

Design Decision:

Our mapper and reducer function is built in python.

First we have preprocessed our data. We have added the line number and a document id. We also have lowercased the words and removed the punctuation marks. Finally the common words are removed using stop words list.

We counted how many times a particular word appeared in a document using wordcount MapReduce program.

If a word appears in a document more than half of the average of occurrence times , then we add it to our stop words list. Working with Shakepear made us realize average was a good enough cut-off point because there were enough words being repeated more than this threshold. But since average can vary if maximum and minimum value has huge difference . Considering this, instead of taking average as threshold, we chose the half value of it.

Now for each file in input directory, we prepend the document id and line number for each line. After that, we block the stop words and remove all the punctuation marks.

Our local query file takes the final output from invertedIndexed file as input. Then it looks up the word from standard input . If the word is present in the file, then it prints out the document ID and line number of the word.

Manual:

Step1-5: Preprocess the Input

1. Run process.sh in your local machine.

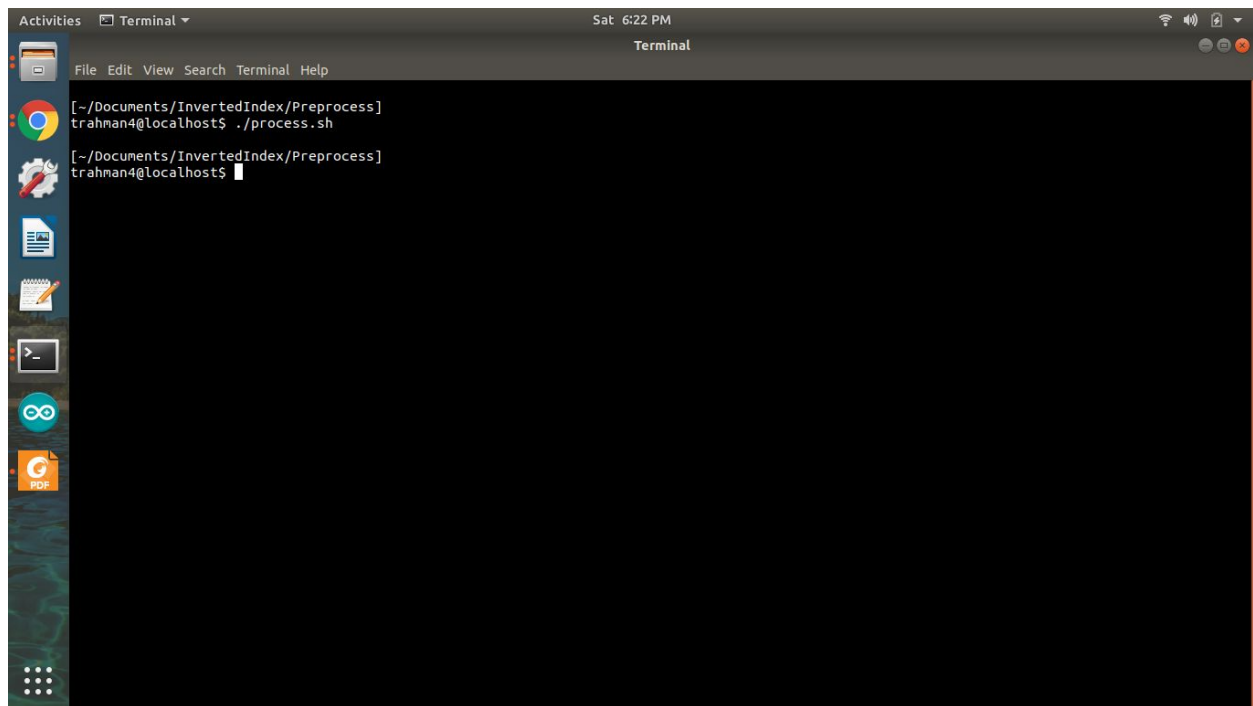


Figure 1: Run Process.sh

First this would remove the punctuation marks, add line number in the raw data and put the initially processed data into 'clean' directory under 'raw' directory.

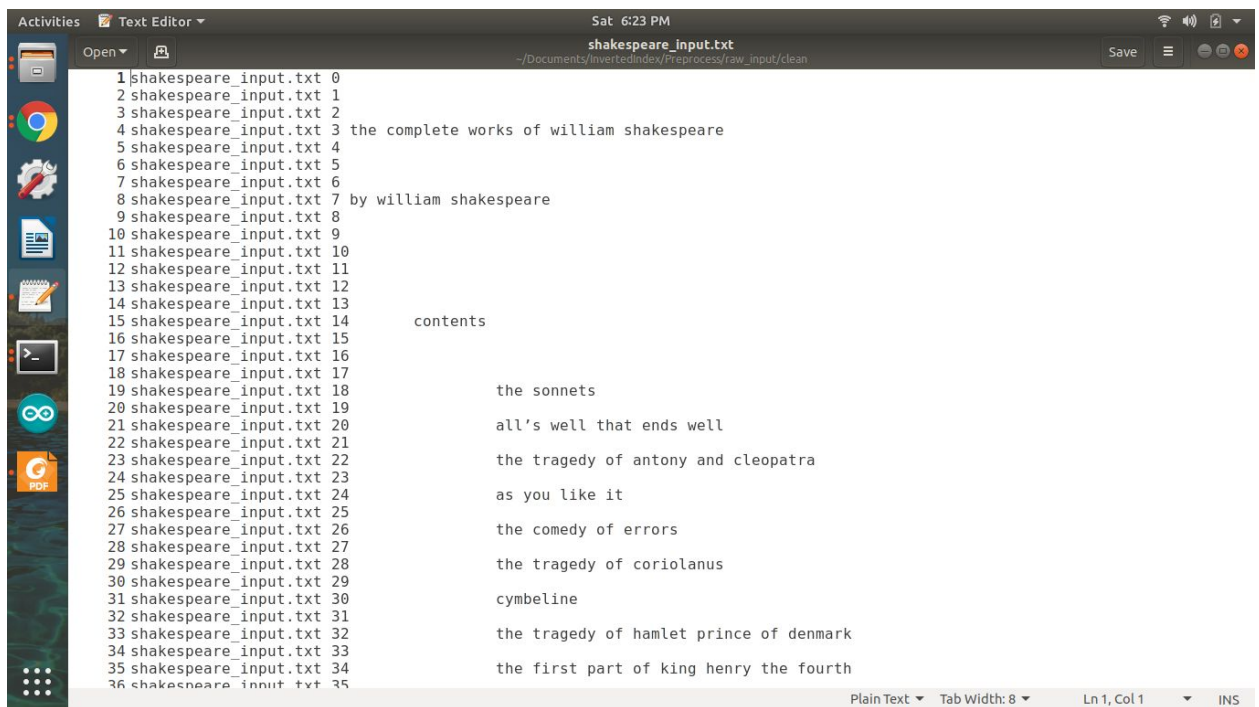


Figure 2: Creates the processed file in clean directory

This initial processed data would be saved under wordCount Mapreduce directory's processed input. Also, we would put this file under generate stop word's directory as a dirty data.

2. Run WordCount on Cloud:

After that, this initially processed data would have to be transferred to cloud as input to wordcount MapReduce program.

