50
    ___  _      __   __  ____         __
   / _ \(_)__ _/ /  / /_/ __/______ _/ /__
  / , _/ / _ `/ _ \/ __/\ \/ __/ _ `/ / -_)
 /_/|_/_/\_, /_//_/\__/___/\__/\_,_/_/\__/
        /___/

Welcome to a virtual machine image brought to you by RightScale!


(27env) [root@ip-172-31-50-135 ~]#
```

C:\Users\linya\Documents\GitHub\w205_exercise_1_and_2\exercise_2

Documents › GitHub › w205_exercise_1_and_2 › exercise_2

Name	Date modified	Type	Size
.gitignore	12/4/2015 5:08 AM	Text Document	1 KB
Exercise-2-Subject-205-Real Time Dat...	12/16/2015 11:19 ...	Adobe Acrobat D...	752 KB
☑ finalresults.py	12/19/2015 6:07 PM	Python source file	1 KB
hello-stream-twitter.py	12/4/2015 5:08 AM	Python source file	2 KB
project2.docx	12/19/2015 6:03 PM	Microsoft Word D...	606 KB
psycopg-sample.py	12/19/2015 5:57 PM	Python source file	3 KB
README.md	12/4/2015 5:08 AM	MD File	1 KB
Twittercredentials.py	12/15/2015 9:18 AM	Python source file	1 KB

11 items   1 item selected 452 bytes

exercise_2 - M3/root@54.152.233.179 - WinSCP

Address /root/exercise_2

Transfer Settings Default

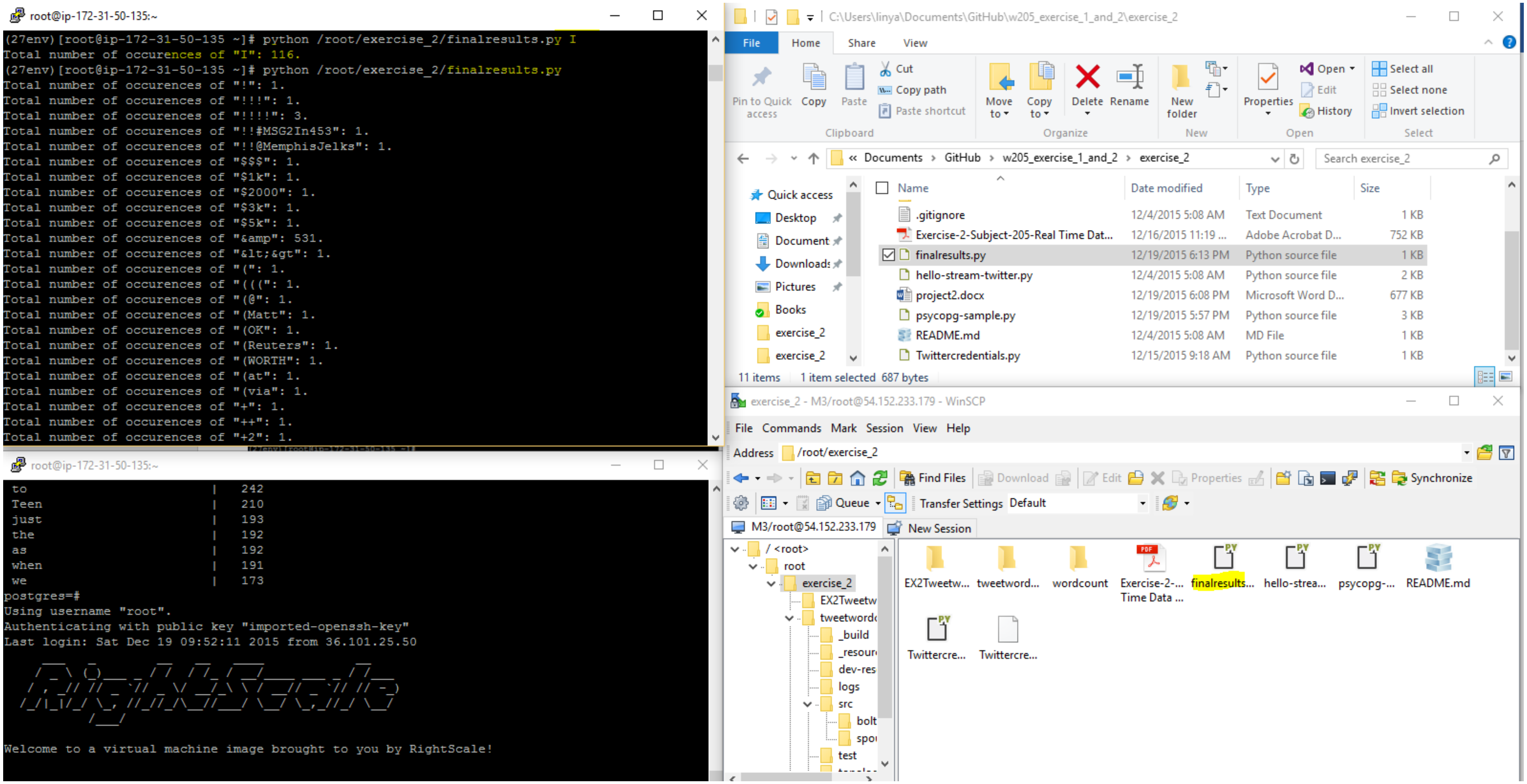
M3/root@54.152.233.179

/ <root>
root
exercise_2
EX2Tweetw
tweetword
_build
_resour
dev-res
logs
src
bolt
spou
test

EX2Tweetw... tweetword... wordcount Exercise-2-... Time Data ... finalresults... hello-strea... psycopg-... README.md Twittercre... Twittercre...

0 B of 758 KB in 0 of 10   1 hidden   SFTP-3   0:16:42

Sample query without supplying arg, it will return all word counts, sorted by word in alphabetical ascending order.

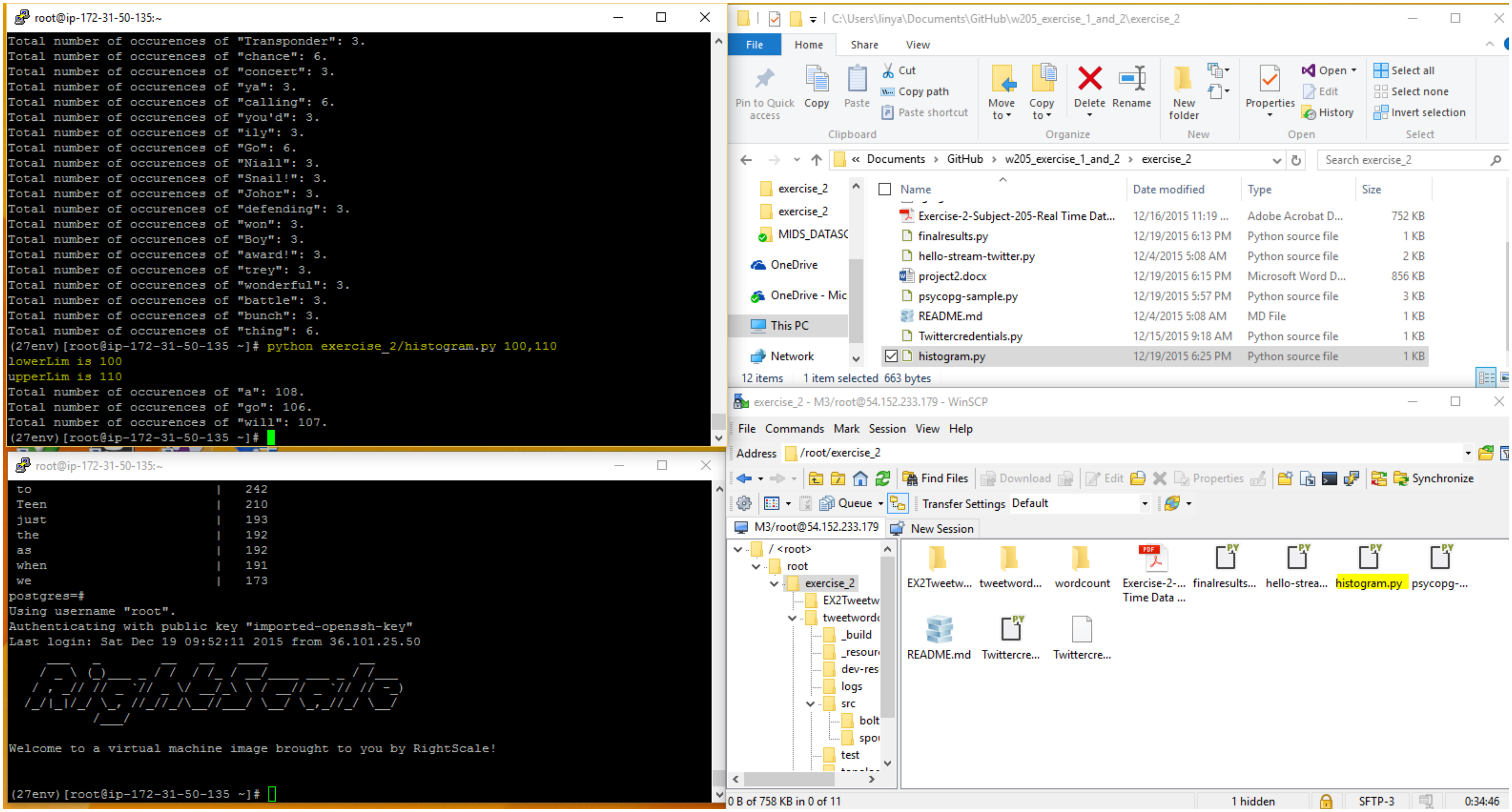


