TOBB University of Economics and Technology
Department of Computer Engineering

BİL401/501 Distributed Data Processing and Analysis
Spring 2015

Homework 1

Due by Jan 26 Mon 24:00

Subject: MapReduce on Hadoop

1) **WordLengthHistogram** (50 points)

Write the MapReduce program to calculate the word length histogram for a given set of text documents. Your program will scan all documents and count the words of length "big" (10+ chars), "medium" (5..9 chars), "small" (2..4 chars), "tiny" (1 char). Download the sample text documents from the lab section of the course in Piazza called **reutersnews.rar**. Run your MapReduce job on the text files and report results. Deliverables:

- WordLengthHistogram.jar, WordLengthHistogramMapper.java, WordLengthHistogramReducer.java, WordLengthHistogramTest.java
- Report1.pdf: Time it took to run the job, number of nodes of your clusters, histogram (big#, medium#, small# and tiny#).

Help: http://hadoop.apache.org/docs/stable1/mapred_tutorial.html

2) **InvertedIndexForTweets** (40 points)

Write the MapReduce program to create an inverted index for tweets. Your program should read all tweets in the data repository (*see part 3 of this assignment*) in the form of (tweetid, tweet) and create an index in the form of (word, tweetid) for all words in tweets. You should make the index case-insensitive, meaning President and president should be treated the same. Deliverables:

- InvertedIndexForTweets.jar, InvertedIndexForTweetsMapper.java, InvertedIndexForTweetsReducer.java, InvertedIndexForTweetsTest.java
- Report2.pdf: Time it took to run the job, number of nodes of your clusters, # tweets, # words.
- tweets.txt: Each line should include: *tweetid, "tweet"*
- inverted_index.txt: Each line should include: *word, tweetid*

3) **Collecting streaming tweets** (10 points)

For the previous task (InvertedIndexForTweets) write a program to collect a large number of tweets into a file in the requested format using Twitter's Streaming API[1] in either Java (hbc) or Python. If you do not want to do this task, find a collection of tweets from the Web for that task (and loose 10 points). Deliverables:

- CollectTweets.rar: Java code only and the script to run the code.

Submit your homework to **edogdu@etu.edu.tr** by email w/ subject header "**[BİL401/501] asg1**" and email attachment containing the zip file. Use the following header in your programs.

[1] https://dev.twitter.com/overview/api/twitter-libraries

BİL425 Software Development for Mobile Apps (Spring 2015)
Homework #
Date: ...
Name: **Firstname Lastname**