**Web Scraping: An Introduction**

**Key Concepts:**

1. **Tags:** HTML uses tags to define elements on a web page.
   * **<tagname>content</tagname>**
2. **Attributes:** Provide additional information about HTML elements.
   * **<a href="https://www.example.com">This is a link</a>**
3. **Classes and IDs:** Used to apply styles or identify elements.
   * **<p class="highlight">This is a highlighted paragraph.</p>**
   * **<div id="header">This is a header</div>**

**CSS Selectors:**

* **CSS (Cascading Style Sheets):** Used to style HTML elements.

**Common CSS Selectors:**

1. **Element Selector:** Selects all elements of a specified type.
   * **p { color: red; }**
2. **Class Selector:** Selects elements with a specific class.
   * **.highlight { background-color: yellow; }**
3. **ID Selector:** Selects a specific element with a given ID.
   * **#header { font-size: 24px; }**
4. **Descendant Selector:** Selects all elements that are descendants of a specified element.
   * **div p { font-style: italic; }**

**Here are explanations for some common HTML tags used in web scraping:**

**<html>:**

Purpose: Defines the root of an HTML document.

Usage: All other HTML elements are nested inside this tag.

**<head>:**

Purpose: Contains meta-information about the HTML document, such as the title, character set, etc.

Usage: Located within the <html> tag.

**<title>:**

Purpose: Sets the title of the HTML document, which is displayed in the browser's title bar or tab.

Usage: Found within the <head> section.

**<body>:**

Purpose: Contains the content of the HTML document, such as text, images, links, etc.

Usage: Located within the <html> tag.

**<h1>, <h2>, ..., <h6>:**

Purpose: Defines headings, with <h1> being the largest and <h6> being the smallest.

Usage: Used to structure and organize content.

**<p>:**

Purpose: Defines a paragraph.

Usage: Used to structure and separate blocks of text.

**<a>:**

Purpose: Defines a hyperlink, linking to another web page or resource.

Usage: The href attribute specifies the URL.

**<img>:**

Purpose: Embeds an image in the document.

Usage: The src attribute specifies the image source.

**<ul>, <ol>, <li>:**

Purpose: Defines unordered lists, ordered lists, and list items.

Usage: <ul> and <ol> contain <li> elements that represent list items.

**<div>:**

Purpose: Defines a division or a container for other HTML elements.

Usage: Used for layout and structuring content.

**<span>:**

Purpose: Defines an inline container and is often used for applying styles to a specific part of text.

Usage: Used for styling or scripting purposes.

**<table>, <tr>, <td>, <th>:**

Purpose: Defines a table, table row, table cell, and table header, respectively.

Usage: Used to structure tabular data.

**<form>:**

Purpose: Defines an HTML form that allows users to input data.

Usage: May contain input fields, buttons, and other form elements.

**<input>:**

Purpose: Defines an input field where the user can enter data.

Usage: The type attribute specifies the type of input (e.g., text, checkbox, radio).

**<script>:**

Purpose: Embeds or references a client-side script, such as JavaScript.

Usage: Used for adding interactivity to web pages.

* **Web Scraping** refers to the extraction of data from websites. It involves fetching the web page and then extracting useful information from it.
* Web scraping is a technique used to extract data from websites. It involves fetching and parsing the HTML code of a web page to extract the information you need. Here's a step-by-step explanation of how web scraping works:
* **HTTP Request:**

Web scraping begins with sending an HTTP request to the target website's server. This request asks the server to return the HTML of a particular webpage.

* **Http status code**

1. **1xx (Informational):** Request received, continuing process.
   * 100: Continue
   * 101: Switching Protocols
2. **2xx (Success):** The request was successfully received, understood, and accepted.
   * 200: OK
   * 201: Created
   * 204: No Content
3. **3xx (Redirection):** Further action needs to be taken in order to complete the request.
   * 301: Moved Permanently
   * 302: Found (previously "Moved Temporarily")
   * 304: Not Modified
4. **4xx (Client Error):** The request contains bad syntax or cannot be fulfilled.
   * 400: Bad Request
   * 401: Unauthorized
   * 403: Forbidden
   * 404: Not Found
5. **5xx (Server Error):** The server failed to fulfill a valid request.
   * 500: Internal Server Error
   * 502: Bad Gateway
   * 503: Service Unavailable
6. In web scraping, the HTTP method used to retrieve data from a web server is typically the **GET** method. The GET method is used to request data from a specified resource. When you make a GET request, you are asking the server to retrieve and return the data associated with a particular URL.
7. The **POST** method is another HTTP method commonly used in web development, but it is not typically used for retrieving data. Instead, the POST method is used to submit data to be processed to a specified resource.

