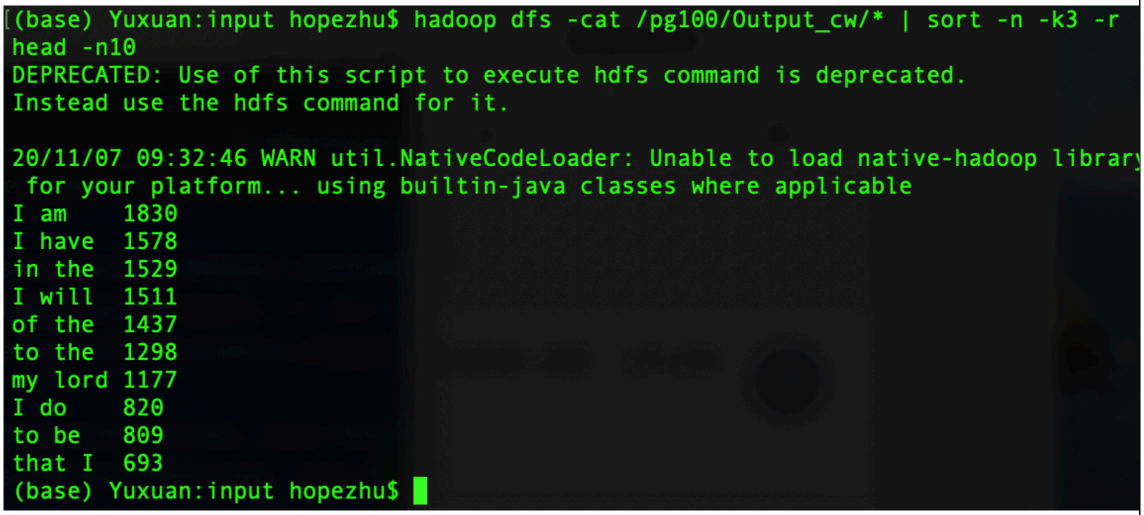
1. My results files:

File 1: top 10 bigram and occurrences （remove punctuations and simple count anything like Let’s, it’s, hasn’t…. as two words）

Bigram	Occurrences
I am	1830
I have	1578
in the	1529
I will	1511
of the	1437
to the	1298
my lord	1177
I do	820

to be	809
that I	693



```
[(base) Yuxuan:input hopezhu$ hadoop dfs -cat /pg100/Output_cw/* | sort -n -k3 -r
 head -n10
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

20/11/07 09:32:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library
 for your platform... using builtin-java classes where applicable
I am    1830
I have  1578
in the  1529
I will  1511
of the  1437
to the  1298
my lord 1177
I do    820
to be   809
that I  693
(base) Yuxuan:input hopezhu$
```

2. My code with comments:

```
import java.io.IOException;
import java.util.*;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Bigram count");
        job.setJarByClass(WordCount.class); //My class name is 'word
count'
        job.setMapperClass(TokenizerMapper.class); // Set my Mapper
class

        job.setReducerClass(TokenizerMapper.IntSumReducer.class); //
Set my Reducer class
        job.setOutputKeyClass(Text.class); // Set the key class for my
output class
        job.setOutputValueClass(IntWritable.class); // Set the value
class for the output data

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }

    public static class TokenizerMapper
            extends Mapper<Object, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        // The method is to extract useful word
        public String deleteNotation(String words) {
            String regEx = "[\\W[0-9]]";
            Pattern p = Pattern.compile(regEx);
            Matcher matcher = p.matcher(words);
            return matcher.replaceAll("").trim();
        }
        public void map(Object key, Text value, Context context
        ) throws IOException, InterruptedException {
            ArrayList<String> bigrams = new ArrayList<String>();
            String[] single_word = value.toString().split("\\s+");

            // create and fill the bigram arraylist
            // the bigram has number of (single_word -1) since it has
two words
            for (int i = 0; i < single_word.length - 1; i++) {
                single_word[i] = deleteNotation(single_word[i]);
                single_word[i + 1] = deleteNotation(single_word[i +
1]);
                if (!(single_word[i].isEmpty()) && !(single_word[i +
1].isEmpty())) {
                    bigrams.add(single_word[i] + " " + single_word[i +
1]);
                }

            }
            for (String token : bigrams) {
//                System.out.print(token);
                word.set(token);
                context.write(word, one);
            }

        }

        public static class IntSumReducer
                extends Reducer<Text, IntWritable, Text, IntWritable>
{
            private IntWritable result = new IntWritable();

            public void reduce(Text key, Iterable<IntWritable> values,
                               Context context
            ) throws IOException, InterruptedException {

                int sum = 0;
                for (IntWritable val : values) {
                    sum += val.get();
                }
                result.set(sum);
```