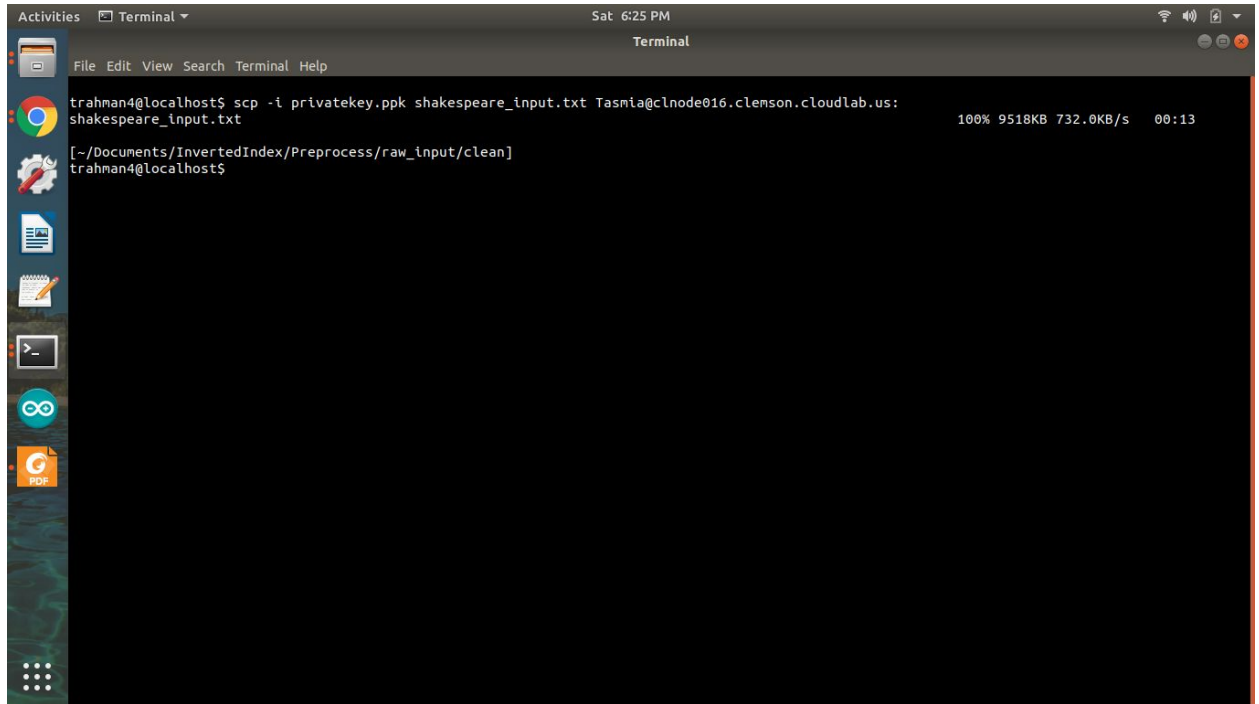


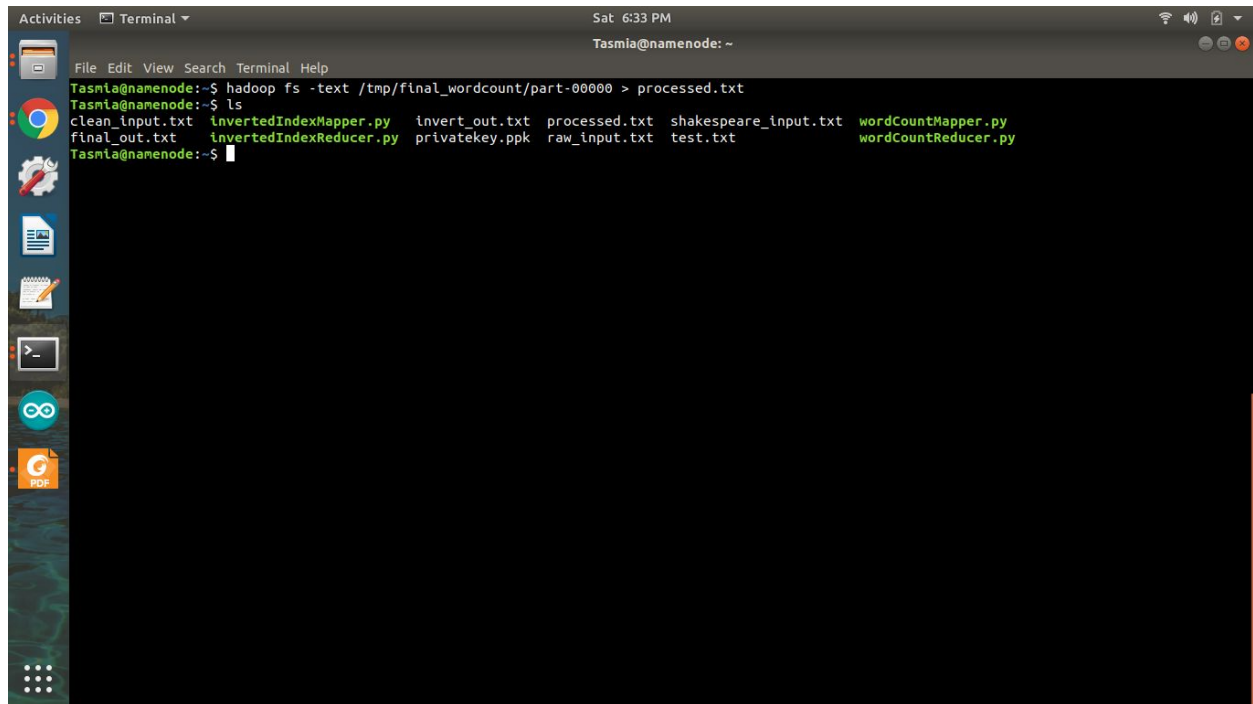
Figure 3: Copy the processes data to hadoop

```
Activities Terminal Sat 6:27 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
Tasmia@namenode:~$ ls
clean_input.txt invertedIndexMapper.py invert_out.txt processed.txt shakespeare_input.txt wordCountMapper.py
final_out.txt invertedIndexReducer.py privatekey.ppk raw_input.txt test.txt wordCountReducer.py
Tasmia@namenode:~$ hadoop fs -put shakespeare_input.txt /tmp/
Tasmia@namenode:~$ hadoop fs -ls /tmp/
Found 28 items
-rw-r--r-- 3 Tasmia supergroup 13828 2018-03-29 23:29 /tmp/before_processing.txt
-rw-r--r-- 3 Tasmia supergroup 148 2018-03-30 13:56 /tmp/clean_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 15:27 /tmp/count_m_r
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 03:23 /tmp/count_out
-rw-r--r-- 3 Tasmia supergroup 43 2018-03-30 13:52 /tmp/dirty_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 16:04 /tmp/final_inverted_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 02:11 /tmp/final_out
drwxrwxrwt - root supergroup 0 2018-03-29 22:57 /tmp/hadoop-yarn
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 08:48 /tmp/invert_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 14:07 /tmp/invert_out_1
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:37 /tmp/map_clean_output
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:32 /tmp/map_dirty_input
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:38 /tmp/map_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 00:03 /tmp/map_red_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:15 /tmp/out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:18 /tmp/out1
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:20 /tmp/out2
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:24 /tmp/out3
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:26 /tmp/out4
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:29 /tmp/out5
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:03 /tmp/out6
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:12 /tmp/out7
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:54 /tmp/out8
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:57 /tmp/out9
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:13 /tmp/output
-rw-r--r-- 3 Tasmia supergroup 2271 2018-03-31 02:10 /tmp/raw_input.txt
-rw-r--r-- 3 Tasmia supergroup 161 2018-03-30 10:12 /tmp/sample_input.txt
-rw-r--r-- 3 Tasmia supergroup 9746146 2018-03-31 16:26 /tmp/shakespeare_input.txt
Tasmia@namenode:~$
```

Figure 4: Put the file into HDFS & ls to see the file

```
Activities Terminal Sat 6:29 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
drwxrwxrwt - root supergroup 0 2018-03-29 22:57 /tmp/hadoop-yarn
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 08:48 /tmp/invert_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 14:07 /tmp/invert_out_1
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:37 /tmp/map_clean_output
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:32 /tmp/map_dirty_input
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:38 /tmp/map_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 00:03 /tmp/map_red_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:15 /tmp/out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:18 /tmp/out1
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:20 /tmp/out2
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:24 /tmp/out3
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:26 /tmp/out4
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:29 /tmp/out5
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:03 /tmp/out6
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:12 /tmp/out7
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:54 /tmp/out8
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:57 /tmp/out9
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:13 /tmp/output
-rw-r--r-- 3 Tasmia supergroup 2271 2018-03-31 02:10 /tmp/raw_input.txt
-rw-r--r-- 3 Tasmia supergroup 161 2018-03-30 10:12 /tmp/sample_input.txt
-rw-r--r-- 3 Tasmia supergroup 9746146 2018-03-31 16:26 /tmp/shakespeare_input.txt
Tasmia@namenode:~$ hadoop jar /usr/local/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -files wordCountMapper.py,wordCountRed
ucer.py -mapper wordCountMapper.py -reducer wordCountReducer.py -input /tmp/shakespeare_input.txt -output /tmp/final_wordcount
packageJobJar: [/tmp/hadoop-unjar285075281522280670/] [] /tmp/streamjob6059689494102125429.jar tmpDir=null
18/03/31 16:29:07 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/10.10.1.2:8032
18/03/31 16:29:07 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/10.10.1.2:8032
18/03/31 16:29:07 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/31 16:29:08 INFO mapreduce.JobSubmitter: number of splits:2
18/03/31 16:29:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522385828397_0021
18/03/31 16:29:08 INFO impl.YarnClientImpl: Submitted application application_1522385828397_0021
18/03/31 16:29:08 INFO mapreduce.Job: The url to track the job: http://resourcemanager.thadoop.educationproject-pg0.clemson.cloudlab.us:8088/p
roxy/application_1522385828397_0021/
18/03/31 16:29:08 INFO mapreduce.Job: Running job: job_1522385828397_0021
18/03/31 16:29:14 INFO mapreduce.Job: Job job_1522385828397_0021 running in uber mode : false
18/03/31 16:29:14 INFO mapreduce.Job: map 0% reduce 0%
18/03/31 16:29:20 INFO mapreduce.Job: map 100% reduce 0%
18/03/31 16:29:28 INFO mapreduce.Job: map 100% reduce 100%
```

Figure 5: WordCount MapReduce 100% successful



The image shows a terminal window titled 'Terminal' with a menu bar (File, Edit, View, Search, Terminal, Help) and a status bar (Sat 6:33 PM, Tasmia@namenode: ~). The terminal content is as follows:

```
Tasmia@namenode:~$ hadoop fs -text /tmp/final_wordcount/part-00000 > processed.txt
Tasmia@namenode:~$ ls
clean_input.txt  invertedIndexMapper.py  invert_out.txt  processed.txt  shakespeare_input.txt  wordCountMapper.py
final_out.txt    invertedIndexReducer.py  privatekey.ppk  raw_input.txt  test.txt               wordCountReducer.py
Tasmia@namenode:~$
```

Figure 6: Copy the WordCount output to Cloud File System

This step counts how many times each word appears in a given input.

