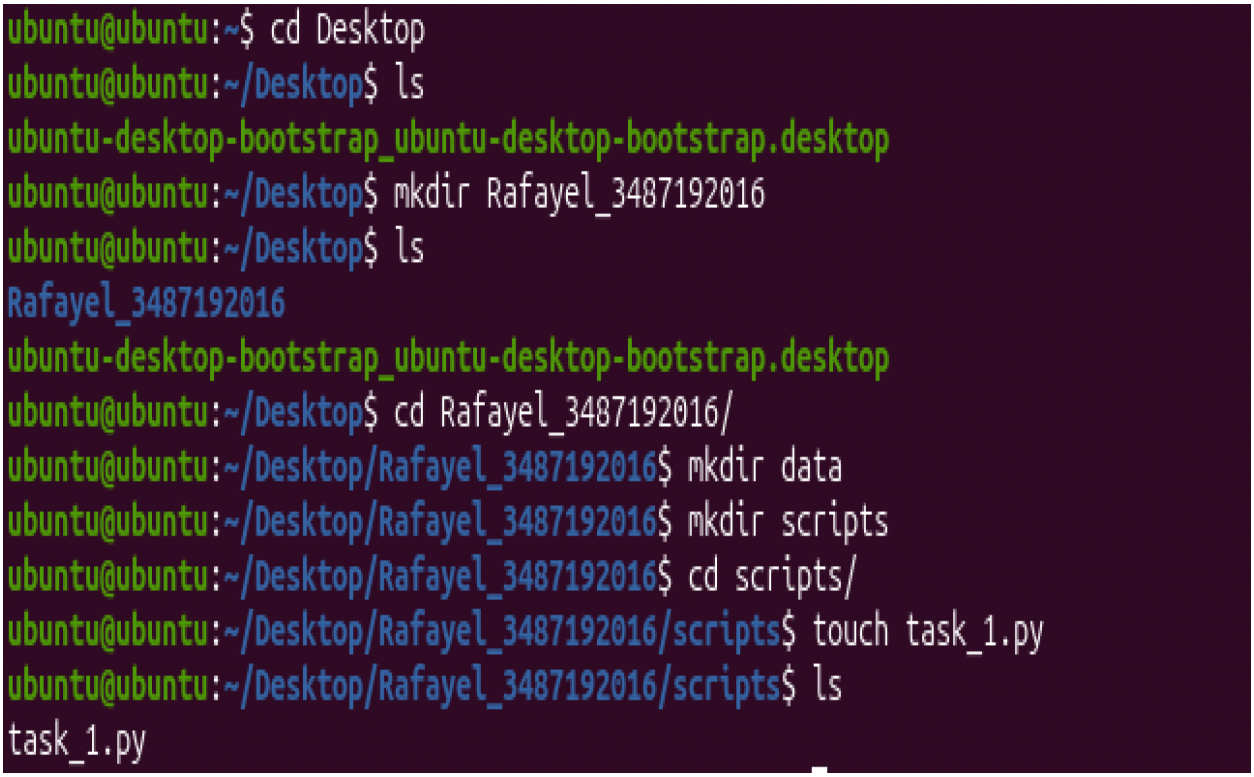
Web Scraping and Data Extraction Report

Author: Rafayel
Environment: Ubuntu Linux
Language: Python 3

This report documents the complete workflow of a web scraping project, including environment setup, script execution, data extraction, and validation. Screenshots are included to demonstrate successful execution.

