Challenge 11 – Data Collection and Web Scraping

Jeremy Hooper

Introduction

Data collection is the systematic process of gathering and measuring information from various sources. It is a crucial step in any data-driven project, providing the foundation for analysis and decision-making. Web scraping, a subset of data collection, automatically extracts data from websites. This technique utilises specialised software to navigate web pages, gather relevant data, and store it for further analysis. Web scraping enables researchers to access vast amounts of information efficiently, from product prices and market trends to social media sentiment and news articles. However, it's essential to consider ethical and legal implications, such as respecting website terms of service and privacy policies, when conducting web scraping activities. Overall, data collection and web scraping play integral roles in harnessing the power of data for insights and decision-making in various fields.

The project consists of two primary components: scraping data from a Martian news website and a Martian weather data website. The project aims to provide insights into Mars's current events and weather patterns by extracting information from these sources. This endeavour facilitates our understanding of Martian conditions and demonstrates the practical application of web scraping techniques in gathering data from unconventional sources for analysis and exploration.

Part 1: Scrape titles and Review Text from the Mars News Website

Part 1 of the project involves scraping data from the Mars News website using automated browsing techniques. By inspecting the webpage and utilising Beautiful Soup, we extract titles and preview text of news articles. This data is stored in Python dictionaries with keys for title and preview, then organised into a list. Optionally, the scraped data can be exported to a JSON file for sharing. This step-by-step process enables us to gather relevant information efficiently, laying the groundwork for further analysis and exploration of Martian news updates. The work went as follows:

- I developed an automated web scraping script to collect the latest news titles and preview texts from a Mars exploration news site. The script uses the Splinter library to navigate the website programmatically, simulating a user browsing experience. Once the target page was accessed, Beautiful Soup was employed to parse the HTML content, enabling the efficient extraction of news titles and their corresponding preview texts.

- The extracted information was meticulously organised into a structured format using a Python list comprised of dictionaries. Each dictionary in the list represents a single news article, with keys for the title and preview text, ensuring the data is accessible and easy to manipulate for future analysis or display purposes.

- Following successfully scraping and structuring the data, the Python list was then serialised into a JSON file, leveraging the **JSON** module. This step facilitates the storage and sharing of the scraped data and provides a standardised format for integrating with other applications or further processing and analysis. The resultant JSON file, **mars_news_scraped_data.json**, was stored in the **11Challenge directory**, completing the data acquisition phase of the project.
- This comprehensive approach to data collection not only automates the extraction process but also significantly enhances the efficiency and reliability of obtaining up-to-date information from the Mars exploration news site.
- All the work with comments is shown in **mars_code/mars_news.ipynb.**

Part 2: Scrape and Analyse Mars Weather Data

- In the second phase of our Mars exploration data project, we extended our data acquisition efforts to include structured data as an HTML table. This table was efficiently scraped and converted into a Pandas DataFrame, a process facilitated by the combined use of Pandas for direct HTML table parsing or Splinter and Beautiful Soup for more complex scraping scenarios. We ensured that the DataFrame was meticulously organised, with accurately labelled columns and appropriate data types for each column, laying a solid foundation for subsequent analysis.
- Our analytical objectives centred around gaining insights into Martian temporal metrics and atmospheric conditions. Specifically, we sought to determine the number of months on Mars, the total count of Martian days (sols) covered by our dataset, and to identify patterns in temperature and atmospheric pressure variations across different Martian months. Through careful analysis, we could ascertain that Mars has 12 months in a year, similar to Earth, but defined within the context of Martian solar orbits.
- To visually represent our findings, we created several data visualisations. These included:
- Bar charts to illustrate the average monthly temperatures and atmospheric pressures, allowing us to easily identify the months with the lowest and highest averages in each category.
- A line chart plotting the minimum temperatures against the number of terrestrial days, providing a visual estimate of the length of a Martian year in terrestrial days, aiming for accuracy within 25%.
- The culmination of our data manipulation and analysis efforts was the exportation of the refined DataFrame into a CSV file. This step preserved data and facilitated easy sharing and further analysis of the Mars weather data. Throughout this project, we adhered to best practices in data science, ensuring the integrity and reliability of our analyses and visualisations, thereby enhancing our understanding of the Martian environment through empirical data.
- All the work is show in **mars_code/mars_weather.ipynb.**

Conclusion

In conclusion, this project has successfully demonstrated the effectiveness of web scraping techniques in extracting valuable data from diverse online sources, particularly Martian news and weather websites. By meticulously navigating through web pages and utilising Beautiful Soup, we collected titles and preview text of news articles, providing insights into current events on Mars. Additionally, the project showcased the potential for automated browsing to streamline data collection processes and enhance efficiency in data-driven analyses. The organised storage of scraped data in Python dictionaries and lists further facilitated data management and sharing. Overall, this project highlights the importance of leveraging technological tools and methodologies to explore new frontiers of information gathering, ultimately contributing to our understanding of distant celestial bodies like Mars.

References

UWA Bootcamp Course Notes

Beautiful Soup 4.12.0 Documentation - https://www.crummy.com/software/BeautifulSoup/bs4/doc/

GifHub Pages – https://tedboy.github.io > bs4_doc

Data Analysis with Pandas and Python – Boris Paskhaver – Udemy – YouTube Tutorial

Python Pandas Data Science Tutorial – Keith Galli – YouToube

Chat GPT