* **HTML Download:**

The server responds to the request by sending the HTML content of the requested webpage. This HTML code contains the structure and content of the page.

* **HTML Parsing:**

Once the HTML is downloaded, the next step is to parse it. Parsing involves breaking down the HTML into its constituent elements, such as headings, paragraphs, tables, and links.

* **Data Extraction:**

After parsing the HTML, the specific data you're interested in needs to be extracted. This is done by identifying the HTML tags and attributes that contain the desired information.

* **Handling Dynamic Content:**

Some websites use JavaScript to load content dynamically after the initial HTML has been loaded. In such cases, a headless browser or an automated tool like Selenium may be required to simulate the interaction with the page and retrieve the dynamically loaded content.

* **Storing the Data:**

Once the data is extracted, it can be stored in various formats, such as a CSV file, database, or any other suitable storage method.

* **Respecting Robots.txt:**

Before scraping a website, it's important to check its **robots.txt** file. This file provides guidelines to web crawlers about which pages can be crawled and which should be avoided. It's essential to respect these rules to avoid legal issues.

* **Handling Rate Limiting and IP Blocking:**

Some websites implement measures to prevent or limit web scraping by monitoring IP addresses and imposing restrictions. To avoid being blocked, developers may implement strategies like using proxies, rotating user agents, and adding delays between requests.

* **Legal and Ethical Considerations:**

Web scraping should be done ethically and legally. Some websites explicitly prohibit scraping in their terms of service. Always review a site's terms of use and privacy policy before scraping, and ensure compliance with applicable laws.

* **Monitoring and Maintenance:**

Websites may change their structure over time, so it's important to monitor and update your web scraping code accordingly. Regular maintenance is necessary to adapt to any changes in the target website.

Popular tools for web scraping include:

* **Beautiful Soup:** A Python library for pulling data out of HTML and XML files.
* **Requests:** A Python library for sending HTTP requests.
* **Selenium:** A browser automation tool that can be used for scraping dynamic websites.
* **Scrapy:** An open-source and collaborative web crawling framework for Python.

**Applications:**

1. **Data Collection:** Collecting data for analysis, research, or reporting.
2. **Price Monitoring:** Scraping e-commerce sites for price changes.
3. **Job Hunting:** Extracting job listings from various websites.
4. **Competitor Analysis:** Gathering information on competitors' products or pricing.
5. **News Aggregation:** Collecting news headlines from different sources.

**HTML Structure:**

* **HTML (Hypertext Markup Language):** The standard markup language for creating web pages.

**Basic HTML Structure:**

htmlCopy code

<!DOCTYPE html>

<html>

<head>

<title>Page Title</title>

</head>

<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

<a href="https://www.example.com">This is a link</a>

</body>

</html>

**Web Scraping Ethics:**

* Always check the website's **robots.txt** file to see if scraping is allowed.
* Don't overload servers with too many requests; use proper timeouts and delays.
* Respect the terms of service of the website.

**Tools for Web Scraping in Python:**

1. **Requests:** For making HTTP requests.
2. **Beautiful Soup:** For parsing HTML and extracting data.
3. **Selenium:** For scraping dynamic websites with JavaScript content.

**Example Python Code:**

import requests from bs4

import BeautifulSoup

url = "https://www.example.com"

response = requests.get(url)

soup = BeautifulSoup(response.text, 'html.parser')

# Extracting all paragraph elements

paragraphs = soup.find\_all('p')

for paragraph in paragraphs:

print(paragraph.text)

**Using Requests for HTTP Requests:**

**1. Introduction to requests library:**

* The **requests** library in Python is a powerful tool for making HTTP requests.
* It simplifies the process of sending HTTP requests and handling responses.

**2. Installation:**

* Before using **requests**, it needs to be installed. You can install it using:

pip install requests

**3. Basic Usage:**

* Import the **requests** module: **import requests**
* Use **requests.get(url)** to make a GET request.
* Use **requests.post(url, data)** to make a POST request.

**4. Handling Responses:**

* The response object has attributes like **status\_code**, **text**, **json()**, etc.
* Example:

import requests

response = requests.get('https://example.com')

print(response.status\_code)

print(response.text)

**Overview of Beautiful Soup for HTML Parsing:**

**1. Introduction to Beautiful Soup:**

* Beautiful Soup is a Python library for pulling data out of HTML and XML files.
* It provides Pythonic idioms for iterating, searching, and modifying the parse tree.

**2. Installation:**

* Install Beautiful Soup using:

pip install beautifulsoup4

**3. Creating a Soup Object:**

* Import Beautiful Soup: **from bs4 import BeautifulSoup**

url = <https://example.com>

req = request.get(url)

content = BeautifulSoup(req.content, 'html.parser')

print(content)