Washington University in St. Louis

College of Arts & Sciences

Introduction to Web Scraping in Python

TRIADS Training Series

Ishita Gopal

March 8, 2024

Today

- Why Web scraping ?
- Webpage and HTML
- Elements and Attributes in HTML
- DOM
- Code walkthrough and 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 transforming it into a structured dataset.

Two different scenarios:

- **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.

We will learn about the first one today!

Hypertext Markup Language (HTML)?

- It is a markup language used to create the content and basic structure of a web page.

- It consists of elements (like headings, paragraphs, images, links, etc.) and attributes (additional information about elements).

```
<html>
<head>
<title>Ishita Gopal</title>
</head>

<body bgcolor="lightblue">
<h1>Ishita Gopal</h1>
<p><strong>Welcome to my Web site!</strong></p>

<h2>Introduction</h2>
<p>This is my place on the Web!</p>

<h2>Education</h2>
<p>I am a postdoc at Washington University.</p>

<h2>Contact Information</h2>
<p>For more information, you can email me at gopali@washu.edu</p>

</body>
</html>
```

Ishita Gopal

Welcome to my Web site!

Introduction

This is my place on the Web!

Education

I am a postdoc at Washington University.

Contact Information

For more information, you can email me at gopali@washu.edu

HTML

- Is structured (hierarchical / tree based)

```
<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		

Components of HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
    <title>Simple HTML Example</title>
</head>
<body>
    <p>This is a <strong>simple</strong> example
    <a href="https://www.example.com" target="_blank">link</a>.</p>
</body>
</html>
```

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>Simple HTML Example</title>
```

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

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

HTML: Attributes

```
<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.

Example

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

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

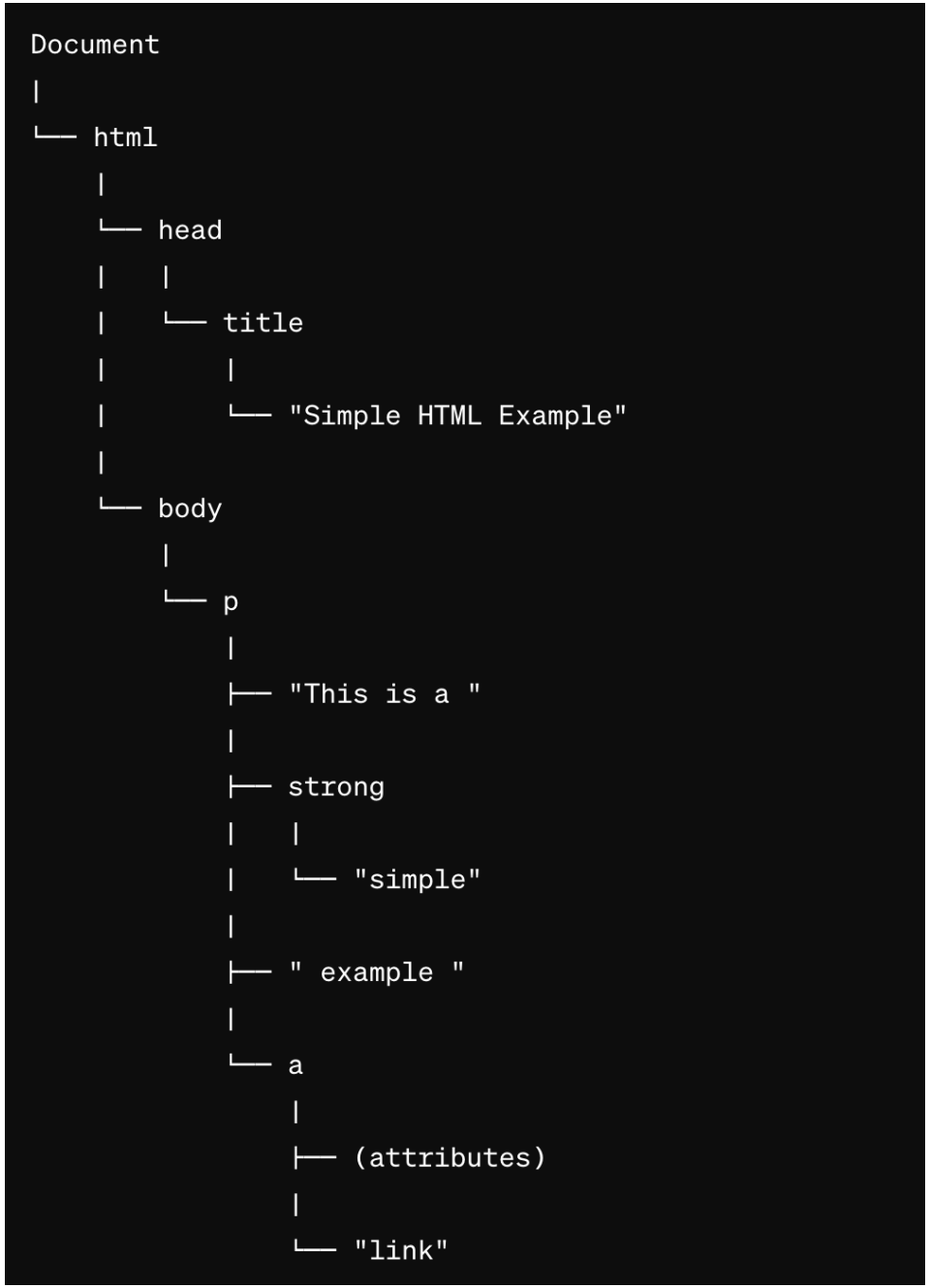
href and target are attributes of the <a> element.

"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.

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 accessed using programming languages like Python.
- Using Python, we will navigate such a DOM and extract data from a website today!



Let’s get started!

Google Collab Notebook:
https://tinyurl.com/TRIADS-web-scraping

Reference:

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