Homework 3
Playing with Hadoop
Due 4/13/18 11:59 pm CT
Early bird deadline: 4/6/18 11:59 pm CT

In this homework, you will prepare an environment to work (Step 1), install Apache Hadoop in this environment (Step 2) and go through two coding tutorial (Steps 3 and 4). There is also a bonus task.

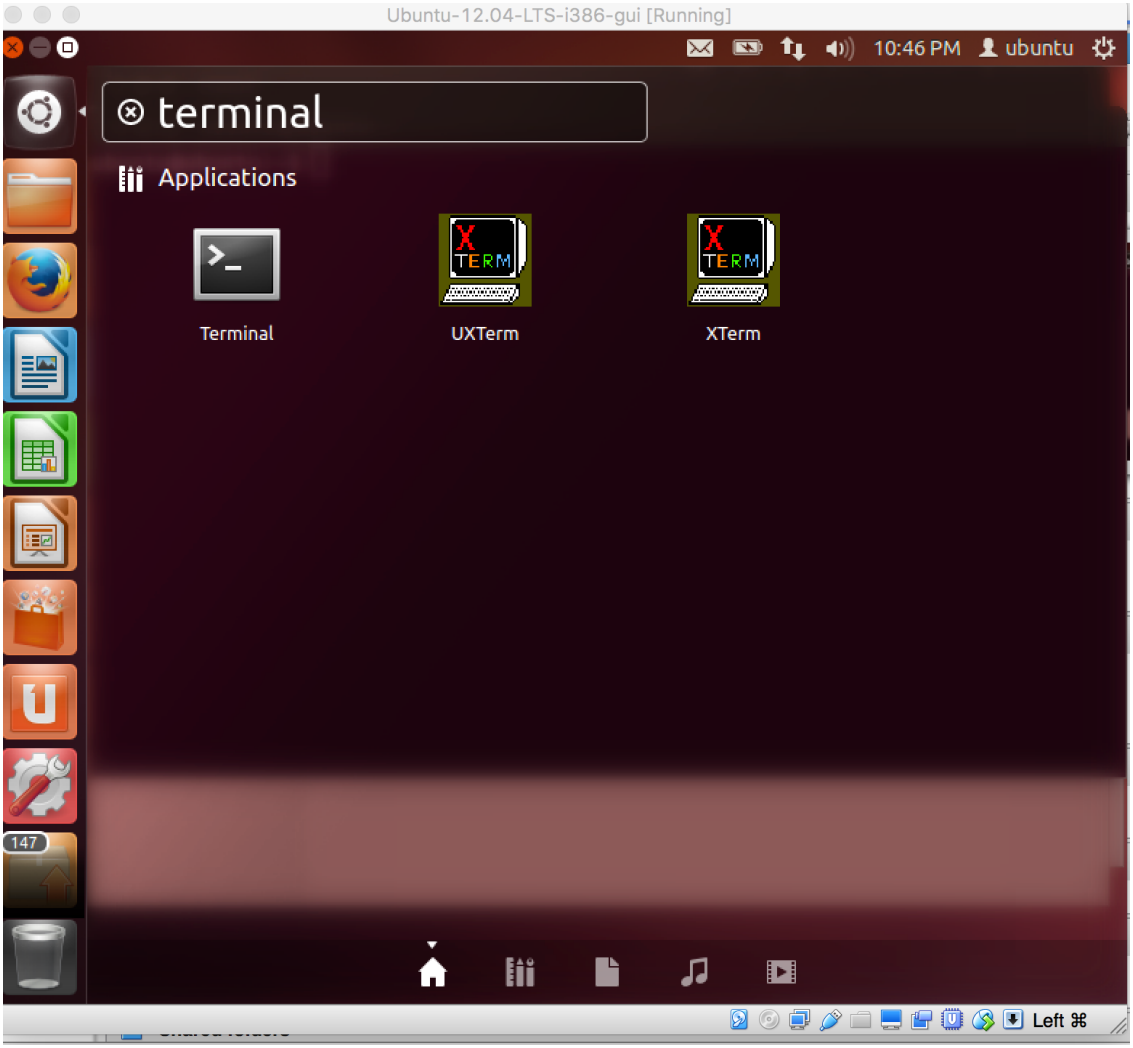
For every step, make sure that you track how long it took you to complete the step. Keep notes of any difficulties or questions you may have.

