Illinois Institute of Technology **School of Applied Technology**

ITMD – 521 **Assignment - 05**

Name : Aneesh Partha **CWID- A20376172**

Table of Contents

PART ONE – *TEXTFILES*..

PART TWO – *INPUT AND OUTPUT FILE PATH IN HADOOP*..

PART THREE – *WORDCOUNT1*..

PART FOUR - *WORDCOUNT2*……….

PART FIVE - *WORDCOUNT1 USED FOR DISPLAYING WORDS OCCURRING MORE THAN OR EQUAL TO 4 TIMES*………..

PART SIX - *WORDCOUNT PROGRAM USING SKIP AND DWORDCOUNT OPTION*…………………………………..

PART SEVEN - *TOP 10 WORDS OF EACH FILE*………………………………………………………………………………….

Name : Aneesh Partha **CWID- A20376172**

WordCount

Text files:

Below is the list of files from which word count will be calculated. The below files are kept under a single folder in Hadoop filesystem.

1796-sotu.txt
1993-sotu.txt
1997-sotu.txt
2001-sotu.txt
2005-sotu.txt
2009-sotu.txt
2013-sotu.txt

Input and output file path in Hadoop:

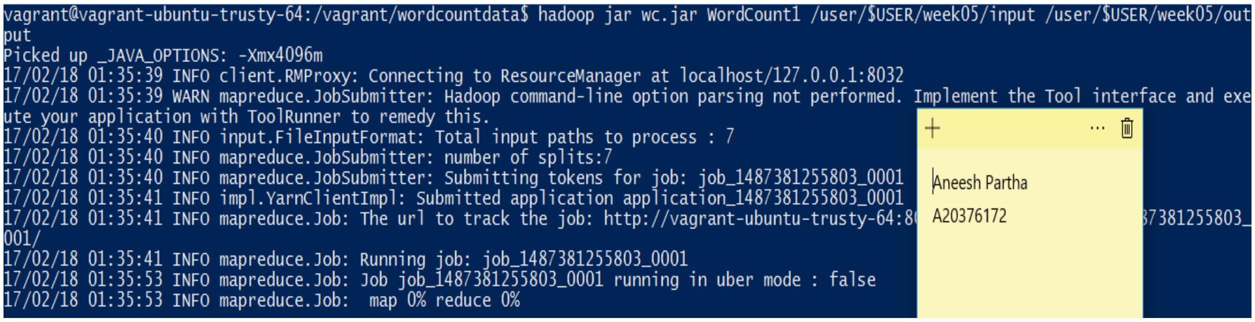
Input: /user/$USER/week05/input

Output:/user/$USER/week05/output*

WordCount1:

WordCount program is renamed as WordCount1 for usability. The WordCount1 is compiled and Jar file is created in the same location. Later the MapReduce program is triggered using the below command

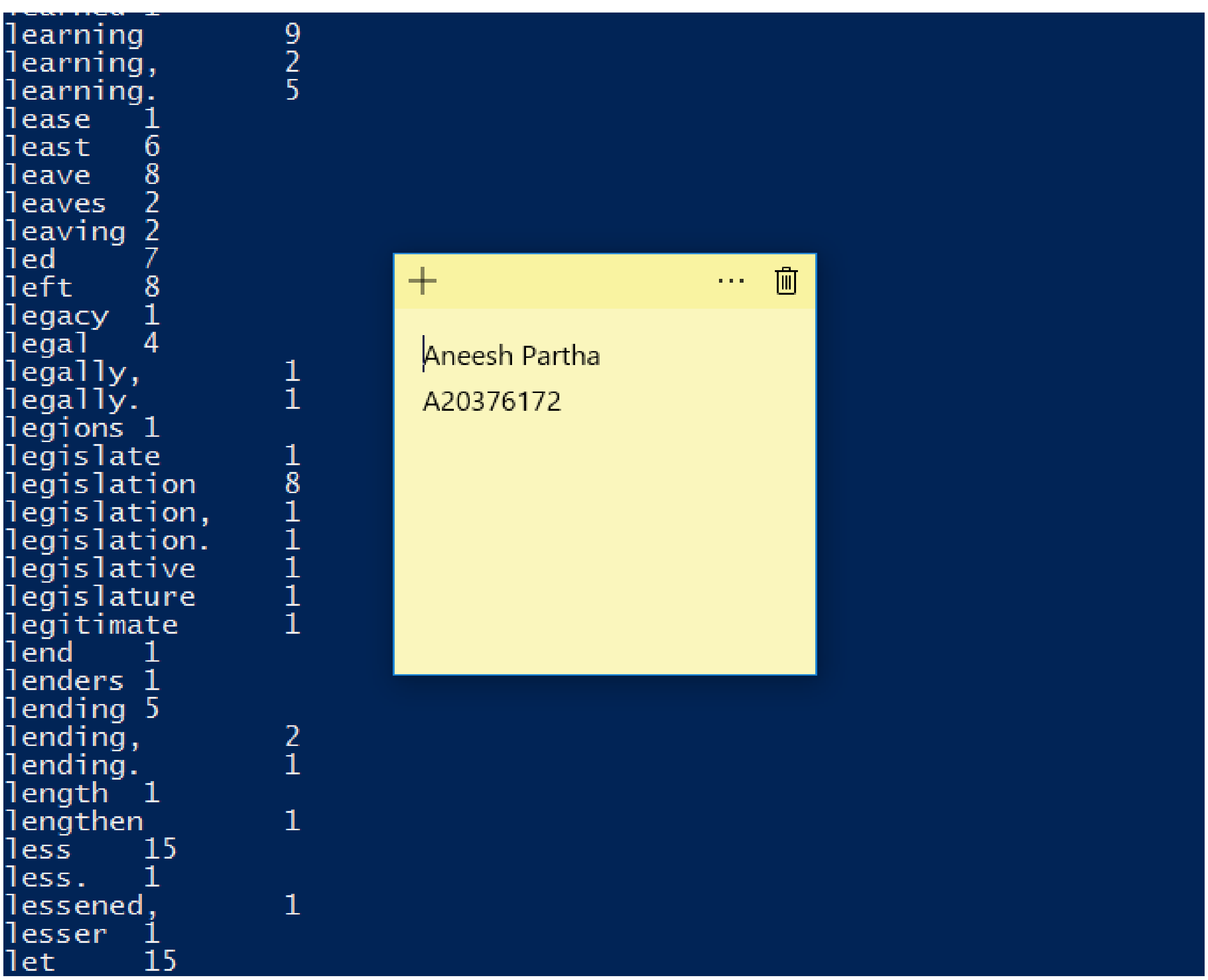
Hadoop jar wc.jar WordCount1 /user/$USER/week05/input /user/$USER/week05/output