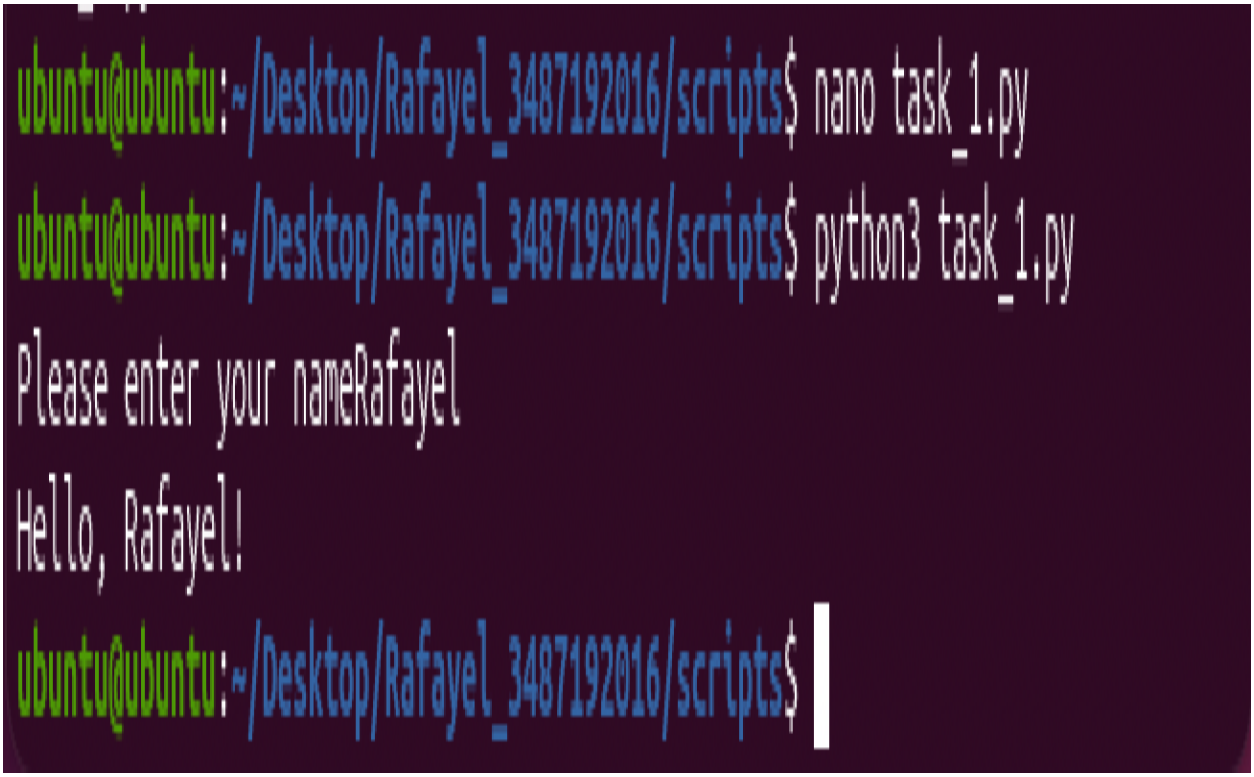
1. Project Directory Setup

This screenshot shows the creation of the project directory and subfolders using Linux terminal commands. It confirms proper navigation, folder creation, and file initialization following best practices.



2. Python Script Creation and Execution

This screenshot demonstrates editing and running a Python script (task_1.py). The successful output confirms that the Python environment is correctly configured.



3. Web Scraper Execution and HTML Storage

This screenshot shows the execution of the web scraper script. A successful HTTP request is made and raw HTML data is saved locally for further processing.

