

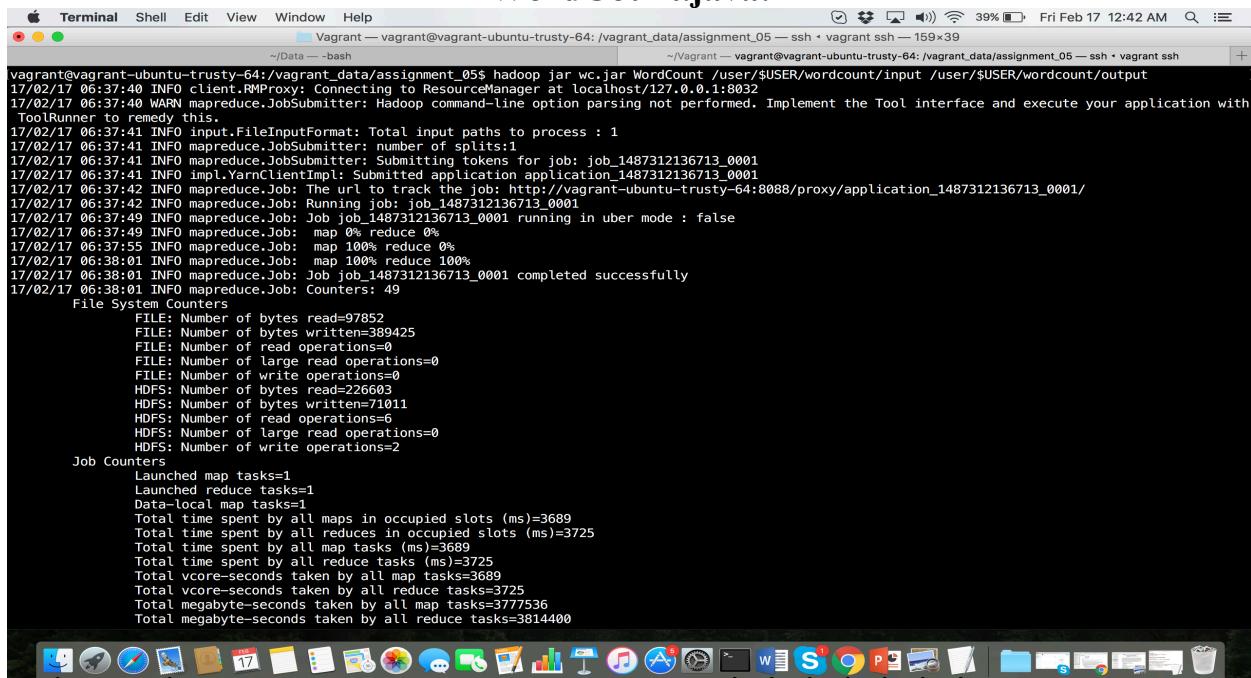
Shrija Chavan
A20381511
ITMD 521
Week 05

Below are the screen shots explaining all the steps of week-05 assignment. The screen shots show the results after executing the following:

