CMPE 282

LAB 2 Hadoop

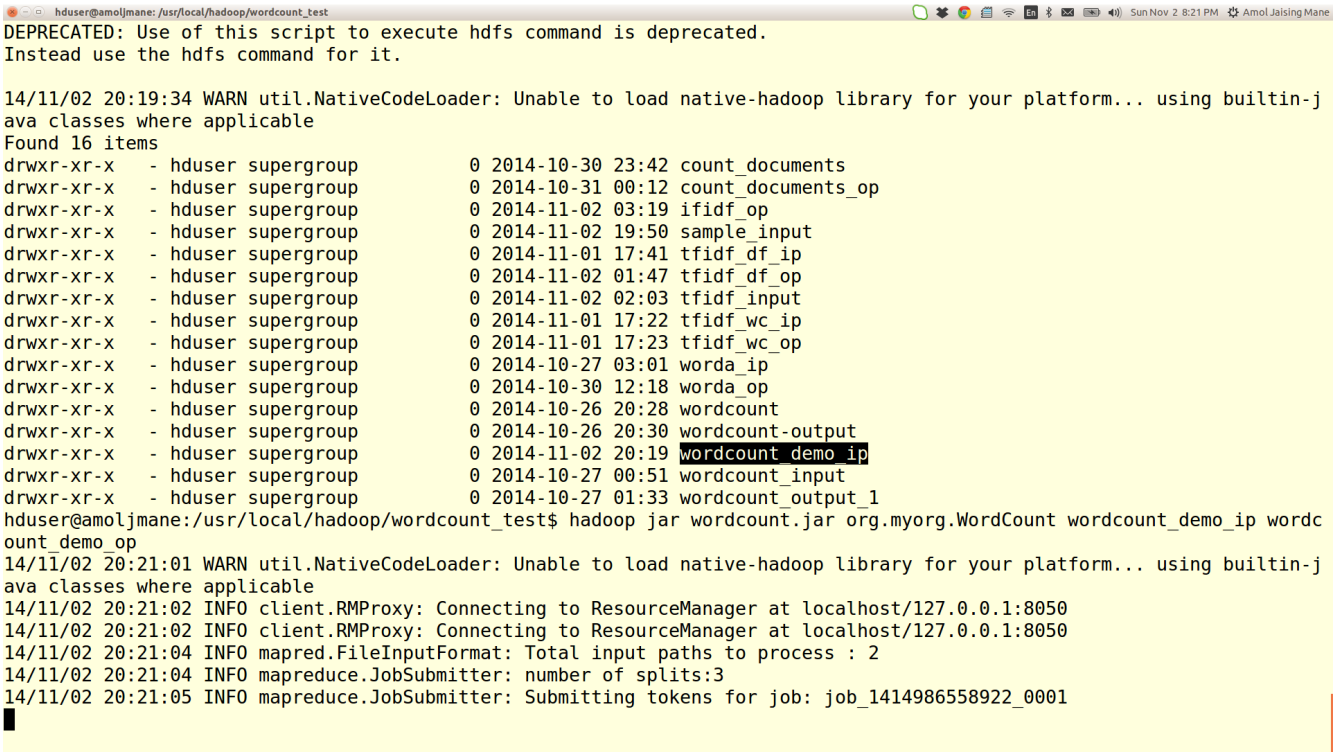
Student
Amol Mane
009270833

Professor
Simon Shim

APPLICATION

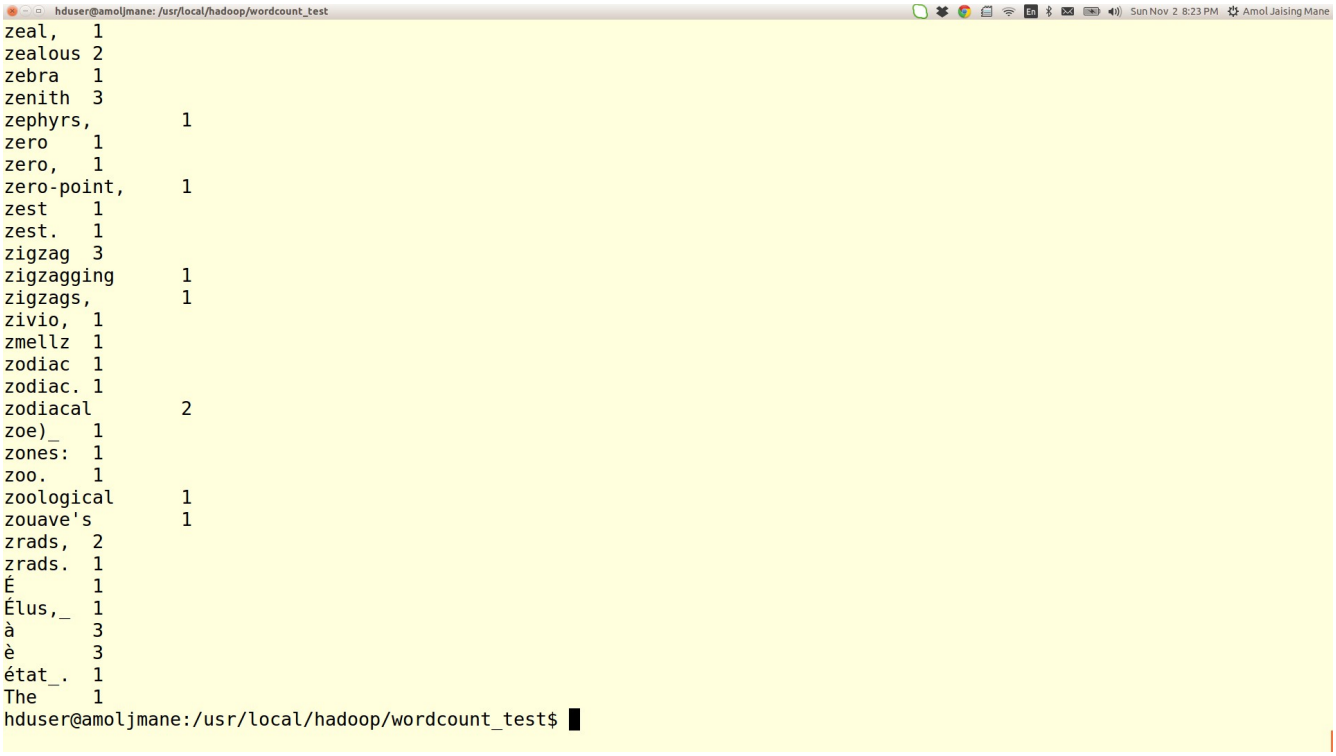
Problem 1 : Sample Program Wordcount

LOCAL

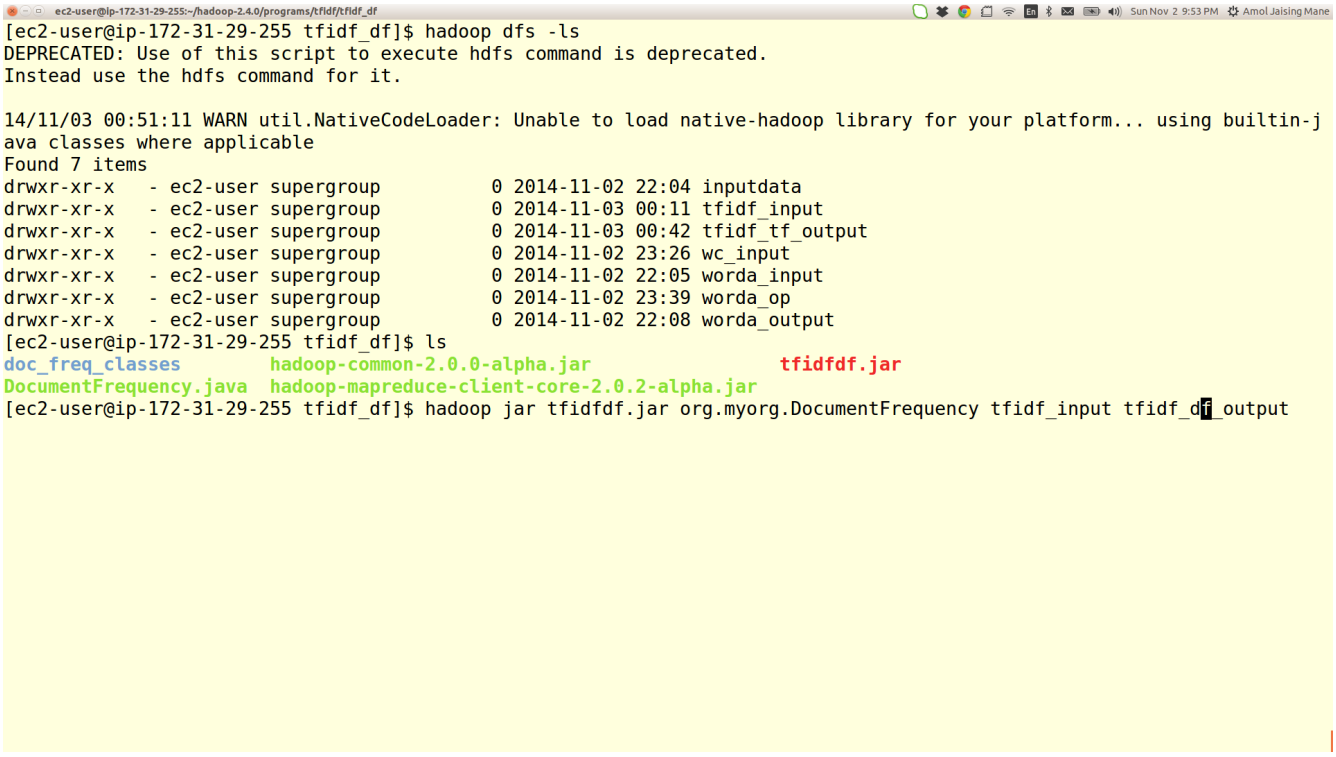


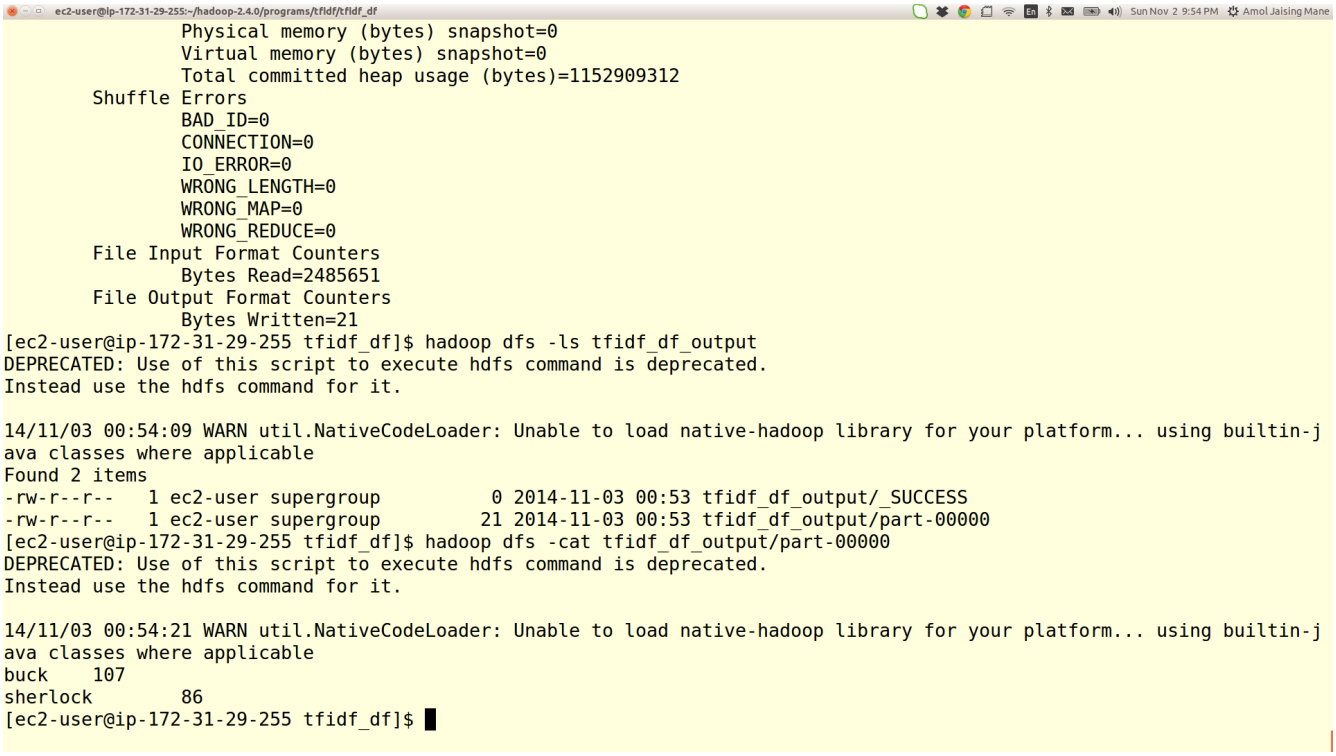
```
hduser@amoljmane: /usr/local/hadoop/wordcount_test                Sun Nov 2 8:22 PM  Amol Jaising Mane
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2167028
        File Output Format Counters
                Bytes Written=612188
hduser@amoljmane:/usr/local/hadoop/wordcount_test$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 20:22:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 17 items
drwxr-xr-x   - hduser supergroup          0 2014-10-30 23:42 count_documents
drwxr-xr-x   - hduser supergroup          0 2014-10-31 00:12 count_documents_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 03:19 ifidf_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 19:50 sample_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:41 tfidf_df_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 01:47 tfidf_df_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:23 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/wordcount_test$
```

```
hduser@amoljmane: /usr/local/hadoop/wordcount_test                Sun Nov 2 8:23 PM  Amol Jaising Mane