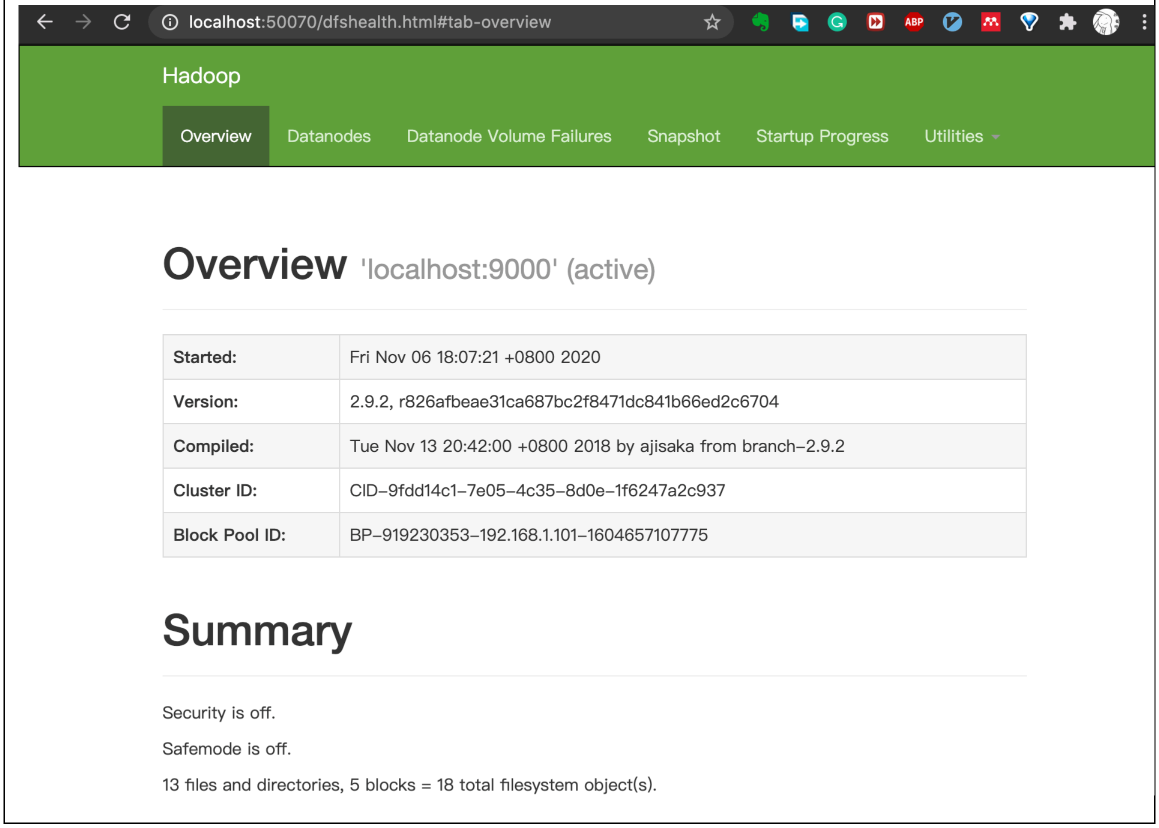
```
                    context.write(key, result);
                }
            }


        }
    }
```

2. Hadoop command list and order you have used to run your job and get your results

I successfully install and configure the Hadoop environment on my own MAC

http://localhost:50070/dfshealth.html#tab-overview



Overview 'localhost:9000' (active)

Started:	Fri Nov 06 18:07:21 +0800 2020
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-9fdd14c1-7e05-4c35-8d0e-1f6247a2c937
Block Pool ID:	BP-919230353-192.168.1.101-1604657107775

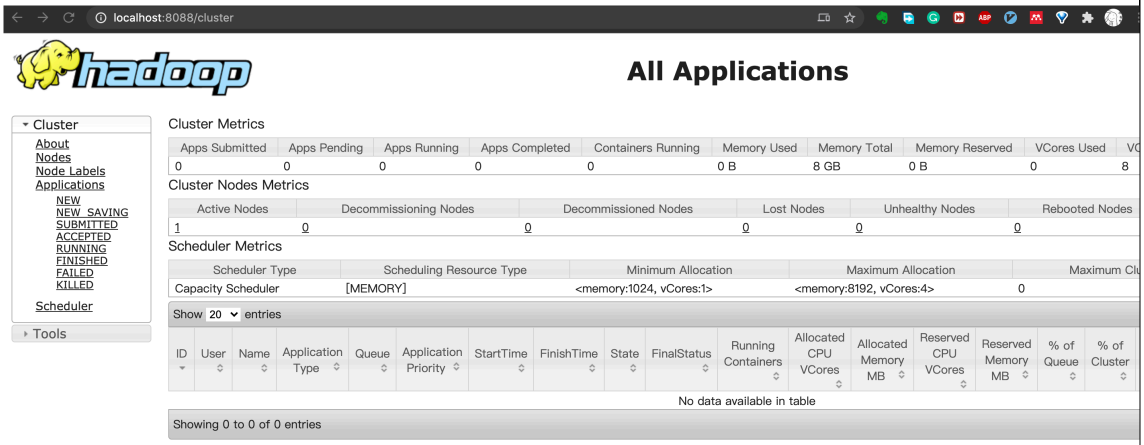
Summary

Security is off.

Safemode is off.