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata$ hadoop jar wc.jar WordCount1 /user/$USER/week05/input /user/$USER/week05/out
put
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/18 01:35:39 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 01:35:39 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
ute your application with ToolRunner to remedy this.
17/02/18 01:35:40 INFO input.FileInputFormat: Total input paths to process : 7
17/02/18 01:35:40 INFO mapreduce.JobSubmitter: number of splits:7
17/02/18 01:35:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487381255803_0001
17/02/18 01:35:41 INFO impl.YarnClientImpl: Submitted application application_1487381255803_0001
17/02/18 01:35:41 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8                    37381255803_
001/
17/02/18 01:35:41 INFO mapreduce.Job: Running job: job_1487381255803_0001
17/02/18 01:35:53 INFO mapreduce.Job: Job job_1487381255803_0001 running in uber mode : false
17/02/18 01:35:53 INFO mapreduce.Job:  map 0% reduce 0%
```

Aneesh Partha
A20376172

Output:



```
learned 1
learning        9
learning,       2
learning.       5
lease   1
least   6
leave   8
leaves  2
leaving 2
led     7
left    8
legacy  1
legal   4
legally,        1
legally.        1
legions 1
legislate       1
legislation     8
legislation,    1
legislation.    1
legislative     1
legislature     1
legitimate      1
lend    1
lenders 1
lending 5
lending,        2
lending.        1
length  1
lengthen        1
less    15
less.   1
lessened,       1
lesser  1
let     15
```

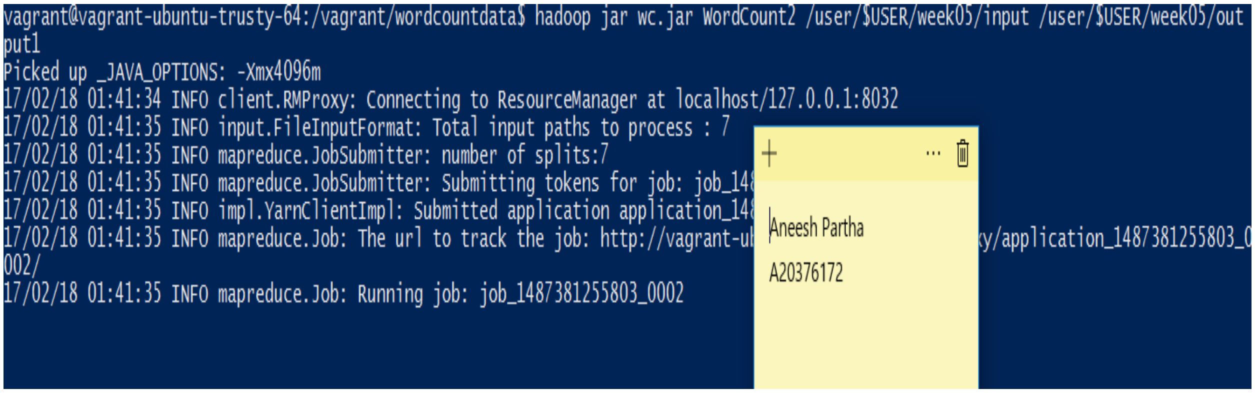
Aneesh Partha

A20376172

WordCount2:

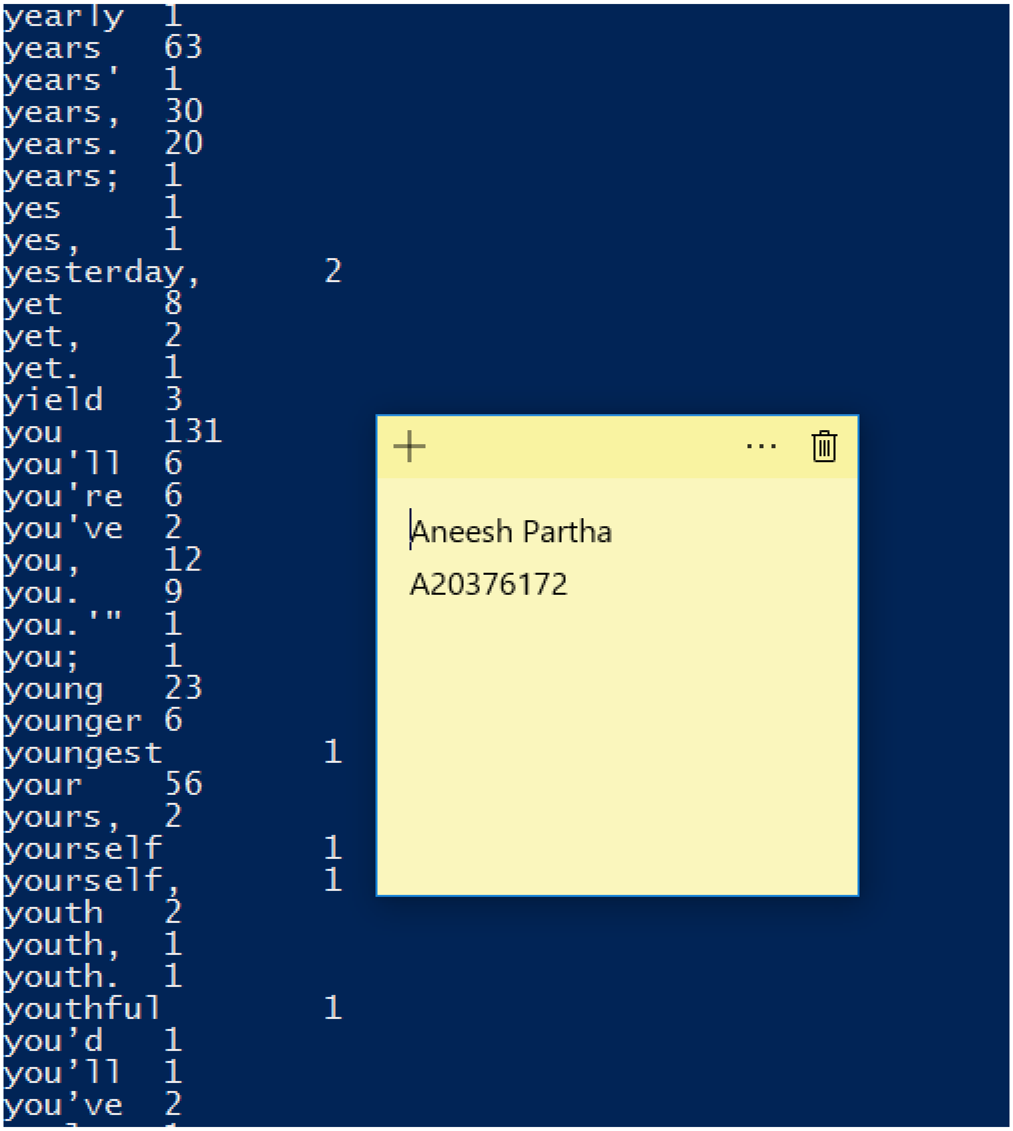
WordCount2 program is saved in the local filesystem and compiled. Jar file is created for the same and the command is executed.

Hadoop jar wc.jar WordCount2 /user/$USER/week05/input /user/$USER/week05/output1



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata$ hadoop jar wc.jar WordCount2 /user/$USER/week05/input /user/$USER/week05/out
put1
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/18 01:41:34 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 01:41:35 INFO input.FileInputFormat: Total input paths to process : 7
17/02/18 01:41:35 INFO mapreduce.JobSubmitter: number of splits:7
17/02/18 01:41:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_148
17/02/18 01:41:35 INFO impl.YarnClientImpl: Submitted application application_148
17/02/18 01:41:35 INFO mapreduce.Job: The url to track the job: http://vagrant-ub                   y/application_1487381255803_0
002/
17/02/18 01:41:35 INFO mapreduce.Job: Running job: job_1487381255803_0002
```

