COMP4434 Big Data Analytics

Assignment 3

PolyU, Hong Kong

Instructor: HUANG Xiao
Logistics: You should submit answers through Learn@PolyU (Blackboard). The deadline is Thursday April 3, 11:55 PM. I will not accept submission from any other channels except Blackboard. These are the best exercises that could help you be well prepared for quizzes. Please work independently.

Problem 1 (3 points)
Assume that a large table is distributed across multiple files, each containing partial rows of the table. Each row is composed of the following data:
(student_name, department, salary).
For example, (Bob, Computing, 30,000) means Bob graduated from the department of Computing and the salary of his first job is 30,000.
The objective is to determine, in each department, the total number of graduated students whose salary is more than 25,000 in their first job.

(a) What are the relationships between MapReduce and Apache Hadoop?

(b) Provide a pseudo-code for the Map workers, specifying the input and output (key, value) pairs.

(c) Provide a pseudo-code for the Reduce workers, specifying the input and output (key, value) pairs.

Problem 2 (2 points)
We have four text files as follows, storing the student grades of four subjects.

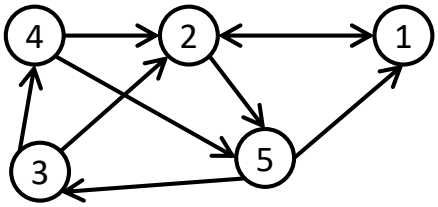
math.txt	physics.txt	chemistry.txt	art.txt
James, 81	James, 57	James, 78	James, 67
John, 83	John, 78	John, 92	John, 89
Robert, 75	Robert, 68	Robert, 68	Robert, 88
Michael, 71	Michael, 71	Michael, 91	Michael, 87
David, 79	David, 79	David, 77	David, 87
Mary, 73	Mary, 69	Mary, 74	Mary, 79
Linda, 83	Linda, 79	Linda, 89	Linda, 94
Susan, 67	Susan, 76	Susan, 87	Susan, 78
Lisa, 76	Lisa, 74	Lisa, 92	Lisa, 91

Our goal is to calculate the total scores of students in all four subjects.

(a) Write a pseudo-code for the Map workers, specifying the input and output (key, value) pairs.

(b) Write a pseudo-code for the Reduce workers, specifying the input and output (key, value) pairs.

Problem 3 (3 points) Assume that the connections among 5 webpages are represented as a graph as follows. We use the PageRank equation (with random teleports) to update the rank value of each webpage. Assume that the initial (iteration 0) rank value of each webpage is 1/5, and the damping factor β is 0.8.



(a) Formulate the PageRank equation for each webpage, ensuring the inclusion of specific weights.

(b) Compute the PageRank values for all five webpages during iterations 1 and 2. Subsequently, arrange the webpages in descending order based on their PageRank values.