3. Saving Output into Local Machine:

Go to generateStopList directory. Save WordCount's output as processed.txt in this directory .

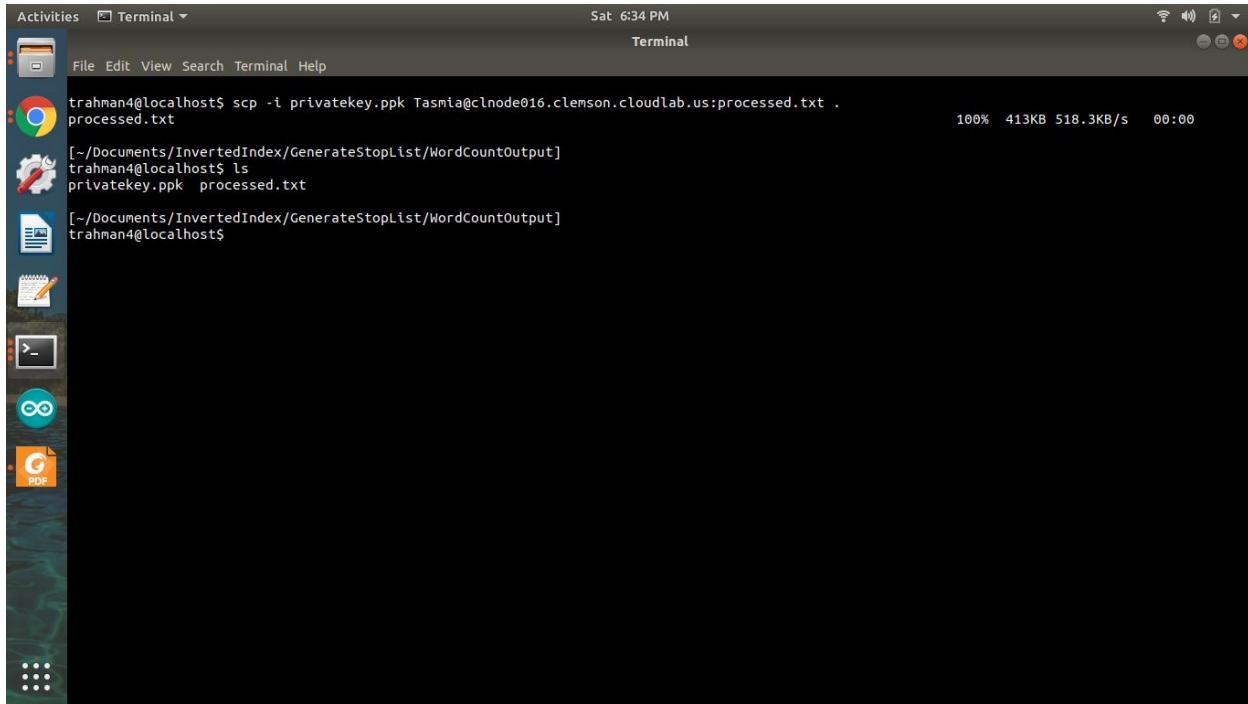


Figure 7: Copying the processed wordcount output from cloud local to WordCountOutput directory
Transfer the output from WordCount Mapreduce to Local File System

4. Generate Stop Word List:

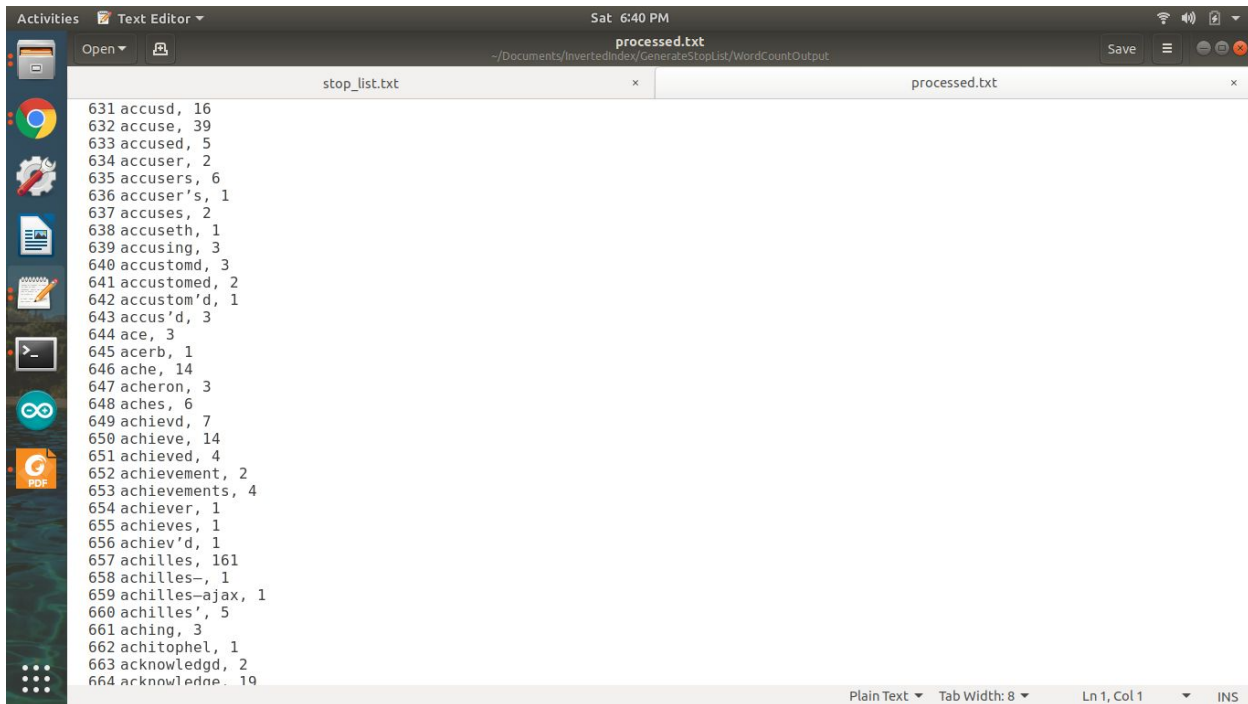
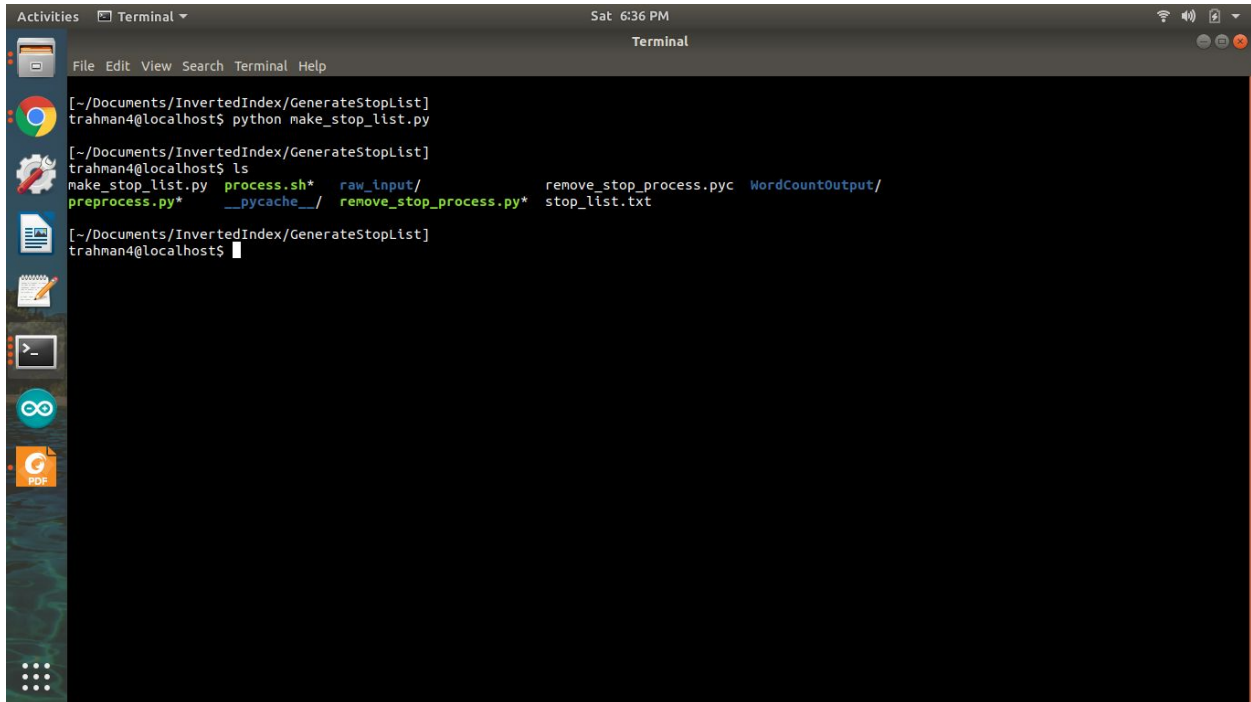


