Webscraping workshop

June 8th 2024, University of Bucharest

A set of minimal rules

- Duration: approximately 3 hours;
- Ask questions/clarifications at any time during the workshop;
- Enjoy!
- Materials available at: https://github.com/MarianNecula/UnibucWebscraping.git

Outline

- HTML
- DevTools
- CSS selectors
- XPATH selectors
- HTTP/HTTPS Verbs
- About storing collected data
- About logging
- About some legal and ethical aspects
- Practical application 1.
- Practical application 2.
- Think about your own project.
- Wrap-up and final remarks.

Introduction to HTML

- What is HTML?

- HTML (HyperText Markup Language) is the standard language for creating web pages.

- It describes the structure of web (content) pages using markup.

- We get static (easy) and dynamic (harder) content.

- History:

Developed by Tim Berners-Lee in 1991.
Continuously evolving; currently at HTML5.

Basic Structure of an HTML Document

```
<!DOCTYPE html>
<html>
  <head>
    <title>Page Title</title>
  </head>
  <body>
    <h1>This is a Heading</h1>
    <p>This is a paragraph.</p>
  </body>
</html>
```

- <!DOCTYPE html>: Declares the document type.
- <html>: Root element.<head>: Contains meta-information about the document.
- <title>: Specifies the title of the document.
- <body>: Contains the content of the document.

HTML Elements and Tags

- Elements:

Consist of a start tag, content, and an end tag.
Example: This is a paragraph.

- Tags:

Tags are the building blocks of HTML.
Example: <h1>, , <a>

Common HTML Tags

- Headings:

<h1> to <h6>: Define headings, <h1> being the highest and <h6> the lowest.

- Paragraph:

<p>: Defines a paragraph.

- Links:

<a href="url">: Defines a hyperlink.

- Images:

<img src="url" alt="description">: Embeds an image.

- Lists:

: Unordered list, <ol>: Ordered list, <li>: List item.

HTML Attributes

- What are Attributes?

Provide additional information about elements.Always included in the opening tag.

Common Attributes:

- href: Specifies the URL for a link.
- src: Specifies the URL for an image.
- alt: Provides alternative text for an image.
- class: Assigns one or more class names for CSS styling.
- id: Specifies a unique id for an element.

HTML Forms

Used to collect user input.

Form Elements:

- <form>: Container for form elements.
- <input>: Defines an input field.
- <label>: Defines a label for an input element.
- <button>: Defines a clickable button.

HTML Tables

- <table>
- <tr>
- <th>Header 1</th>
- <th>Header 2</th>
- </tr>
- <tr>
- <td>Data 1</td>
- <td>Data 2</td>
- </tr>
- </table>

Table Elements:

- <table>: Defines a table.
- <tr>: Defines a table row.
- <td>: Defines a table cell.
- <th>: Defines a table header.

HTML Semantic Elements

- Definition:Elements that clearly describe their meaning in a human- and machine-readable way.

Examples:

- <header>: Defines a header for a document or section.<nav>: Defines a set of navigation links.
- <section>: Defines a section in a document.
- <article>: Defines an independent, self-contained content.
- <footer>: Defines a footer for a document or section.

DevTools

- Developer tools, commonly referred to as DevTools, are built into modern web browsers and provide a powerful set of utilities for web development and debugging.

- Press F12 key when using any mainstream modern browser.

CSS selectors

- CSS (Cascade Style Sheets) selectors are used to target HTML elements based on their attributes and relationships.
- Essential for extracting specific data from web pages.
- Some examples with the rvest package to navigate and extract data.

CSS Selectors

Selector	Example	Example description
.class	.intro	Selects all elements with class="intro"
.class1.class2	.name1.name2	Selects all elements with both *name1* and *name2* set within its class attribute
.class1 .class2	.name1 .name2	Selects all elements with *name2* that is a descendant of an element with *name1*
#id	#firstname	Selects the element with id="firstname"
*	*	Selects all elements
element	p	Selects all <p> elements

Source W3Schools.com
https://www.w3schools.com/cssref/css_selectors.php

Xpath selectors

- XPath (XML Path Language) is a query language for selecting nodes from an XML document, which can also be used to navigate HTML documents.

- XPath expressions use a path-like syntax to navigate through the document.Example: //tagname[@attribute='value’]

- //tagname: Selects all elements with the specified tag name.
- //tagname[@attribute='value']: Selects all elements with the specified tag name and attribute value.
- //tagname[text()]: Selects all elements with the specified tag name containing text.
- //tagname[contains(text(),'substring')]: Selects all elements with the specified tag name containing the specified substring in the text.

Xpath selectors

Expression	Description
nodename	Selects all nodes with the name "*nodename*"
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

Source W3Schools.com
https://www.w3schools.com/xml/xpath_syntax.asp

Introduction to HTTP/HTTPS Verbs

- HTTP(S) (HyperText Transfer Protocol Secure) verbs (or methods) are standardized methods used by web clients to communicate with web servers.
- They define the type of action to be performed on the server.
- GET, POST, PUT, DELETE, PATCH, HEAD, OPTIONS

GET Verb

- Retrieve data from the server.
- Does not modify the resource.
- Multiple identical requests have the same effect as a single request.

POST Verb

- Submit data to be processed to the server.
- Modifies the resource.
- Multiple identical requests can have different effects.
- Responses are not typically cached.

PUT and DELETE Verbs

- Update/create or delete a resource on the server.
- Multiple identical requests have the same effect as a single request.
- Modifies the resource.

About storing collected data

- Most common type of storage – text files.
- But for larger projects a database storage system is recommended.
- SQL databases: MySQL, sqlite, etc.
- NoSQL databases: MongoDB, Couchbase, etc

About logging

- Logging represent the action of recording each step from a process.
- Debugging: Helps identify and fix issues in the scraping process.
- Monitoring: Tracks the progress and performance of the scraper.
- Error Handling: Captures errors and exceptions for later analysis.
- Compliance: Maintains records of requests and responses for legal or ethical compliance.

About some legal and ethical aspects

- Legal Considerations:
  - Terms of Service of websites
  - Copyright and data privacy laws

- Ethical Practices:
  - Respecting website terms (check robots.txt)
  - Avoiding overload on servers

Practical application 1.

- Romanian MPs activity

Practical application 2

- Pupils results at Romanian National Evaluation 2023.

Think about your own project.

Wrap-up and final remarks.

- Webscraping is a useful tool for data collection.
- Data selectivity!!! (representativity???)
- No need to do yourself the webscraping (can be very cumbersome.)

Thank you!