Homework 4

COMP 7150/8150

FALL 2016

Instructor: Deepak Venugopal

Due Date: November 28 2016 (submit to ecourseware)

Spark: In this assignment you will work with Apache Spark.

Part 1: The easiest way to get Spark working is to use a virtual machine as follows.

a. Download VMWare player from http://www.vmware.com/products/player/playerpro-evaluation.html
b. Download the Cloudera Quickstart VM from http://www.cloudera.com/downloads/quickstart_vms/5-8.html
   Here choose the VMWare option for "Select a platform"
c. Start the VMWare player and load the Cloudera VM.
d. The Cloudera VM is packaged with Spark, so you can directly start using Spark. Go to terminal and execute pyspark. It will take you to the spark shell.
e. Work on the pyspark shell to answer the following questions. For your submission, simply submit a .txt file that contains the code that you wrote on the pyspark command line.

If you don't want to use a virtual machine, you need to follow instructions in http://spark.apache.org/docs/latest/building-spark.html

2. Write a program to print out the number of lines of an input file that contain more than 10 words. You can create your own input file to test your program. You don't need to submit these test files. (30 points)
3. Given two files, find all the common non-article (not "a", "an" and "the") words between the files. Create your own files to test your program. You don't need to submit these test files. (30 points)
4. Create an inverted index for a document. Specifically, for every word, output the line numbers in the document that word occurs in. Create your own document to test your program. You don't need to submit these test files. (40 points)