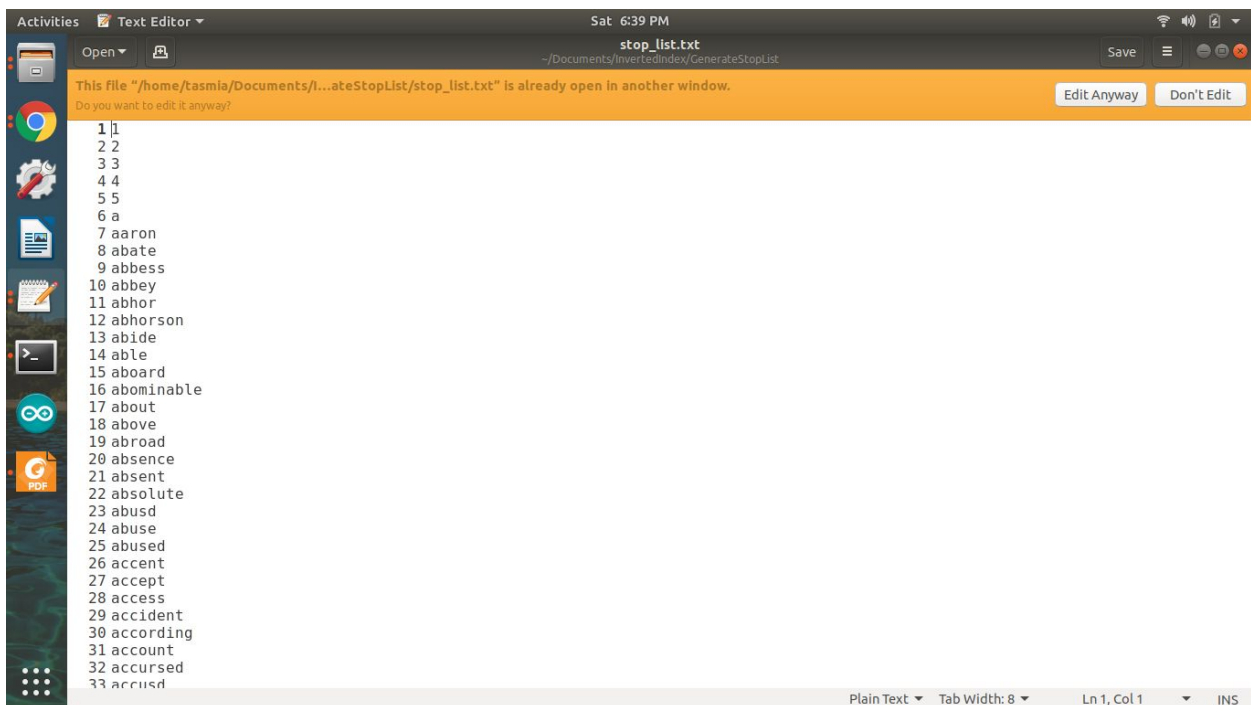
Figure 8: WordCount MapReduce Output



```
trahman4@localhost: ~/Documents/InvertedIndex/GenerateStopList
trahman4@localhost$ python make_stop_list.py

trahman4@localhost: ~/Documents/InvertedIndex/GenerateStopList
trahman4@localhost$ ls
make_stop_list.py  process.sh*  raw_input/  remove_stop_process.pyc  WordCountOutput/
preprocess.py*    __pycache__  remove_stop_process.py*  stop_list.txt
```

Figure 9: Run make_stop_list file to generate stop words
Run make_stop_list.py. The output will be saved as stop_list.txt .

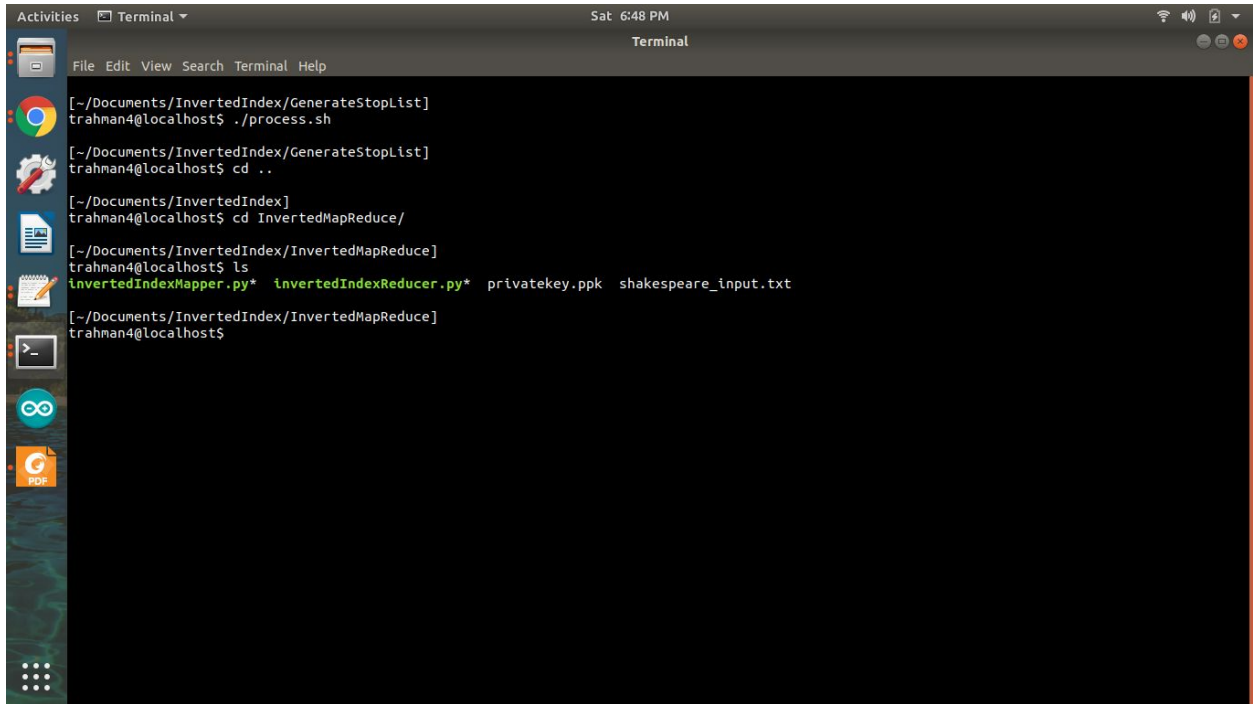


```
1 l
2 2
3 3
4 4
5 5
6 a
7 aaron
8 abate
9 abess
10 abbey
11 abhor
12 abhorson
13 abide
14 able
15 aboard
16 abominable
17 about
18 above
19 abroad
20 absence
21 absent
22 absolute
23 abusd
24 abuse
25 abused
26 accent
27 accept
28 access
29 accident
30 according
31 account
32 accursed
33 accusd
```

