

Lab 1

Daniel Li
2248244297

1. Installation and Setup

1.1. VMware

I followed the instruction downloading the VMware Fusion 13.6.2 from broadroom.com and installed it accordingly

VMware Fusion (for Intel-based and Apple silicon Macs)

←

13.6.2

Primary Downloads Open Source

13.6.2 52667


VMware Fusion (for Intel-based and Apple silicon Macs)	Release 13.6.2	Release Level Info 526671
--	----------------	---------------------------

1.2. Ubuntu Setup

- Download

I downloaded the lasted version of 24.04.1

Ubuntu 24.04.1 LTS



The latest [LTS](#) version of Ubuntu, for desktop PCs and laptops. LTS stands for long-term support — which means five years of free security and maintenance updates, extended up to 12 years with [Ubuntu Pro](#).

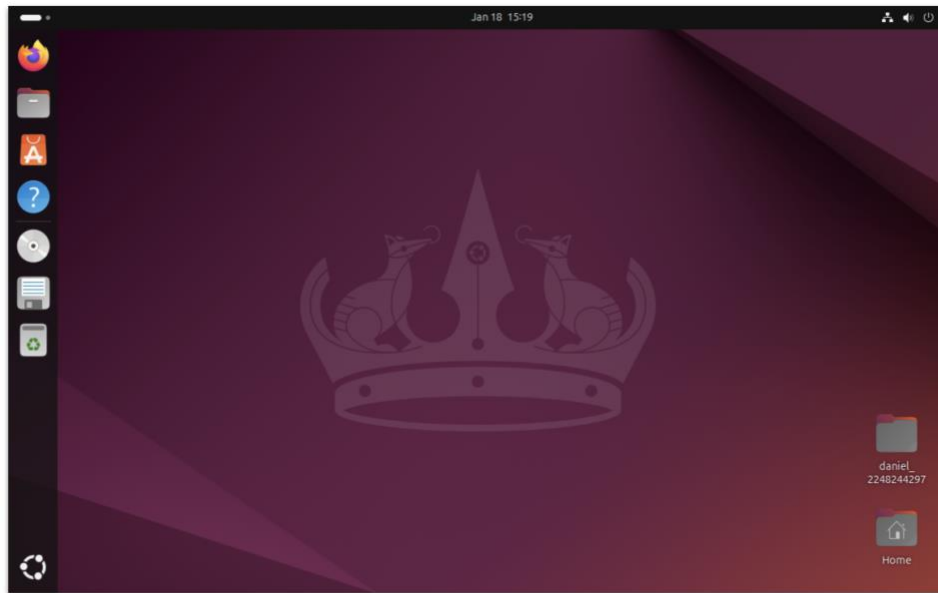
[Download 24.04.1 LTS](#) 5.8GB

For other versions of Ubuntu Desktop including torrents, the network installer, a list of local mirrors and past releases [check out our alternative downloads](#).

[What's new](#) [System requirements](#) [How to install](#)

- ✓ New Desktop installer with support for autoinstall
- ✓ New App Center and Firmware Updater applications
- ✓ GNOME 46 with support for quarter screen tiling
- ✓ Advanced Active Directory Group Policy Object support for Ubuntu Pro users
- ✓ Experimental support for TPM-backed Full Disc Encryption and ZFS encryption

- Setting up a virtual machine using Ubuntu
I am using 4 GB RAM and 20 GB disk space



Ubuntu 64-bit 24.04.1

Ubuntu 64-bit

[Add a note here](#)

2 Processor Cores ⓘ

4096 MB Memory



1.3. Python Installation

- Python 3 is automatically installed on Ubuntu

```
daniel@daniel-VMware-Virtual-Platform:~$ python3 - version
Python 3.12.3 (main, Nov  6 2024, 18:32:19) [GCC 13.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
```

- PIP installed

```
daniel@daniel-VMware-Virtual-Platform:~$ python3 --version
Python 3.12.3
daniel@daniel-VMware-Virtual-Platform:~$ pip --version
pip 24.0 from /usr/lib/python3/dist-packages/pip (python 3.12)
```

2. Get familiar with Python and Linux

2.1. Linux

Go to Desktop and create a directory (I already created before)

```
daniel@daniel-VMware-Virtual-Platform:~$ ls
Desktop Documents Downloads Music Pictures Public snap Templates Videos
daniel@daniel-VMware-Virtual-Platform:~$ cd Desktop
daniel@daniel-VMware-Virtual-Platform:~/Desktop$ mkdir daniel_2248244297
mkdir: cannot create directory 'daniel_2248244297': File exists
```

