Introduction to Web scraping in Python

Ishita Gopal

8th March

https://www2.stat.duke.edu/courses/Fall19/sta199.001/slides/lec-slides/06a-web-scrape.html#28

- Why Web scraping
- Html and Components of a webpage
- Code walkthrough how to use these components to collect data

Scraping the web: what? why?

- Increasing amount of data is available on the web.
- These data are provided in an unstructured format: you can always copy & paste, but it's time-consuming and prone to errors.
- Web scraping is the process of extracting this information automatically and transform it into a structured dataset.

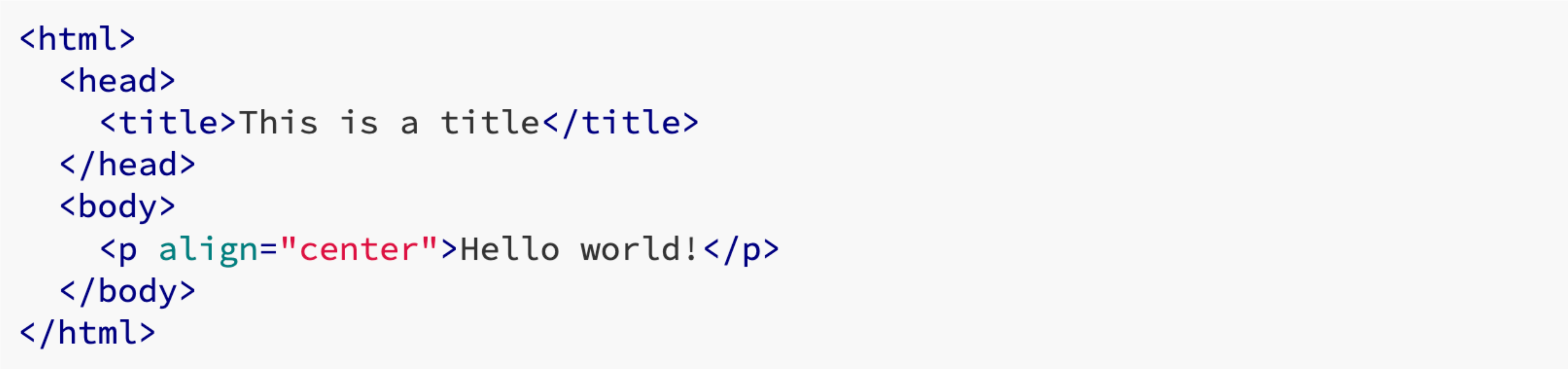
- Two different scenarios:
  - **Screen scraping**: extract data from source code of website, with html parser (easy) or regular expression matching (less easy).

  - **Web APIs (application programming interface)**: website offers a set of structured http requests that return JSON or XML files.

Hypertext Markup Language (HTML)?

- It is a markup language used to create the structure of a web page.
- It consists of elements (like headings, paragraphs, images, links, etc.) and attributes (additional information about elements).
- HTML provides the basic structure and content of a web page.

- Is structured (hierarchical / tree based)



```html
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p align="center">Hello world!</p>
  </body>
</html>
```

- But not in a form useful for analysis (flat / tidy).

	Attribute 1	Attribute 2
Record 1		
Record 2		

Document Object Model (DOM)

Document Object Model (DOM)

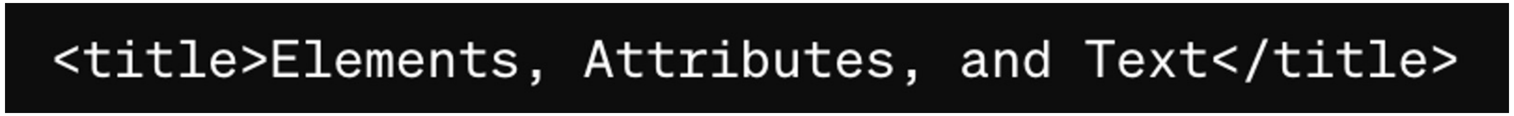
- Programming interface that represents the structured document as a tree of objects.

- Each HTML element becomes a "node" in the tree, and these nodes can be manipulated using programming languages like Python.

HTML Elements:

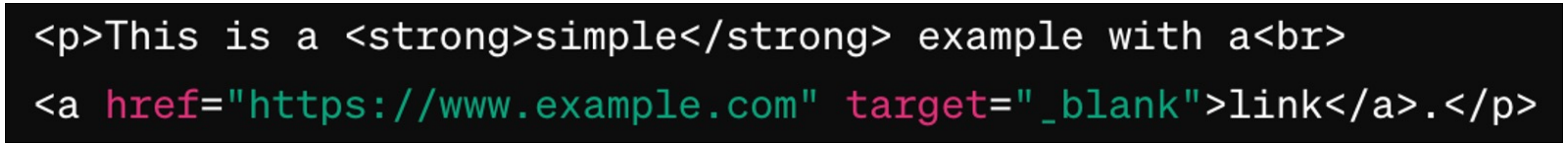
- Fundamental building block in HTML that defines the structure of a document.

- Composed of a start tag, content, and an end tag.



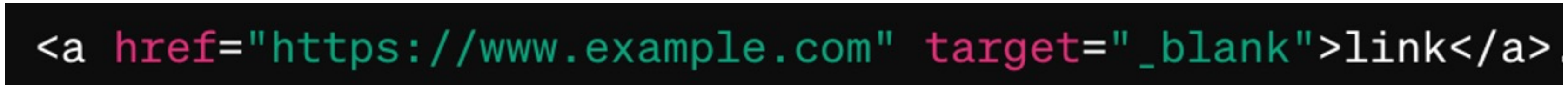
```
<title>Elements, Attributes, and Text</title>
```

- Elements can contain other elements, forming a hierarchical structure in the Document Object Model (DOM).



```
<p>This is a <strong>simple</strong> example with a<br>
<a href="https://www.example.com" target="_blank">link</a>.</p>
```

HTML: Attributes

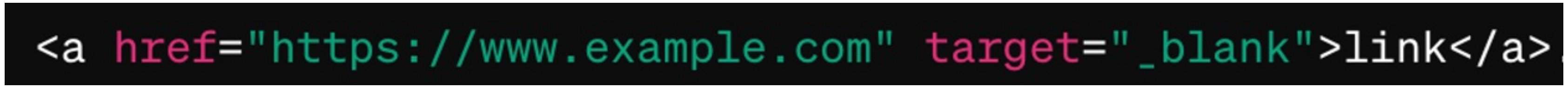


```html
<a href="https://www.example.com" target="_blank">link</a>
```

- Provide additional information about HTML elements

- Are added to the opening tag of the element

- Consist of a name and a value, separated by an equals sign (=) and enclosed in double or single quotes.



```html
<a href="https://www.example.com" target="_blank">link</a>
```

- <a> is the anchor element (used for creating hyperlinks).

- href, target, and are attributes of the <a> element.

- "https://www.example.com" is the value assigned to the href attribute. It specifies the URL the link points to.

- "_blank" is the value assigned to the target attribute. It specifies that the link should open in a new browser tab or window.

- This example demonstrates the integration of elements, attributes, and text within an HTML document. Each plays a distinct role in defining the structure, behavior, and content of the web page.