13 files and directories, 5 blocks = 18 total filesystem object(s).

http://localhost:8088/cluster



All Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VC
0	0	0	0	0	0 B	8 GB	0 B	0	8

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Clu
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

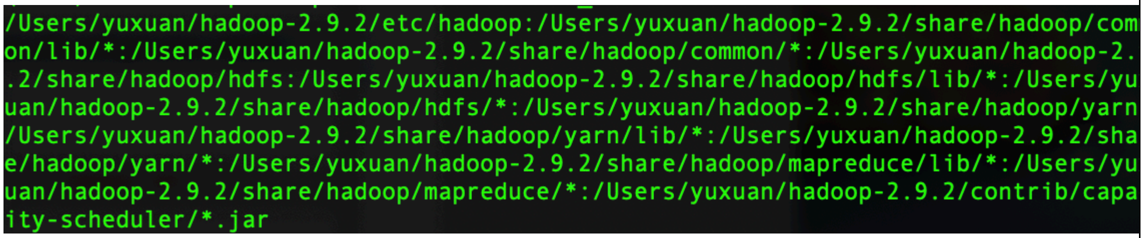
Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster
No data available in table																

Showing 0 to 0 of 0 entries

First, you need to configure the Hadoop path and print to check it

```
export HADOOP_CLASSPATH=$(hadoop classpath)
echo $HADOOP_CLASSPATH
```

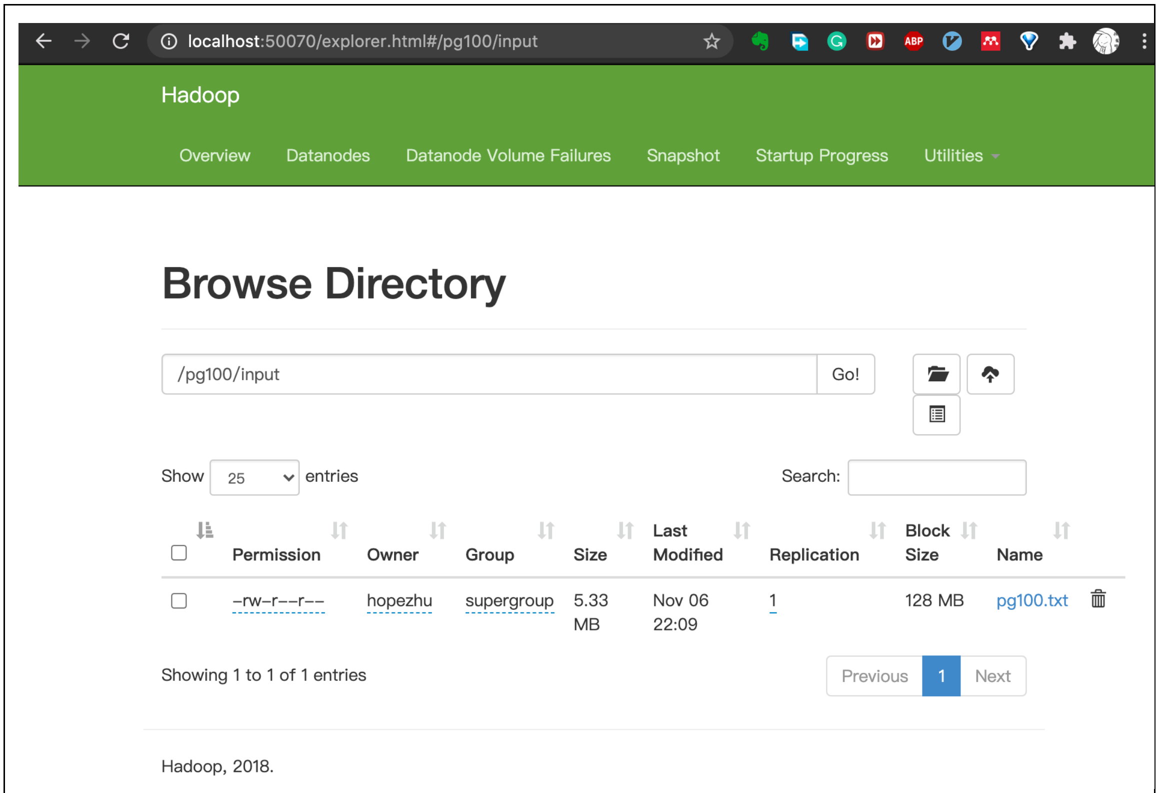


```
/Users/yuxuan/hadoop-2.9.2/etc/hadoop:/Users/yuxuan/hadoop-2.9.2/share/hadoop/com
on/lib/*:/Users/yuxuan/hadoop-2.9.2/share/hadoop/common/*:/Users/yuxuan/hadoop-2.
.2/share/hadoop/hdfs:/Users/yuxuan/hadoop-2.9.2/share/hadoop/hdfs/lib/*:/Users/yu
uan/hadoop-2.9.2/share/hadoop/hdfs/*:/Users/yuxuan/hadoop-2.9.2/share/hadoop/yarn
/Users/yuxuan/hadoop-2.9.2/share/hadoop/yarn/lib/*:/Users/yuxuan/hadoop-2.9.2/sha
e/hadoop/yarn/*:/Users/yuxuan/hadoop-2.9.2/share/hadoop/mapreduce/lib/*:/Users/yu
uan/hadoop-2.9.2/share/hadoop/mapreduce/*:/Users/yuxuan/hadoop-2.9.2/contrib/capa
ity-scheduler/*.jar
```

Make a direction on HDFS and a file inside the directory

```
hadoop fs -mkdir /pg100
hadoop fs -mkdir /pg100/input
hadoop fs -put '/Users/yuxuan/Desktop/pg100.txt' /pg100/input
```

check the file on http://localhost:50070/dfshealth.html#tab-overview



Browse Directory

/pg100/input

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	hopezhu	supergroup	5.33 MB	Nov 06 22:09	1	128 MB	pg100.txt	🗑

Showing 1 to 1 of 1 entries

Hadoop, 2018.

