Assignment 1
ITCS-6010/8010: Cloud Computing for Data Analysis

Due by 11:59:59pm on Thursday, February 2, 2012

The goal of this programming assignment is to become familiar with Hadoop. Please see the Assignment 1 web page (`http://www.cs.uncc.edu/~sakella/courses/cloud/assign1/`) for details and links to resources.

1. You should first install and get Hadoop running (in pseudo-distributed mode) on your computer. The Assignment 1 web page has links to installation resources.

2. You should next run the example WordCount program on the data files provided on the Assignment web page, and submit the resulting output files. The WordCount program counts the total number of times each distinct word appears in the input set of files. So the output is of the form (word, count).

   You should submit the output when you run the program on each individual data file, and when you run the program on the set of all data files.

3. You should then modify the WordCount program so it outputs the wordcount for each distinct word in each file. So the output of this DocWordCount program should be of the form ((word, filename), count). Submit your source code (named `DocWordCount.java`), and the output file from running your program when the input is the set of all example data files.

Assignments are due by 11:59:59pm on Thursday, February 2, 2012. Submission will be on Moodle and instructions will be posted on the Assignment web page.

Assignments are to be done individually. See course syllabus for late submission policy and academic integrity guidelines.