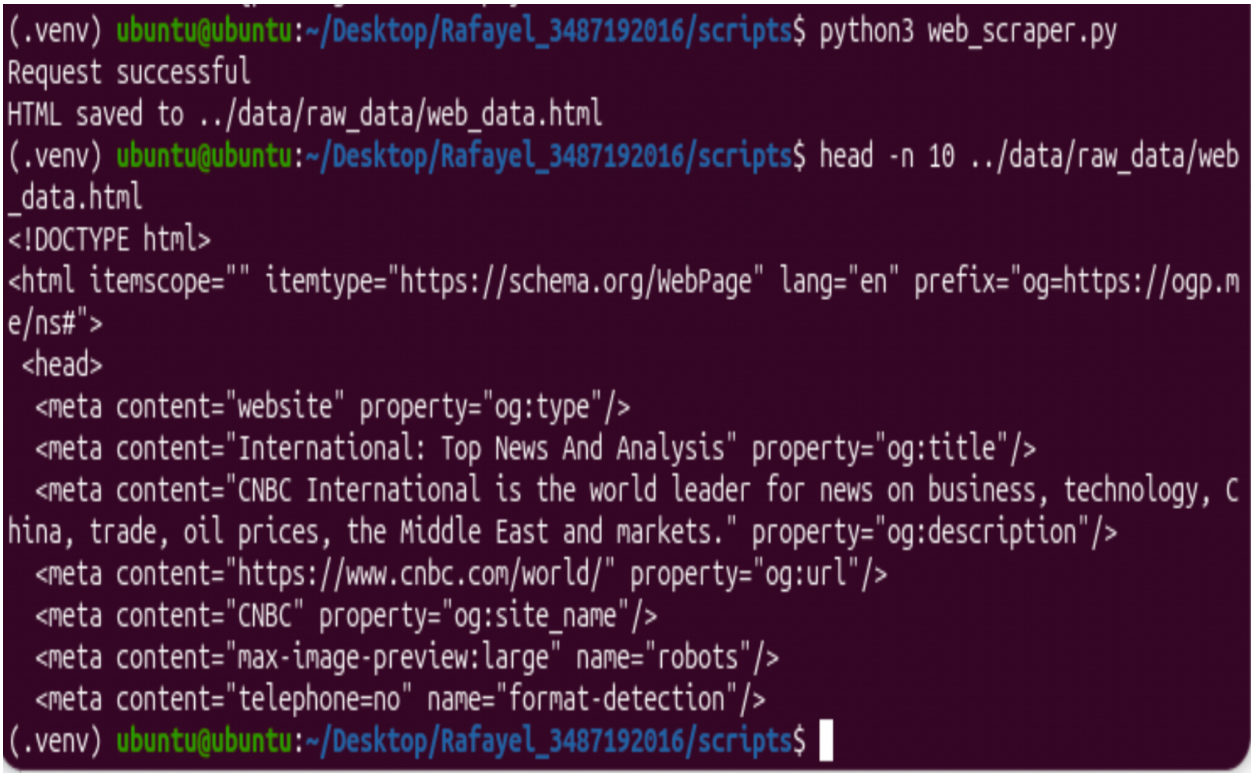
4. HTML Content Verification

This screenshot verifies that the downloaded HTML file contains valid webpage content. The presence of metadata and page structure confirms successful data retrieval.

```
(.venv) ubuntu@ubuntu:~/Desktop/Rafayel_3487192016/scripts$ pip install requests
Collecting requests
  Downloading requests-2.32.5-py3-none-any.whl.metadata (4.9 kB)
Collecting charset_normalizer<4,>=2 (from requests)
  Downloading charset_normalizer-3.4.4-cp313-cp313-manylinux2014_aarch64.manylinux_2_17_aarch64.manylinux_2_
28_aarch64.whl.metadata (37 kB)
Collecting idna<4,>=2.5 (from requests)
  Downloading idna-3.11-py3-none-any.whl.metadata (8.4 kB)
Collecting urllib3<3,>=1.21.1 (from requests)
  Downloading urllib3-2.6.3-py3-none-any.whl.metadata (6.9 kB)
Collecting certifi>=2017.4.17 (from requests)
  Downloading certifi-2026.1.4-py3-none-any.whl.metadata (2.5 kB)
Downloading requests-2.32.5-py3-none-any.whl (64 kB)
Downloading charset_normalizer-3.4.4-cp313-cp313-manylinux2014_aarch64.manylinux_2_17_aarch64.manylinux_2_28
_aarch64.whl (147 kB)
Downloading idna-3.11-py3-none-any.whl (71 kB)
Downloading urllib3-2.6.3-py3-none-any.whl (131 kB)
Downloading certifi-2026.1.4-py3-none-any.whl (152 kB)
Installing collected packages: urllib3, idna, charset_normalizer, certifi, requests
Successfully installed certifi-2026.1.4 charset_normalizer-3.4.4 idna-3.11 requests-2.32.5 urllib3-2.6.3
(.venv) ubuntu@ubuntu:~/Desktop/Rafayel_3487192016/scripts$ pip install beautifulsoup4
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.14.3-py3-none-any.whl.metadata (3.8 kB)
Collecting soupsieve>=1.6.1 (from beautifulsoup4)
  Downloading soupsieve-2.8.1-py3-none-any.whl.metadata (4.6 kB)
Collecting typing-extensions>=4.0.0 (from beautifulsoup4)
  Downloading typing_extensions-4.15.0-py3-none-any.whl.metadata (3.3 kB)
Downloading beautifulsoup4-4.14.3-py3-none-any.whl (107 kB)
Downloading soupsieve-2.8.1-py3-none-any.whl (36 kB)
Downloading typing_extensions-4.15.0-py3-none-any.whl (44 kB)
Installing collected packages: typing-extensions, soupsieve, beautifulsoup4
Successfully installed beautifulsoup4-4.14.3 soupsieve-2.8.1 typing-extensions-4.15.0
```

5. Data Extraction and CSV Generation

This screenshot shows the parsing process using BeautifulSoup and the creation of structured CSV files. Both market and news data are extracted successfully.



