Data Science Professional Practicum (DSCI 560)
Laboratory Assignment 1
Instructor: Young Cho, Ph.D.

NOTE: You must use GitHub to store all source code and reports associated with the lab assignments. You must commit/push all of the modifications for the assignments with detailed descriptions at least once daily when working on the lab assignment. Please submit the GitHub link and record of your commits and detailed commit descriptions to Coursistant before the due date.

This is the first assignment of the course that focuses on helping you get your systems set up, finish required installations, and run a few basic tasks. We will use virtual machines (VM) or containers with Linux as the operating system and Python as the programming language for the entire course.

In this assignment, you will install and set up the Ubuntu Linux VM and install the required software packages for Python. You will run a few tasks on Linux Terminal and Python tasks to get hands-on experience. You are expected to submit a document with snapshots and a few lines of description for each task performed as a part of the assignment. Please ensure that the script files you submit have proper comments.

This lab must be completed individually.

1. Installation and Setup

1.1. Register at Broadcom and then download and install VMware

- If you do not already have an account, register here: https://support.broadcom.com/
- Then download VMWare Fusion from here https://support.broadcom.com/group/ecx/free-downloads

1.2. Download Ubuntu ISO Image

We would be using Ubuntu throughout the semester. Follow the instructions below to download the Ubuntu image and set up the virtual machine.

- Go to https://ubuntu.com/download and download the Ubuntu Desktop ISO image.
- Open VMware
- Create a new virtual machine.
- Name the virtual machine, select ***"Linux"*** as the type, and ***"Ubuntu (64-bit)"*** as the version.
- Allocate memory (RAM) for the virtual machine ***(at least 2GB is recommended***).
- Allocate disk space for the virtual machine (***at least 20GB is recommended***).
- Create the virtual machine.

- When prompted, choose the Ubuntu ISO image you downloaded earlier as the installation medium.
- The Ubuntu installer will launch. Follow the installation wizard to install Ubuntu on the virtual machine.

1.3. Install Python on Linux

Once you’ve installed Linux VM, install Python and some Python packages within Linux.

- Open a terminal in your Ubuntu virtual machine.
- Update the packages list using: `sudo apt update`
- Ubuntu usually comes with Python 3 pre-installed. To ensure it is installed, run the following command: `sudo apt install python3`
- To verify that Python3 has been installed successfully, run: `python 3 —-version`
- Install pip (Python Package Manager) - pip is a package manager for Python that allows you to easily install Python packages from the Python Package Index (PyPI): `sudo apt install python3-pip`
- To verify that pip has been installed successfully, run: `pip3 –version`

1.4. Tutorials

If you are new to Linux or Python, spend some time reading the documentation and tutorials below to understand the basic concepts and commands.

Linux:

- Start with https://linuxjourney.com/ tutorial by Linux Journey. This comprehensive tutorial covers basic Linux commands and concepts.
- Refer to https://www.cheatography.com/davechild/cheat-sheets/linux-command-line/ for quick access to essential Linux commands.

Python:

- Begin with https://docs.python.org/3/tutorial/ which provides a solid foundation in Python programming.
- For practical examples and exercises, follow https://realpython.com/python-for-beginners/ by Real Python.

2. Get Familiar with Linux and Python

In this section, we perform a few different tasks using both the AWS Management Console and the CLI to provide you with hands-on experience of how to set up services and tasks on AWS and get you prepared for the upcoming assignments.

2.1. Playing around with Linux Terminal

- Open the Linux terminal.
- Create a new directory named ***"<your name>_<your USC ID>"*** on the desktop.
- Inside the folder, create two subdirectories named ***"data"*** and ***"scripts"***
- Create an empty Python file inside the scripts folder named ***"task_1.py"***
- Use the list command to view the created script file.

2.2. A basic Python Script

- Open the task_1.py Python file you created in the previous step using vim / nano.
- Write a Python script that reads a user's name as input and greets the user with "Hello, [name]!".
- The script should prompt the user for input and display the greeting in the terminal.
- Save and exit the editor.
- Run the Python code.
- Feel free to complete the Python tutorials and tasks till you feel comfortable using the editors and get accustomed to the Python syntax before moving to the next step.

2.3. Python Web-scraping Task

- Create a new file ***"web_scraper.py"*** in the scripts folder.
- Install the required libraries (***Requests and BeautifulSoup4***) using pip: `pip install requests beautifulsou4`
- Open this website https://www.cnbc.com/world/?region=world
- Analyze the HTML structure by inspecting the elements on the Page.
- Find the corresponding tags for the Market banner on the top and the Section titled Latest News on the page.
- Create two new folders in the ***"data"*** folder called ***"raw_data"*** and ***"processed_data"***
- Write a Python script that uses Requests and BeautifulSoup to collect data from the provided link.
- Save the collected data in the ***"raw_data"*** folder to a file named ***"web_data.html"***
- Using the terminal print the first 10 lines of the created html file on the terminal.

2.4. Data Filtering Task

- Write a python script called ***"data_filter.py"***. Read the "web_data.html" file into a Python list, extracting specific elements of interest from the data.
- Store the (***marketCard_symbol, marketCard_stockPosition*** and ***marketCard-changePct***) from the market banner and the (***LatestNews-timestamp, title*** and ***link***) for each entry in the LatestNews list.
- Store the market banner data in a CSV named ***"market_data.csv"*** in the ***processed_data*** folder.
- Store the new data in a CSV named ***"news_data.csv"*** in the ***processed_data*** folder.
- Print appropriate messages to the console. For example: Filtering fields, storing Market data, CSV created, etc.

3. Team Formation

3.1. You must make teams of 3 students by the end of this week

4. Submission

4.1. Please submit a ZIP archive of all of the source code to Coursistant. https://usc.xlearnedu.com

4.2. Take screenshots of your work and results and put them in a report with the descriptions.

4.3. Please include your ***Name and USC ID*** in the documents.

4.4. There will be a 50% penalty for all late submissions.