drwxr-xr-x   - hduser supergroup          0 2014-11-02 01:47 tfidf_df_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:23 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/wordcount_test$ hadoop dfs -la wordcount_demo_op
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

-la: Unknown command
hduser@amoljmane:/usr/local/hadoop/wordcount_test$ hadoop dfs -ll wordcount_demo_op
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

-ll: Unknown command
hduser@amoljmane:/usr/local/hadoop/wordcount_test$ hadoop dfs -ls wordcount_demo_op
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 20:23:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   3 hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op/_SUCCESS
-rw-r--r--   3 hduser supergroup     612188 2014-11-02 20:22 wordcount_demo_op/part-00000
hduser@amoljmane:/usr/local/hadoop/wordcount_test$
```

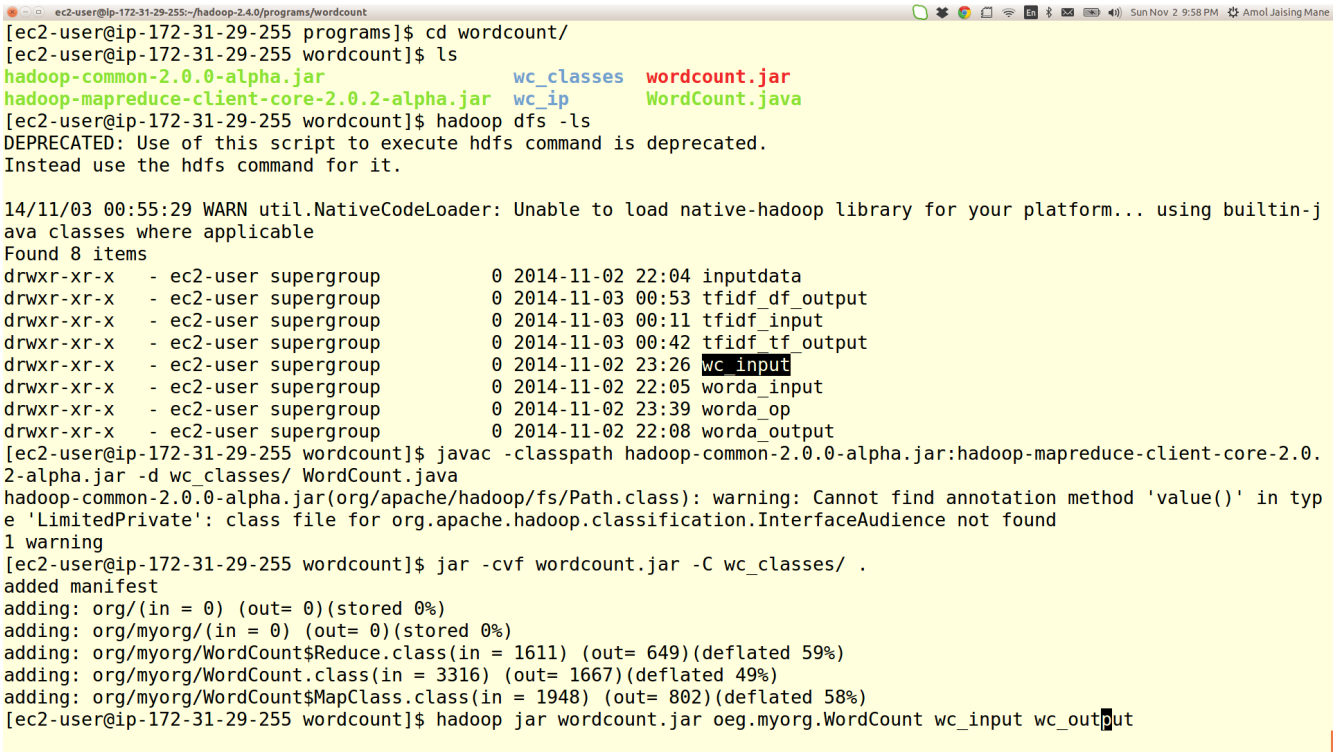

hduser@amoljmane: /usr/local/hadoop/wordcount_test
Sun Nov 2 8:23 PM
Amol Jaising Mane
zeal, 1
zealous 2
zebra 1
zenith 3
zephyrs, 1
zero 1
zero, 1
zero-point, 1
zest 1
zest. 1
zigzag 3
zigzagging 1
zigzags, 1
zivio, 1
zmellz 1
zodiac 1
zodiac. 1
zodiacal 2
zoe)_ 1
zones: 1
zoo. 1
zoological 1
zouave's 1
zrads, 2
zrads. 1
É 1
Élus,_ 1
à 3
è 3
état_. 1
The 1
hduser@amoljmane:/usr/local/hadoop/wordcount_test$