Step 1: Prepare a Linux environment to work

You need a Linux environment to work on, one in which you have administrative rights. You can use any Linux distribution installation that you may prefer.

If you do not have a Linux environment readily available, we recommend that you install a virtual machine on your computer (Windows or Mac) and run an Ubuntu distribution:

1. Download and install Oracle VirtualBox
   Instructions available at https://www.virtualbox.org/wiki/Downloads
2. Download an Ubuntu virtual machine image for VirtualBox
   There are many available on the web, for example in Sourceforge: https://sourceforge.net/projects/virtualappliances/files/Linux/Ubuntu/
   We used the image Ubuntu-12.04-LTS-i386-gui.ova
   This image is large (around 1.4 GB), so start the download when you are not in a hurry (for example, while watching a video for our course)
3. Import the image into VirtualBox as described in
   https://docs.oracle.com/cd/E26217_01/E26796/html/qs-import-vm.html
   The image we suggested in item (3) is configured to use 1 GB of RAM. If your computer is constrained in resources (e.g. only 2 GB of RAM), please e-mail the instructors as soon as possible so that we can make another platform available to you.
4. Start your Linux instance by selecting the virtual machine you just imported and in the toolbar click the “Start” button.
   The username is “Ubuntu”, and its password is “nimda”.
   The root password is toor.
   In the GUI, you can get a terminal window by searching for “terminal” as in the following figure:



Step 2: Install Hadoop

Only start this step after you go through the material on Hadoop (videos 1, 2, and 3 in weeks 7 and 8.) Use the instructions available at https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html#Standalone_Operation.

Step 3: Go through a simple Hadoop application

Only start this step after you watched the video that goes through an example (video-4).
Go through the Hadoop Tutorial available from Cloudera:
https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1.html

You need to go through all the three parts:

- WordCount v1.0 Source Code
- Running WordCount v1.0
- Walkthrough of WordCount v1.0

Step 4: Go through a simple Hadoop application using Pig

Only start this step after you watched the Pig video (video-5).

Go through the Pig Tutorial in http://pig.apache.org/docs/r0.10.0/start.html#tutorial
If you encounter a difficulty that you cannot solve after working on it for 15 minutes, post your problem on and Piazza and move on.

Final step: Writing the report and submitting it

Your report has two parts:

1. For every one of the four steps, it documents the following:
    a. how much time you needed to conclude each step
    b. list of difficulties, questions, comments for each step. If everything went smoothly, just state "no problems or questions".
    c. snapshot of the step:
        i. for Step 1, snapshot – picture(s) of your screen – of your virtual machine running or of your Linux environment
        ii. for Step 2, snapshot indicating completion
        iii. for Step 3, snapshot indicating completion
        iv. for Step 4, snapshot indicating how far you got
2. (**Bonus Task**) A short description (6-10 sentences may suffice) describing how you could use Hadoop to build an application for analyzing Tweets feeds as follows:
    - using the Twitter API (https://dev.twitter.com/overview/api), the application retrieves the tweets for the last hours and the associated attributes for users, places (geo-location that may be added to tweets), and entities (URLs, hashtags, user mentions)
    - the application processes this retrieved information to identify the following:
        - most popular hashtags in the 50 largest cities in the world;
        - most popular words in tweets that include a URL for a popular newspaper, magazine, or TV show
        - most popular video

    Think of this part as an initial direction you would give to your intern(s) on how to approach the problem.

Submission instructions

Submit your report as a pdf file on eCampus (Projects and Homeworks → HW3)
There is no need to use github in this homework, given that you are not developing new code.

Rubric

Steps 1, 2, and 3: 25 points each
Step 4:

- partial attempt well documented on the roadblock: 17 points;
- complete: 25 points

Bonus Task: 15 points

Early bird bonus: 10 points