Make two directories named “scripts” and “data”

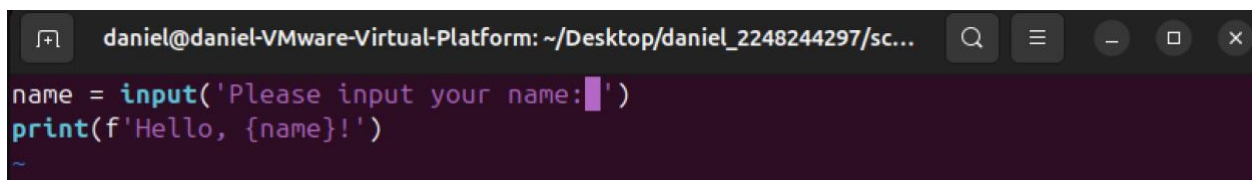
```
daniel@daniel-VMware-Virtual-Platform:~/Desktop$ ls
daniel_2248244297
daniel@daniel-VMware-Virtual-Platform:~/Desktop$ cd daniel_2248244297/
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297$ ls
data scripts
```

Create a file named “task_1.py”

```
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297$ ls
data scripts
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297$ cd scripts/
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ ls
task_1.py
```

2.2. Basic Python Script

I installed the vim to write python script below is the screenshot

A screenshot of a terminal window with a dark background. The title bar shows the path ~/Desktop/daniel_2248244297/sc... and standard window controls. The terminal content shows a Python script being edited in vim: name = input('Please input your name:') followed by print(f'Hello, {name}!'). A tilde symbol is at the bottom of the screen.

```
daniel@daniel-VMware-Virtual-Platform: ~/Desktop/daniel_2248244297/sc...
name = input('Please input your name:')
print(f'Hello, {name}!')
~
```

Run it

```
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ python3 task_1.py
Please input your name: Daniel
Hello, Daniel!
```

2.3. Web-scraping

- Inspecting the html

Finding that the market banner is in <div> MarketsBanner-main

The screenshot shows the CNBC MarketsBanner-main with a table of market indices and their performance. Below the table, a headline reads: "Dow surges more than 300 points, S&P 500 posts best week since period following Trump's election". The HTML structure is displayed on the right, showing the <div class="MarketsBanner-main"> container and its sub-elements, including market data cards for DJIA, S&P 500, NASDAQ, RUSSELL 2000, and VIX.

Index	Value	Change	% Change
DJIA	43,487.83	+334.70	+0.78%
S&P 500	5,996.66	+59.32	+1.00%
NASDAQ	19,630.20	+291.91	+1.51%
RUSSELL 2000	2,275.88	+9.09	+0.40%
VIX	15.97	-0.63	-3.80%

```
<div class="MarketsBanner-main">
  <div id="market-data-scroll-container" class="MarketsBanner-marketData">
    <a href="//www.cnbc.com/quotes/.DJIA" class="MarketCard-container MarketCard-up MarketCard-wrap">
    <a href="//www.cnbc.com/quotes/.SPX" class="MarketCard-container MarketCard-up">
    <a href="//www.cnbc.com/quotes/.IXIC" class="MarketCard-container MarketCard-up MarketCard-wrap">
    <a href="//www.cnbc.com/quotes/.RUT" class="MarketCard-container MarketCard-up">
    <a href="//www.cnbc.com/quotes/.VIX" class="MarketCard-container MarketCard-down">
  <div class="MarketsBanner-ad">
</div>
```

The Latest News is in <div> LatestNews-isHomePage LatestNews-isIntlHomepage

The screenshot shows the CNBC Latest News section with a list of news items. The HTML structure is displayed on the right, showing the <div class="LatestNews-isHomePage LatestNews-isIntlHomepage"> container and its sub-elements, including a header and a list of news items.

LATEST NEWS

- falls
- 11 HOURS AGO
What to expect from travel prices in 2025, and which spots have the best deals
- 12 HOURS AGO
Consumer protection agencies at risk in Trump's second term: What it means for you
- 13 HOURS AGO
Why the gold boom is causing a surge in illegal mining
- 13 HOURS AGO
Google Maps is turning 20 — mapping more countries and adding AI capabilities

```
<div class="LatestNews-isHomePage LatestNews-isIntlHomepage" data-test="latestNews-0" data-analytics="HomePageInternational-latestNews-7-0">
  <header class="LatestNews-header">
  <ul class="LatestNews-list">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-0">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-1">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-2">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-3">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-4">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-5">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-6">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-7">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-8">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-9">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-10">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-11">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-12">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-13">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-14">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-15">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-16">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-17">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-18">
    <li class="LatestNews-item" id="HomePageInternational-latestNews-7-19">
  </ul>
</div>
```