EC2


ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/tfidf/tfidf_df
Sun Nov 2 9:53 PM
Amol Jaising Mane
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/03 00:51:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 7 items
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:04 inputdata
drwxr-xr-x - ec2-user supergroup 0 2014-11-03 00:11 tfidf_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-03 00:42 tfidf_tf_output
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 23:26 wc_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:05 worda_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 23:39 worda_op
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:08 worda_output
[ec2-user@ip-172-31-29-255 tfidf_df]$ ls
doc_freq_classes hadoop-common-2.0.0-alpha.jar tfidfdf.jar
DocumentFrequency.java hadoop-mapreduce-client-core-2.0.2-alpha.jar
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop jar tfidfdf.jar org.myorg.DocumentFrequency tfidf_input tfidf_df_output




```
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=1152909312
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2485651
        File Output Format Counters
                Bytes Written=21
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop dfs -ls tfidf_df_output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 00:54:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   1 ec2-user supergroup          0 2014-11-03 00:53 tfidf_df_output/_SUCCESS
-rw-r--r--   1 ec2-user supergroup         21 2014-11-03 00:53 tfidf_df_output/part-00000
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop dfs -cat tfidf_df_output/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 00:54:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck    107
sherlock        86
[ec2-user@ip-172-31-29-255 tfidf_df]$
```



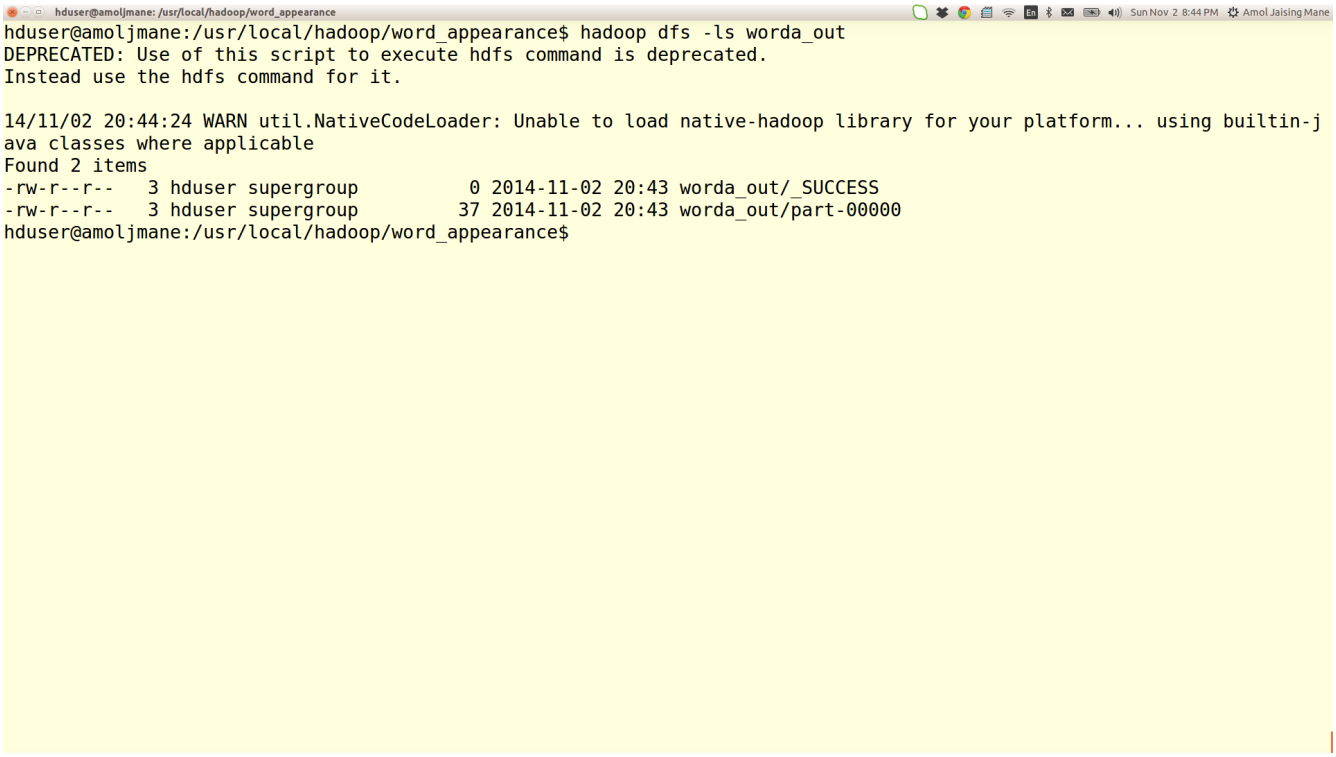

```
[ec2-user@ip-172-31-29-255 programs]$ cd wordcount/
[ec2-user@ip-172-31-29-255 wordcount]$ ls
hadoop-common-2.0.0-alpha.jar                wc_classes  wordcount.jar
hadoop-mapreduce-client-core-2.0.2-alpha.jar  wc_ip       WordCount.java
[ec2-user@ip-172-31-29-255 wordcount]$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 00:55:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 8 items
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:04 inputdata
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:53 tfidf_df_output
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:11 tfidf_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:42 tfidf_tf_output
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 23:26 wc_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:05 worda_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 23:39 worda_op
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:08 worda_output
[ec2-user@ip-172-31-29-255 wordcount]$ javac -classpath hadoop-common-2.0.0-alpha.jar:hadoop-mapreduce-client-core-2.0.
2-alpha.jar -d wc_classes/ WordCount.java
hadoop-common-2.0.0-alpha.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find annotation method 'value()' in typ
e 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not found
1 warning
[ec2-user@ip-172-31-29-255 wordcount]$ jar -cvf wordcount.jar -C wc_classes/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCount$Reduce.class(in = 1611) (out= 649)(deflated 59%)
adding: org/myorg/WordCount.class(in = 3316) (out= 1667)(deflated 49%)
adding: org/myorg/WordCount$MapClass.class(in = 1948) (out= 802)(deflated 58%)
[ec2-user@ip-172-31-29-255 wordcount]$ hadoop jar wordcount.jar oeg.myorg.WordCount wc_input wc_output
```

