TOBB University of Economics and Technology
Department of Computer Engineering

BİL401/501 Distributed Data Processing and Analysis
Spring 2015

Homework 3

Due by Feb 17 Tue 24:00

Subject: "Secondary Sorting" and "Order Inversion" patterns

Your input is the same reutersnews document set you used in the first assignment. You will develop two programs:

1) Develop a MapReduce program for computing **tf-idf** values for word-document pairs using "*order inversion*" pattern. Please read section 3.3 of the textbook (Lin&Dyer book) for reference.

Your output should look like the following:

(word, doc): tf df tf-idf

for all words and documents in increasing order. *tf*: term frequency, *df*: document frequency, *tf-idf*: term frequency-inverse document frequency.

2) Develop a MapReduce program for generating the inverted index for all words in all documents. Document list for each word should be listed in increasing order. To accomplish this use "*secondary sorting*" and "order inversion" patterns. Read section 4.4 of the textbook (Lin&Dyer book) for reference. Your output should look like this:

word: 4 9 104 189 567 987

For both programs (a) transform words to lower case, (b) ignore words less than 5 characters long.

In your programs add comments in lines where you used the above requested patterns (order inversion and secondary sorting).

Report the running times in a report for both programs.

Submit your homework via google drive. Use the following header in your programs.

BİL401/501 Distributed Data Processing and Analysis (Spring 2015)
Homework #
Date: ...
Name: **Firstname Lastname**