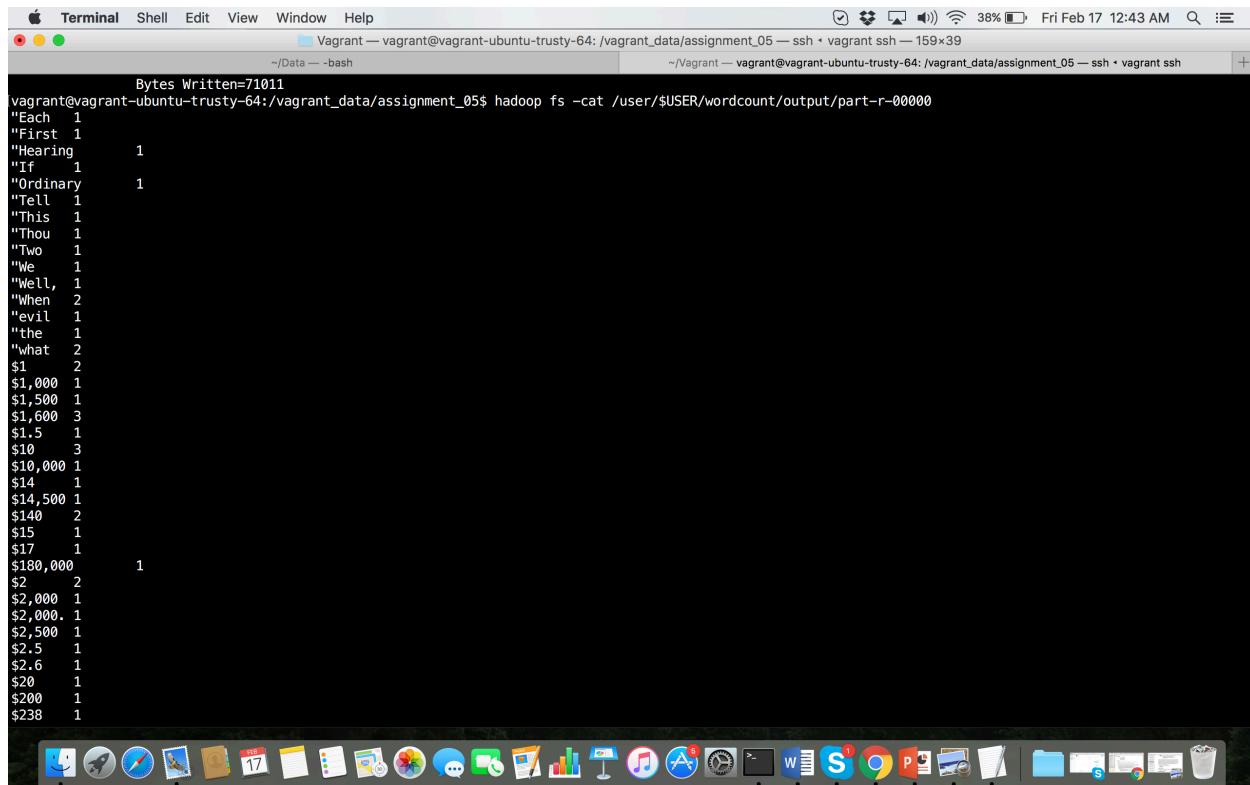
1. WordCount.java
2. Modified WordCount.java to show the list of words which appear more than 4 times
3. WordCount2.java
4. Modified WordCount2.java with case sensitive= true and all the punctuations and preposition removed by giving the skip command from the pattern.txt file.
5. The top ten words for all the above 4 steps.

Note: All the numeric and non-numeric symbols were included in the pattern.txt.

**Screen shots:
WordCount.java:**



```
vagrant@vagrant-ubuntu-trusty-64:~/vagrant_data/assignment_05$ hadoop jar wc.jar WordCount /user/$USER/wordcount/input /user/$USER/wordcount/output
17/02/17 06:37:40 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/17 06:37:40 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/17 06:37:41 INFO input.FileInputFormat: Total input paths to process : 1
17/02/17 06:37:41 INFO mapreduce.JobSubmitter: number of splits:1
17/02/17 06:37:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487312136713_0001
17/02/17 06:37:42 INFO impl.YarnClientImpl: Submitted application application_1487312136713_0001
17/02/17 06:37:42 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1487312136713_0001/
17/02/17 06:37:49 INFO mapreduce.Job: Job job_1487312136713_0001 running in uber mode : false
17/02/17 06:37:49 INFO mapreduce.Job: map 100% reduce 0%
17/02/17 06:37:55 INFO mapreduce.Job: map 100% reduce 0%
17/02/17 06:38:01 INFO mapreduce.Job: map 100% reduce 100%
17/02/17 06:38:01 INFO mapreduce.Job: Job job_1487312136713_0001 completed successfully
17/02/17 06:38:01 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=97852
  FILE: Number of bytes written=389425
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=226603
  HDFS: Number of bytes written=71011
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=3689
  Total time spent by all reduces in occupied slots (ms)=3725
  Total time spent by all map tasks (ms)=3699
  Total time spent by all reduce tasks (ms)=3725
  Total vcore-seconds taken by all map tasks=3689
  Total vcore-seconds taken by all reduce tasks=3725
  Total megabyte-seconds taken by all map tasks=3777536
  Total megabyte-seconds taken by all reduce tasks=3814400
```