Figure 10: Generated Stop List

5. Run Process.sh on WordCount :

This step will remove all the stop words from raw input file. Now save the output as clean input under raw directory.



```
Activities Terminal Sat 6:48 PM
File Edit View Search Terminal Help

[~/Documents/InvertedIndex/GenerateStopList]
trahman4@localhost$ ./process.sh

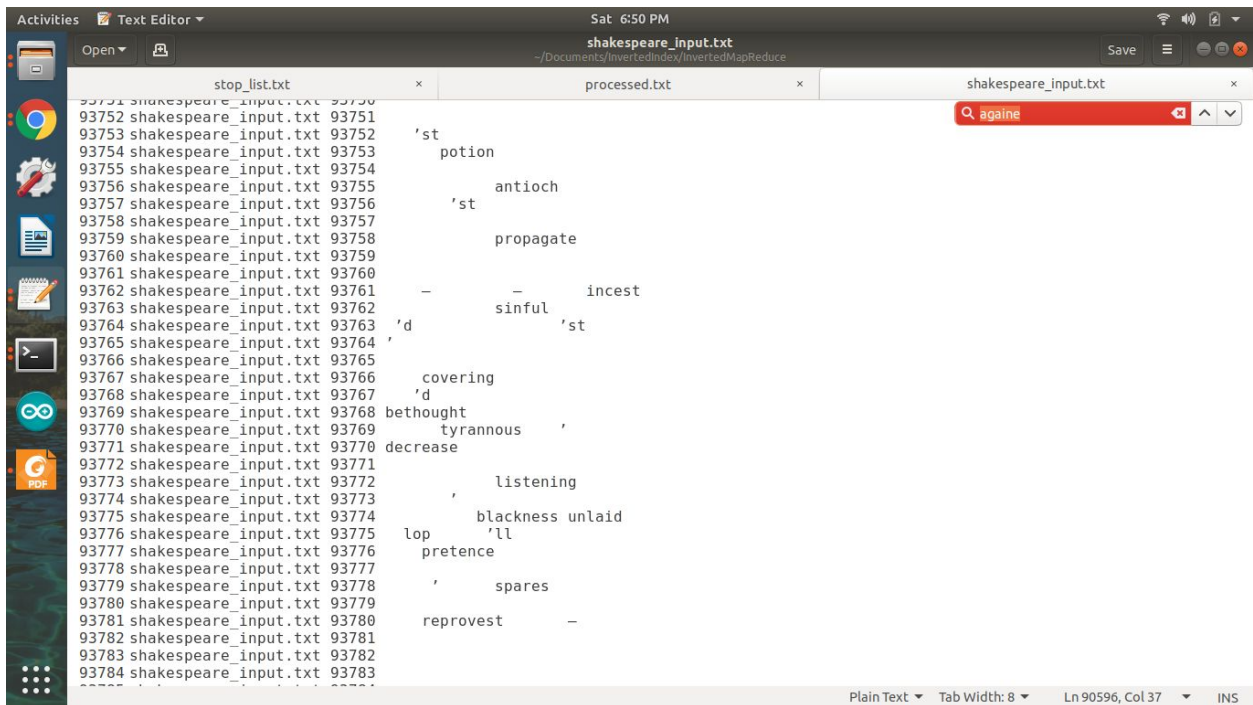
[~/Documents/InvertedIndex/GenerateStopList]
trahman4@localhost$ cd ..

[~/Documents/InvertedIndex]
trahman4@localhost$ cd InvertedMapReduce/

[~/Documents/InvertedIndex/InvertedMapReduce]
trahman4@localhost$ ls
invertedIndexMapper.py* invertedIndexReducer.py* privatekey.ppk shakespeare_input.txt

[~/Documents/InvertedIndex/InvertedMapReduce]
trahman4@localhost$
```

Figure 11: Run Process.sh again to remove the stop words



```
Activities Text Editor Sat 6:50 PM
shakespeare_input.txt
~/Documents/InvertedIndex/InvertedMapReduce
Save

stop_list.txt x processed.txt x shakespeare_input.txt x
93751 shakespeare_input.txt 93750
93752 shakespeare_input.txt 93751
93753 shakespeare_input.txt 93752
93754 shakespeare_input.txt 93753
93755 shakespeare_input.txt 93754
93756 shakespeare_input.txt 93755
93757 shakespeare_input.txt 93756
93758 shakespeare_input.txt 93757
93759 shakespeare_input.txt 93758
93760 shakespeare_input.txt 93759
93761 shakespeare_input.txt 93760
93762 shakespeare_input.txt 93761
93763 shakespeare_input.txt 93762
93764 shakespeare_input.txt 93763
93765 shakespeare_input.txt 93764
93766 shakespeare_input.txt 93765
93767 shakespeare_input.txt 93766
93768 shakespeare_input.txt 93767
93769 shakespeare_input.txt 93768
93770 shakespeare_input.txt 93769
93771 shakespeare_input.txt 93770
93772 shakespeare_input.txt 93771
93773 shakespeare_input.txt 93772
93774 shakespeare_input.txt 93773
93775 shakespeare_input.txt 93774
93776 shakespeare_input.txt 93775
93777 shakespeare_input.txt 93776
93778 shakespeare_input.txt 93777
93779 shakespeare_input.txt 93778
93780 shakespeare_input.txt 93779
93781 shakespeare_input.txt 93780
93782 shakespeare_input.txt 93781
93783 shakespeare_input.txt 93782
93784 shakespeare_input.txt 93783

'st
potion
antioch
'st
propagate
- - incest
sinful 'st
'd
covering
'd
bethought
tyrannous
decrease
listening
blackness unlaidd
lop 'll
pretence
spares
reprovest -
```

Figure 12: Proof that Stop Words are removed

Also, transfer the clean data to cloud. This clean file is our input for inverted Index MapReduce program.


```
Activities Terminal Sat 6:53 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
Tasmia@namenode:~$ ls
clean_input.txt  invertedIndexMapper.py  invert_out.txt  processed.txt  shakespeare_input.txt  wordCountMapper.py
final_out.txt    invertedIndexReducer.py  privatekey.ppk  raw_input.txt  test.txt               wordCountReducer.py
Tasmia@namenode:~$ rm shakespeare_input.txt
Tasmia@namenode:~$ ls
clean_input.txt  invertedIndexMapper.py  invert_out.txt  processed.txt  test.txt               wordCountReducer.py
final_out.txt    invertedIndexReducer.py  privatekey.ppk  raw_input.txt  wordCountMapper.py
Tasmia@namenode:~$
```

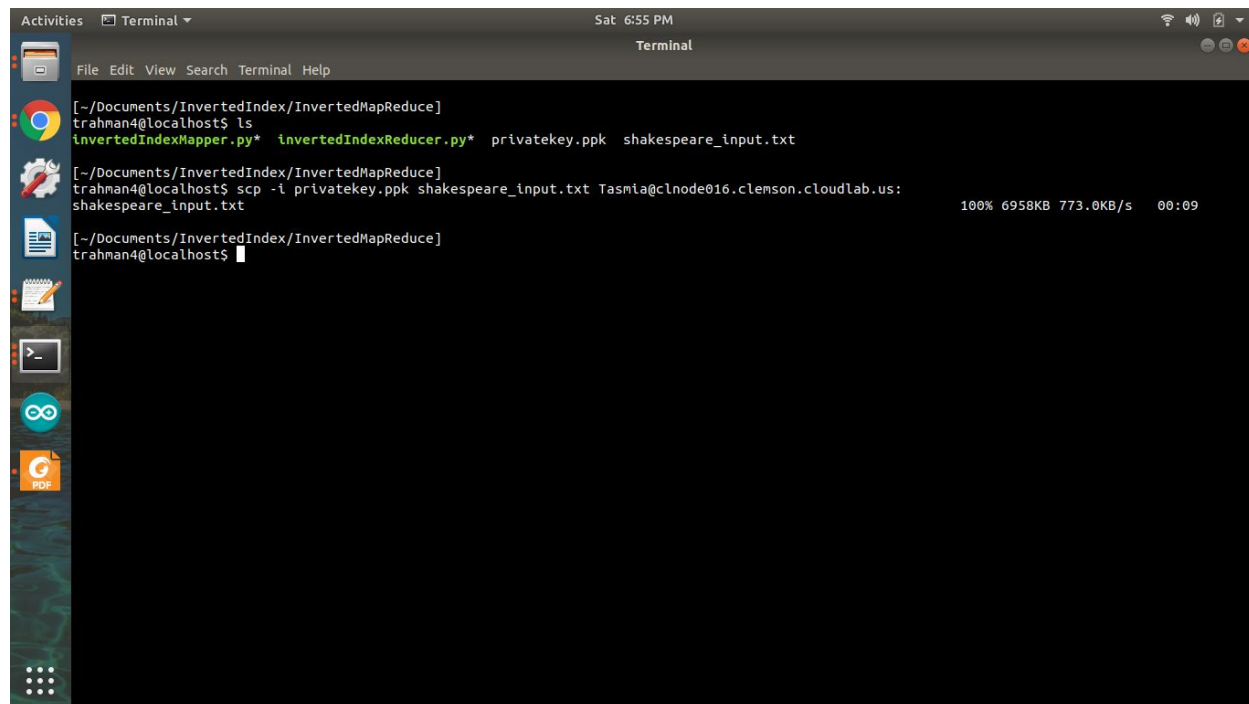
Figure 13: Delete the previous shakespeare_input text file from cloud file system

```
Activities Terminal Sat 6:54 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
Tasmia@namenode:~$ hadoop fs -rm /tmp/shakespeare_input.txt
18/03/31 16:54:03 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /tmp/shakespeare_input.txt
Tasmia@namenode:~$ ls
clean_input.txt  invertedIndexMapper.py  invert_out.txt  processed.txt  test.txt               wordCountReducer.py
final_out.txt    invertedIndexReducer.py  privatekey.ppk  raw_input.txt  wordCountMapper.py
Tasmia@namenode:~$ hadoop fs -ls /tmp/
Found 28 items
-rw-r--r-- 3 Tasmia supergroup 13828 2018-03-29 23:29 /tmp/before_processing.txt
-rw-r--r-- 3 Tasmia supergroup 148 2018-03-30 13:56 /tmp/clean_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 15:27 /tmp/count_m_r
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 03:23 /tmp/count_out
-rw-r--r-- 3 Tasmia supergroup 43 2018-03-30 13:52 /tmp/dirty_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 16:04 /tmp/final_inverted_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 02:11 /tmp/final_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 16:29 /tmp/final_wordcount
drwxrwxrwt - root supergroup 0 2018-03-29 22:57 /tmp/hadoop-yarn
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 08:48 /tmp/invert_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 14:07 /tmp/invert_out_1
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:37 /tmp/map_clean_output
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:32 /tmp/map_dirty_input
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:38 /tmp/map_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 00:03 /tmp/map_red_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:15 /tmp/out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:18 /tmp/out1
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:20 /tmp/out2
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:24 /tmp/out3
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:26 /tmp/out4
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:29 /tmp/out5
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:03 /tmp/out6
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:12 /tmp/out7
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:54 /tmp/out8
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:57 /tmp/out9
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:13 /tmp/output
-rw-r--r-- 3 Tasmia supergroup 2271 2018-03-31 02:10 /tmp/raw_input.txt
-rw-r--r-- 3 Tasmia supergroup 161 2018-03-30 10:12 /tmp/sample_input.txt
Tasmia@namenode:~$
```