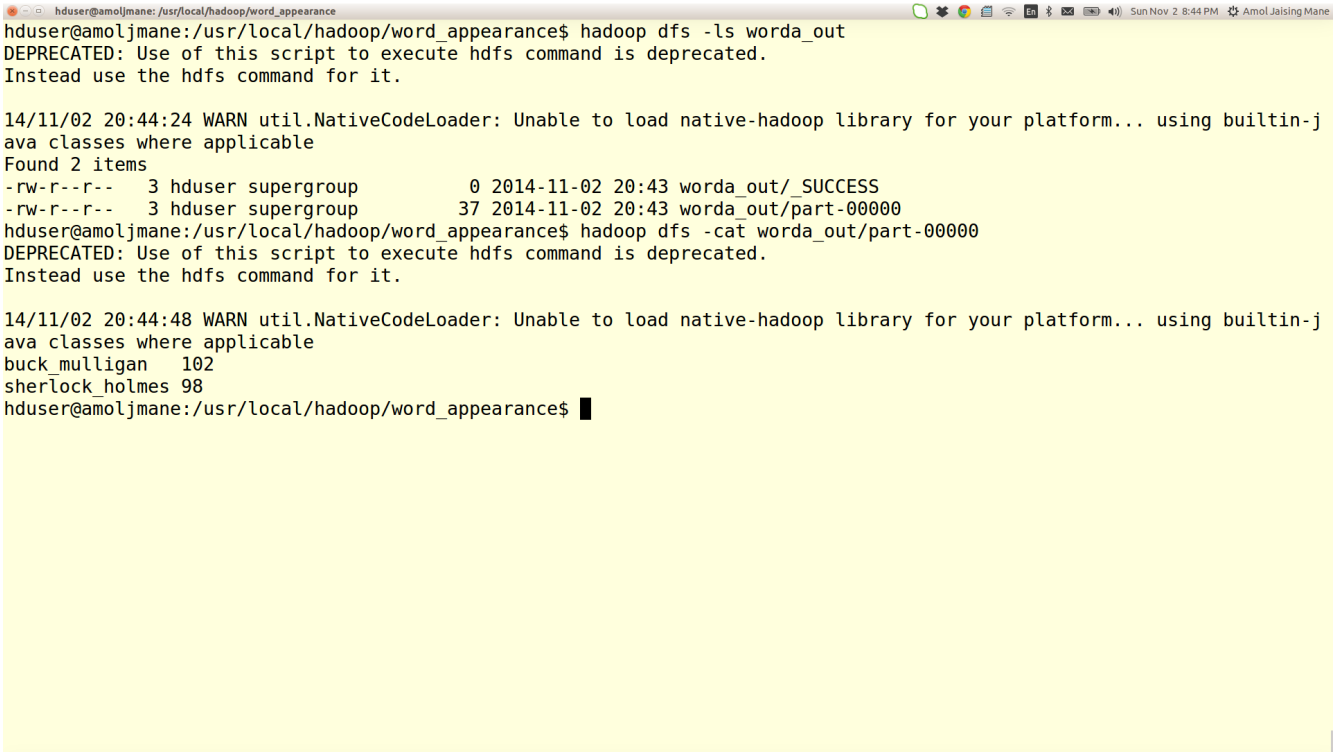
```
ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/wordcount
zeal,   1
zealous 2
zebra   1
zenith  3
zephyrs,        1
zero    1
zero,   1
zero-point,     1
zest    1
zest.   1
zigzag  3
zigzagging      1
zigzags,        1
zivio,  1
zmellz  1
zodiac  1
zodiac. 1
zodiacal        2
zoe)_   1
zones:  1
zoo.    1
zoological      1
zouave's        1
zrads,  2
zrads.  1
É       1
Élus,_  1
à       3
è       3
état_.  1
The     1
[ec2-user@ip-172-31-29-255 wordcount]$
```

Problem 2 Word Appearance

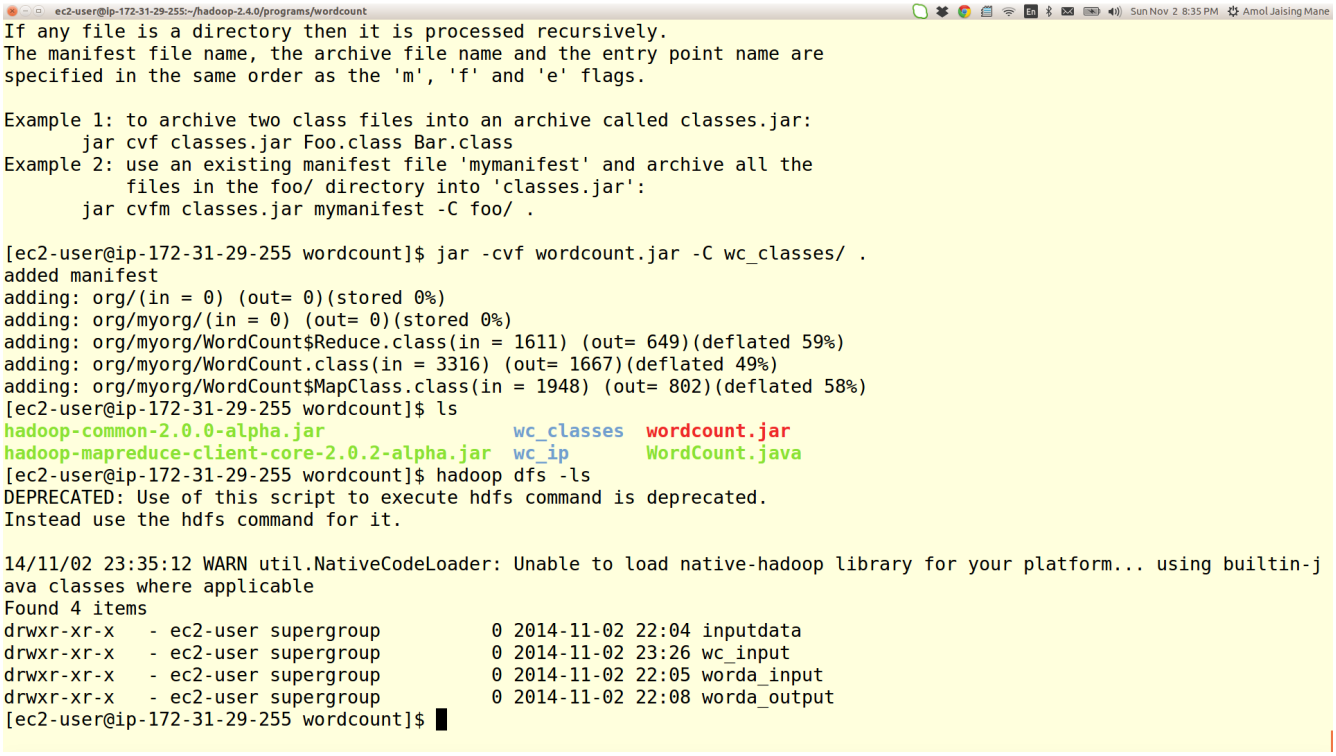
LOCAL

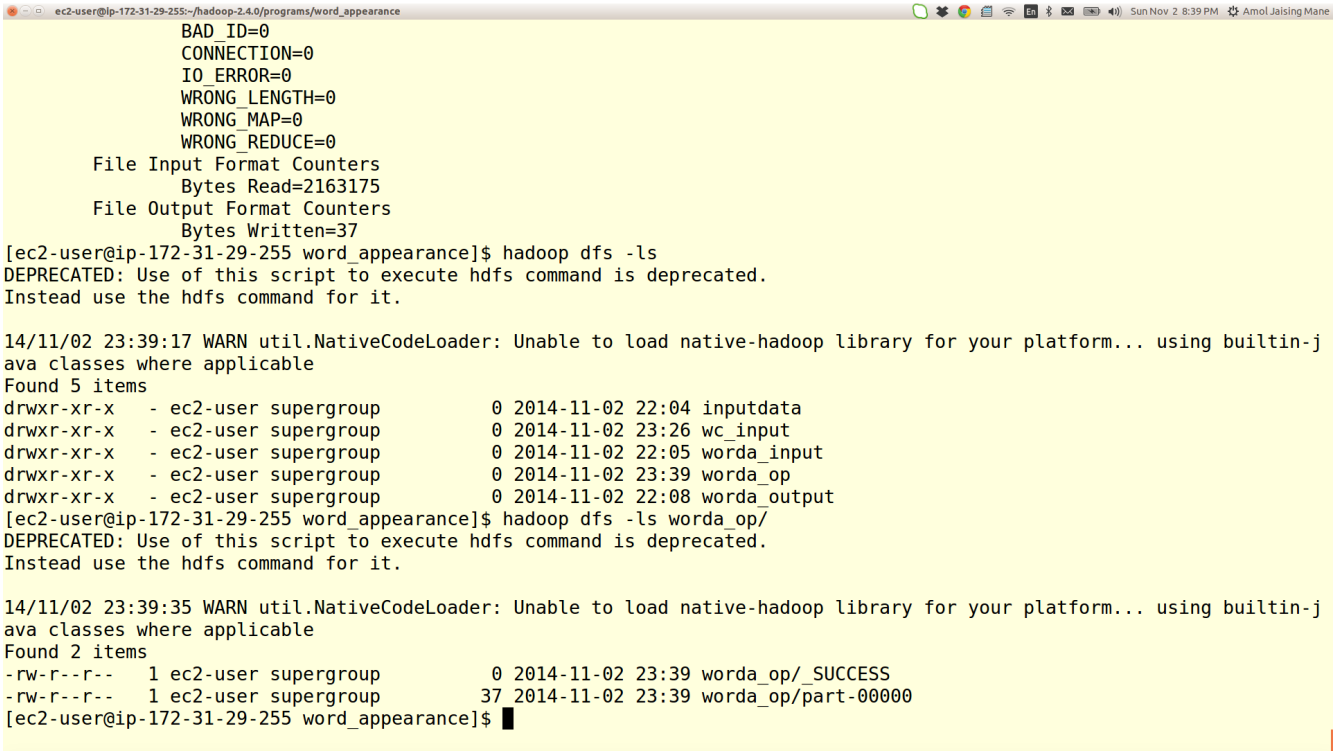
```
hduser@amoljmane: /usr/local/hadoop/word_appearance
hduser@amoljmane:/usr/local/hadoop/word_appearance$ la
hadoop-common-2.0.0-alpha.jar               wa_classes  worda_op           WordAppearance.java
hadoop-mapreduce-client-core-2.0.2-alpha.jar  wa_input    wordappearance.jar
hduser@amoljmane:/usr/local/hadoop/word_appearance$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 20:41:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 17 items
drwxr-xr-x   - hduser supergroup          0 2014-10-30 23:42 count_documents
drwxr-xr-x   - hduser supergroup          0 2014-10-31 00:12 count_documents_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 03:19 ifidf_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 19:50 sample_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:41 tfidf_df_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 01:47 tfidf_df_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:23 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/word_appearance$ ls
hadoop-common-2.0.0-alpha.jar               wa_classes  worda_op           WordAppearance.java
hadoop-mapreduce-client-core-2.0.2-alpha.jar  wa_input    wordappearance.jar
hduser@amoljmane:/usr/local/hadoop/word_appearance$ hadoop jar wordappearance.jar org.myorg.WordAppearance worda_ip wor
da_out
```

