Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 1: MapReduce (Chapter 2)

Due: Mar 23, 09:30 AM

Reminders

- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Jinhong Jung (montecast9@gmail.com)
- Please submit one zip file to eTL. The zip file should include the report (in pdf), Makefile, README, jar files (*.jar), and source files (*.java). Also, the zip file's name should be formed in "HW1_StudentID.zip" (e.g., HW1_2015-12345.zip).
- Please turn in the hardcopy of your report to T.A. at the classroom on Mar 23.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date. That is, after Mar 30, we will NOT receive your submission for this assignment.

1. Questions

Question 1: Implement the in-degree counting algorithm in '*IndegreeCounter.java*'. Make sure that when we type '*make in*', Hadoop should execute your program, and print the in-degree of each node in the graph, given by the '*problem.edge*' file included in the homework package. Report the screenshot after printing the in-degree of each node. (To look the content of files in HDFS, use the *'hadoop fs –cat'* command.) [25 points]

Solution:

In-degree of 'problem.edge'
1 3
2 1
3 2
4 2
6 1
7 1
8 1
9 1

Question 2: Implement the degree counting algorithm for counting out-degree in '*OutdegreeCounter.java*'. Make sure that when we type 'make out', Hadoop should execute your program, and the program should print the out-degree of each node in the graph, which is represented in the '*problem.edge*' file. Report the screenshot after printing the out-degree of each node. [25 points]

Solution:

Out-degree of 'problem.edge'
0 2
1 3
2 2
4 2
5 2
6 1

Question 3: Implement a MapReduce code for computing the degree distribution in '*DegreeDistribution.java*'. Make sure that the following commands perform the corresponding tasks for the graph in the '*problem.edge*' file:

- '*make in_dist*': compute the in-degree distribution on Hadoop, and print the in-degree distribution.
- '*make out_dist*': compute the out-degree distribution on Hadoop, and print the out-degree distribution.

Report the screenshots after printing the in-degree and the out-degree distributions. [25 points]

Solution:

In-degree distribution of 'problem.edge'
1 5 2 2 3 1

Out-degree distribution of 'problem.edge'
1 1 2 4 3 1

Question 4: Using your implementations, compute the in-degree and the out-degree distributions of the *LiveJournal* dataset, which is one of real-world graphs. Report the plots of both degree distributions for the graph in log-log scale. [25 points]

Solution: The left figure is the in-degree distribution, and the right one is the out-degree distribution of the *LiverJournal* dataset.