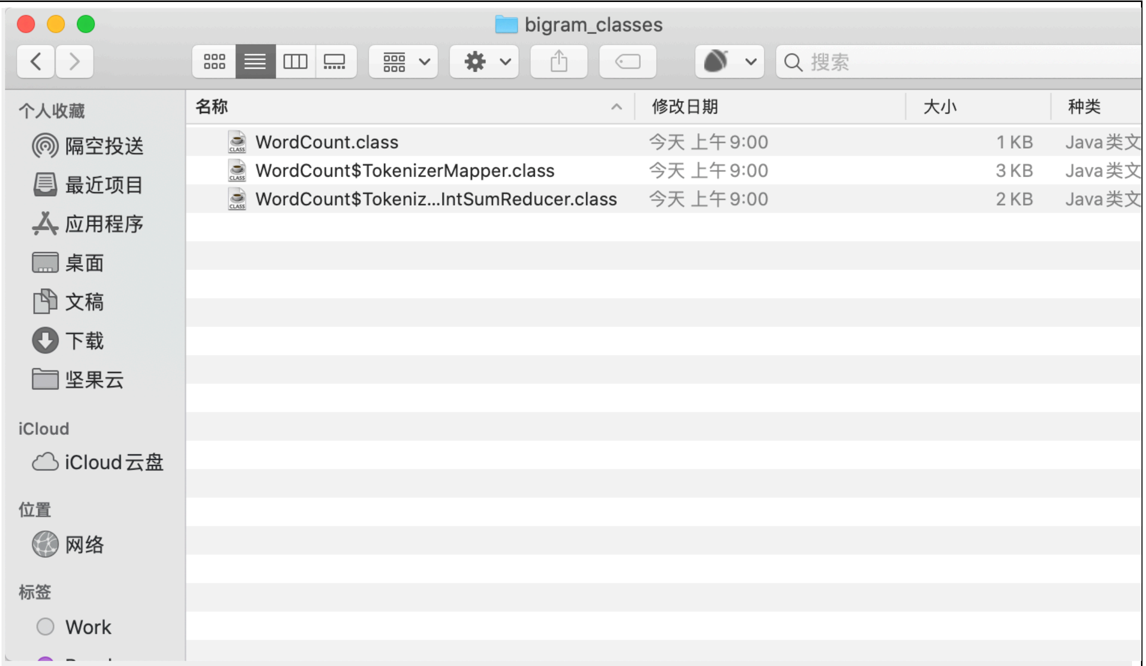
Now, I successfully transferred the pg100.txt file on HDFS

Enter my own directory

Create a directory called bigram_classes to store the classes of WordCount.java, and use javac to compile the java file

```
javac -classpath ${HADOOP_CLASSPATH} -d '/Users/yuxuan/hadoop-
2.9.2/input/bigram_classes'
'/Users/yuxuan/coding/IDEA/Hadoop_cw2/src/WordCount.java'
```

It will generate three classes in the bigram_classes file



名称	修改日期	大小	种类
WordCount.class	今天 上午9:00	1 KB	Java类文
WordCount$TokenizerMapper.class	今天 上午9:00	3 KB	Java类文
WordCount$Tokeniz...IntSumReducer.class	今天 上午9:00	2 KB	Java类文

Combine these classes into one jar file called bigram.jar

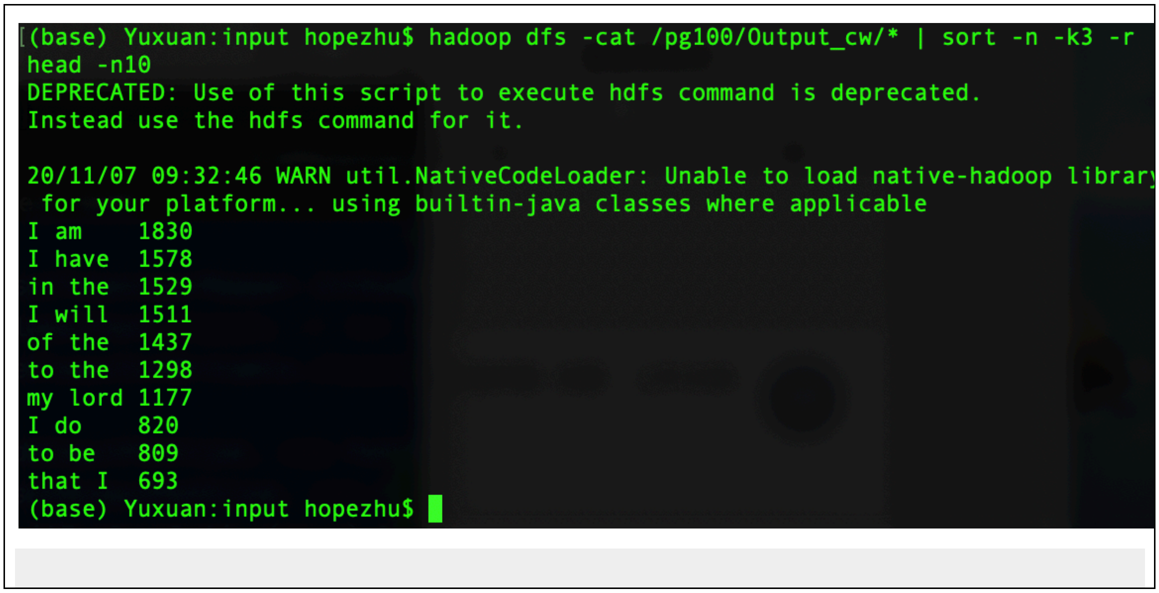
```
jar -cvf bigram.jar -C bigram_classes/ .
```

