Web Scraping: Takeaways

Syntax

- Importing BeautifulSoup:

```
from bs4 import BeautifulSoup
```

- Initializing the HTML parser:

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

- Getting the inside text of a tag:

```
title_text = title.text
```

- Returning a list of all occurrences of a tag:

```
head.find_all("title")
```

- Getting the first instance of a tag:

```
title=head[0].find_all("title")
```

- Creating an example page using HTML:

```
<html>
    <head>
        <title>
    </head>
    <body>
        <p>Here is some simple content for this page.<p>
    </body>
</html>
```

- Using CSS to make all of the text inside all paragraphs red:

```
p{
    color: red
 }
```

- Using CSS selectors to style all elements with the class "inner-text" red:

```
.inner-text{
    color: red
  }
```

- Working with CSS selectors:

```
parsar.select(".first-item")
```

Concepts

- A lot of data is not accessible through data sets or APIs; they exist on the Internet as Web pages. We can use a technique called web scraping to access the data without waiting for the provider to create an API.
- We can use the `requests` library to download a web page, and `Beautifulsoup` to extract the relevant parts of the web page.
- Web pages use HyperText Markup Language (HTML) as the foundation for the content on the page, and browsers such as Google Chrome and Mozilla Firefox reads the HTML to determine how to render and display the page.
- The `head` tag in HTML contains information that's useful to the Web browser that's rendering the page. The `body` section contains the bulk of the content the user interacts with on the page. The `title` tag tells the Web browser what page title to display in the toolbar.
- HTML allows elements to have IDs so we can use them to refer to specific elements since IDs are unique.
- **Cascading Style Sheets**, or **CSS**, is a language for adding styles to HTML pages.
- We can also use CSS selectors to select elements when we do web scraping.

Resources

- HTML basics
- HTML element
- BeautifulSoup Documentation