

# DSCI 560 – Laboratory Assignment 1

Name: Tianyi Ge

USC ID: 5804514679

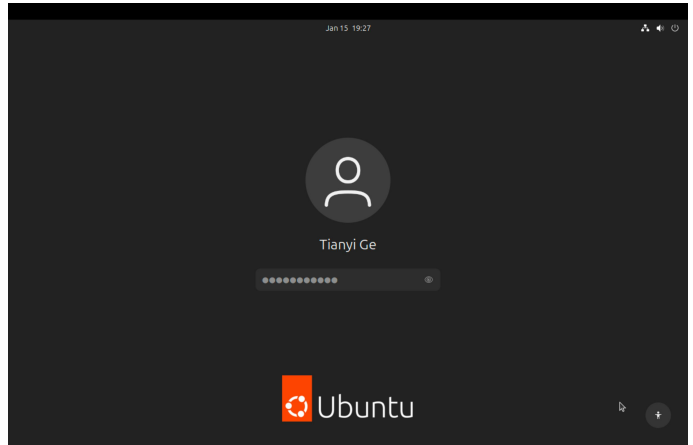
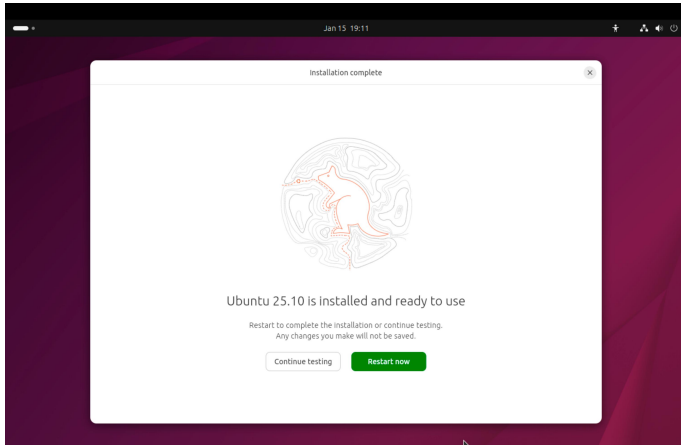
Github Link: <https://github.com/Mia0403/DSCI560-Lab1.git>

Youtube Link: <https://youtu.be/4peSgunoNGo>

## 1. Installation and Setup

### ● VMware & Ubuntu Installation

I installed VMware Fusion on macOS and created an Ubuntu 64-bit virtual machine. The virtual machine was successfully launched and used throughout this lab.



### ● Python Installation

Python 3 and pip were installed and verified in the Ubuntu virtual machine.

```
tianyi-ge@tianyi-ge-VMware20-1:~$ sudo apt update
[sudo: authenticate] Password:
Hit:1 http://ports.ubuntu.com/ubuntu-ports InRelease
Hit:2 http://ports.ubuntu.com/ubuntu-ports questing InRelease
Hit:3 http://ports.ubuntu.com/ubuntu-ports questing-backports InRelease
Hit:4 http://ports.ubuntu.com/ubuntu-ports questing-security InRelease
72 packages can be upgraded. Run 'apt list --upgradable' to see them.
tianyi-ge@tianyi-ge-VMware20-1:~$ sudo apt install python3
python3 is already the newest version (3.13.7-1).
python3 set to manually installed.
The following packages were automatically installed and are no longer required:
  linux-headers-6.17.0-5 linux-tools-6.17.0-5
  linux-headers-6.17.0-5-generic linux-tools-6.17.0-5-generic
  linux-modules-6.17.0-5-generic
Use 'sudo apt autoremove' to remove them.

Summary:
  Upgrading: 0, Installing: 0, Removing: 0, Not Upgrading: 72
tianyi-ge@tianyi-ge-VMware20-1:~$ python3 --version
python3 --version: command not found
tianyi-ge@tianyi-ge-VMware20-1:~$ python3 --version
Python 3.13.7
tianyi-ge@tianyi-ge-VMware20-1:~$
```

```
tianyi-ge@tianyi-ge-VMware20-1:~$
Setting up dpkg-dev (1.22.21ubuntu3.1) ...
Setting up libstdc++-15-dev:arm64 (15.2.0-4ubuntu4) ...
Setting up libpython3.13-dev:arm64 (3.13.7-1ubuntu0.2) ...
Setting up python3-pip (25.1.1+dfsg-1ubuntu2) ...
Setting up libjs-sphinxdoc (8.2.3-1ubuntu2) ...
Setting up gcc-15-aarch64-linux-gnu (15.2.0-4ubuntu4) ...
Setting up libpython3-dev:arm64 (3.13.7-1) ...
Setting up python3.13-dev (3.13.7-1ubuntu0.2) ...
Setting up gcc-15 (15.2.0-4ubuntu4) ...
Setting up g++-15-aarch64-linux-gnu (15.2.0-4ubuntu4) ...
Setting up python3-dev (3.13.7-1) ...
Setting up g++-15 (15.2.0-4ubuntu4) ...
Setting up gcc-aarch64-linux-gnu (4:15.2.0-4ubuntu1) ...
Setting up gcc (4:15.2.0-4ubuntu1) ...
Setting up g++-aarch64-linux-gnu (4:15.2.0-4ubuntu1) ...
Setting up g++ (4:15.2.0-4ubuntu1) ...
update-alternatives: using /usr/bin/g++ to provide /usr/bin/c++ (c++) in auto mode
Setting up build-essential (12.12ubuntu1) ...
Processing triggers for man-db (2.13.1-1) ...
Processing triggers for libc-bin (2.42-0ubuntu3) ...
tianyi-ge@tianyi-ge-VMware20-1:~$ pip3 --version
pip 25.1.1 from /usr/lib/python3/dist-packages/pip (python 3.13)
```