Run the jar file on Hadoop and output to Output_cw on hadoop clusters

```
hadoop jar '/Users/yuxuan/hadoop-2.9.2/input/bigram.jar' WordCount
/pg100/input_cw /pg100/Output_cw
```

Return the Top 10 frequency bigram words

```
hadoop dfs -cat /pg100/Output_cw/* | sort -n -k3 -r |head -n10
```



```
[(base) Yuxuan:input hopezhu$ hadoop dfs -cat /pg100/Output_cw/* | sort -n -k3 -r
head -n10
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

20/11/07 09:32:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library
 for your platform... using builtin-java classes where applicable
I am    1830
I have  1578
in the  1529
I will  1511
of the  1437
to the  1298
my lord 1177
I do    820
to be   809
that I  693
(base) Yuxuan:input hopezhu$ 
```

1. My results files:

File 2: the single line contains “torture”

All length is torture. Since the torch is out,

And torture him with grievous ling'ring death.

But purgatory, torture, hell itself.

By a sharp torture.

Drawn on with torture.

For torturers ingenious. It is I

From thee to die were torture more than death.

He may at pleasure whip or hang or torture,

I play the torturer, by small and small

In leads or oils? What old or newer torture

Let hell want pains enough to torture me!

O happy torment, when my torturer

O, torture me no more! I will confess.

On pain of torture, from those bloody hands

On thy soul's peril and thy body's torture,

Refuse me, hate me, torture me to death!

Say he be taken, rack'd, and tortured;

Strange tortures for offenders, never heard of,

Than on the torture of the mind to lie

That same Berowne I'll torture ere I go.

That so her torture may be shortened.

The time, the place, the torture. O, enforce it!

This torture should be roar'd in dismal hell.

To torture thee the more, being what thou art.

Turning dispiteous torture out of door!

Which is our honour, bitter torture shall

Which to be spoke would torture thee.

With vilest torture let my life be ended.

You go about to torture me in vain.

curses he shall have, the tortures he shall feel, will break the

hand, to th' infernal deep, with Erebus and tortures vile also.

then torture my wife, pluck the borrowed veil of modesty

Do in consent shake hands to torture me,

FIRST SOLDIER. He calls for the tortures. What will you say without

GEORGE. While we devise fell tortures for thy faults.