Figure 14: Delete the previous shakespeare_input text file from hdfs as well

Step 6-7: Building the Inverted Index:

6.Move invertedIndexMapReduce code to cloud



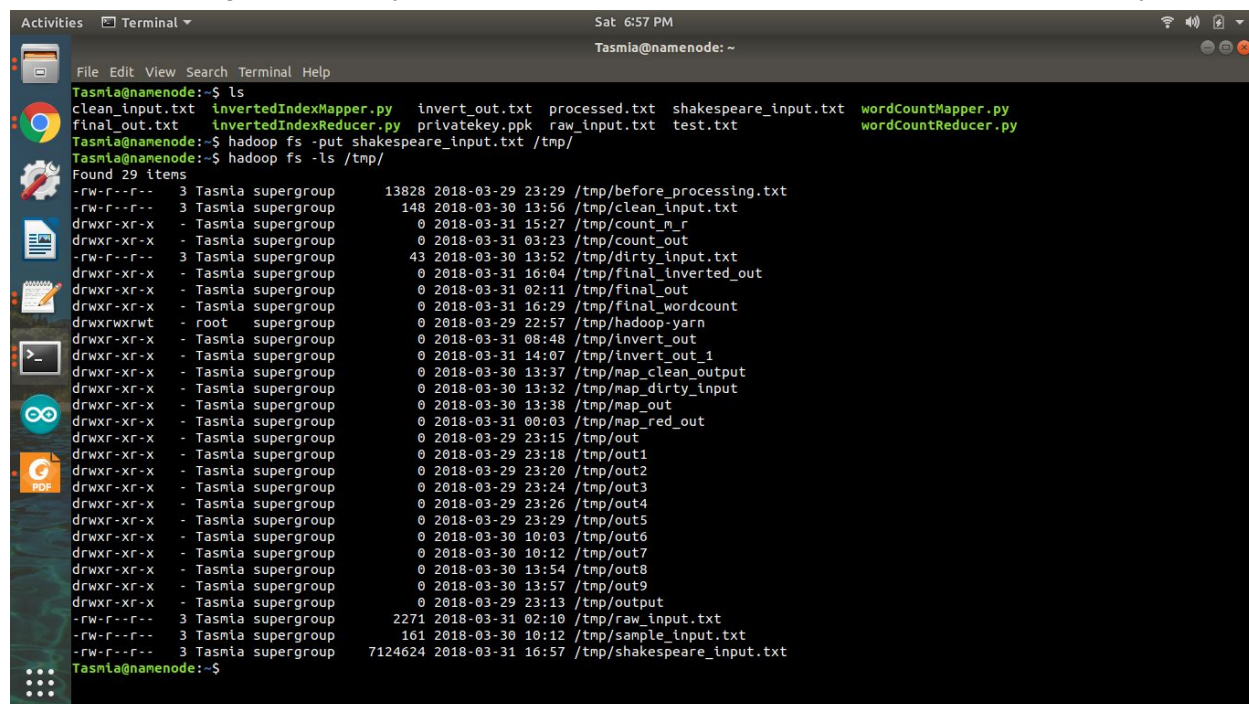
A terminal window showing the process of copying files to a cloud node. The user is in the directory `~/Documents/InvertedIndex/InvertedMapReduce`. They list the files: `trahman4@localhost$ ls`, showing `invertedIndexMapper.py*`, `invertedIndexReducer.py*`, `privatekey.ppk`, and `shakespeare_input.txt`. Then, they execute `trahman4@localhost$ scp -i privatekey.ppk shakespeare_input.txt Tasmia@clnode016.clemson.cloudlab.us:shakespeare_input.txt`. The progress bar shows 100% transfer of 6958KB at 773.0KB/s, taking 00:09.

```
trahman4@localhost$ ls
invertedIndexMapper.py*  invertedIndexReducer.py*  privatekey.ppk  shakespeare_input.txt

trahman4@localhost$ scp -i privatekey.ppk shakespeare_input.txt Tasmia@clnode016.clemson.cloudlab.us:
shakespeare_input.txt 100% 6958KB 773.0KB/s 00:09

trahman4@localhost$
```