## 2. Get Familiar with Linux and Python

### ● Playing around with Linux Terminal

A new directory named `d "<your name>_<your USC ID>"` (A new directory named `"TianyiGe_5804514679"` was created on the desktop.) was created on the desktop. Inside the directory, two subdirectories named data and scripts were created, and an empty Python file named `task_1.py` was added to the scripts folder.

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ cd ~/Desktop
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop$ mv Yourname_YourUSCID TianyiGe_5804514679
```

The task\_1.py file was edited using a terminal text editor. The script prompts the user to enter their name and prints a greeting message in the format “Hello, [name]!”. The script was executed successfully in the terminal.



The screenshot shows a terminal window with a dark purple background. The title bar at the top reads "tianyige@tianyige-VMware20-1: ~/Desktop/Yourname...". Below the title bar, the terminal shows the command prompt "~/.Desktop/Yourname\_YourUSCID/scripts" followed by the command "cat task\_1.py". The output of the command is displayed in a white box, showing the contents of the file: "name=input("Please enter your name: ")", "print("Hello, ", name + "!)".

```
tianyige@tianyige-VMware20-1: ~/Desktop/Yourname...  
~/.Desktop/Yourname_YourUSCID/scripts  
cat task_1.py  
name=input("Please enter your name: ")  
print("Hello, ", name + "!")
```

(1) Installed required Python libraries (requests and BeautifulSoup4) using the Ubuntu package manager due to system-managed Python environment.

```

tianyi1-ge@tianyi1-ge-VNware20-1:~/Desktop/Yourname_YourUSCID/scripts:
$ sudo apt install python3-requests python3-bs4
[sudo: authenticate] Password:
python3-requests is already the newest version (2.32.3+dfsg-5ubuntu1).
python3-requests set to manually installed.
The following packages were automatically installed and are no longer
needed:
  linux-headers-6.17.0-5      linux-tools-6.17.0-5
  linux-headers-6.17.0-5-generic  linux-tools-6.17.0-5-generic
  linux-modules-6.17.0-5-generic
Use 'sudo apt autoremove' to remove them.

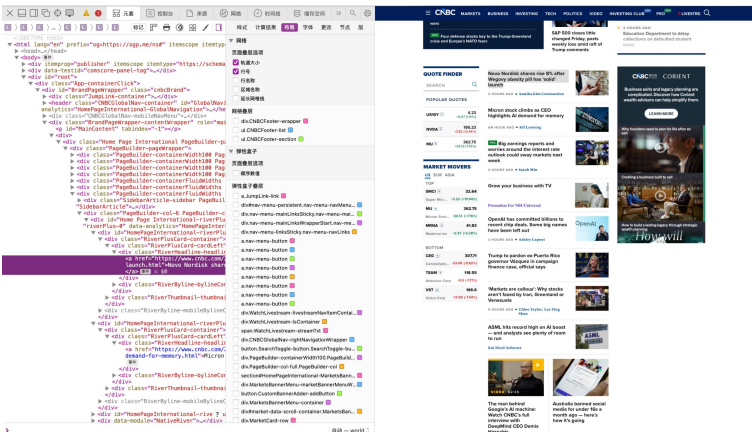
Installing:
python3-bs4

Installing dependencies:
python3-cssselect  python3-lxml  python3-webencodings
python3-html5lib  python3-soupsieve

Suggested packages:
python3-genshi  python-lxml-doc

```

(2) A web scraping script was written to collect HTML data from the CNBC World News webpage.



(3) Two directories named raw\_data and processed\_data were created inside the data folder to store raw and processed data separately

```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ cd ..
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID$ cd data
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/data$ mkdir raw_data
processed_data
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/data$ ls
processed_data raw_data
```

(4) A Python script named web\_scraper.py was written using Requests and BeautifulSoup to send an HTTP request to the target URL and retrieve the HTML content of the webpage. The collected HTML content was saved into a file named web\_data.html inside the data/raw\_data directory for further processing.