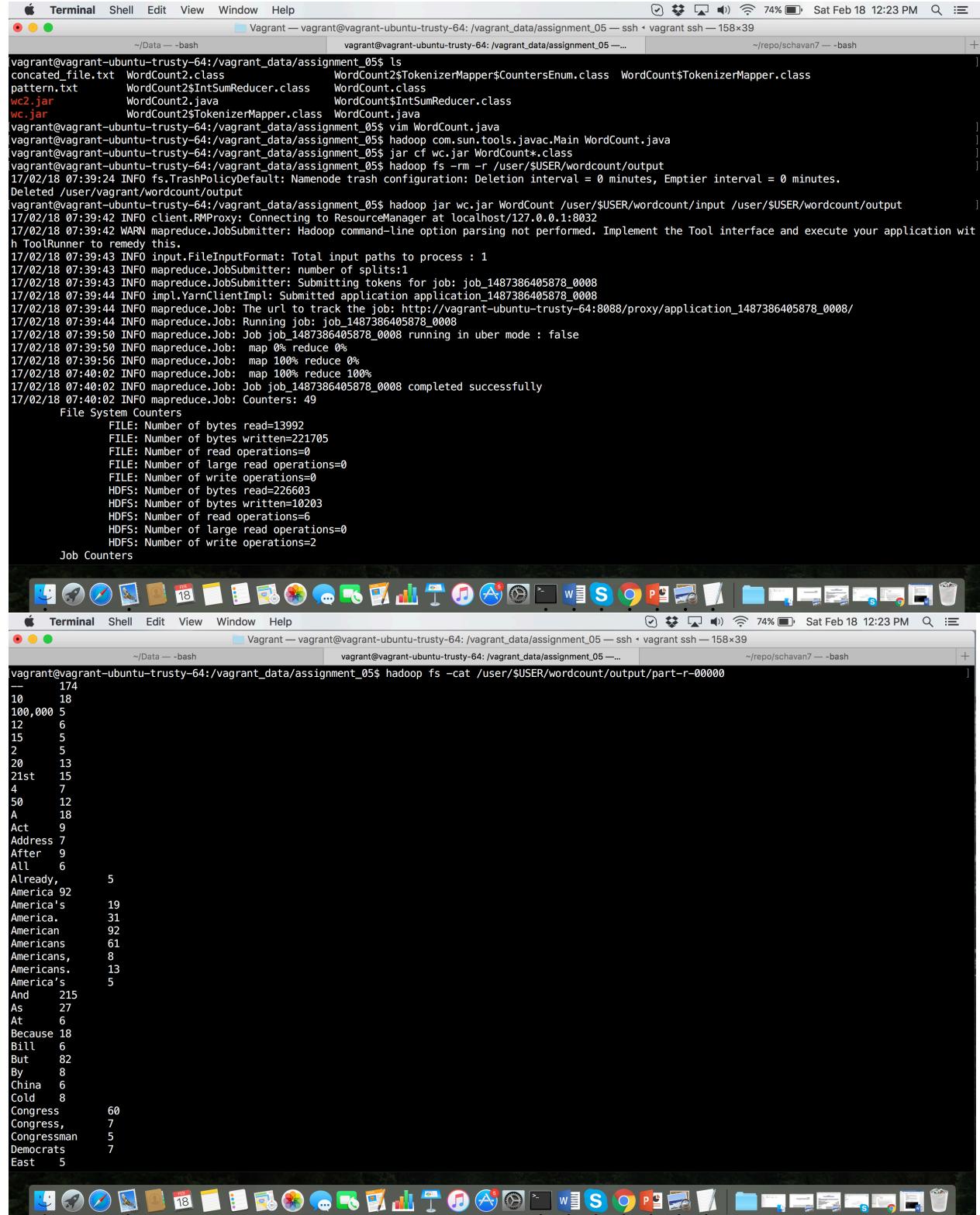


```
Bytes Written=71011
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
'Each 1
'First 1
'Hearing 1
'If 1
'Ordinary 1
'Tell 1
'This 1
'Thou 1
'Two 1
'We 1
'Well, 1
'When 2
'evil 1
'the 1
'what 2
'$1 2
'$1,000 1
'$1,500 1
'$1,600 3
'$1.5 1
'$10 3
'$10,000 1
'$14 1
'$14,500 1
'$140 2
'$15 1
'$17 1
'$180,000 1
'$2 2
'$2,000 1
'$2,000. 1
'$2,500 1
'$2.5 1
'$2.6 1
'$20 1
'$200 1
'$238 1
```