6. News Data Validation

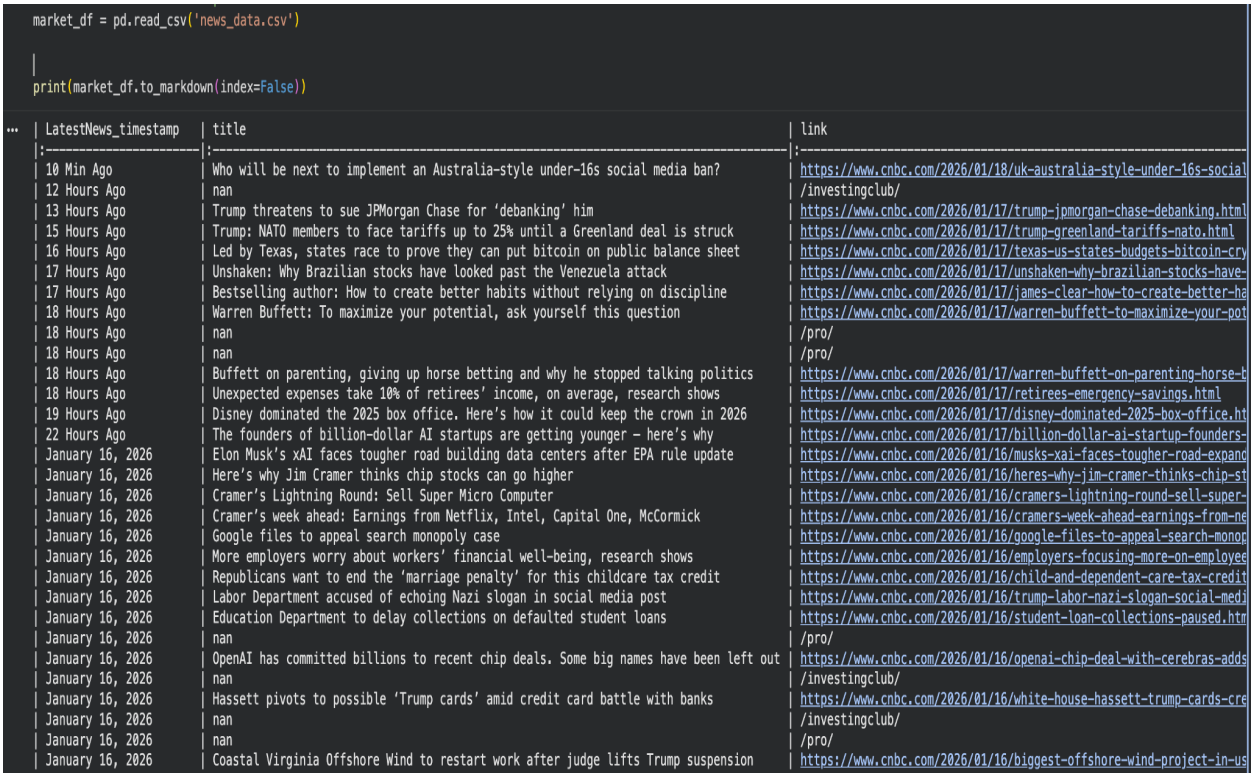
This screenshot displays the news_data.csv file loaded into pandas. It confirms accurate extraction of timestamps, headlines, and links.

```
        print(f"Storing data into {filepath}...")
        with open(filepath, "w", newline="", encoding="utf-8") as f:
            writer = csv.DictWriter(f, fieldnames=data[0].keys())
            writer.writeheader()
            writer.writerows(data)

        print("CSV created successfully")


    def main():
        html = read_html_file(HTML_PATH)
        soup = BeautifulSoup(html, "html.parser")

        market_data = extract_market_data(soup)
        news_data = extract_latest_news(soup)

        write_csv(MARKET_CSV_PATH, market_data)
        write_csv(NEWS_CSV_PATH, news_data)

        print("Data filtering and storage completed.")


    if __name__ == "__main__":
        main()
```

```
...  Reading raw HTML data...
     Filtering market banner fields...
     Extracted 5 market records
     Filtering latest news fields...
     Extracted 30 news records
     Storing data into market_data.csv...
     CSV created successfully
     Storing data into news_data.csv...
     CSV created successfully
     Data filtering and storage completed.
```

