## HU Extension School E-63 Big Data Analytics

## Assignment 03

### Handed out: 02/12/2016 Due by 11:30PM on Friday, 02/19/2016

Download and Install VMWare Workstation or Fusion product from HU VMWare store.

**Problem 1)** Modify attached class WordCount.java so that its result excludes the following stop words:

|  |  |  |
| --- | --- | --- |
| I  a  about  an  are  as  at  be  by  com  for  from how | in  is  it  of  on  or  that the  this to  was  what  when where | who  will  with the www |

as well as special characters (dashes, parentases, etc). Stop word lists could be much lomger than this. You do not have to be extremely thoroughful. You would like to get a more or less clean list of ordinary words with the numbers of their occurances. Do not fret. Be reasonable. Perform analysis on te text of James Joyce’s Ulysis.

**Problem 2)** Write another MapReduce program which would read the “word count” output of the previous job and order the results by the declaining number of occurances.

**Problem 3)** Create a program that will find out how many words appear only once, how many twice, three times, four times and so on in James Joyce’s Ulysis

**Problem 4)** Combine operations of two MapReduce programs in Problems 1 and 3 above into a single program with chained MapReduce jobs.

**Problem 5)** Move attached Inverter.java class from old MapReduce API to the new API. Demonstrate that new and old clas produce the same result. Use patent data set to demonstrate your work.

If you have your own working VM with installed CDH5.5.1, do this assignment on that VM. If, for what ever reason, you do not posses a working VM with CDH5.5.1, please be free to download Clouder’s Getting started VM and do you assignment on that VM.

Capture all steps of your implementation with comments indicating what is it you are accomplishing with every step in an MS Word document. Upload to the class site. Please upload your working Java files as well. Please post comments and questions to the class Discussion Board on the Canvas site.