```
(27env)[root@ip-172-31-50-135 ~]# python /root/exercise_2/finalresults.py I
Total number of occurences of "I": 116.
(27env)[root@ip-172-31-50-135 ~]# python /root/exercise_2/finalresults.py
Total number of occurences of "!": 1.
Total number of occurences of "!!!": 1.
Total number of occurences of "!!!!": 3.
Total number of occurences of "!!#MSG2In453": 1.
Total number of occurences of "!!@MemphisJelks": 1.
Total number of occurences of "$$$": 1.
Total number of occurences of "$1k": 1.
Total number of occurences of "$2000": 1.
Total number of occurences of "$3k": 1.
Total number of occurences of "$5k": 1.
Total number of occurences of "&amp": 531.
Total number of occurences of "&lt;&gt": 1.
Total number of occurences of "(": 1.
Total number of occurences of "(((": 1.
Total number of occurences of "(@": 1.
Total number of occurences of "(Matt": 1.
Total number of occurences of "(OK": 1.
Total number of occurences of "(Reuters": 1.
Total number of occurences of "(WORTH": 1.
Total number of occurences of "(at": 1.
Total number of occurences of "(via": 1.
Total number of occurences of "+": 1.
Total number of occurences of "++": 1.
Total number of occurences of "+2": 1.
```

```
 to                    |   242
 Teen                  |   210
 just                  |   193
 the                   |   192
 as                    |   192
 when                  |   191
 we                    |   173
postgres=#
Using username "root".
Authenticating with public key "imported-openssh-key"
Last login: Sat Dec 19 09:52:11 2015 from 36.101.25.50

Welcome to a virtual machine image brought to you by RightScale!
```

Name	Date modified	Type	Size
.gitignore	12/4/2015 5:08 AM	Text Document	1 KB
Exercise-2-Subject-205-Real Time Dat...	12/16/2015 11:19 ...	Adobe Acrobat D...	752 KB
☑ finalresults.py	12/19/2015 6:13 PM	Python source file	1 KB
hello-stream-twitter.py	12/4/2015 5:08 AM	Python source file	2 KB
project2.docx	12/19/2015 6:08 PM	Microsoft Word D...	677 KB
psycopg-sample.py	12/19/2015 5:57 PM	Python source file	3 KB
README.md	12/4/2015 5:08 AM	MD File	1 KB
Twittercredentials.py	12/15/2015 9:18 AM	Python source file	1 KB

Query histogram, using lower limit and upper limit, inclusive on both sides. Parameter format is: lower,upper, such as 3,8



```
Total number of occurences of "Transponder": 3.
Total number of occurences of "chance": 6.
Total number of occurences of "concert": 3.
Total number of occurences of "ya": 3.
Total number of occurences of "calling": 6.
Total number of occurences of "you'd": 3.
Total number of occurences of "ily": 3.
Total number of occurences of "Go": 6.
Total number of occurences of "Niall": 3.
Total number of occurences of "Snail!": 3.
Total number of occurences of "Johor": 3.
Total number of occurences of "defending": 3.
Total number of occurences of "won": 3.
Total number of occurences of "Boy": 3.
Total number of occurences of "award!": 3.
Total number of occurences of "trey": 3.
Total number of occurences of "wonderful": 3.
Total number of occurences of "battle": 3.
Total number of occurences of "bunch": 3.
Total number of occurences of "thing": 6.
(27env)[root@ip-172-31-50-135 ~]# python exercise_2/histogram.py 100,110
lowerLim is 100
upperLim is 110
Total number of occurences of "a": 108.
Total number of occurences of "go": 106.
Total number of occurences of "will": 107.
(27env)[root@ip-172-31-50-135 ~]#
```

```
 to                         |   242
 Teen                       |   210
 just                       |   193
 the                        |   192
 as                         |   192
 when                       |   191
 we                         |   173
postgres=#
Using username "root".
Authenticating with public key "imported-openssh-key"
Last login: Sat Dec 19 09:52:11 2015 from 36.101.25.50
    ___   _          __    __   ____           __
   / _ \ (_)___ _ / /   / /_ / __/____ ___ _ / /___
  / , _// // _ `// _ \ / __/_\ \ / __// _ `// // -_)
 /_/|_|/_/ \_, //_//_/ \__//___/ \__/ \_,_//_/ \__/
          /___/

Welcome to a virtual machine image brought to you by RightScale!


(27env)[root@ip-172-31-50-135 ~]#
```

Name	Date modified	Type	Size
Exercise-2-Subject-205-Real Time Dat...	12/16/2015 11:19 ...	Adobe Acrobat D...	752 KB
finalresults.py	12/19/2015 6:13 PM	Python source file	1 KB
hello-stream-twitter.py	12/4/2015 5:08 AM	Python source file	2 KB
project2.docx	12/19/2015 6:15 PM	Microsoft Word D...	856 KB
psycopg-sample.py	12/19/2015 5:57 PM	Python source file	3 KB
README.md	12/4/2015 5:08 AM	MD File	1 KB
Twittercredentials.py	12/15/2015 9:18 AM	Python source file	1 KB
☑ histogram.py	12/19/2015 6:25 PM	Python source file	1 KB

12 items 1 item selected 663 bytes

Address /root/exercise_2

Transfer Settings Default

0 B of 758 KB in 0 of 11