Top 10 words:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
'Each 1
'First 1
'Hearing 1
'If 1
'Ordinary 1
'Tell 1
'This 1
'Thou 1
'Two 1
'We 1
```

Modified WordCount.java to show the list of words which appear more than 4 times:



```
vagrant@vagrant-ubuntu-trusty-64:~/vagrant_data/assignment_05$ ls
concatenated_file.txt  WordCount2.class          WordCount$TokenizerMapper$CountersEnum.class  WordCount$TokenizerMapper.class
pattern.txt           WordCount$IntSumReducer.class  WordCount.class
wc2.jar               WordCount2.java          WordCountsIntSumReducer.class
wc.jar                WordCount$TokenizerMapper.class  WordCount.java
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ vim WordCount.java
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop com.sun.tools.javac.Main WordCount.java
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ jar cf wc.jar WordCount*.class
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -rm -r /user/$USER/wordcount/output
17/02/18 07:39:24 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/vagrant/wordcount/output
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop jar wc.jar WordCount /user/$USER/wordcount/input /user/$USER/wordcount/output
17/02/18 07:39:42 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 07:39:42 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/02/18 07:39:43 INFO input.FileInputFormat: Total input paths to process : 1
17/02/18 07:39:43 INFO mapreduce.JobSubmitter: number of splits:1
17/02/18 07:39:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487386405878_0008
17/02/18 07:39:44 INFO impl.YarnClientImpl: Submitted application application_1487386405878_0008
17/02/18 07:39:44 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1487386405878_0008/
17/02/18 07:39:44 INFO mapreduce.Job: Running job: job_1487386405878_0008
17/02/18 07:39:50 INFO mapreduce.Job: Job job_1487386405878_0008 running in uber mode : false
17/02/18 07:39:50 INFO mapreduce.Job: map 0% reduce 0%
17/02/18 07:39:56 INFO mapreduce.Job: map 100% reduce 0%
17/02/18 07:40:02 INFO mapreduce.Job: map 100% reduce 100%
17/02/18 07:40:02 INFO mapreduce.Job: Job job_1487386405878_0008 completed successfully
17/02/18 07:40:02 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=13992
    FILE: Number of bytes written=221705
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=226603
    HDFS: Number of bytes written=10203
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
vagrant@vagrant-ubuntu-trusty-64:~/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
--      174
10      18
100,000  5
12      6
15      5
2       5
20      13
21st    15
4       7
50      12
A       18
Act     9
Address 7
After   9
All     6
Already  5
America  92
America's 19
America. 31
American 92
Americans 61
Americans, 8
Americans. 13
America's 5
And     215
As      27
At      6
Because 18
Bill    6
But     82
By      8
China   6
Cold    8
Congress 60
Congress, 7
Congressman 5
Democrats 7
East    5
```

Top 10 words:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
-
174
1    4
10   18
100,000 5
12   6
15   5
2    5
20   13
21st 15
4    7
```

WordCount2.java

```
Terminal Shell Edit View Window Help
Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh — 159x39
~/Data — bash ~/Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh +
```

