Programming Assignment 2
Map-Reduce Optimization

Due on Tuesday March 3 before midnight

Description

The purpose of this project is to improve the performance of the program that creates histograms of pixels developed in Project 1 by using a combiner and in-mapper combining.

This project must be done individually. No copying is permitted. **Note: We will use a system for detecting software plagiarism, called Moss, which is an automatic system for determining the similarity of programs.** That is, your program will be compared with the programs of the other students in class as well as with the programs submitted in previous years. This program will find similarities even if you rename variables, move code, change code structure, etc.

Note that, if you use a Search Engine to find similar programs on the web, we will find these programs too. So don't do it because you will get caught and you will get an F in the course (this is cheating). Don't look for code to use for your project on the web or from other students (current or past). Just do your project alone using the help given in this project description and from your instructor and GTA only.

Platform

As in Project 1, you will develop your program on SDSC Comet. Optionally, you may use your laptop/PC to develop your program first and then, after you make sure that it works there, you transfer it and test it to Comet. You may also use IntelliJ IDEA or Eclipse to help you develop your program, if you have done so in Project 1. Note that it is required that you test your programs on Comet before you submit them.

Setting up your Project

Login to comet or your laptop and copy project1 to project2:

```
cd
cp -a project1 project2
```

Your project is to improve your Java code in `project2/src/main/java/Histogram.java`.

Project Description

In this project, you are asked to improve the performance of your code developed in Project 1 in two different ways: 1) using a combiner and 2) using in-mapper combining. See pages 14-18 in the class notes. You need

to write two Hadoop Map-Reduce jobs in the same file `project2/src/main/java/Histogram.java`. Each one of this Map-Reduce jobs will read the same input file but will produce output to a different output directory: in your Java main program, both Map-Reduce jobs will read args[0] as the input file with the pixels (which is pixels-small.txt or pixels-large.txt). The first Map-Reduce job will write on the output directory args[1] (which is `output`) while the second Map-Reduce job will write on the output directory args[1]+"2" (which is `output2`). If you want to look at the results of the second Map-Reduce job in distributed mode, you can add the following statement in `project2/histogram.distr.run`:

```
hdfs dfs -get /user/$USER/output2/part* output2-distr
```

Note: The hash table in your second Map-Reduce job must have at most 3*256=768 entries. If you get more, it means that either your Color hashCode() method is wrong (it should return the same int for the same Color) or your Color compareTo method is wrong (it should rteturn 0 for two equal Colors).

Instructions for Comet: Go inside the directory `project2`.

```
run histogram.build
```

and you can run Histogram.java in standalone mode over a small dataset using:

```
sbatch histogram.local.run
```

The file `histogram.local.out` will contain the trace log of the Map-Reduce evaluation while the file `output/part-r-00000` will contain the results. These results must be equal to `solution-small.txt`. To run Histogram.java in distributed mode over a larger dataset, use:

```
sbatch histogram.distr.run
```

The file `histogram.distr.out` will contain the trace log of the Map-Reduce evaluation while the file `output-distr/part-r-00000` will contain the results. These results must be equal to `solution-large.txt`. Note that running in distributed mode will waste at least 10 of your SUs so do this after you make sure that your program works correctly for small data (it produces the same results as the solution).

What to Submit

You need to submit the following files only:

```
project2/src/main/java/Histogram.java
project2/histogram.local.out
project2/output-distr/part-r-00000       (the output of the first Map-Reduce job)
project2/histogram.distr.out
```

Submit Programming Assignment #2:

Select a file: Choose File no file selected

Send File

Last modified: 02/25/2020 by Leonidas Fegaras