XI'AN JIAOTONG-LIVERPOOL UNIVERSITY 西交利物浦大学

YEAR 4 COURSE WORK SUBMISSION

Name	Yuxuan	Wu
ID Number	1716309	
Programme	Bioinformatics	
Module Title	Big Data Analytics	
Module Code	INT303	
Assignment Title	Lab2	
Submission Deadline	20 th November	
Lecturer Responsible	Gangmin Li	

I certify that:

• I have read and understood the University's definitions of COLLUSION and PLAGIARISM (available in the Student Handbook of Xi'an Jiaotong-Liverpool University).

With reference to these definitions, I certify that:

- I have not colluded with any other student in the preparation and production of this work;
- this document has been written solely by me and in my own words except where I
 have clearly indicated and acknowledged that I have quoted or used figures from
 published or unpublished sources (including the web);
- where appropriate, I have provided an honest statement of the contributions made to my work by other people including technical and other support staff.

I understand that unauthorised collusion and the incorporation of material from other works without acknowledgement (plagiarism) are serious disciplinary offences.

Signature 吳字轩		Date2020.11.11		
For Academic Office use:	Date Received	Days Late	Penalty	

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 ll	1742
I have	1581
in the	1528
I will	1509
of the	1437
to the	1298
my lord	1189
I do	820

to be	809

```
I am 1830
I ll 1742
I have 1581
in the 1528
I will 1509
of the 1437
to the 1298
my lord 1189
I do 820
to be 809
```

2. My code with comments:

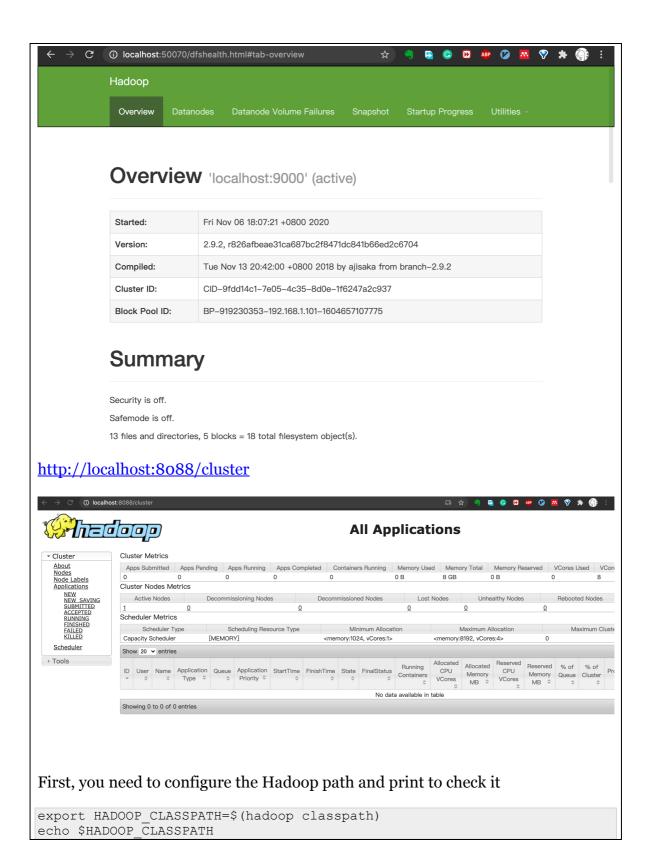
```
import java.io.IOException;
import java.util.*;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
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; <math>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> {
```

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



/Users/yuxuan/hadoop-2.9.2/etc/hadoop:/Users/yuxuan/hadoop-2.9.2/share/hadoop/comm on/lib/*:/Users/yuxuan/hadoop-2.9.2/share/hadoop/common/*:/Users/yuxuan/hadoop-2.9 .2/share/hadoop/hdfs:/Users/yuxuan/hadoop-2.9.2/share/hadoop/hdfs/lib/*:/Users/yux 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/shar e/hadoop/yarn/*:/Users/yuxuan/hadoop-2.9.2/share/hadoop/mapreduce/lib/*:/Users/yux uan/hadoop-2.9.2/share/hadoop/mapreduce/*:/Users/yuxuan/hadoop-2.9.2/contrib/capac 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 ① localhost:50070/explorer.html#/pg100/input Hadoop **Browse Directory** /pg100/input Show entries Search: Block ↓↑ ↓↑ Last Replication Permission Group Size Modified Size Name pg100.txt 5.33 Nov 06 128 MB hopezhu supergroup 22:09 Showing 1 to 1 of 1 entries Next 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 bigram_classes < > **Ⅲ** □ □ **₩ ∨ ♦ ∨** ♠ **○** ■ V Q 搜索 个人收藏 ^ 修改日期 大小 种类 今天 上午 9:00 WordCount.class 1 KB Java 类文件 @ 隔空投送 WordCount\$TokenizerMapper.class 今天 上午9:00 3 KB Java 类文件 ■ 最近项目 WordCount\$Tokeniz...IntSumReducer.class 今天 上午 9:00 2 KB Java 类文件 🚣 应用程序 🔙 桌面 1 文稿 ● 下载 □ 坚果云 iCloud △ iCloud 云盘 位置 网络 网络 标签 Work 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

```
I am 1830
I ll 1742
I have 1581
in the 1528
I will 1509
of the 1437
to the 1298
my lord 1189
I do 820
to be 809
```

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

```
Description of the second control of the sec
```

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> {
       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(" "));
```

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
```

```
JOHING 14:30 MM or note, Pergethangerian, Reging 5 files, 2315 bytes from memory reto reduce

JOHING 24:00 MM or note, Pergethangerian, Reging 6 sepents, 0 % places from memory reto reduce

JOHING 14:30 MM owers Reging 2000 magnetic Local states and the second states of the second
```

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

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

```
District Shufflesd

Regard May computed

Regard May
```