🧑‍💻 Lesson 2: Understanding HTML for Web Scraping

Welcome back! In this second lesson of our web scraping series, we dive into a foundational topic: **HTML** — the language behind every web page. Understanding HTML is *essential* if you want to scrape data from websites accurately and efficiently.

🌐 Why Learn HTML?

When scraping a website, you're pulling data from the structure of its web pages. Since websites are built using HTML (HyperText Markup Language), knowing how to read this structure helps you find and extract exactly what you need.

🔍 Basic Structure of HTML

HTML is made up of elements enclosed in **angle brackets**. Here's a simplified breakdown:

- `<html>` ... `</html>`: The root of the page, wrapping all content.
- `<head>` ... `</head>`: Contains metadata and titles.
- `<body>` ... `</body>`: Holds the visible content of the page.
- Tags like `<p>` (paragraph), `<title>`, `<a>` (links), `<table>`, `<tr>`, `<td>` are commonly used.
- Tags are often paired — an opening tag and a closing tag (with a `/` in front).

Inside these tags, you'll find:

- **Text content** – the actual data we might want to scrape.
- **Attributes** – such as `class`, `id`, and `href`, which help us target specific elements.

🖱️ Inspecting a Real Web Page

The lesson then moves to a hands-on example using a practice site: **scrapethesite.com**.

Here’s how you explore the site:

1. **Right-click** on the page and choose **Inspect** to open the browser's developer tools.
2. You’ll see the real HTML of the website — it’s more complex than a simple example, but the structure is the same.
3. Use the **select tool (mouse icon)** to click on elements on the page — like a title, name, or link — and it will highlight the corresponding HTML code.

This makes it *super easy* to locate exactly where the data lives on the page!

📊 Spotting a Table in the HTML

Tables are often full of structured data — great for scraping!

In the practice site, there's a table containing hockey team names. Here’s how to explore it:

- The table is enclosed in a `<table>` tag, with `class="table"`.
- Inside are rows (`<tr>`), header cells (`<th>`), and data cells (`<td>`).
- By clicking on an element in the table, the inspector shows you the full structure, so you know exactly which tags and classes to target later.

🔗 Understanding Links and Other Elements

You’ll also come across:

- `<a href="...">` — anchor tags with **hyperlinks**.
- `<p>` — paragraph tags, usually holding plain text.

These details matter because you can choose whether to scrape the text itself, the link, or both.

💡 Key Takeaways

- HTML is the blueprint of a webpage.
- You can inspect any page using your browser to understand the HTML structure.
- Knowing how elements are nested and labeled (like tags, classes, attributes) helps you extract only what you need.
- This sets the stage for **targeted scraping** in future lessons.