Output:

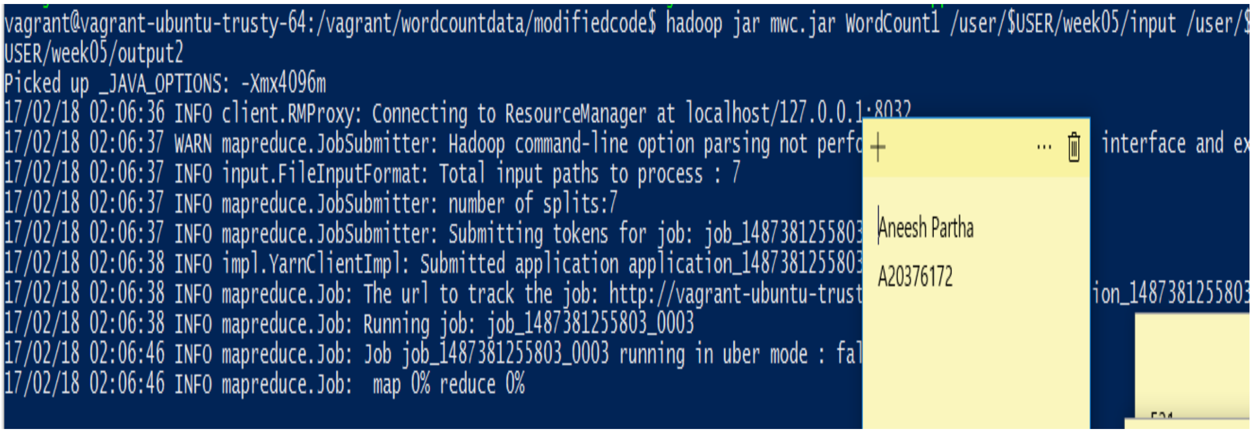


```
yearly  1
years   63
years'  1
years,  30
years.  20
years;  1
yes     1
yes,    1
yesterday,      2
yet     8
yet,    2
yet.    1
yield   3
you     131
you'll  6
you're  6
you've  2
you,    12
you.    9
you.'"  1
you;    1
young   23
younger 6
youngest        1
your    56
yours,  2
yourself        1
yourself,       1
youth   2
youth,  1
youth.  1
youthful        1
you’d   1
you’ll  1
you’ve  2
```