7. Market Data Validation

This screenshot shows market_data.csv loaded into pandas. The table confirms correct extraction of market indices, values, and percentage changes.



```
market_df = pd.read_csv('news_data.csv')

print(market_df.to_markdown(index=False))
```

LatestNews_timestamp	title	link
10 Min Ago	Who will be next to implement an Australia-style under-16s social media ban?	https://www.cnbc.com/2026/01/18/uk-australia-style-under-16s-social
12 Hours Ago	nan	/investingclub/
13 Hours Ago	Trump threatens to sue JPMorgan Chase for 'debanking' him	https://www.cnbc.com/2026/01/17/trump-jpmorgan-chase-debanking.html
15 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
16 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-cry
17 Hours Ago	Unshaken: Why Brazilian stocks have looked past the Venezuela attack	https://www.cnbc.com/2026/01/17/unshaken-why-brazilian-stocks-have-
17 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-ha
18 Hours Ago	Warren Buffett: To maximize your potential, ask yourself this question	https://www.cnbc.com/2026/01/17/warren-buffett-to-maximize-your-pot
18 Hours Ago	nan	/pro/
18 Hours Ago	nan	/pro/
18 Hours Ago	Buffett on parenting, giving up horse betting and why he stopped talking politics	https://www.cnbc.com/2026/01/17/warren-buffett-on-parenting-horse-b
18 Hours Ago	Unexpected expenses take 10% of retirees' income, on average, research shows	https://www.cnbc.com/2026/01/17/retirees-emergency-savings.html
19 Hours Ago	Disney dominated the 2025 box office. Here's how it could keep the crown in 2026	https://www.cnbc.com/2026/01/17/disney-dominated-2025-box-office.ht
22 Hours Ago	The founders of billion-dollar AI startups are getting younger — here's why	https://www.cnbc.com/2026/01/17/billion-dollar-ai-startup-founders-
January 16, 2026	Elon Musk's xAI faces tougher road building data centers after EPA rule update	https://www.cnbc.com/2026/01/16/musks-xai-faces-tougher-road-expanc
January 16, 2026	Here's why Jim Cramer thinks chip stocks can go higher	https://www.cnbc.com/2026/01/16/heres-why-jim-cramer-thinks-chip-st
January 16, 2026	Cramer's Lightning Round: Sell Super Micro Computer	https://www.cnbc.com/2026/01/16/cramers-lightning-round-sell-super-
January 16, 2026	Cramer's week ahead: Earnings from Netflix, Intel, Capital One, McCormick	https://www.cnbc.com/2026/01/16/cramers-week-ahead-earnings-from-ne
January 16, 2026	Google files to appeal search monopoly case	https://www.cnbc.com/2026/01/16/google-files-to-appeal-search-monop
January 16, 2026	More employers worry about workers' financial well-being, research shows	https://www.cnbc.com/2026/01/16/employers-focusing-more-on-employee
January 16, 2026	Republicans want to end the 'marriage penalty' for this childcare tax credit	https://www.cnbc.com/2026/01/16/child-and-dependent-care-tax-credit
January 16, 2026	Labor Department accused of echoing Nazi slogan in social media post	https://www.cnbc.com/2026/01/16/trump-labor-nazi-slogan-social-medi
January 16, 2026	Education Department to delay collections on defaulted student loans	https://www.cnbc.com/2026/01/16/student-loan-collections-paused.htm
January 16, 2026	nan	/pro/
January 16, 2026	OpenAI has committed billions to recent chip deals. Some big names have been left out	https://www.cnbc.com/2026/01/16/openai-chip-deal-with-cerebras-adds
January 16, 2026	nan	/investingclub/
January 16, 2026	Hassett pivots to possible 'Trump cards' amid credit card battle with banks	https://www.cnbc.com/2026/01/16/white-house-hassett-trump-cards-cre
January 16, 2026	nan	/investingclub/
January 16, 2026	nan	/pro/
January 16, 2026	Coastal Virginia Offshore Wind to restart work after judge lifts Trump suspension	https://www.cnbc.com/2026/01/16/biggest-offshore-wind-project-in-us

Conclusion

This project successfully demonstrates an end-to-end web scraping pipeline using Python. By capturing live web data, storing raw HTML, extracting structured information, and validating outputs, the project highlights practical data engineering and automation skills.