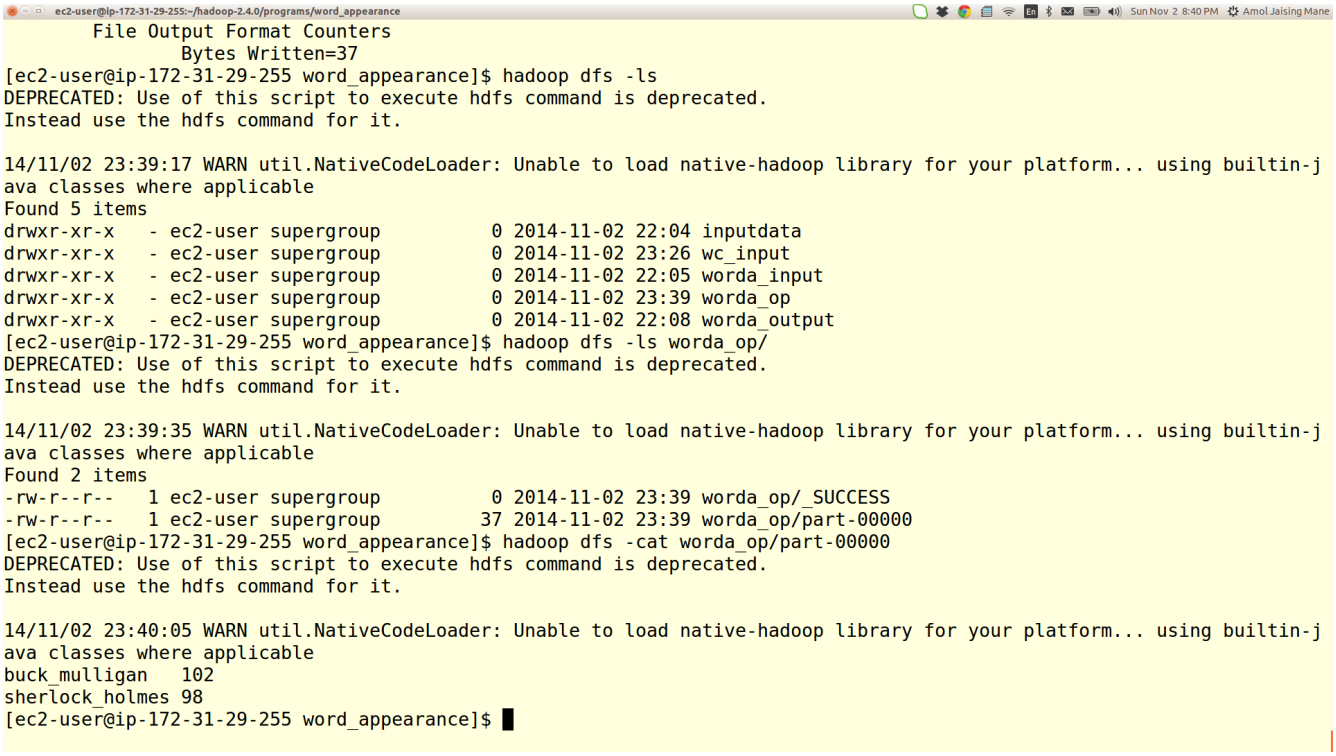

hduser@amoljmane: /usr/local/hadoop/word_appearance
Sun Nov 2 8:44 PM
Amol Jaising Mane
hduser@amoljmane:/usr/local/hadoop/word_appearance$ hadoop dfs -ls worda_out
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 20:44:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   3 hduser supergroup   0 2014-11-02 20:43 worda_out/_SUCCESS
-rw-r--r--   3 hduser supergroup   37 2014-11-02 20:43 worda_out/part-00000
hduser@amoljmane:/usr/local/hadoop/word_appearance$


hduser@amoljmane: /usr/local/hadoop/word_appearance
Sun Nov 2 8:44 PM
Amol Jaising Mane
hduser@amoljmane:/usr/local/hadoop/word_appearance$ hadoop dfs -ls worda_out
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 20:44:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r-- 3 hduser supergroup 0 2014-11-02 20:43 worda_out/_SUCCESS
-rw-r--r-- 3 hduser supergroup 37 2014-11-02 20:43 worda_out/part-00000
hduser@amoljmane:/usr/local/hadoop/word_appearance$ hadoop dfs -cat worda_out/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 20:44:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck_mulligan 102
sherlock_holmes 98
hduser@amoljmane:/usr/local/hadoop/word_appearance$

EC2


ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/wordcount
Sun Nov 2 8:35 PM
Amol Jaising Mane
If any file is a directory then it is processed recursively.
The manifest file name, the archive file name and the entry point name are
specified in the same order as the 'm', 'f' and 'e' flags.
Example 1: to archive two class files into an archive called classes.jar:
jar cvf classes.jar Foo.class Bar.class
Example 2: use an existing manifest file 'mymanifest' and archive all the
files in the foo/ directory into 'classes.jar':
jar cvfm classes.jar mymanifest -C foo/ .
[ec2-user@ip-172-31-29-255 wordcount]$ jar -cvf wordcount.jar -C wc_classes/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCount$Reduce.class(in = 1611) (out= 649)(deflated 59%)
adding: org/myorg/WordCount.class(in = 3316) (out= 1667)(deflated 49%)
adding: org/myorg/WordCount$MapClass.class(in = 1948) (out= 802)(deflated 58%)
[ec2-user@ip-172-31-29-255 wordcount]$ ls
hadoop-common-2.0.0-alpha.jar
wc_classes
wordcount.jar
hadoop-mapreduce-client-core-2.0.2-alpha.jar
wc_ip
WordCount.java
[ec2-user@ip-172-31-29-255 wordcount]$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 23:35:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 4 items
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:04 inputdata
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 23:26 wc_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:05 worda_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:08 worda_output
[ec2-user@ip-172-31-29-255 wordcount]$


ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/word_appearance
Sun Nov 2 8:39 PM
Amol Jaising Mane
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=2163175
File Output Format Counters
Bytes Written=37
[ec2-user@ip-172-31-29-255 word_appearance]$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 23:39:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 5 items
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:04 inputdata
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 23:26 wc_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:05 worda_input
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 23:39 worda_op
drwxr-xr-x - ec2-user supergroup 0 2014-11-02 22:08 worda_output
[ec2-user@ip-172-31-29-255 word_appearance]$ hadoop dfs -ls worda_op/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/02 23:39:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r-- 1 ec2-user supergroup 0 2014-11-02 23:39 worda_op/_SUCCESS
-rw-r--r-- 1 ec2-user supergroup 37 2014-11-02 23:39 worda_op/part-00000
[ec2-user@ip-172-31-29-255 word_appearance]$



sherlock and Holmes are together for 98
buck and mulligan are together for 102

Problem 3: tfidf

A) Term Frequency
Have removed all the preposition and stop words.

