COMP4434 Big Data Analytics

Assignment 3

PolyU, Hong Kong

Instructor: HUANG Xiao
Logistics: You should submit your solutions through Learn@PolyU (Blackboard). The deadline is Monday April 8, 11:55 PM. I will no accept submission from any other channels except Blackboard. These are the best exercises that could help you be well prepared for quizzes. Thus, please work independently.

Problem 1 (3 points)
Assume that we have a large number of files that store the weighted edges of a huge directed graph. Each one of them contains content in the format as follows.

$(1, 5, 0.3)$,	$(2, 3, 0.7)$,	$(6, 3, 0.5)$,	$(2, 6, 0.8), ...$
$(4, 7, 0.2)$,	$(7, 2, 0.1)$,	$(9, 3, 0.9)$,	$(1, 7, 0.4), ...$
$(8, 3, 0.8)$,	$(9, 1, 0.3)$,	$(3, 4, 0.4)$,	$(4, 5, 0.2), ...$

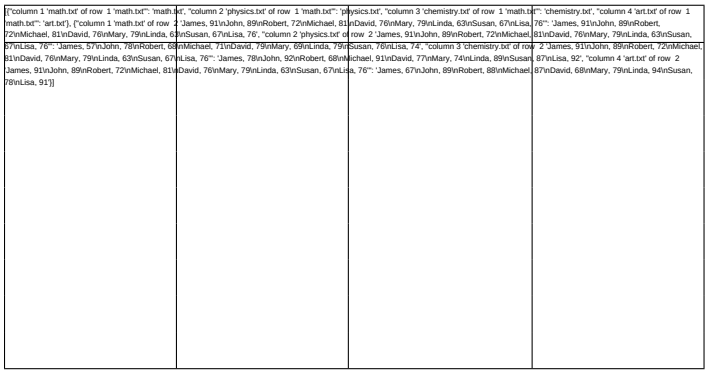
...

As we could see, each file contains many lines. Each line contains a set of triples, i.e., (source id, target id, weight). Our goal is to calculate the sum of weights of all edges of each vertex (including incoming and outgoing edges).

(a) Write pseudo-code for map worker, including the (key, value) pairs of the input and output.

(b) Write pseudo-code for reduce worker, including the (key, value) pairs of the input and output.

Problem 2 (5 points)
We have four text files as follows, storing the student grades of four subjects.



Our goal is to calculate the total scores of students in all four subjects. (In practice, we could have more students and more subjects.)

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

(b) Write pseudo-code for map worker, including the (key, value) pairs of the input and output.

(c) Write pseudo-code for reduce worker, including the (key, value) pairs of the input and output.

(d) What are the concrete inputs and outputs of your implemented mapper and reducer when processing the above four text files?