Figure 15: Copy the input file for InvertedIndexMapReduce to cloud file system



A terminal window showing the contents of the `/tmp` directory on the cloud node. The user runs `Tasmia@namenode:~$ ls`, listing various files including `clean_input.txt`, `invertedIndexMapper.py`, `invert_out.txt`, `processed.txt`, `shakespeare_input.txt`, `wordCountMapper.py`, `final_out.txt`, `invertedIndexReducer.py`, `privatekey.ppk`, `raw_input.txt`, `test.txt`, and `wordCountReducer.py`. They then run `Tasmia@namenode:~$ hadoop fs -put shakespeare_input.txt /tmp/`. Finally, they run `Tasmia@namenode:~$ hadoop fs -ls /tmp/`, which displays a long list of files and their metadata, including `shakespeare_input.txt` with a size of 7124624 bytes.

```
Tasmia@namenode:~$ ls
clean_input.txt  invertedIndexMapper.py  invert_out.txt  processed.txt  shakespeare_input.txt  wordCountMapper.py
final_out.txt   invertedIndexReducer.py  privatekey.ppk  raw_input.txt  test.txt              wordCountReducer.py

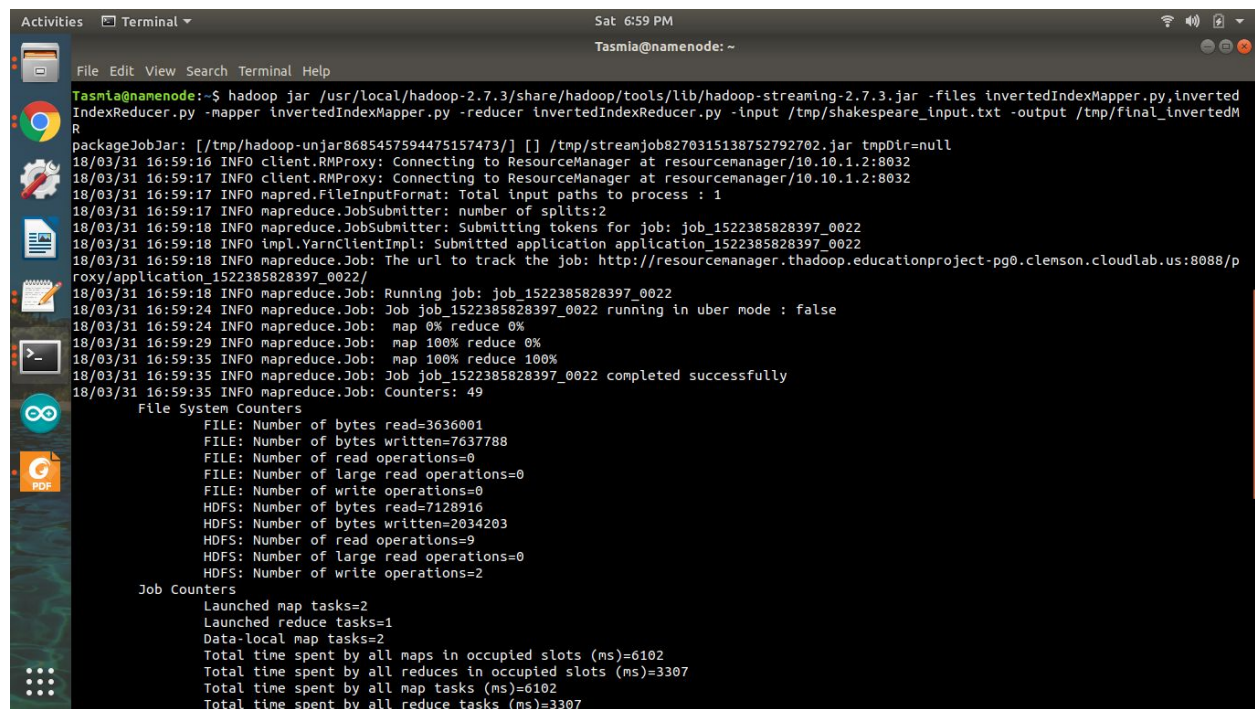
Tasmia@namenode:~$ hadoop fs -put shakespeare_input.txt /tmp/

Tasmia@namenode:~$ hadoop fs -ls /tmp/
Found 29 items
-rw-r--r-- 3 Tasmia supergroup 13828 2018-03-29 23:29 /tmp/before_processing.txt
-rw-r--r-- 3 Tasmia supergroup 148 2018-03-30 13:56 /tmp/clean_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 15:27 /tmp/count_m_r
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 03:23 /tmp/count_out
-rw-r--r-- 3 Tasmia supergroup 43 2018-03-30 13:52 /tmp/dirty_input.txt
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 16:04 /tmp/final_inverted_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 02:11 /tmp/final_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 16:29 /tmp/final_wordcount
drwxrwxrwt - root supergroup 0 2018-03-29 22:57 /tmp/hadoop-yarn
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 08:48 /tmp/invert_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 14:07 /tmp/invert_out_1
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:37 /tmp/map_clean_output
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:32 /tmp/map_dirty_input
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:38 /tmp/map_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-31 00:03 /tmp/map_red_out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:15 /tmp/out
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:18 /tmp/out1
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:20 /tmp/out2
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:24 /tmp/out3
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:26 /tmp/out4
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:29 /tmp/out5
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:03 /tmp/out6
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 10:12 /tmp/out7
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:54 /tmp/out8
drwxr-xr-x - Tasmia supergroup 0 2018-03-30 13:57 /tmp/out9
drwxr-xr-x - Tasmia supergroup 0 2018-03-29 23:13 /tmp/output
-rw-r--r-- 3 Tasmia supergroup 2271 2018-03-31 02:10 /tmp/raw_input.txt
-rw-r--r-- 3 Tasmia supergroup 161 2018-03-30 10:12 /tmp/sample_input.txt
-rw-r--r-- 3 Tasmia supergroup 7124624 2018-03-31 16:57 /tmp/shakespeare_input.txt

Tasmia@namenode:~$
```

Figure 16. Copying the new shakespeare_input text file from cloud file system to hdfs

7 Run invertedIndexMapreduce

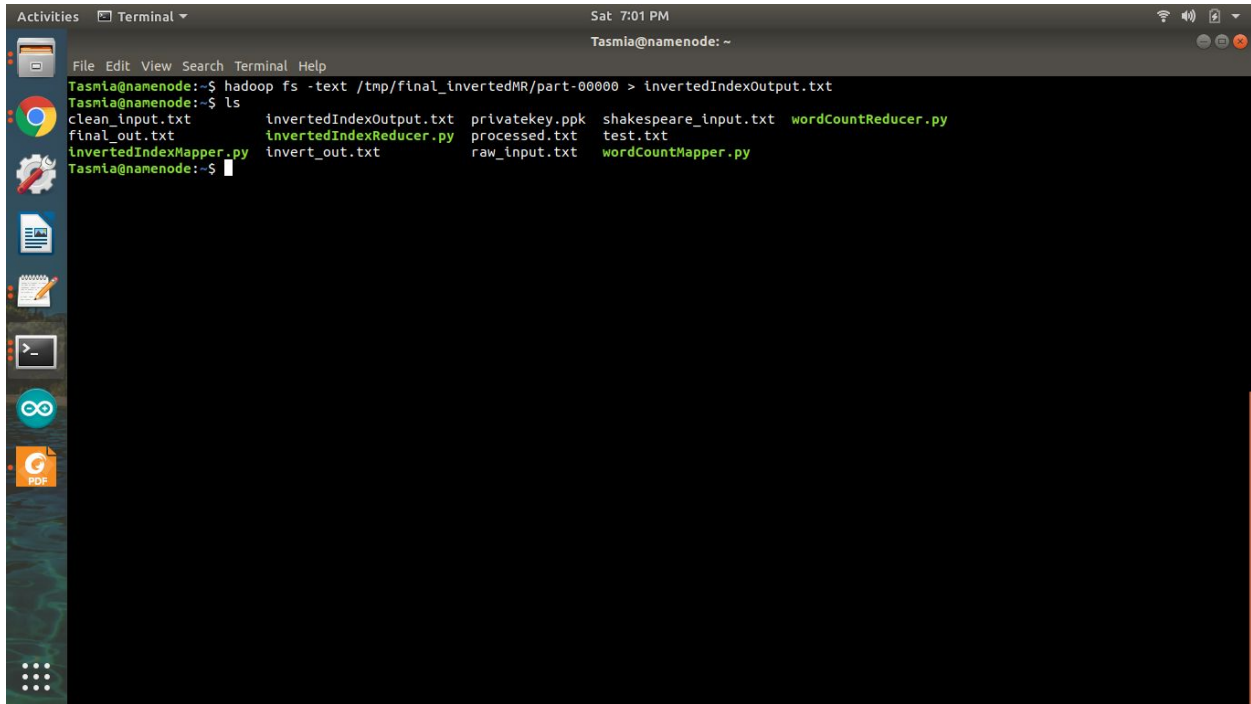


```
Activities Terminal Sat 6:59 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
Tasmia@namenode:~$ hadoop jar /usr/local/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -files invertedIndexMapper.py,invertedIndexReducer.py -mapper invertedIndexMapper.py -reducer invertedIndexReducer.py -input /tmp/shakespeare_input.txt -output /tmp/final_invertedM
R
packageJobJar: [/tmp/hadoop-unjar8685457594475157473/] [] /tmp/streamjob8270315138752792702.jar tmpDir=null
18/03/31 16:59:16 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/10.10.1.2:8032
18/03/31 16:59:17 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/10.10.1.2:8032
18/03/31 16:59:17 INFO mapred.FileInputFormat: Total input paths to process : 1
18/03/31 16:59:17 INFO mapreduce.JobSubmitter: number of splits:2
18/03/31 16:59:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522385828397_0022
18/03/31 16:59:18 INFO impl.YarnClientImpl: Submitted application application_1522385828397_0022
18/03/31 16:59:18 INFO mapreduce.Job: The url to track the job: http://resourcemanager.thadoop.educationproject-pg0.clemson.cloudlab.us:8088/p
roxy/application_1522385828397_0022/
18/03/31 16:59:18 INFO mapreduce.Job: Running job: job_1522385828397_0022
18/03/31 16:59:24 INFO mapreduce.Job: Job job_1522385828397_0022 running in uber mode : false
18/03/31 16:59:24 INFO mapreduce.Job: map 0% reduce 0%
18/03/31 16:59:29 INFO mapreduce.Job: map 100% reduce 0%
18/03/31 16:59:35 INFO mapreduce.Job: map 100% reduce 100%
18/03/31 16:59:35 INFO mapreduce.Job: Job job_1522385828397_0022 completed successfully
18/03/31 16:59:35 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=3636001
FILE: Number of bytes written=7637788
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=7128916
HDFS: Number of bytes written=2034203
HDFS: Number of read operations=9
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=2
Total time spent by all maps in occupied slots (ms)=6102
Total time spent by all reduces in occupied slots (ms)=3307
Total time spent by all map tasks (ms)=6102
Total time spent by all reduce tasks (ms)=3307
```