Local

```
hduser@amoljmane: /usr/local/hadoop/tfidf/wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-31 00:12 count_documents_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 03:19 ifidf_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 19:50 sample_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:41 tfidf_df_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 01:47 tfidf_df_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 21:46 tfidf_df_op2
drwxr-xr-x   - hduser supergroup          0 2014-11-02 22:39 tfidf_df_op3
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 21:30 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:43 worda_out
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/tfidf/wordcount$ hadoop jar tfidfwc.jar org.myorg.WordCount tfidf_df_ip tfidf_tf_op
14/11/02 22:45:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
14/11/02 22:45:38 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8050
14/11/02 22:45:39 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8050
14/11/02 22:45:39 INFO mapred.FileInputFormat: Total input paths to process : 2
14/11/02 22:45:39 INFO mapreduce.JobSubmitter: number of splits:3
14/11/02 22:45:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1414986558922_0006
14/11/02 22:45:40 INFO impl.YarnClientImpl: Submitted application application_1414986558922_0006
14/11/02 22:45:40 INFO mapreduce.Job: The url to track the job: http://amoljmane:8088/proxy/application_1414986558922_0
006/
14/11/02 22:45:40 INFO mapreduce.Job: Running job: job_1414986558922_0006
```

```
hduser@amoljmane: /usr/local/hadoop/tfidf/wordcount
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-02 22:46 tfidf_tf_op
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 21:30 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:43 worda_out
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/tfidf/wordcount$ hadoop dfs -ls tfidf_tf_op
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 22:46:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   3 hduser supergroup          0 2014-11-02 22:46 tfidf_tf_op/_SUCCESS
-rw-r--r--   3 hduser supergroup         35 2014-11-02 22:46 tfidf_tf_op/part-00000
hduser@amoljmane:/usr/local/hadoop/tfidf/wordcount$ hadoop dfs -cat tfidf_tf_op/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 22:47:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck    169
mulligan        165
total   199801
hduser@amoljmane:/usr/local/hadoop/tfidf/wordcount$
```

EC2

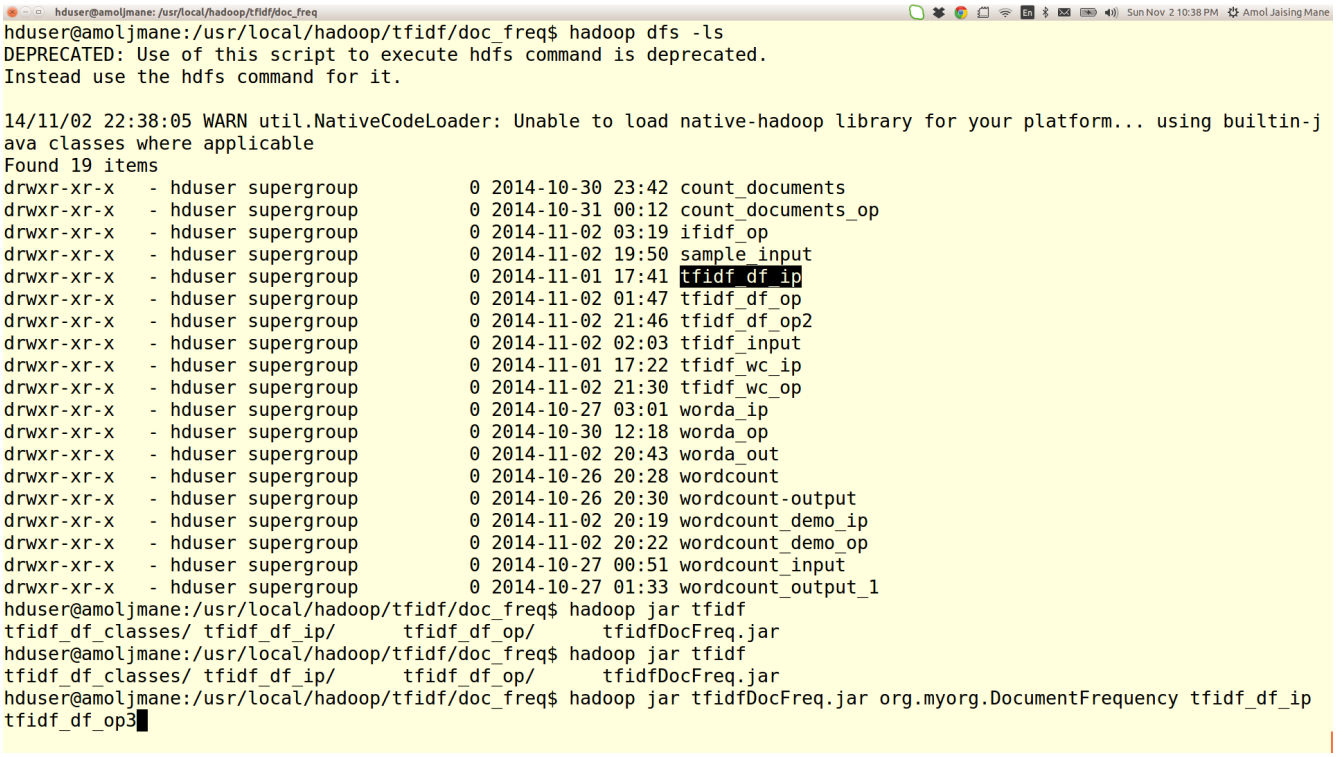
```
ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/tfidf/tfidf_tf
-alpha.jar -d tfidf_tf_classes/ WordCount.java
hadoop-common-2.0.0-alpha.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find annotation method 'value()' in typ
e 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not found
1 warning
[ec2-user@ip-172-31-29-255 tfidf_tf]$ jar -cvf tfidftf.jar -C tfidf_tf_classes/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCount$Reduce.class(in = 1611) (out= 651)(deflated 59%)
adding: org/myorg/WordCount.class(in = 3316) (out= 1681)(deflated 49%)
adding: org/myorg/TermFrequency$ReduceClass.class(in = 1687) (out= 682)(deflated 59%)
adding: org/myorg/TermFrequency.class(in = 3310) (out= 1655)(deflated 50%)
adding: org/myorg/TermFrequency$MapClass.class(in = 2525) (out= 1095)(deflated 56%)
adding: org/myorg/WordCount$MapClass.class(in = 5553) (out= 2947)(deflated 46%)
[ec2-user@ip-172-31-29-255 tfidf_tf]$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 01:09:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 9 items
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:04 inputdata
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:53 tfidf_df_output
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:11 tfidf_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:42 tfidf_tf_output
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 23:26 wc_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-03 00:59 wc_output
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:05 worda_input
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 23:39 worda_op
drwxr-xr-x   - ec2-user supergroup          0 2014-11-02 22:08 worda_output
[ec2-user@ip-172-31-29-255 tfidf_tf]$ hadoop jar tfidftf.jar org.myorg.WordCount ^C
[ec2-user@ip-172-31-29-255 tfidf_tf]$ hadoop jar tfidftf.jar org.myorg.WordCount tfidf_input tfidf_tf_output_1
```

```
ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/tfidf/tfidf_tf
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=1204813824
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2485651
        File Output Format Counters
                Bytes Written=35
[ec2-user@ip-172-31-29-255 tfidf_tf]$ hadoop dfs -ls tfidf_tf_output_1
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 01:10:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   1 ec2-user supergroup          0 2014-11-03 01:10 tfidf_tf_output_1/_SUCCESS
-rw-r--r--   1 ec2-user supergroup         35 2014-11-03 01:10 tfidf_tf_output_1/part-00000
[ec2-user@ip-172-31-29-255 tfidf_tf]$ hadoop dfs -cat tfidf_tf_output_1/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/03 01:10:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck    169
mulligan        165
total   199801
[ec2-user@ip-172-31-29-255 tfidf_tf]$
```

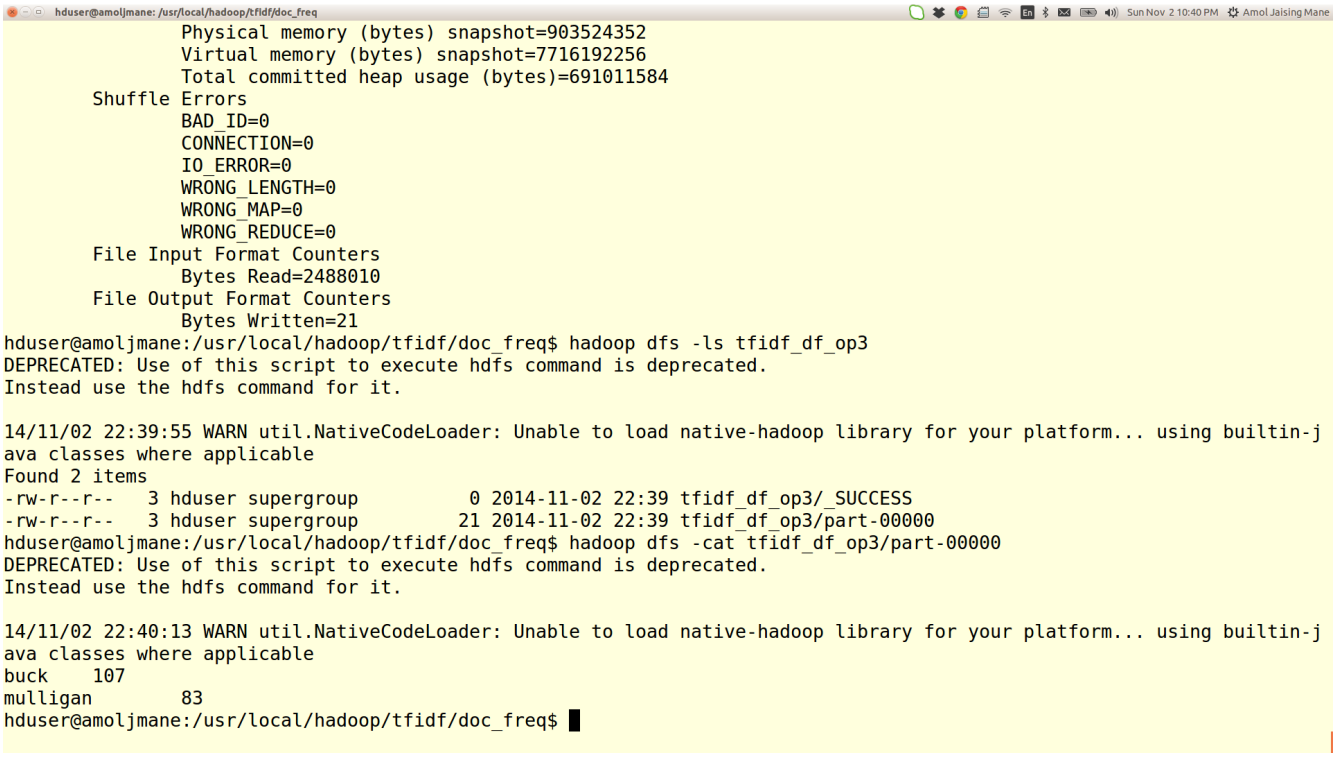
Term Frequency
Buck 169
Mulligan 165
Total words 200135