- Dynamic Scraping

Due to the fact that the page contains Javascript that makes the html loads contents subsequent to entering the page. I choose to use selenium with chrome webdriver.

a), install chromium

```
daniel@daniel-VMware-Virtual-Platform:~$ sudo apt install -y chromium-browser
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  chromium-browser
0 upgraded, 1 newly installed, 0 to remove and 135 not upgraded.
Need to get 50.0 kB of archives.
After this operation, 105 kB of additional disk space will be used.
Get:1 http://us.archive.ubuntu.com/ubuntu noble/universe amd64 chromium-browser
amd64 2:1snap1-0ubuntu2 [50.0 kB]
Fetched 50.0 kB in 1s (59.9 kB/s)
Preconfiguring packages ...
Selecting previously unselected package chromium-browser.
(Reading database ... 156833 files and directories currently installed.)
```

```
daniel@daniel-VMware-Virtual-Platform:~$ chromium-browser --version
Chromium 132.0.6834.83 snap
```

b), install chromedriver

```
daniel@daniel-VMware-Virtual-Platform:~$ sudo apt install -y chromium-chromedr
er
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  chromium-chromedriver
0 upgraded, 1 newly installed, 0 to remove and 135 not upgraded.
Need to get 2,308 B of archives.
After this operation, 18.4 kB of additional disk space will be used.
Get:1 http://us.archive.ubuntu.com/ubuntu noble/universe amd64 chromium-chromedr
```

```
daniel@daniel-VMware-Virtual-Platform:~$ /usr/bin/chromedriver --version
ChromeDriver 132.0.6834.83 (03d59cf5ecf1d8444838ff9a1e96231304d4ff9c-refs/branch
-heads/6834@{#3390})
```

c), Install packages

Firstly, I create a venv to hold this lab

```
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ pytho
n3 -m venv myenv
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ sourc
e myenv/bin/activate
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scip
daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ pytho
n3 -m venv myenv
```

Second, use pip to install package

```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scip
ts$ pip install beautifulsoup4
```

```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scip
ts$ pip install selenium
```

d), code

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from bs4 import BeautifulSoup
import time

# set up ChromeDriver and Chromium
options = Options()
options.add_argument("--headless")
service = Service('/usr/bin/chromedriver')

# WebDriver
driver = webdriver.Chrome(service=service, options=options)

# Load
driver.get("https://www.cnn.com/world/?region=world")
time.sleep(5)

# get page
page_source = driver.page_source

soup = BeautifulSoup(page_source, 'html.parser')
```

```
if market_banner and latest_news:
    with open("../data/raw_data/web_data.html", "w", encoding="utf-8") as file:
        file.write(str(market_banner))
        file.write(str(latest_news))
    print("raw_data.html saved successfully!")
else:
    print("No matching element found.")

driver.quit()
```

e) run it


```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ python3 web_scrape.py
raw_data.html saved successfully!
```

f) test it

```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/data/raw_data$ head -n 10 web_data.html
<div class="MarketsBanner-marketData" id="market-data-scroll-container"><a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.DJI"><div class="MarketCard-row"><span class="MarketCard-symbol">DJIA</span><span class="MarketCard-stockPosition">43,487.83</span></div><div class="MarketCard-row"><span aria-hidden="true" class="MarketCard-triangle-up"></span><div class="MarketCard-changeData"><span class="MarketCard-changesPts">+334.70</span><span class="MarketCard-changesPct">+0.78%</span></div></div><div class="MarketCard-row"><span class="MarketCard-lastTime">LAST | 1/17/25 EST</span></div></a><a class="MarketCard-container MarketCard-up" href="//www.cnbc.com/quotes/.SPX"><div class="MarketCard-row"><span class="MarketCard-symbol">S&amp;P 500</span><span class="MarketCard-stockPosition">5,996.66</span></div><div class="MarketCard-row"><span aria-hidden="true" class="MarketCard-triangle-up"></span><div class="MarketCard-changeData"><span class="MarketCard-changesPts">+59.32</span><span class="MarketCard-changesPct">+1.00%</span></div></div><div class="MarketCard-row"><span class="MarketCard-lastTime">LAST | 1/17/25 EST</span></div></a><a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.IXIC"><div class="MarketCard-row"><span class="MarketCard-symbol">NASDAQ</span><span class="MarketCard-stockPosition">19,630.20</span></div><div class="MarketCard-row"><span aria-hidden="true" class="MarketCard-triangle-up"></span><div class="MarketCard-changeData"><span class="MarketCard-changesPts">+291.91</span><span class="MarketCard-changesPct">+1.51%</span></div></div><div class="
```

2.4. Data Filtering

