INF 551 – Spring 2019

Homework #5

Due: April 28, Sunday
100 points

TA handling this homework: Mingyi Lin linmingy@usc.edu

Task: Write a Hadoop MapReduce program to compute frequencies (number of occurrences) of bigrams that appear in the lines of a collection of documents.

A bigram consists of two consecutive words in a line of text. For example, consider a line :

"usc data informatics students". This line will only have three bigrams: "usc data", "data informatics" and "informatics students" and their count will be 1 respectively (You do not need to consider the reverse order). You need to count the frequencies of bigrams from the collection of documents. Your two output files should look like the following picture (bigrams and frequencies) :

```
1   data informatics    4
2   data scientists 1
3   have the    1
4   informatics degree  1
5   informatics provides    1
6   informatics skills  1
7   knowledge and   1
8   skills required 2
9   skills to   1
10  technical challenges    1
```

You can modify the example WordCount.java to complete the above task. For this homework, you need to **set the number of reduce tasks to be 2**. You can use the provided **bigram.txt** file to test your program.

Notes for your reference before you start the task :

1. For Hadoop installation, you can refer to Hadoop slides p.76 – p.81
2. For compile & run MapReduce programs, you can refer to Hadoop slides p.82

Execution format

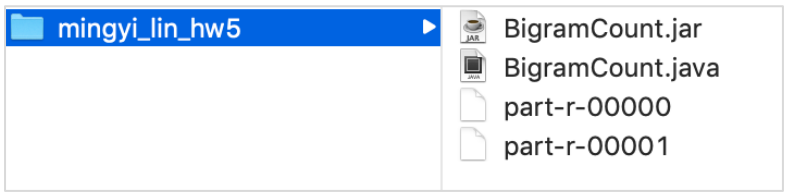
1. Hadoop jar BigramCount.jar BigramCount *input-directory* *output-directory*
2. *input-directory* : any folder name that contains a number of text files
3. *output-directory* : any folder name that contains your output results

(Note: input-directory and output-directory should be any name you assigned in the command)

Submissions

1. Name your files as below and submit to Blackboard by the due time. Your submission **must be a zip file** with the naming convention: <FirstName>_<LastName>_hw5.zip.
   - BigramCount.java
   - BigramCount.jar
   - part-r-00000 (output files)
   - part-r-00001 (output files)

2. Example :
   - **Your unzipped folder should be like the following picture:**



Grading Policy:

1. Homework assignments are due at 11:59pm on the due date and should be submitted on Blackboard. Late homework will be deducted by 10% of its points for every 24 hours that it is late. No credit will be given after 72 hours of its due time.
2. If your code cannot be run with the commands as above, there will be a 40 point penalty.
3. You can only use **built-in java library** and **org.apache.hadoop** for this homework. If you use non -standard packages, there will be a 30 point penalty.
4. If your code takes more than 3 minutes to finish, there will be a 20 point penalty.
5. Please make sure your code can run with hadoop on EC2.