Instruction on setup your local machine or login onto class server

For the implementation exercise in homework 1, we recommend you use jupyter notebook with pyspark package. There are 2 choices we provide : 1) use your local machine; 2) use class server. I will give you some insturction on how to do it for both options.

1. If you want to use class server

1. Be in the NEU network. Unfortunately the class server cannot be accessed using off-campus network.
2. Open any browser and try to access https://ds5230.ccs.neu.edu:8000/. Ignore what browser warns you and allow the unsecure access.
3. Use the account info I sent to you by email to login.
4. Once you successfully login, you are at the root directory,i.e. `/home/yourusername/`. There is a folder called `assignment_dataset` which stores the dataset you need for homework assignments. For homework 1, you can read dataset from the path `/home/yourusername/assignment_dataset/homework1/`. Please do NOT upload the dataset to server yourself, which is entirely not necessary.

2. If you want to use your own laptop/desktop

a. Windows Pro/Windows Home

It is easy to get wrong if you manually set up all dependencies from scratch on Windows system. So we recommend you use docker to run jupyter and pyspark.

1). Docker Installation

1. Windows Pro.

   Install Docker by following instruction on

   https://docs.docker.com/docker-for-windows/install/

2. Windows Home.

   Unfortunately, Windows Home doesn't support HyperV, which is required for docker installation. But we can install docker toolbox instead. Please follow the instruction on

https://docs.docker.com/toolbox/toolbox_install_windows/

2). Start up jupyter with pyspark instance

After Docker installation, we can start up a `jupyter/pyspark` instance by the following commmand

```
docker run -it --rm -p 8888:8888 \
  -v "$PWD":/home/jovyan/work
  jupyter/pyspark-notebook
```

b. Linux/Unix based system, or Mac OS

If you are using either Mac OS, or any linux/unix system (e.g. Ubuntu, Debian, Fedora), please read this section and follow these instructions.

Install Anaconda

Anaconda is a python package manager which we will use to install any package we want to use. Download and install anaconda by the insturction on

https://www.anaconda.com/download/#linux for linux system,

or https://www.anaconda.com/download/#macos for Mac OS.

Type 'yes' when you are asked whether to add anaconda to path, or you can also add it manually after installation.

To verify the installation, open up a new terminal and type

```
$ conda
```

to see if it's a valid command now.

Install jupyter notebook

Now you are ready to install the jupyter notebook. Follow the instruction on https://test-jupyter.readthedocs.io/en/rtd-theme/install.html

Install pySpark

Install pySpark by running the following command line in terminal :

```
$ conda install -c conda-forge pyspark
```

Now you should have everything you need to do homework 1. To verify pySpark installation, open up a jupyter instance and try to see if you can import *pyspark*.