

Lab 1 Report

Name: Sen Pang

USC ID: 8598139533

Operating System: Ubuntu 24.04.3 LTS (ARM64)

Environment: VMware Fusion on macOS

GitHub Link: <https://github.com/senpang-coder/26SP-DSCI560>

YouTube Link: <https://youtu.be/wCh3s8ZFVM0>

1. Introduction

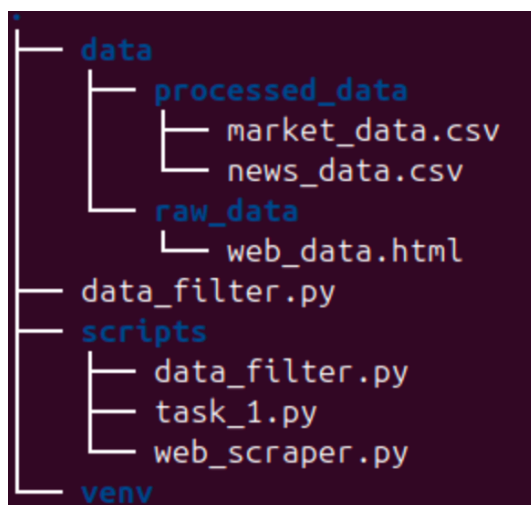
This assignment focuses on gaining hands-on experience with Linux terminal commands and basic Python scripting. The tasks include directory management, script creation, user input handling, and simple web data processing.

2. Playing Around with Linux Terminal

The Linux terminal was opened to perform basic file system operations.

A project directory `SenPang_8598139533` was created on the Desktop, with `data` and `scripts` as subdirectories. An empty Python file `task_1.py` was added to the `scripts` folder. Directory structure was verified using commands such as `ls` and `tree`.

Screenshot 1: Directory Structure Creation



3. A Basic Python Script

The task_1.py file was opened in a terminal editor (nano). A short Python script was added to prompt the user for their name and print a greeting message..

The script was executed successfully from the terminal and displayed the correct output.

Screenshot 2: Python Script Execution

```
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533$ nano scripts/task_1.py
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533$ python3 scripts/task_1.py
Please enter your name: sen
Hello, sen!
```

4. Web Scraping and Data Processing

A Python script fetched a webpage and saved the HTML content locally, returning an HTTP 200 status code. The raw HTML file was stored in `data/raw` data.

A second script parsed the HTML and extracted the required fields into CSV files saved in data/processed data.

Screenshot 3: Web Scraper Running Successfully

```
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533/scripts$ python3 web_scraper.py
Fetching webpage...
HTTP status code: 200
Saved HTML to: ../data/raw_data/web_data.html
Found 'Latest News' text?: True
```

Screenshot 4: Printing first 10 lines of HTML file

```
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533/scripts$ head -n 10 ../data/raw_data/web_data.html
<!DOCTYPE html><html lang="en" prefix="og=https://ogp.me/ns#" itemscope="" itemtype="https://schema.org/WebPage"><head><meta property="og:type" content="website"/><meta property="og:title" content="International: Top News And Analysis"/><meta property="og:description" content="CNBC International is the world leader for news on business, technology, China, trade, oil prices, the Middle East and markets."/><meta property="og:url" content="https://www.cnn.com/world/"><meta property="og:site_name" content="CNBC"/><meta name="robots" content="max-image-preview:large"/><meta name="format-detection" content="telephone=no"/><meta property="og:image" itemprop="image" content="https://fm.cnn.com/applications/cnn.com/staticcontent/img/versant/cnn_share_versant.png?v=1524171804&w=1920&h=1080"/>
```

Screenshot 5: Data Filtering and CSV Generation

```
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533/scripts$ python3 data_fetcher.py
Reading HTML: ../data/raw_data/web_data.html
Filtering Market banner fields...
Market rows found: 0
Market CSV created: ../data/processed_data/market_data.csv
Filtering Latest News fields...
News rows found: 30
News CSV created: ../data/processed_data/news_data.csv
Done.
```

The market banner returned zero rows due to dynamic content loading on the CNBC webpage at runtime.

Screenshot 6: Generated Files Verification

```
(venv) sen@sen-VMware20-1:~/Desktop/SenPang_8598139533/scripts$ ls ../data/processed_data
market_data.csv  news_data.csv
```

5. Conclusion

This assignment offered practical exposure to Linux terminal commands, Python scripting, and basic data processing workflows. Completing the tasks improved my familiarity with the Linux environment and strengthened my ability to write Python programs for automation and data handling.