TOBB University of Economics and Technology
Department of Computer Engineering

BİL401/501 Distributed Data Processing and Analysis
Spring 2016

Homework 1

Due by Jan 25 (Monday) 09:00am

Subject: Hadoop environment

In this assignment you will setup the standalone Hadoop environment and test Wordcount example.

1) Download and install **VirtualBox** or VMware on your pc.
2) Download and run **HortonWorks Sandbox** on the virtual machine (step1) from this site: http://hortonworks.com/products/hortonworks-sandbox/#install.
3) Follow this tutorial http://hortonworks.com/hadoop-tutorial/introducing-apache-hadoop-developers/ and execute wordcount program on the following text file: http://www.gutenberg.org/cache/epub/100/pg100.txt (The Complete Works of William Shakespeare).
4) Open a GitHub or BitBucket private account (so that you share only with the instructor) to submit your assignments and project codes. GitHub allows private account for educational use only; apply for one with your school email address at: https://education.github.com/discount_requests/new
5) Upload and submit your assignment via your git repository account. Name the repo **asg1**, and upload the sample text file (blahblah.txt), execution results (output.txt), program (WordCount.java). Upload dates/times should be before the submission date/time to get full credit.
6) Share your git account with the instructor (edogdu@github, erdogandogdu@bitbucket).

Extra:

1) Download WordCount.java and pom.xml files (piazza).
2) Create a Java project (bigdata2016s/asg1) in Eclipse
3) Run maven on the project (mvn clean)
4) Upload jar and class files to Sandbox
5) Run hadoop job with your WordCount and submit the screenshot of successful execution (execution.png) to your git repo.