IACHIMO. Thou'lt torture me to leave unspoken that

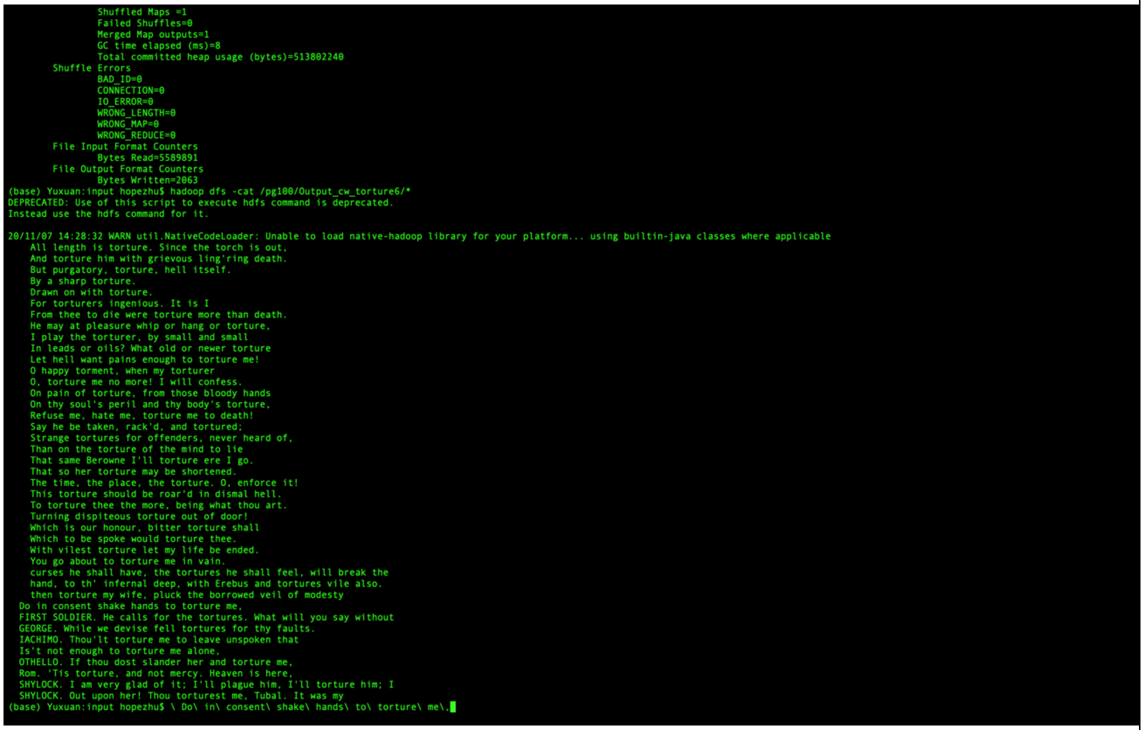
Is't not enough to torture me alone,

OTHELLO. If thou dost slander her and torture me,

Rom. 'Tis torture, and not mercy. Heaven is here,

SHYLOCK. I am very glad of it; I'll plague him, I'll torture him; I

SHYLOCK. Out upon her! Thou torturest me, Tubal. It was my



```
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=513802240
        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=5589891
        File Output Format Counters
                Bytes Written=2063
(base) Yuxuan:input hopezhu$ hadoop dfs -cat /pg100/Output_cw_torture6/*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

20/11/07 14:28:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
    All length is torture. Since the torch is out,
    And torture him with grievous ling'ring death.
    But purgatory, torture, hell itself.
    By a sharp torture.
    Drawn on with torture.
    For torturers ingenious. It is I
    From thee to die were torture more than death.
    He may at pleasure whip or hang or torture,
    I play the torturer, by small and small
    In leads or oils? What old or newer torture
    Let hell want pains enough to torture me!
    O happy torment, when my torturer
    O, torture me no more! I will confess.
    On pain of torture, from those bloody hands
    On thy soul's peril and thy body's torture,
    Refuse me, hate me, torture me to death!
    Say he be taken, rack'd, and tortured;
    Strange tortures for offenders, never heard of,
    Than on the torture of the mind to lie
    That same Berowne I'll torture ere I go.
    That so her torture may be shortened.
    The time, the place, the torture. O, enforce it!
    This torture should be roar'd in dismal hell.
    To torture thee the more, being what thou art.
    Turning dispiteous torture out of door!
    Which is our honour, bitter torture shall
    Which to be spoke would torture thee.
    With vilest torture let my life be ended.
    You go about to torture me in vain.
    curses he shall have, the tortures he shall feel, will break the
    hand, to th' infernal deep, with Erebus and tortures vile also.
    then torture my wife, pluck the borrowed veil of modesty
  Do in consent shake hands to torture me,
  FIRST SOLDIER. He calls for the tortures. What will you say without
  GEORGE. While we devise fell tortures for thy faults.
  IACHIMO. Thou'lt torture me to leave unspoken that
  Is't not enough to torture me alone,
  OTHELLO. If thou dost slander her and torture me,
  Rom. 'Tis torture, and not mercy. Heaven is here,
  SHYLOCK. I am very glad of it; I'll plague him, I'll torture him; I
  SHYLOCK. Out upon her! Thou torturest me, Tubal. It was my
(base) Yuxuan:input hopezhu$ \ Do\ in\ consent\ shake\ hands\ to\ torture\ me\,
```

2. My code with comments:

```
import java.io.IOException;
import java.util.*;


import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordCount_torture {

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "torture count");
        job.setJarByClass(WordCount_torture.class); //My class name is
'word count'

        job.setMapperClass(TokenizerMapper.class); // Set my Mapper
class
        job.setReducerClass(IntSumReducer.class); // Set my Reducer
class

        job.setOutputKeyClass(Text.class);  // Set the key class for
my output class
        job.setOutputValueClass(Text.class); // Set the value class
for the output data

        job.setInputFormatClass(TextInputFormat.class);// Set input
format for this job
        job.setOutputFormatClass(TextOutputFormat.class); // Set
```

```
output format for this job

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }

    public static class TokenizerMapper
            extends Mapper<LongWritable, Text, Text, Text> {

        private Text word = new Text();

        public void map(LongWritable key, Text value, Context context
        ) throws IOException, InterruptedException {

            String line[] = value.toString().split("\\n{1,}");
            ArrayList<String> results = new ArrayList<>();

            for (int i = 0; i < line.length; i++) {
                if (line[i].contains("torture")) {
                    results.add(line[i]);
                }
            }


            for (String result : results) {
                word.set(result);
                context.write(word, new Text(" "));
            }
        }
    }

    public static class IntSumReducer
            extends Reducer<Text, IntWritable, Text, Text> {
        public void reduce(Text key, Iterable<IntWritable> values,
                           Context context
        ) throws IOException, InterruptedException {

            context.write(new Text(key), new Text(" "));

        }
    }


}
```

3. Hadoop command list and order you have used to run your job and get your results

Since we have uploaded the pg100.txt file to our HDFS, so we simply ignore the procedure.