Name : Aneesh Partha CWID- A20376172

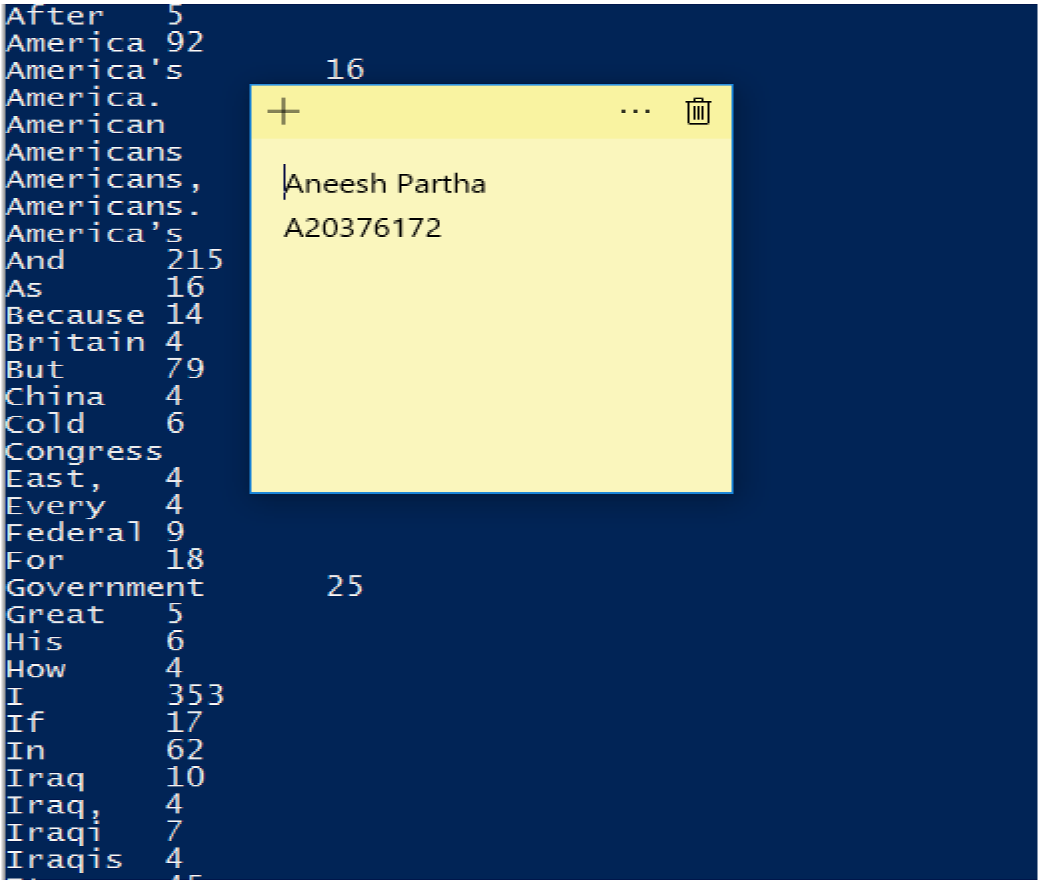
WordCount1 used for displaying words occurring more than or equal to 4 times

Hadoop jar mwc.jar WordCount1 /user/$USER/week05/input /user/$USER/week05/output2



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata/modifiedcode$ hadoop jar mwc.jar WordCount1 /user/$USER/week05/input /user/$
USER/week05/output2
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/18 02:06:36 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 02:06:37 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not perfo          interface and ex
17/02/18 02:06:37 INFO input.FileInputFormat: Total input paths to process : 7
17/02/18 02:06:37 INFO mapreduce.JobSubmitter: number of splits:7
17/02/18 02:06:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487381255803
17/02/18 02:06:38 INFO impl.YarnClientImpl: Submitted application application_1487381255803
17/02/18 02:06:38 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trust          ion_1487381255803
17/02/18 02:06:38 INFO mapreduce.Job: Running job: job_1487381255803_0003
17/02/18 02:06:46 INFO mapreduce.Job: Job job_1487381255803_0003 running in uber mode : fal
17/02/18 02:06:46 INFO mapreduce.Job:  map 0% reduce 0%
```

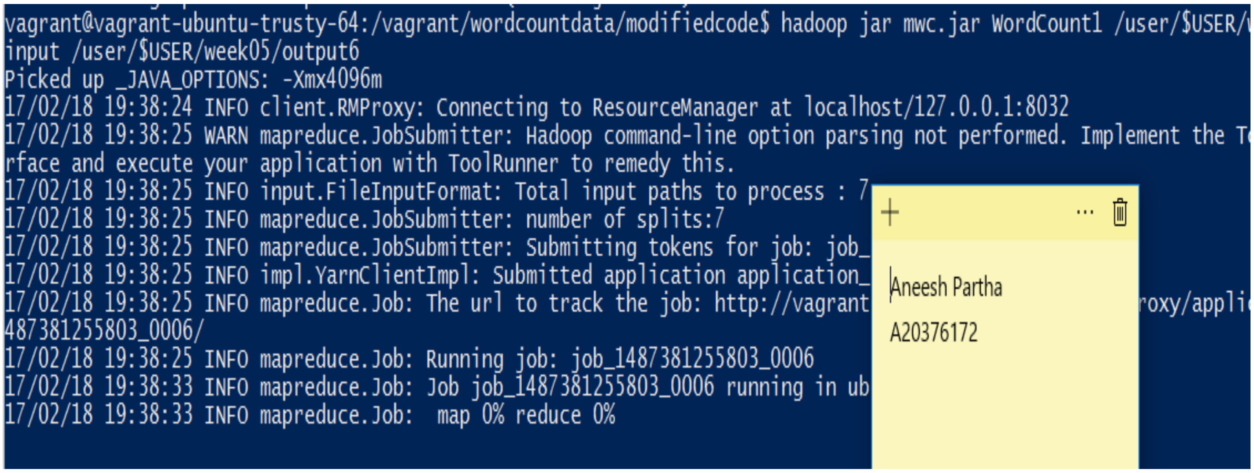
Output:



```
After   5
America 92
America's       16
America.
American
Americans
Americans,
Americans.
America’s
And     215
As      16
Because 14
Britain 4
But     79
China   4
Cold    6
Congress
East,   4
Every   4
Federal 9
For     18
Government      25
Great   5
His     6
How     4
I       353
If      17
In      62
Iraq    10
Iraq,   4
Iraqi   7
Iraqis  4
```

WordCount1 used for displaying words more than 4 times

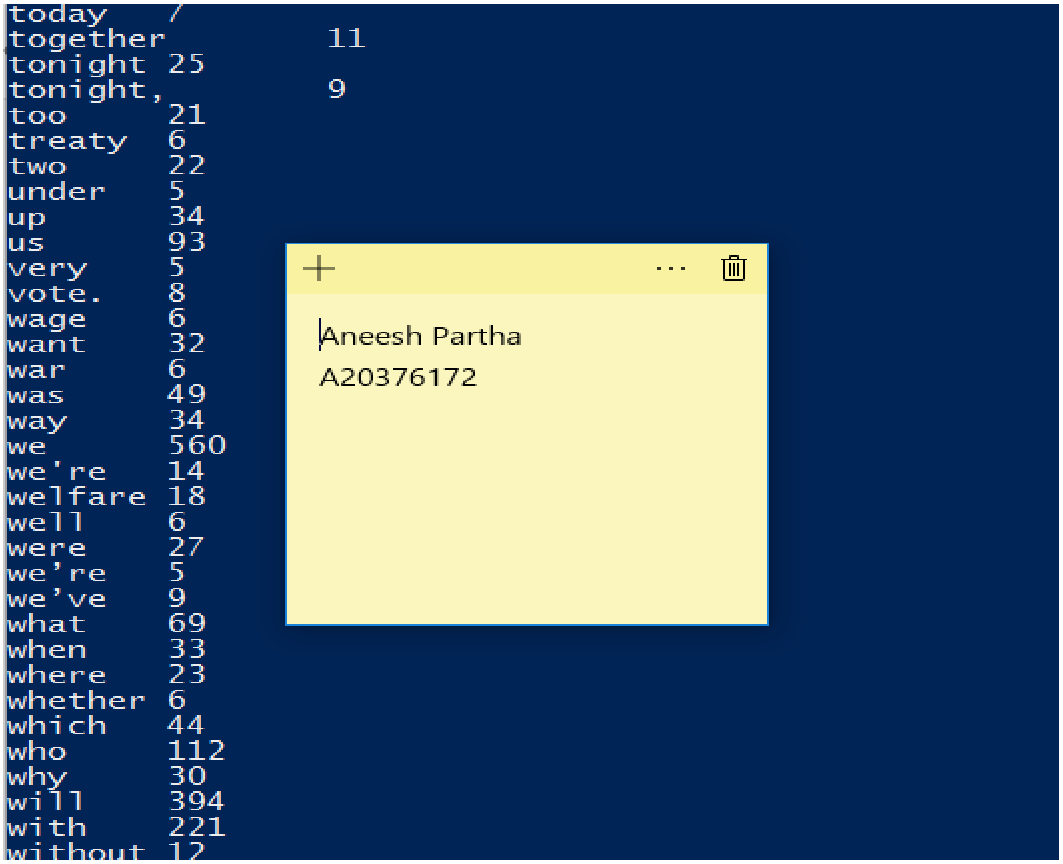
Hadoop jar mwc.jar WordCount1 /user/$USER/week05/input /user/$USER/week05/output6



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata/modifiedcode$ hadoop jar mwc.jar WordCount1 /user/$USER/w
input /user/$USER/week05/output6
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/18 19:38:24 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 19:38:25 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the T
rface and execute your application with ToolRunner to remedy this.
17/02/18 19:38:25 INFO input.FileInputFormat: Total input paths to process : 7
17/02/18 19:38:25 INFO mapreduce.JobSubmitter: number of splits:7
17/02/18 19:38:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_
17/02/18 19:38:25 INFO impl.YarnClientImpl: Submitted application application_
17/02/18 19:38:25 INFO mapreduce.Job: The url to track the job: http://vagrant                                 roxy/appli
487381255803_0006/
17/02/18 19:38:25 INFO mapreduce.Job: Running job: job_1487381255803_0006
17/02/18 19:38:33 INFO mapreduce.Job: Job job_1487381255803_0006 running in ub
17/02/18 19:38:33 INFO mapreduce.Job:  map 0% reduce 0%
```

Aneesh Partha
A20376172

Output:



```
today   7
together        11
tonight 25
tonight,        9
too     21
treaty  6
two     22
under   5
up      34
us      93
very    5
vote.   8
wage    6
want    32
war     6
was     49
way     34
we      560
we're   14
welfare 18
well    6
were    27
we’re   5
we’ve   9
what    69
when    33
where   23
whether 6
which   44
who     112
why     30
will    394
with    221
without 12
```

Aneesh Partha
A20376172

WordCount program using skip and Dwordcount option:

Pattern.txt:

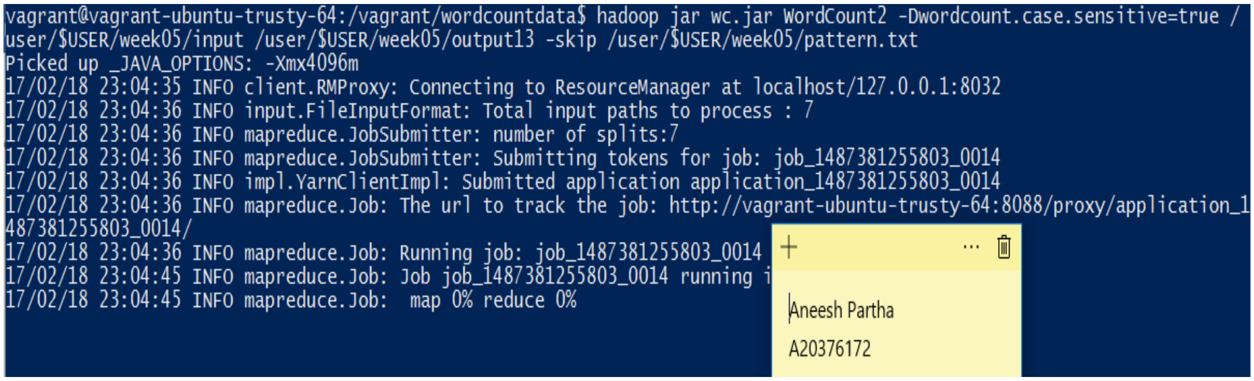
File contains prepositions and punctuations which must be ignored during the output.

Note : Apostrophe between a word is not included in pattern as this would decrease the readability of the word.