B) Document Frequency

Local

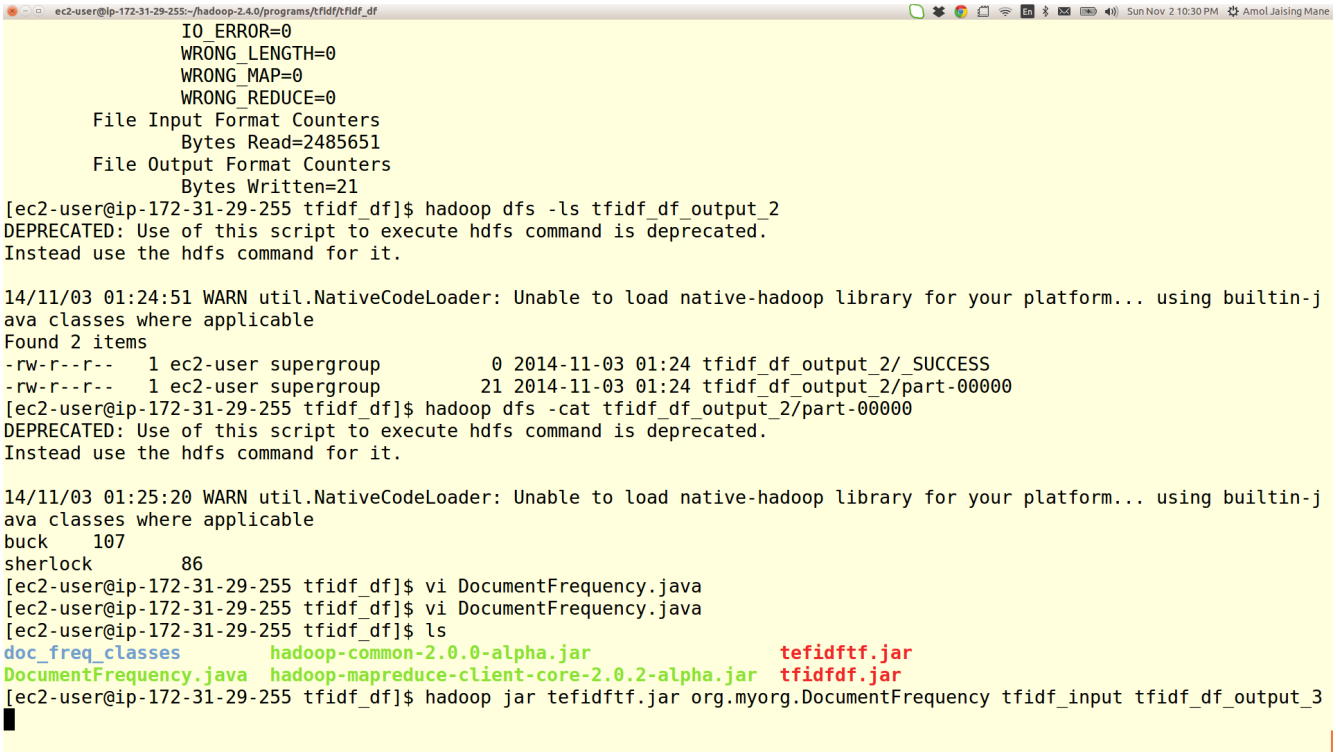


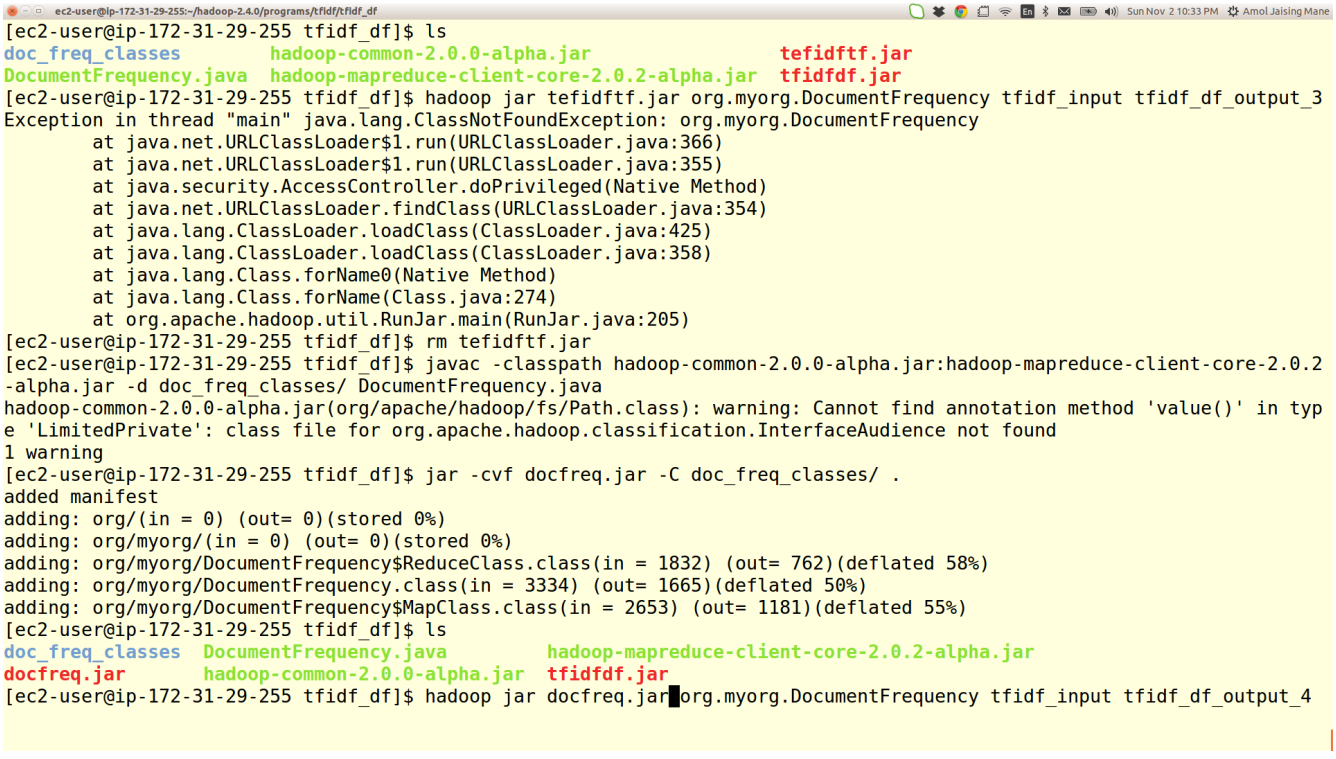

```
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 22:38:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 19 items
drwxr-xr-x   - hduser supergroup          0 2014-10-30 23:42 count_documents
drwxr-xr-x   - hduser supergroup          0 2014-10-31 00:12 count_documents_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 03:19 ifidf_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 19:50 sample_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:41 tfidf_df_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 01:47 tfidf_df_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 21:46 tfidf_df_op2
drwxr-xr-x   - hduser supergroup          0 2014-11-02 02:03 tfidf_input
drwxr-xr-x   - hduser supergroup          0 2014-11-01 17:22 tfidf_wc_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 21:30 tfidf_wc_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 03:01 worda_ip
drwxr-xr-x   - hduser supergroup          0 2014-10-30 12:18 worda_op
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:43 worda_out
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:28 wordcount
drwxr-xr-x   - hduser supergroup          0 2014-10-26 20:30 wordcount-output
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:19 wordcount_demo_ip
drwxr-xr-x   - hduser supergroup          0 2014-11-02 20:22 wordcount_demo_op
drwxr-xr-x   - hduser supergroup          0 2014-10-27 00:51 wordcount_input
drwxr-xr-x   - hduser supergroup          0 2014-10-27 01:33 wordcount_output_1
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop jar tfidf
tfidf_df_classes/ tfidf_df_ip/      tfidf_df_op/      tfidfDocFreq.jar
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop jar tfidf
tfidf_df_classes/ tfidf_df_ip/      tfidf_df_op/      tfidfDocFreq.jar
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop jar tfidfDocFreq.jar org.myorg.DocumentFrequency tfidf_df_ip
tfidf_df_op3
```