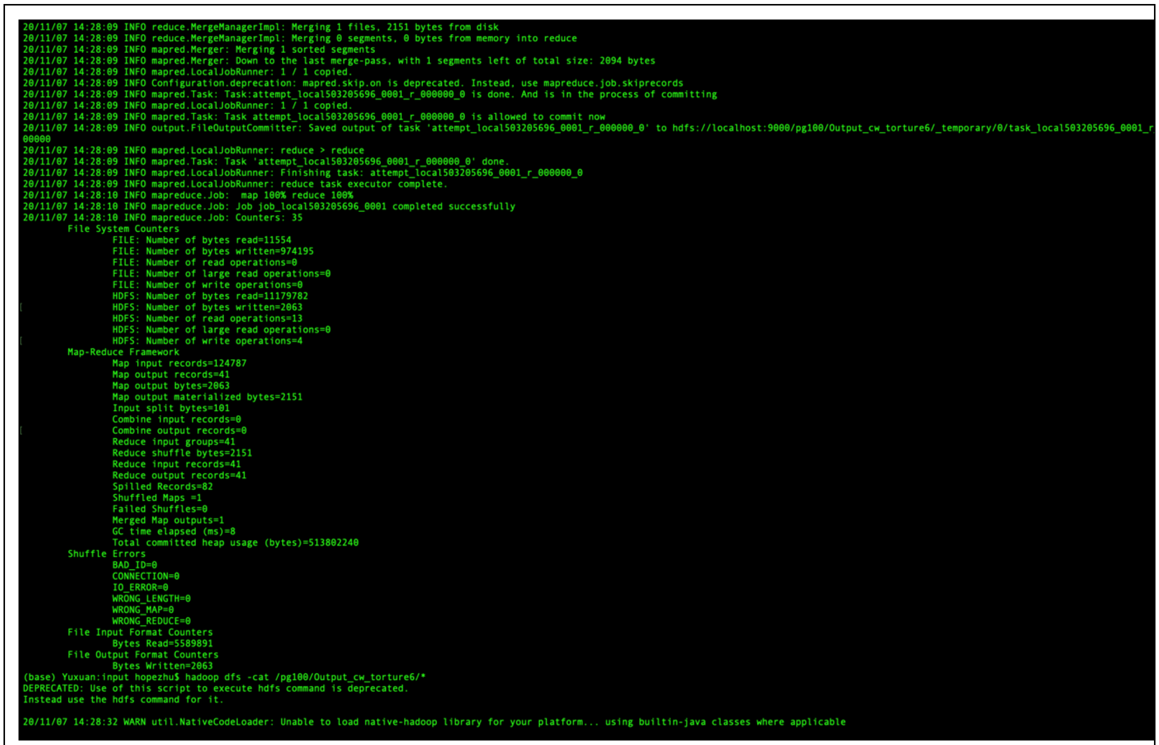
We create a file called torture_classes in my directory to store the classes and create a jar file by javac command

```
 javac -classpath ${HADOOP_CLASSPATH} -d '/Users/yuxuan/hadoop-
2.9.2/input/torture_classes'
'/Users/yuxuan/coding/IDEA/Hadoop_cw2/src/WordCount_torture.java'

jar -cvf torture.jar -C torture_classes/ .
```

We compile the jar file and obtain the results

```
hadoop jar '/Users/yuxuan/hadoop-2.9.2/input/torture.jar'
WordCount_torture /pg100/input_cw /pg100/Output_cw_torture
```