For ex – We’re

With Dwordcount.case.sensitive= true option

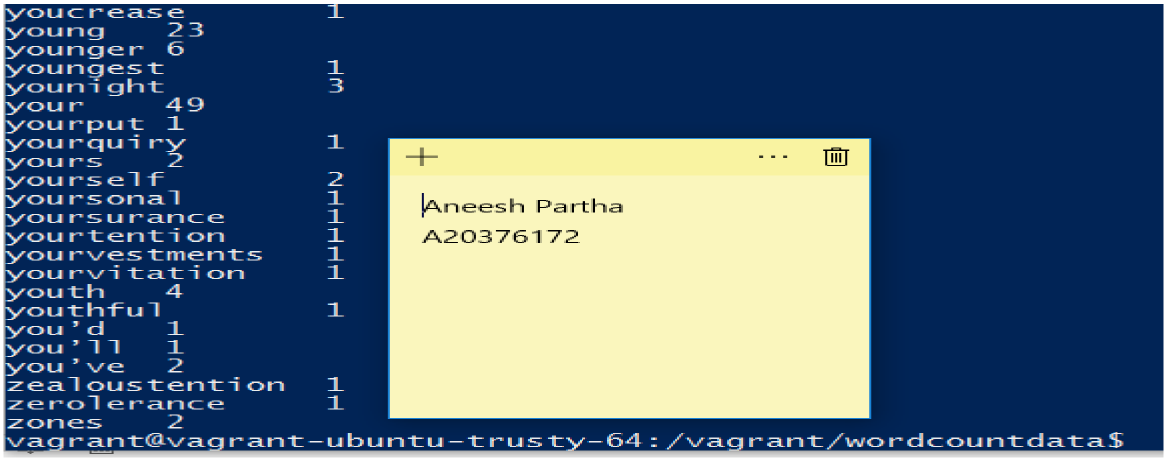
Hadoop jar wc.jar WordCount2 -Dwordcount.case.sensitive=true /user/$USER/week05/input /user/$USER/week05/output13 -skip /user/$USER/week05/pattern.txt



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata$ hadoop jar wc.jar WordCount2 -Dwordcount.case.sensitive=true /
user/$USER/week05/input /user/$USER/week05/output13 -skip /user/$USER/week05/pattern.txt
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/18 23:04:35 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
17/02/18 23:04:36 INFO input.FileInputFormat: Total input paths to process : 7
17/02/18 23:04:36 INFO mapreduce.JobSubmitter: number of splits:7
17/02/18 23:04:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1487381255803_0014
17/02/18 23:04:36 INFO impl.YarnClientImpl: Submitted application application_1487381255803_0014
17/02/18 23:04:36 INFO mapreduce.Job: The url to track the job: http://vagrant-ubuntu-trusty-64:8088/proxy/application_1
487381255803_0014/
17/02/18 23:04:36 INFO mapreduce.Job: Running job: job_1487381255803_0014
17/02/18 23:04:45 INFO mapreduce.Job: Job job_1487381255803_0014 running i
17/02/18 23:04:45 INFO mapreduce.Job:  map 0% reduce 0%
```

Aneesh Partha
A20376172

Output:



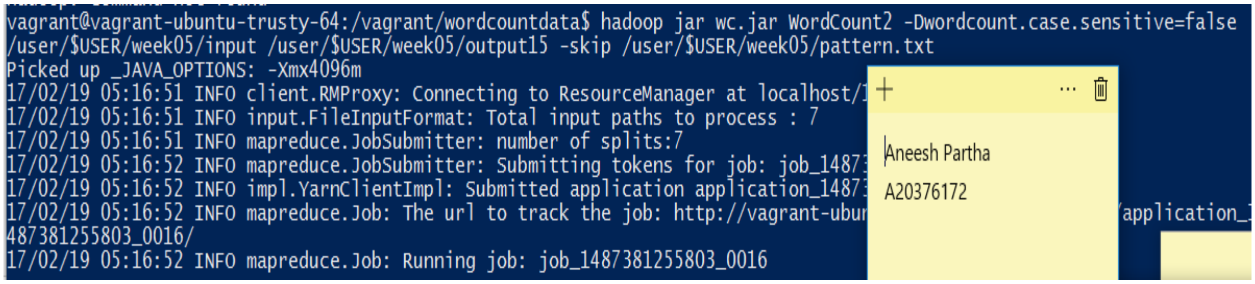
```
youcrease       1
young   23
younger 6
youngest        1
younight        3
your    49
yourput 1
yourquiry       1
yours   2
yourself        2
yoursonal       1
yoursurance     1
yourtention     1
yourvestments   1
yourvitation    1
youth   4
youthful        1
you’d   1
you’ll  1
you’ve  2
zealoustention  1
zerolerance     1
zones   2
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata$
```

Aneesh Partha
A20376172

Name : Aneesh Partha CWID- A20376172

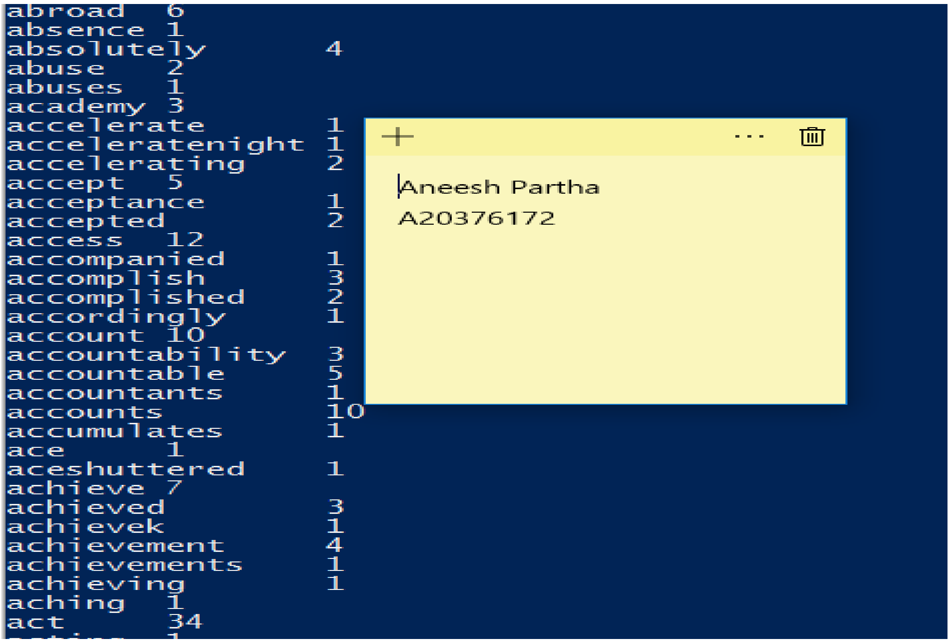
With Dwordcount.case.sensitive= false option

Hadoop jar wc.jar WordCount2 -Dwordcount.case.sensitive=true /user/$USER/week05/input /user/$USER/week05/output15 -skip /user/$USER/week05/pattern.txt



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata$ hadoop jar wc.jar WordCount2 -Dwordcount.case.sensitive=false
/user/$USER/week05/input /user/$USER/week05/output15 -skip /user/$USER/week05/pattern.txt
Picked up _JAVA_OPTIONS: -Xmx4096m
17/02/19 05:16:51 INFO client.RMProxy: Connecting to ResourceManager at localhost/1
17/02/19 05:16:51 INFO input.FileInputFormat: Total input paths to process : 7
17/02/19 05:16:51 INFO mapreduce.JobSubmitter: number of splits:7
17/02/19 05:16:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14873
17/02/19 05:16:52 INFO impl.YarnClientImpl: Submitted application application_14873
17/02/19 05:16:52 INFO mapreduce.Job: The url to track the job: http://vagrant-ubun                    /application_1
487381255803_0016/
17/02/19 05:16:52 INFO mapreduce.Job: Running job: job_1487381255803_0016
```