Figure 17: InvertedIndex MapReduce 100%

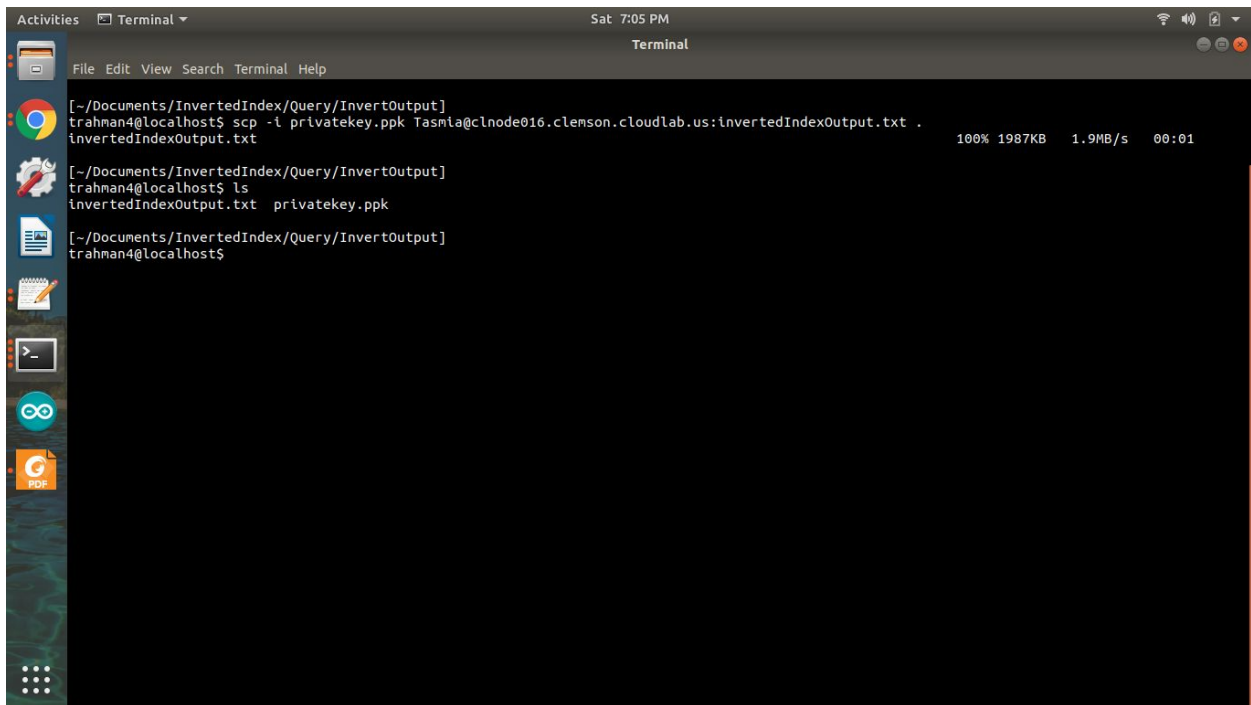
Step 8-9:Query the Inverted Index:

8.Save the output of step 7 into local machine's query directory.



```
Activities Terminal Sat 7:01 PM
Tasmia@namenode: ~
File Edit View Search Terminal Help
Tasmia@namenode:~$ hadoop fs -text /tmp/final_invertedMR/part-00000 > invertedIndexOutput.txt
Tasmia@namenode:~$ ls
clean_input.txt      invertedIndexOutput.txt  privatekey.ppk  shakespeare_input.txt  wordCountReducer.py
final_out.txt         invertedIndexReducer.py  processed.txt    test.txt
invertedIndexMapper.py invert_out.txt           raw_input.txt   wordCountMapper.py
Tasmia@namenode:~$
```

Figure 18: Copying the InvertedIndex MapReduce output to cloud file system



```
Activities Terminal Sat 7:05 PM
Terminal
[~/Documents/InvertedIndex/Query/InvertOutput]
trahman4@localhost$ scp -i privatekey.ppk Tasmia@clnode016.clemson.cloudlab.us:invertedIndexOutput.txt .
invertedIndexOutput.txt 100% 1987KB 1.9MB/s 00:01
[~/Documents/InvertedIndex/Query/InvertOutput]
trahman4@localhost$ ls
invertedIndexOutput.txt privatekey.ppk
[~/Documents/InvertedIndex/Query/InvertOutput]
trahman4@localhost$
```

Figure 19: Copying the InvertedIndex output to query invertoutput directory so that we can make query

9.Run the query.py file to find out the location of a word .


```
1 {'10': {'shakespeare_input.txt': {136792: [2], 275: [0], 54020: [0]}}}
2 {'100': {'shakespeare_input.txt': {1896: [0], 146194: [2]}}}
3 {'1000': {'shakespeare_input.txt': {147244: [1]}}}
4 {'1004': {'shakespeare_input.txt': {147249: [2]}}}
5 {'1009': {'shakespeare_input.txt': {147255: [1]}}}
6 {'101': {'shakespeare_input.txt': {1914: [0]}}}
7 {'1012': {'shakespeare_input.txt': {147258: [1]}}}
8 {'1016': {'shakespeare_input.txt': {147263: [0]}}}
9 {'102': {'shakespeare_input.txt': {1932: [0]}}}
10 {'1020': {'shakespeare_input.txt': {147267: [1]}}}
11 {'1024': {'shakespeare_input.txt': {147272: [2]}}}
12 {'1028': {'shakespeare_input.txt': {147277: [2]}}}
13 {'103': {'shakespeare_input.txt': {1950: [0]}}}
14 {'1033': {'shakespeare_input.txt': {147283: [1]}}}
15 {'1036': {'shakespeare_input.txt': {147286: [0]}}}
16 {'104': {'shakespeare_input.txt': {1968: [0], 146199: [2]}}}
17 {'1040': {'shakespeare_input.txt': {147291: [1]}}}
18 {'1044': {'shakespeare_input.txt': {147295: [1]}}}
19 {'1049': {'shakespeare_input.txt': {147301: [1]}}}
20 {'105': {'shakespeare_input.txt': {1986: [0]}}}
21 {'1053': {'shakespeare_input.txt': {147306: [2]}}}
22 {'1057': {'shakespeare_input.txt': {147311: [2]}}}
23 {'106': {'shakespeare_input.txt': {2004: [0]}}}
24 {'1060': {'shakespeare_input.txt': {147314: [0]}}}
25 {'1065': {'shakespeare_input.txt': {147320: [2]}}}
26 {'1069': {'shakespeare_input.txt': {147325: [1]}}}
27 {'107': {'shakespeare_input.txt': {2022: [0]}}}
28 {'1072': {'shakespeare_input.txt': {147328: [1]}}}
29 {'1078': {'shakespeare_input.txt': {147335: [1]}}}
30 {'108': {'shakespeare_input.txt': {2040: [0], 146203: [0]}}}
31 {'1081': {'shakespeare_input.txt': {147339: [2]}}}
32 {'1085': {'shakespeare_input.txt': {147343: [1]}}}
33 {'1088': {'shakespeare_input.txt': {147347: [2]}}}
34 {'109': {'shakespeare_input.txt': {2058: [0]}}}
35 {'1093': {'shakespeare_input.txt': {147353: [2]}}}
36 {'1096': {'shakespeare_input.txt': {147356: [0]}}}
```

Figure 20. InvertedIndex MapReduce Output

```
[~/Documents/InvertedIndex/Query]
trahman4@localhost$ python query.py InvertOutput/invertedIndexOutput.txt
Type ":quit" to quit.
Query: adding
Query Word: adding
Document: shakespeare_input.txt
Line: 52039, 105292, 67725, 67823, 68177, 99507, 468, 67828, 69340
Query: advance
Query: advanced
Query Word: advanced
Document: shakespeare_input.txt
Line: 127824, 122154, 109812, 96197, 55670
Query: :quit
[~/Documents/InvertedIndex/Query]
trahman4@localhost$
```

Figure 21: Different forms of Query

Directory Description:

- InvertedIndex
 - Preprocess
 - Raw_input
 - Clean
 - dirty
 - process.sh
 - GenerateStopList
 - Raw_input
 - Clean
 - dirty
 - WordCount_Output
 - processed.txt
 - process.sh
 - WordCountMapReduce_Stop
 - Processed_input
 - WordCountMapper.py
 - WordCountReducer.py
 - InvertedMapReduce
 - Shakespear_input.txt
 - invertedIndexMapper.py
 - invertedIndexReducer.py
 - Query
 - InvertedOutput
 - Query.py

Reference:

1. <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
2. <https://stackoverflow.com/questions/28294352/mapreduce-inverted-index-program>
3. https://github.com/kiran4399/inverted-index_hadoop
4. <http://stdatalabs.blogspot.com/2017/03/mapreduce-vs-spark-inverted-index.html>
5. https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm