Programming Project 3: Web Graph Processing Using Map-Reduce

(This Project is for grads and undergrads)

The goal of this assignment is to provide you with exposure to Map-Reduce (Hadoop) programming. Assume that you are provided with a file containing the edges of a web graph. Each line of the file contains a pair of document ids (each id is an integer between 000 and 999) signifying an edge in the web graph from the first document to the second document of the pair (i.e., the first vertex is the source of the edge and the second is the destination of the edge). An example input file is shown below.

000 001
002 001
000 001
001 002
000 002
002 000
001 002

Given such an input file, you need to write two map-reduce programs. The first program will compute the in-degree and out-degree of each document. For the above input file the output of your program should be

000 2 1
001 1 2
002 2 2

The second program should produce an output consisting of a sequence of lines. Each line consists of a document ID (say "ABC") followed by a unique list of document IDS that are sources of edges that are directed towards ABC. For the above input file, the output should be

000 [002]
001 [000 002]
002 [001 000]

Important points to note:

1. Your program should conform to the Map-Reduce programming paradigm. Mappers should process each line and output appropriate key-value pairs. Reducers should process all the values corresponding to each key value. Submissions that do not conform to the map-reduce paradigm will automatically get zero points.
2. You can assume that each document is the destination to at least one edge.

3. The input file may contain duplicate edges. However, the list of source vertices in the output file must be unique (in the example above, although the edge 000 001 is repeated, the list corresponding to 001 contains 000 only once).
4. This project is to be done in groups of two (both for grads and undergrads). Projects done individually will get 15% bonus credits.
5. Graduate students have to install Hadoop on their own. If you have a laptop on which YOU have installed Hadoop, you can use that set up as well.However, by using such a set up you are certifying that you installed Hadoop on the laptop and not someone else. Undergrads can request the TA to install Hadoop on their VMs. Hadoop: The Definitive Guide e-book is available here (http://it-ebooks.info/book/635/).
6. A sample test file is provided here (http://www.cs.uga.edu/~laks/DCS-2014-Sp/PA3GS-test-file.txt).