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop jar wc2.jar WordCount2 /user/$USER/wordcount/input /user/$USER/wordcount/output
17/02/18 03:00:18 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 03:00:19 INFO input.FileInputFormat: Total input paths to process : 1
17/02/18 03:00:19 INFO mapreduce.JobSubmitter: number of splits=1
17/02/18 03:00:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487386405878_0001
17/02/18 03:00:19 INFO impl.YarnClientImpl: Submitted application application_1487386405878_0001
17/02/18 03:00:19 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1487386405878_0001/
17/02/18 03:00:19 INFO mapreduce.Job: Running job: job_1487386405878_0001
17/02/18 03:00:28 INFO mapreduce.Job: Job job_1487386405878_0001 running in uber mode : false
17/02/18 03:00:28 INFO mapreduce.Job: map 0% reduce 0%
17/02/18 03:00:34 INFO mapreduce.Job: map 100% reduce 0%
17/02/18 03:00:41 INFO mapreduce.Job: map 100% reduce 100%
17/02/18 03:00:41 INFO mapreduce.Job: Job job_1487386405878_0001 completed successfully
17/02/18 03:00:41 INFO mapreduce.Job: Counters: 50
File System Counters
FILE: Number of bytes read=97852
FILE: Number of bytes written=389735
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=226603
HDFS: Number of bytes written=71011
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=4291
Total time spent by all reduces in occupied slots (ms)=3912
Total time spent by all map tasks (ms)=4291
Total time spent by all reduce tasks (ms)=3912
Total vcore-seconds taken by all map tasks=4291
Total vcore-seconds taken by all reduce tasks=3912
Total megabyte-seconds taken by all map tasks=4393984
Total megabyte-seconds taken by all reduce tasks=4005888
Map-Reduce Framework
Map input records=1096
```



```
Terminal Shell Edit View Window Help
Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh — 159x39
~/Data — -bash ~/Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh +
```

```
File Output Format Counters
Bytes Written=71011
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
"Each 1
"First 1
"Hearing 1
"If 1
"Ordinary 1
"Tell 1
"This 1
"Thou 1
"Two 1
"We 1
"Well, 1
"when 1
"evil 1
"the 1
"what 2
$1 2
$1,000 1
$1,500 1
$1,600 3
$1.5 1
$10 3
$10,000 1
$14 1
$14,500 1
$140 2
$15 1
$17 1
$180,000 1
$2 2
$2,000 1
$2,000. 1
$2,500 1
$2.5 1
$2.6 1
$20 1
$200 1
```

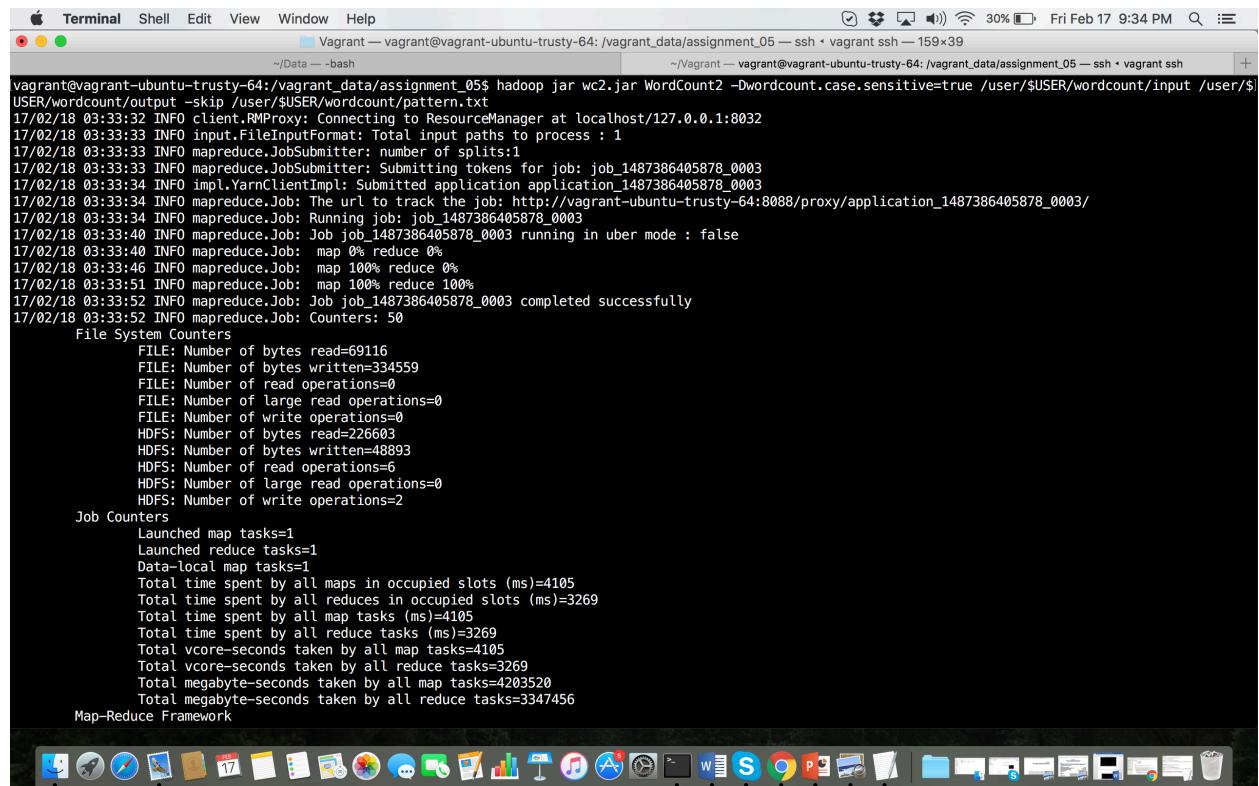
```
Terminal Shell Edit View Window Help
Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh — 159x39
~/Data — -bash ~/Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh ✘ vagrant ssh +
```