```
import os
import csv
from bs4 import BeautifulSoup

# Create a directory for processed data
output_dir = "../data/processed_data"
if not os.path.exists(output_dir):
    os.makedirs(output_dir)

# Read the web_data.html file
input_file = "../data/raw_data/web_data.html"
if not os.path.exists(input_file):
    print(f"Error: {input_file} not found.")
    exit()

with open(input_file, "r", encoding="utf-8") as file:
    soup = BeautifulSoup(file, "html.parser")

# Filter and extract market banner data
print("Filtering Market Banner data...")
market_data = []
market_cards = soup.find_all("a", class_="MarketCard-container")
```

```

for card in market_cards:
    symbol = card.find("span", class_="MarketCard-symbol").text.strip() if card.find("span", class_="MarketCard-symbol") else None
    stock_position = card.find("span", class_="MarketCard-stockPosition").text.strip() if card.find("span", class_="MarketCard-stockPosition") else None
    change_pct = card.find("span", class_="MarketCard-changesPct").text.strip() if card.find("span", class_="MarketCard-changesPct") else None
    if symbol and stock_position and change_pct:
        market_data.append([symbol, stock_position, change_pct])

if market_data:
    market_data_file = os.path.join(output_dir, "market_data.csv")
    with open(market_data_file, "w", encoding="utf-8", newline="") as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(["Symbol", "Stock Position", "Change Percentage"])
        writer.writerows(market_data)
    print(f"Market data stored in {market_data_file}")
else:
    print("No market data found.")

```

```

# Filter and extract Latest News data
print("Filtering Latest News data...")
news_data = []
news_items = soup.find_all("li", class_="LatestNews-item")

for item in news_items:
    timestamp = item.find("time", class_="LatestNews-timestamp").text.strip() if item.find("time", class_="LatestNews-timestamp") else None
    title = item.find("a", class_="LatestNews-headline").text.strip() if item.find("a", class_="LatestNews-headline") else None
    link = item.find("a", class_="LatestNews-headline")["href"] if item.find("a", class_="LatestNews-headline") else None
    if timestamp and title and link:
        news_data.append([timestamp, title, link])

if news_data:
    news_data_file = os.path.join(output_dir, "news_data.csv")
    with open(news_data_file, "w", encoding="utf-8", newline="") as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(["Timestamp", "Title", "Link"])
        writer.writerows(news_data)
    print(f"News data stored in {news_data_file}")
else:
    print("No news data found.")

print("Processing complete.")

```

- Run it

```

(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/scripts$ python3 data_filter.py
Filtering Market Banner data...
Market data stored in ../data/processed_data/market_data.csv
Filtering Latest News data...
News data stored in ../data/processed_data/news_data.csv
Processing complete.

```


Check the output

```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/data/processed_data$ cat market_data.csv
Symbol,Stock Position,Change Percentage
DJIA,"43,487.83",+0.78%
S&P 500,"5,996.66",+1.00%
NASDAQ,"19,630.20",+1.51%
RUSS 2K*,"2,275.88",+0.40%
VIX,15.97,-3.80%
```

```
(myvenv) daniel@daniel-VMware-Virtual-Platform:~/Desktop/daniel_2248244297/data/processed_data$ cat news_data.csv
Timestamp,Title,Link
3 Hours Ago,"Apple, Google remove TikTok from stores as app halts service in U.S.",https://www.cnbc.com/2025/01/18/apple-google-remove-tiktok-from-stores-as-app-halts-service-in-us.html
11 Hours Ago,Perplexity AI makes a bid to merge with TikTok U.S.,https://www.cnbc.com/2025/01/18/perplexity-ai-makes-a-bid-to-merge-with-tiktok-us.html
14 Hours Ago,"Solana surges 12% on launch of Trump-themed meme coin, ether falls",https://www.cnbc.com/2025/01/18/crypto-market-today.html
15 Hours Ago,"What to expect from travel prices in 2025, and which spots have the best deals",https://www.cnbc.com/2025/01/18/what-to-expect-from-travel-prices-in-2025.html
16 Hours Ago,Consumer protection agencies at risk in Trump's second term: What it means for you,https://www.cnbc.com/2025/01/18/how-trumps-second-term-could-mean-the-downfall-of-the-fdic-cfpb.html
17 Hours Ago,Why the gold boom is causing a surge in illegal mining,https://www.cnbc.com/2025/01/18/why-the-gold-boom-is-causing-a-surge-in-illegal-mining.html
```