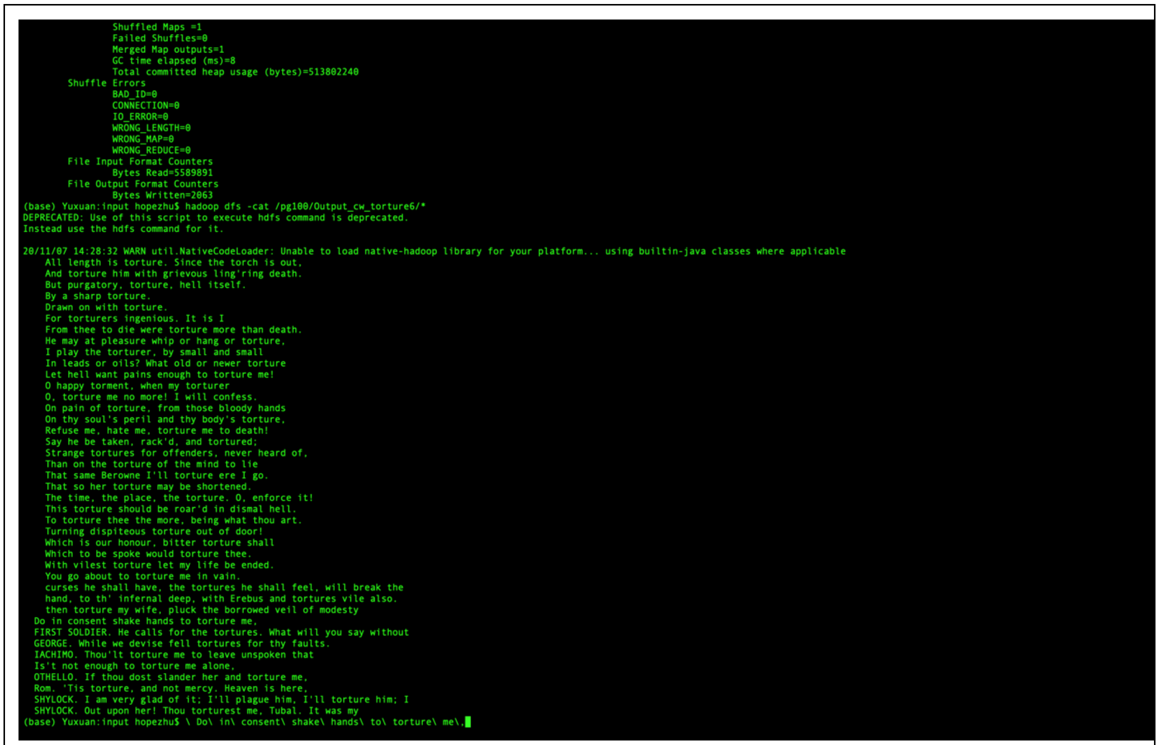


```
20/11/07 14:28:09 INFO reduce.MergeManagerImpl: Merging 1 files, 2151 bytes from disk
20/11/07 14:28:09 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
20/11/07 14:28:09 INFO mapred.Merger: Merging 1 sorted segments
20/11/07 14:28:09 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 2094 bytes
20/11/07 14:28:09 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/07 14:28:09 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
20/11/07 14:28:09 INFO mapred.Task: Task:attempt_local503205696_0001_r_000000_0 is done. And is in the process of committing
20/11/07 14:28:09 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/07 14:28:09 INFO mapred.Task: Task attempt_local503205696_0001_r_000000_0 is allowed to commit now
20/11/07 14:28:09 INFO output.FileOutputCommitter: Saved output of task 'attempt_local503205696_0001_r_000000_0' to hdfs://localhost:9000/pg100/Output_cw_torture6/_temporary/0/task_local503205696_0001_r
00000
20/11/07 14:28:09 INFO mapred.LocalJobRunner: reduce > reduce
20/11/07 14:28:09 INFO mapred.Task: Task 'attempt_local503205696_0001_r_000000_0' done.
20/11/07 14:28:09 INFO mapred.LocalJobRunner: Finishing task: attempt_local503205696_0001_r_000000_0
20/11/07 14:28:09 INFO mapred.LocalJobRunner: reduce task executor complete.
20/11/07 14:28:10 INFO mapreduce.Job:  map 100% reduce 100%
20/11/07 14:28:10 INFO mapreduce.Job: Job job_local503205696_0001 completed successfully
20/11/07 14:28:10 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=11554
                FILE: Number of bytes written=974195
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=11179782
                HDFS: Number of bytes written=2063
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=124787
                Map output records=41
                Map output bytes=2063
                Map output materialized bytes=2151
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=41
                Reduce shuffle bytes=2151
                Reduce input records=41
                Reduce output records=41
                Spilled Records=82
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=513802240
        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=5589891
        File Output Format Counters
                Bytes Written=2063
(base) Yuxuan:input hopezhu$ hadoop dfs -cat /pg100/Output_cw_torture6/*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

20/11/07 14:28:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Here, we successfully run the file and we need to output the file and check the results

```
hadoop dfs -cat /pg100/Output_cw_torture/*
```



```
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=513802240
        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=5589891
        File Output Format Counters
                Bytes Written=2063
(base) Yuxuan:input hopezhu$ hadoop dfs -cat /pg100/Output_cw_torture6/*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

20/11/07 14:28:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
    All length is torture. Since the torch is out,
    And torture him with grievous ling'ring death.
    But purgatory, torture, hell itself.
    By a sharp torture.
    Drawn on with torture.
    For torturers ingenious. It is I
    From thee to die were torture more than death.
    He may at pleasure whip or hang or torture,
    I play the torturer, by small and small
    In leads or oils? What old or newer torture
    Let hell want pains enough to torture me!
    O happy torment, when my torturer
    O, torture me no more! I will confess.
    On pain of torture, from those bloody hands
    On thy soul's peril and thy body's torture,
    Refuse me, hate me, torture me to death!
    Say he be taken, rack'd, and tortured;
    Strange tortures for offenders, never heard of,
    Than on the torture of the mind to lie
    That same Berowne I'll torture ere I go.
    That so her torture may be shortened.
    The time, the place, the torture. O, enforce it!
    This torture should be roar'd in dismal hell.
    To torture thee the more, being what thou art.
    Turning dispiteous torture out of door!
    Which is our honour, bitter torture shall
    Which to be spoke would torture thee.
    With vilest torture let my life be ended.
    You go about to torture me in vain.
    curses he shall have, the tortures he shall feel, will break the
    hand, to th' infernal deep, with Erebus and tortures vile also.
    then torture my wife, pluck the borrowed veil of modesty
  Do in consent shake hands to torture me,
  FIRST SOLDIER. He calls for the tortures. What will you say without
  GEORGE. While we devise fell tortures for thy faults.
  IACHIMO. Thou'lt torture me to leave unspoken that
  Is't not enough to torture me alone,
  OTHELLO. If thou dost slander her and torture me,
  Rom. 'Tis torture, and not mercy. Heaven is here,
  SHYLOCK. I am very glad of it; I'll plague him, I'll torture him; I
  SHYLOCK. Out upon her! Thou torturest me, Tubal. It was my
(base) Yuxuan:input hopezhu$ \ Do\ in\ consent\ shake\ hands\ to\ torture\ me\,
```