Map Reduce Assignment **DUE DATE: 03/20/2017 by 11:59 p.m.**

PROBLEM 1:

Preparation:

1. Launch your VM.
2. Unzip the file called Shakespeare.tar.gz – you should have done this in your HDFS hands-on exercise.
3. Copy the shakespeare directory to HDFS – verify that it is there.

d. The contents of the /user/training/shakespeare HDFS directory should contain the files comedies, glossary, histories, poems, and tragedies.

Requirements: Use the shakespeare directory in HDFS to do the following:

1. Write a map-reduce program that creates an inverted list of words and the files that contain them. That is, for each word, you will display a list of files that contain that word. A sample output is given below:

‘anger’ [‘histories’, ‘tragedies’]

‘laugh’ [‘comedies’, ‘poems’, ‘histories’]

…….

1. Modify your map-reduce program (in 1) to display the number of times the word occurs in each file. A sample output is shown below:

‘anger’ {‘histories’: 3, ‘tragedies’: 8}

‘laugh’ {‘comedies’: 7, ‘poems’: 2, ‘histories’: 15}

………

In order to determine the file from which a line is being streamed, you could use the following command:

filename = os.getenv(‘map\_input\_file’)

Note that the filename will include the complete path (e.g., hdfs://user/training/shakespeare/comedies). You will have to extract just filename (‘comedies’ in this example).

PROBLEM 2:

(Source: <https://class.coursera.org/bigdata-003/wiki/ProgrammingAssignmentsHW3>

**Note that this link is no longer available**)

Download data files bundled as a .zip file from blackboard.

Each file in this archive contains entries that look like:

journals/cl/SantoNR90:::Michele Di Santo::Libero Nigro::Wilma Russo:::Programmer-Defined Control Abstractions in Modula-2.

that represent bibliographic information about publications, formatted as follows:

paper-id:::author1::author2::…. ::authorN:::title

**Your task is to compute how many times every term occurs across titles, for *each* author.**

For example, the author Alberto Pettorossi has the following terms occur in titles with the indicated cumulative frequencies (across all his papers): program:3, transformation:2, transforming:2, using:2, programs:2, and logic:2.

Remember that an author might have written multiple papers, which might be listed in multiple files. Further notice that ‘terms’ must **exclude** common stop-words, such as prepositions etc. For the purpose of this assignment, the stop-words that need to be omitted are listed in the script [stopwords.py](http://spark-public.s3.amazonaws.com/bigdata/Assignment%20Data/stopwords.py). In addition, single letter words, such as "a" can be ignored; also hyphens can be ignored (i.e., deleted). Lastly, periods, commas, etc. need to be ignored; in other words, only alphabets and numbers can be part of a title term: Thus, “program” and “program.” should both be counted as the term ‘program’, and "map-reduce" should be taken as 'mapreduce'. Note: You do *not* need to do stemming, i.e. "algorithm" and "algorithms" can be treated as separate terms.

The assignment is to write a map-reduce program for the above task using either Java or Python Hadoop Streaming.

**PLEASE UPLOAD YOUR PROGRAMS AND OUTPUTS TO BLACKBOARD. ALSO, DON’T UPLOAD THE ENTIRE OUTPUTS FILE. JUST THE FIRST 50 LINES SHOULD DO**