Aneesh Partha
A20376172

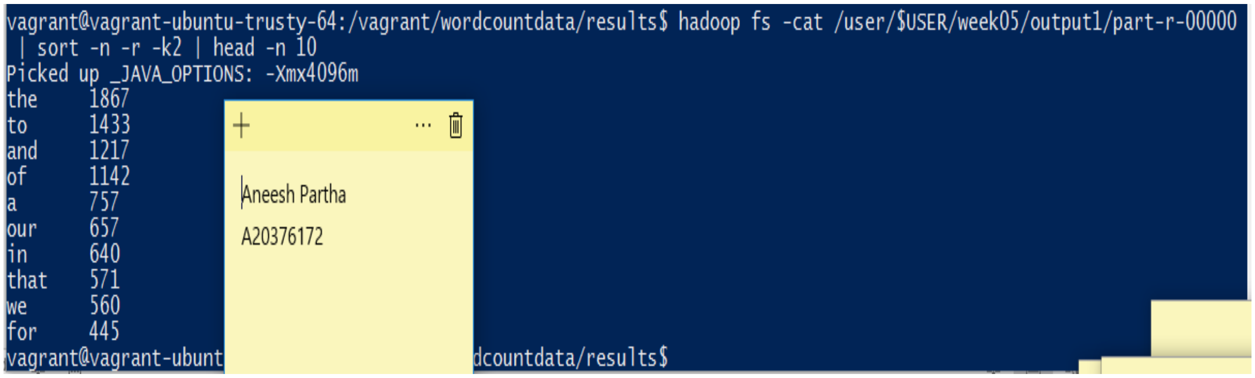


```
abroad   6
absence  1
absolutely       4
abuse    2
abuses   1
academy  3
accelerate       1
acceleratenight  1
accelerating     2
accept   5
acceptance       1
accepted         2
access   12
accompanied      1
accomplish       3
accomplished     2
accordingly      1
account  10
accountability   3
accountable      5
accountants      1
accounts         10
accumulates      1
ace      1
aceshuttered     1
achieve  7
achieved         3
achievek         1
achievement      4
achievements     1
achieving        1
aching   1
act      34
```

