CSC 4760/6760 Big Data Programming

Assignment 2

Due Date: 11:59 pm, Wednesday, Feb 24, 2021

1. (100 points) (PageRank)

Dataset:

The toy dataset is the following graph. The PageRank values are already known. We can use it to check your program.

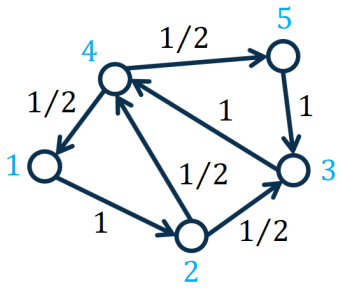

Figure 1: A toy graph for computing PageRank. The number on the edge represents the transition probability from one node to another.

The PageRank values are given in the following table (given that the decay factor $c = 0.85$):

Nodes	PageRank Values
1	0.1556
2	0.1622
3	0.2312
4	0.2955
5	0.1556

PageRank:

Compute the PageRank value of each node in the graph. Please refer to the slides for more details about the PageRank method. The key PageRank equation is as follows.

$$\mathbf{r} = c\mathbf{P}^{\top}\mathbf{r} + (1-c)\,\mathbf{1}/n$$

where $\mathbf{r}$ represents the $n \times 1$ PageRank vector with each element $\mathbf{r}_i$ representing the PageRank value of node i, n represents the number of nodes in the graph, $\mathbf{P}$ represents the $n \times n$ transition probability matrix with each element $\mathbf{P}_{i,j} = p_{i,j} = \frac{1}{d_i}$ representing the transition probability from node i to node j, d_i represents the degree of node i, $\mathbf{P}^{\top}$ represents the transpose of $\mathbf{P}$, $c \in (0,1)$ represents a decay factor, $\mathbf{1}$ represents a $n \times 1$ vector of all 1's, and n represents the number of nodes in the graph.

Please see the slides for more details.

In this assignment, we set the decay factor $c = 0.85$ and set the number of iterations to 30.

Implementation:

Design and implement a MapReduce program to compute the PageRank values.

You are encouraged to implement the PageRank algorithm from scratch without using the provided "PageRankIncomplete.java" file.

The provided "PageRankIncomplete.java" file is incomplete. It will help you start programming with Hadoop. You need to understand the existing code and basic structure in order to complete the file.

Example command:

hadoop jar PageRank.jar file:///home/rob/pagerank/01InitialPRValues.txt file:///home/rob/pagerank/02AdjacencyList.txt file:///home/rob/pagerank/output 30

Report:

Please write a report illustrating your experiments. You need to explain your basic idea about how to design the computing algorithm. You may add comments to the source code such that the source code can be read and understood by the graders.

In the report, you should include the answers to the following questions.

1) Explanation of the source code:

1.1) How is the Mapper function defined? Which kind of intermediate results are generated?

1.2) How is the Reducer function defined? How do you aggregate the intermediate results and get the final outputs?

1.3) Do you use a Combiner function? Why or why not?

2) Experimental Results

2.1) Screenshots of the key steps. For example, the screenshot for the outputs in the terminal when you run "Hadoop jar YourJarFile" command. It will demonstrate that your program has no bug.

2.2) Explain your results. Does your implementation give the exact PageRank values? How large are the errors?

Submission Materials:

a) Your report

b) Source code (.java file) and the .jar file for the Hadoop

c) The output file of your program.