HyperionDev

Web Scraping

21 January 2024

Tech Talks Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions.
- If you have any questions outside of this session, or that are not answered during this session, please do submit these for upcoming Tech Talks Sessions. You can submit these questions here: **https://forms.gle/MomSYvUWiSfKgMaZ9**

Tech Talks Session Housekeeping cont.

- For all **non-academic questions**, please submit a query:

  www.hyperiondev.com/support
- We would love your **feedback**. Please fill in the feedback form after the session.
- If you are hearing impaired, please kindly use your computer's function through Google chrome to enable captions.

Safeguarding & Welfare

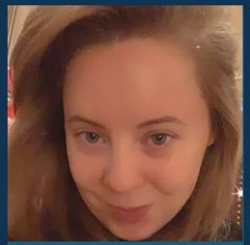
We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:

Ian Wyles
Designated Safeguarding Lead

Simone Botes

Rafiq Manan



Charlotte Witcher

Nurhaan Snyman

Ronald Munodawafa

Tevin Pitts

Scan to report a safeguarding concern

or email the Designated Safeguarding Lead:
Ian Wyles
safeguarding@hyperiondev.com

HyperionDev

Learning Outcomes

1. Understand the fundamentals of web scraping.

   Learn how to retrieve data from websites using HTTP requests and understand the role of tools like BeautifulSoup in parsing HTML content.

2. Extract and process targeted data from web pages.

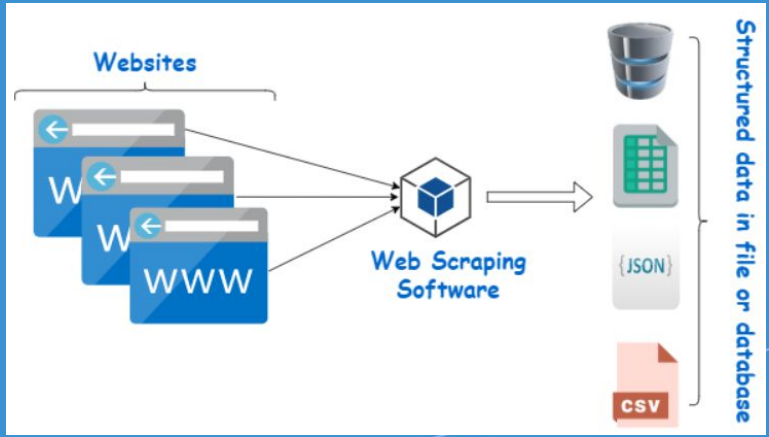
   Gain practical experience in locating and extracting specific elements (e.g., quotes, authors) from web pages using methods like .find, .find_all, and .get_text(strip=True).

3. Apply ethical and legal practices in web scraping:

   Understand how to responsibly use web scraping by adhering to website policies and avoiding misuse of data.

Introduction to Web Scraping

- Web scraping is the process of automatically extracting data from websites.
- Most websites do not allow you to save or download this data. If you need the data, the only option is to manually copy and paste the data - a very tedious job which can take many hours or days to complete.
- Web Scraping is the technique of automating this process, so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.



Source: https://www.webharvy.com/articles/what-is-web-scraping.html

Importance of Web Scraping

- Web scraping provides access to data that might not be available via APIs or datasets
- It can provide real-time or frequently updated data, which can be crucial for many types of analysis
- Scraped data can be used to train Machine Learning models for tasks like natural language processing (NLP) or image recognition

Source: https://www.mindbowser.com/how-does-web-scraping-work/

Why It's Useful

Data Scientists	Software Engineers	Web Developers
• Data collection: extracting stock market prices, analysing sentiment on social media (scraping reviews), building datasets for machine learning models. • Data extraction automation.	• Testing and automation: building backend services that aggregate data, automation of repetitive tasks such as extracting data from dashboards or monitoring websites. • Integrating web scraping with backend systems to streamline workflows	• Scraping design elements or layouts from other websites for inspiration. • Improving site performance or features. • Gathering data on competitors' websites for search engine optimisation.

HTML

- Before we get into any web scraping, it's important to know a few things about HTML (Hypertext Markup Language).
- Familiarity with HTML is crucial for effective web scraping because it defines the structure and content of web pages.
- This knowledge allows you to accurately identify, navigate, and extract data from websites and other online sources using web scraping tools.

Source:
https://www.toptal.com/python/web-scraping-with-python

HTML

- HTML consists of a **set of elements** (tags) that define the structure and content of a web page.
- Each tag represents a different structure such as headings, paragraphs, images, tables, etc.
- Tags are enclosed in angle brackets (`<>`) and they typically come in pairs e.g., `<p>` denotes the beginning of a paragraph and `</p>` denotes the end.
- HTML elements are nested within each other forming a **hierarchical** structure known as the Document Object Model (DOM).

```
1   <!DOCTYPE html>
2   <html>
3       <head>
4       </head>
5       <body>
6           <h1>My First Page</h1>
7           <p>This is my first page.</p>
8           <h2>A secondary header.</h2>
9           <p>Some more text.</p>
10      </body>
11  </html>
```

Ethics

When engaging in web scraping, it is essential to adhere to ethical guidelines:

- Websites often have a **robots.txt** file that specifies which pages can be crawled. It's ethical to abide by these rules.
- Ensure compliance with a website's terms of service regarding data usage and scraping.
- Scraping should be done responsibly to avoid overloading servers and causing disruption.

Source: https://techjury.net/blog/is-web-scraping-legal/

Questions and Answers

Questions around Web Scraping

Thank you!