```
Abbas. 1
Abess, 1
About 1
Abraham 1
Accountability 1
Across 1
Act 9
Act, 1
Action 3
Address 7
Administration. 1
Affordable 1
Afghan 3
Afghanistan 4
Afghanistan, 3
Afghanistan. 2
Africa. 2
Africa; 1
African 1
After 9
Against 1
Age 1
Age, 1
Agreement 1
Airlines 1
Algiers 1
Algiers, 1
Alice, 1
All 6
Allawi 1
Almost 1
Along 1
Already, 5
Alzheimer's. 1
Alzheimer's. 1
Amendment 2
AmeriCorps 1
AmeriCorps, 1
America 92
```

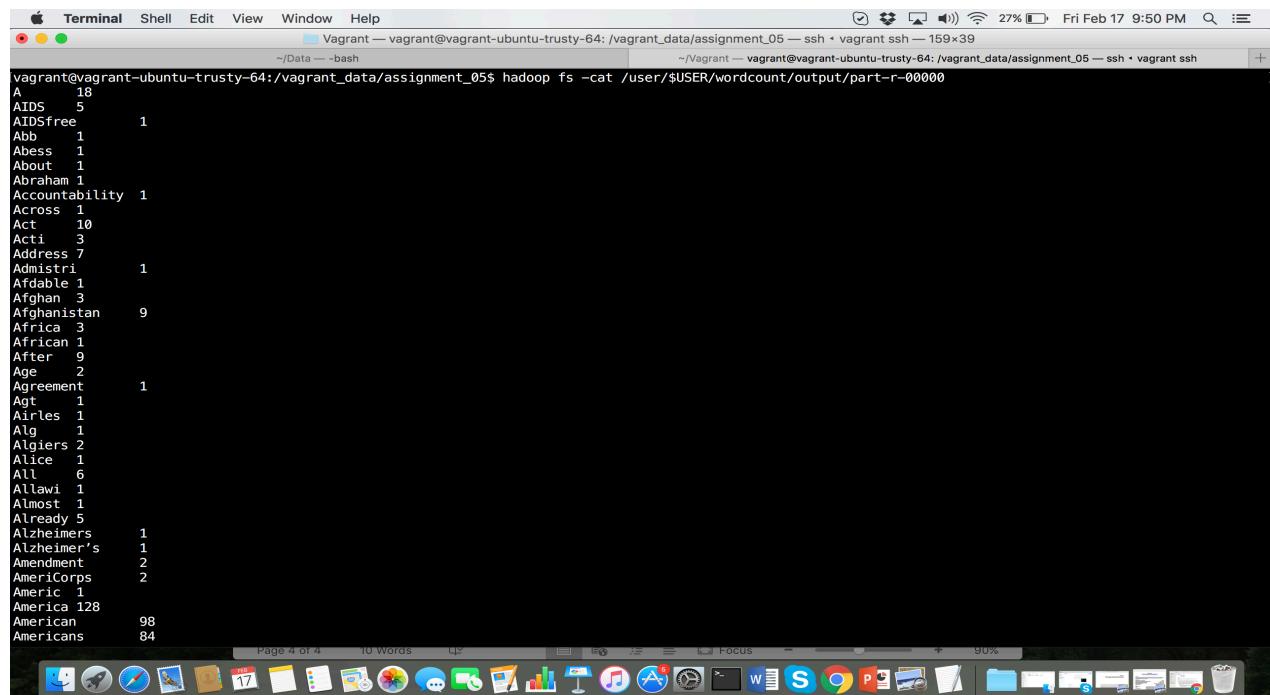
Top 10 words:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
'Each 1
'First 1
'Hearing 1
'If 1
'Ordinary 1
'Tell 1
'This 1
'Thou 1
'Two 1
'We 1
```