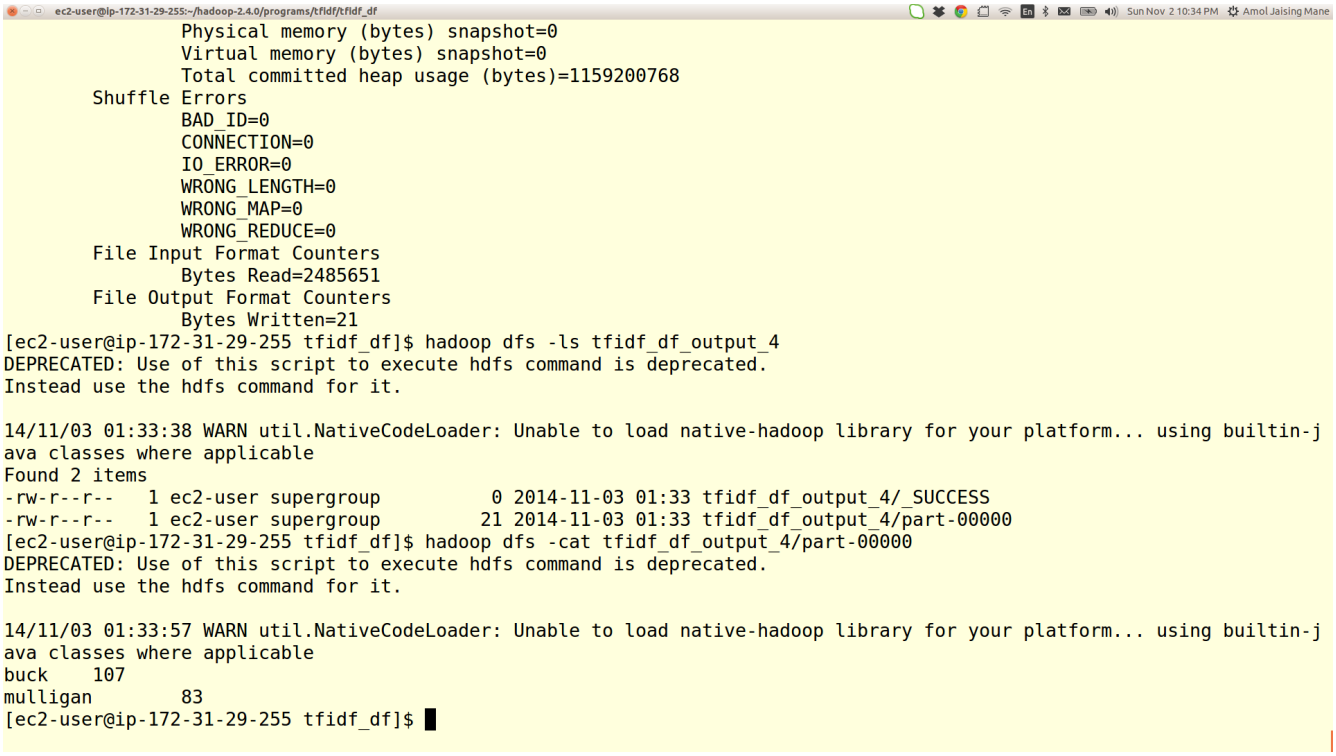



```
                Physical memory (bytes) snapshot=903524352
                Virtual memory (bytes) snapshot=7716192256
                Total committed heap usage (bytes)=691011584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2488010
        File Output Format Counters
                Bytes Written=21
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop dfs -ls tfidf_df_op3
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 22:39:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r--   3 hduser supergroup          0 2014-11-02 22:39 tfidf_df_op3/_SUCCESS
-rw-r--r--   3 hduser supergroup         21 2014-11-02 22:39 tfidf_df_op3/part-00000
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$ hadoop dfs -cat tfidf_df_op3/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

14/11/02 22:40:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck    107
mulligan        83
hduser@amoljmane:/usr/local/hadoop/tfidf/doc_freq$
```

EC2


ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/tfidf/tfidf_df
Sun Nov 2 10:30 PM
Amol Jaising Mane
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=2485651
File Output Format Counters
Bytes Written=21
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop dfs -ls tfidf_df_output_2
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/03 01:24:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
Found 2 items
-rw-r--r-- 1 ec2-user supergroup 0 2014-11-03 01:24 tfidf_df_output_2/_SUCCESS
-rw-r--r-- 1 ec2-user supergroup 21 2014-11-03 01:24 tfidf_df_output_2/part-00000
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop dfs -cat tfidf_df_output_2/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
14/11/03 01:25:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
ava classes where applicable
buck 107
sherlock 86
[ec2-user@ip-172-31-29-255 tfidf_df]$ vi DocumentFrequency.java
[ec2-user@ip-172-31-29-255 tfidf_df]$ vi DocumentFrequency.java
[ec2-user@ip-172-31-29-255 tfidf_df]$ ls
doc_freq_classes
hadoop-common-2.0.0-alpha.jar
tefidftf.jar
DocumentFrequency.java
hadoop-mapreduce-client-core-2.0.2-alpha.jar
tfidfdf.jar
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop jar tefidftf.jar org.myorg.DocumentFrequency tfidf_input tfidf_df_output_3


ec2-user@ip-172-31-29-255:~/hadoop-2.4.0/programs/tfidf/tfidf_df
Sun Nov 2 10:33 PM
Amol Jaising Mane
[ec2-user@ip-172-31-29-255 tfidf_df]$ ls
doc_freq_classes
hadoop-common-2.0.0-alpha.jar
tefidftf.jar
DocumentFrequency.java
hadoop-mapreduce-client-core-2.0.2-alpha.jar
tfidfdf.jar
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop jar tefidftf.jar org.myorg.DocumentFrequency tfidf_input tfidf_df_output_3
Exception in thread "main" java.lang.ClassNotFoundException: org.myorg.DocumentFrequency
at java.net.URLClassLoader$1.run(URLClassLoader.java:366)
at java.net.URLClassLoader$1.run(URLClassLoader.java:355)
at java.security.AccessController.doPrivileged(Native Method)
at java.net.URLClassLoader.findClass(URLClassLoader.java:354)
at java.lang.ClassLoader.loadClass(ClassLoader.java:425)
at java.lang.ClassLoader.loadClass(ClassLoader.java:358)
at java.lang.Class.forName0(Native Method)
at java.lang.Class.forName(Class.java:274)
at org.apache.hadoop.util.RunJar.main(RunJar.java:205)
[ec2-user@ip-172-31-29-255 tfidf_df]$ rm tefidftf.jar
[ec2-user@ip-172-31-29-255 tfidf_df]$ javac -classpath hadoop-common-2.0.0-alpha.jar:hadoop-mapreduce-client-core-2.0.2
-alpha.jar -d doc_freq_classes/ DocumentFrequency.java
hadoop-common-2.0.0-alpha.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find annotation method 'value()' in typ
e 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not found
1 warning
[ec2-user@ip-172-31-29-255 tfidf_df]$ jar -cvf docfreq.jar -C doc_freq_classes/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/DocumentFrequency$ReduceClass.class(in = 1832) (out= 762)(deflated 58%)
adding: org/myorg/DocumentFrequency.class(in = 3334) (out= 1665)(deflated 50%)
adding: org/myorg/DocumentFrequency$MapClass.class(in = 2653) (out= 1181)(deflated 55%)
[ec2-user@ip-172-31-29-255 tfidf_df]$ ls
doc_freq_classes
DocumentFrequency.java
hadoop-mapreduce-client-core-2.0.2-alpha.jar
docfreq.jar
hadoop-common-2.0.0-alpha.jar
tfidfdf.jar
[ec2-user@ip-172-31-29-255 tfidf_df]$ hadoop jar docfreq.jar org.myorg.DocumentFrequency tfidf_input tfidf_df_output_4



Document Frequency =
Buck = 107
mulligan = 83

TDIDF ==

TF::
buck = 169
malligan = 165
Total Words = 200135
hence

TF for buck = 169/200135 ==> 0.00084443
TF for mulligan = 165/200135 ==> 000824444

IDF ::
Buck = 107
mulligan = 83
Total Docs = 1317

Hence IDF =
IDF for buck = log(1317/107) ==> 1.090201997
IDF for mulligan = log(1317/83) ==> 1.200507683

TDIDF ==
for buck ==>> TF for buck * IDF for buck ==> 0.00084443 * 1.090201997 = 0.000920599

for mulligan ==>> TF for mulligan * IDF for mulligan ==> 0.00084443 * 1.200507683 = 0.001013745