```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ nano web_scraper.py
import requests
from bs4 import BeautifulSoup
url="https://www.cnbc.com/world/?region=world"
headers={
    "User-Agent":"Mozilla/5.0"
}
html=requests.get(url,headers=headers)
with open("../data/raw_data/web_data.html", "w", encoding="utf-8") as f:
    f.write(html.text)
print("HTML saved")

#make it lool pretty
beautifulsoup=BeautifulSoup(html.text,"html parser")
with open("../data/raw_data/web_data.html", "w", encoding="utf-8") as f:
    f.write(beautifulsoup.prettify())
print("HTML saved and prettified")

tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ python3 web_scraper.py
HTML saved
HTML saved and prettified
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/Yourname_YourUSCID/scripts$ head ../data/raw_data/web_data.html
<!DOCTYPE html>
<html itemscope="" itemtype="https://schema.org/WebPage" lang="en" prefix="og:https://ogp.me/ns#">
  <head>
    <meta content="website" property="og:type"/>
    <meta content="International: Top News And Analysis" property="og:title"/>
    <meta content="CNBC International is the world leader for news on business, technology, China, trade, oil prices, the Middle East and markets." property="og:description"/>
    <meta content="https://www.cnbc.com/world/" property="og:url"/>
    <meta content="CNBC" property="og:site_name"/>
    <meta content="max-image-preview:large" name="robots"/>
    <meta content="telephone=no" name="format-detection"/>
```

## ● Data Filtering Task

A Python script named data\_filter.py was written to read the saved HTML file and extract relevant information using BeautifulSoup. Market data and latest news data were parsed from the webpage and stored into two CSV files named market\_data.csv and news\_data.csv inside the data/processed\_data directory. The script was executed successfully and printed messages to the terminal during processing. This screenshot shows the successful execution of the data filtering script.



```
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ python3 data_filter.py
Filtering field
Store the market data
CSV created

tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/data$ ls processed_data
market_data.csv news_data.csv

tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ head ../data/processed_data/market_data.csv
symbol,stockposition,pct
tianyi-ge@tianyi-ge-VMware20-1:~/Desktop/TianyiGe_5804514679/scripts$ head ../data/processed_data/news_data.csv
timestamp,title,link
2 Hours Ago,Week in review: Stocks battled a flood of news and we booked some profits,https://www.cnbc.com/2026/01/17/week-in-review-stocks-battled-a-flood-of-news-and-we-booked-some-profits.html
3 Hours Ago,Trump threatens to sue JPMorgan Chase for 'debanking' him,https://www.cnbc.com/2026/01/17/trump-jpmorgan-chase-debanking.html
5 Hours Ago,Trump: NATO members to face tariffs up to 25% until a Greenland deal is struck,https://www.cnbc.com/2026/01/17/trump-greenland-tariffs-nato.html
6 Hours Ago,"Led by Texas, states race to prove they can put bitcoin on public balance sheet",https://www.cnbc.com/2026/01/17/texas-us-states-budgets-bitcoin-crypto-strategic-reserve.html
7 Hours Ago,Unshaken: Why Brazilian stocks have looked past the Venezuela attack,https://www.cnbc.com/2026/01/17/unshaken-why-brazilian-stocks-have-looked-past-the-venezuela-attack.html
7 Hours Ago,Bestselling author: How to create better habits without relying on discipline,https://www.cnbc.com/2026/01/17/james-clear-how-to-create-better-habits-without-relying-on-discipline.html
8 Hours Ago,"Warren Buffett: To maximize your potential, ask yourself this question",https://www.cnbc.com/2026/01/17/warren-buffett-to-maximize-your-potential-ask-yourself-this-question.html
8 Hours Ago,"Buy these five stocks ahead of earnings, Bank of America says",https://www.cnbc.com/2026/01/17/stocks-to-buy-ahead-of-earnings-bank-of-america-says.html
8 Hours Ago,This week's most overbought names include Darden Restaurants and Target,https://www.cnbc.com/2026/01/17/this-weeks-most-overbought-names-include-darden-restaurants-and-target.html
```

!!!! The CNBC webpage contains certain elements, such as the market banner at the top of the page, that are dynamically rendered using JavaScript. Since this script uses the Requests library, which only retrieves the initial HTML response from the server and does not execute JavaScript, these dynamically generated elements are not present in the saved HTML file. The script therefore collects and saves the raw HTML content returned by the server.