Modified WordCount2.java with the case sensitive= true and all the and punctuations and preposition removed by giving the skip command for the pattern.txt file.



```
Terminal Shell Edit View Window Help
Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh + vagrant ssh — 159x39
~/Data — bash ~[Vagrant] — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh + vagrant ssh
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop jar wc2.jar WordCount2 --wordcount.case.sensitive=true /user/$USER/wordcount/input /user/$USER/wordcount/output --skip /user/$USER/wordcount/pattern.txt
17/02/18 03:33:32 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 03:33:33 INFO input.FileInputFormat: Total input paths to process : 1
17/02/18 03:33:33 INFO mapreduce.JobSubmitter: number of splits:1
17/02/18 03:33:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487386405878_0003
17/02/18 03:33:34 INFO impl.YarnClientImpl: Submitted application application_1487386405878_0003
17/02/18 03:33:34 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1487386405878_0003/
17/02/18 03:33:34 INFO mapreduce.Job: Running job: job_1487386405878_0003
17/02/18 03:33:40 INFO mapreduce.Job: Job job_1487386405878_0003 running in uber mode : false
17/02/18 03:33:40 INFO mapreduce.Job: map 0% reduce 0%
17/02/18 03:33:46 INFO mapreduce.Job: map 100% reduce 0%
17/02/18 03:33:51 INFO mapreduce.Job: map 100% reduce 100%
17/02/18 03:33:52 INFO mapreduce.Job: Job job_1487386405878_0003 completed successfully
17/02/18 03:33:52 INFO mapreduce.Job: Counters: 50
  File System Counters
    FILE: Number of bytes read=69116
    FILE: Number of bytes written=334559
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=226603
    HDFS: Number of bytes written=48893
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=4105
    Total time spent by all reduces in occupied slots (ms)=3269
    Total time spent by all map tasks (ms)=4105
    Total time spent by all reduce tasks (ms)=3269
    Total vcore-seconds taken by all map tasks=4105
    Total vcore-seconds taken by all reduce tasks=3269
    Total megabyte-seconds taken by all map tasks=4203520
    Total megabyte-seconds taken by all reduce tasks=3347456
Map-Reduce Framework
```



A screenshot of a Mac OS X desktop environment. In the center is a Terminal window titled "Vagrant — vagrant@vagrant-ubuntu-trusty-64: /vagrant_data/assignment_05 — ssh + vagrant.ssh — 159x39 ~/Data — bash". The window displays a list of words and their counts from a Hadoop word count operation. The output is as follows:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
A      18
AIDS    5
AIDSfree   1
Abb     1
Abess   1
About   1
Abraham 1
Accountability 1
Across   1
Act     10
Acti    3
Address  7
Admistri 1
Afdbable 1
Afghan   3
Afghanistan 9
Africa   3
African  1
After    9
Age     2
Agreement 1
Agt     1
Airlines 1
Alg     1
Algiers  2
Alice   1
All     6
Allawi   1
Almost   1
Already  5
Alzheimers 1
Alzheimer's 1
Amendment 2
Americorps 2
Americ 1
America 128
American 98
Americans 84
```

The desktop bar at the bottom shows various application icons, and the Dock below contains more application icons.

Top 10 words:

```
vagrant@vagrant-ubuntu-trusty-64:/vagrant_data/assignment_05$ hadoop fs -cat /user/$USER/wordcount/output/part-r-00000
A      18
AIDS    5
AIDSfree   1
Abb     1
Abess   1
About   1
Abraham 1
Accountability 1
Across   1
Act     10
```