Aneesh Partha
A20376172

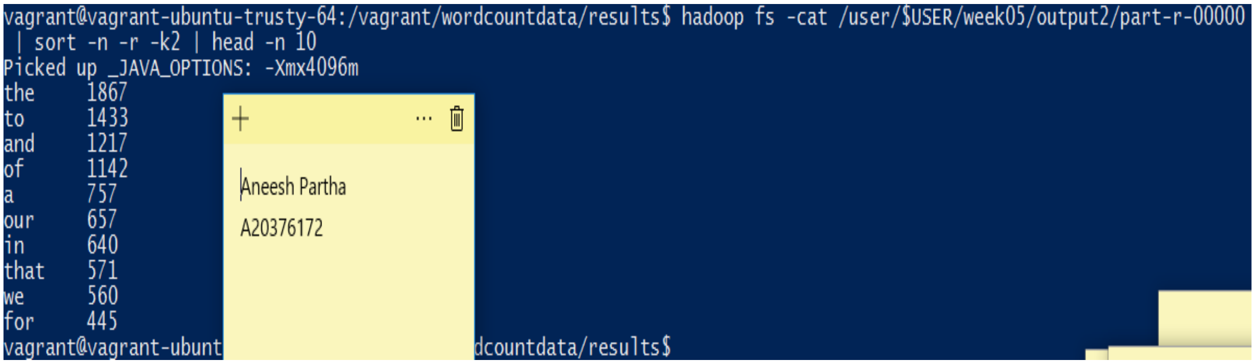
Top 10 words of each file

hadoop fs -cat /user/$USER/week05/output1/part-r-00000 | sort -k2 -r | head



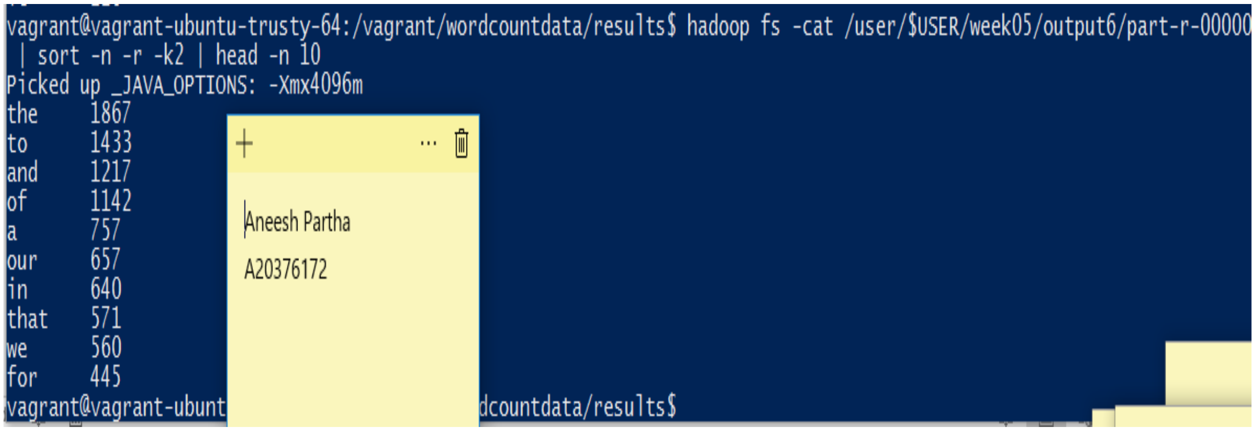
```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata/results$ hadoop fs -cat /user/$USER/week05/output1/part-r-00000
 | sort -n -r -k2 | head -n 10
Picked up _JAVA_OPTIONS: -Xmx4096m
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
vagrant@vagrant-ubunt                         dcountdata/results$
```

hadoop fs -cat /user/$USER/week05/output2/part-r-00000 | sort -k2 -r | head



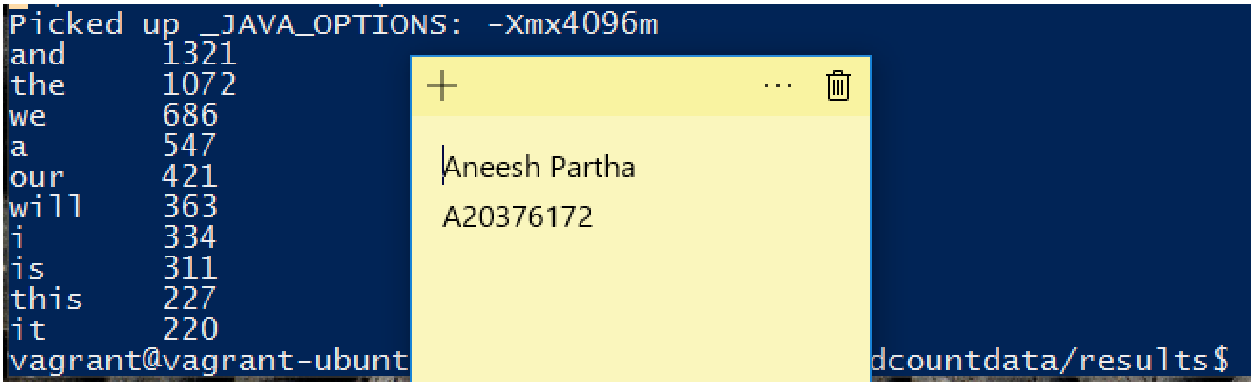
```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata/results$ hadoop fs -cat /user/$USER/week05/output2/part-r-00000
 | sort -n -r -k2 | head -n 10
Picked up _JAVA_OPTIONS: -Xmx4096m
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
vagrant@vagrant-ubunt                         dcountdata/results$
```

hadoop fs -cat /user/$USER/week05/output6/part-r-00000 | sort -k2 -r | head



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/wordcountdata/results$ hadoop fs -cat /user/$USER/week05/output6/part-r-00000
 | sort -n -r -k2 | head -n 10
Picked up _JAVA_OPTIONS: -Xmx4096m
the     1867
to      1433
and     1217
of      1142
a       757
our     657
in      640
that    571
we      560
for     445
vagrant@vagrant-ubunt                         dcountdata/results$
```

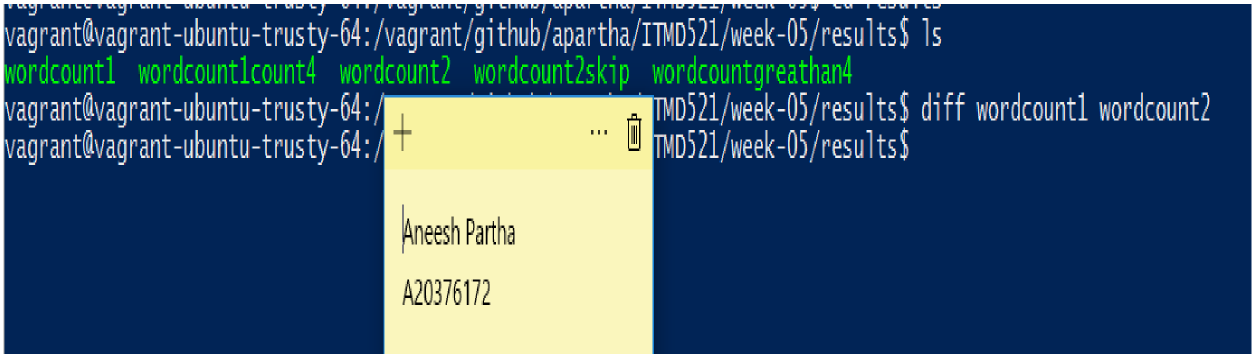
hadoop fs -cat /user/$USER/week05/output13/part-r-00000 | sort -k2 -r | head



```
Picked up _JAVA_OPTIONS: -Xmx4096m
and      1321
the      1072
we       686
a        547
our      421
will     363
i        334
is       311
this     227
it       220
vagrant@vagrant-ubunt                        dcountdata/results$
```

Difference between WordCount1 and WordCount2

Using diff keyword you can determine if there are any differences between the output of both programs which was triggered as part of wordcount1 and wordcount2.



```
vagrant@vagrant-ubuntu-trusty-64:/vagrant/github/apartha/ITMD521/week-05$ cd results
vagrant@vagrant-ubuntu-trusty-64:/vagrant/github/apartha/ITMD521/week-05/results$ ls
wordcount1  wordcount1count4  wordcount2  wordcount2skip  wordcountgreathan4
vagrant@vagrant-ubuntu-trusty-64:/              TMD521/week-05/results$ diff wordcount1 wordcount2
vagrant@vagrant-ubuntu-trusty-64:/              TMD521/week-05/results$
```