CCL accredited event

Introduction to Web Scraping in Python



Speaker: Ayazhan Kadessova

BSc in Business Computing and Data Analytics, Department of Computer Science, HKBU

Dr. Eric Chow

Digital Scholarship Manager, HKBU Library

March 7, 2023 3:30 - 6:00pm

Online via Zoom



Sign-up / Sign-in for Google Account



https://colab.research.google.com/

Expectations

- This is a hands-on workshop; follow along with our coding demo!
- Sign-up / sign-in to a Google account we will be using Google CoLab
- Assumes you already have basic Python programming knowledge (LIB03 - Introduction to Python Programming)

Run-Down

- Warm Up, Google CoLab, review of basic Python programming (20 min)
- Use of Python + BeautifulSoup to scrape <u>BookDepository.com</u> (120 min)
- Wrap up, ethics of web scraping, Q&A (10 min)

Unveiling the Basics of HTML, CSS, and Javascript for Web Scraping with Python

Ayazhan Kadessova - Year 2 Business Computing and Data Analytics Student

March, 7

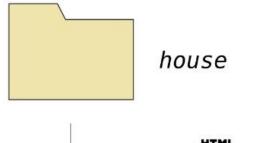
HTML, CSS, Javascript Overview
How to Obtain HTML using Requests
Parsing Data with BeautifulSoup



HTML, CSS, Javascript

HTML elements have attributes such as class or id which are used for styling with CSS and adding interactivity with Javascript.

- HTML text file with a syntax that will tell the browser what content to paint, what text to show, and what resources to download.
- CSS will format and style the content (i.e., colors, fonts, and many more.
- Javascript adds functionality and behavior, such as opening a pop-up window.





HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
 <meta charset="UTF-8">
 __meta name="viewport" content="width=device-width, initial-scale=1.0">
 <meta http-equiv="X-UA-Compatible" content="ie=edge">
 <link rel="stylesheet" href="./styles.css">
 <title>Document</title>
</head>
<body>
 <h1>This is a first level heading in HTML. With CSS, I will turn this into red co
 <h2>This is a second level heading in HTML. With CSS, I will turn this into blue
 <h3>This is a third level heading in HTML. With CSS, I will turn this into green
 This is a <em>paragragh</em> As you can see, I placed an empahisis on the word
   the background color of the word "paragraph" to black, and its text color to a
 The main essence of this tutorial is to:
   <l
      Show you how to format a web document with HTML
      Show you how to design a web page with CSS
      Show you how to program a web document with JavaScript
   Next, I am going to add the following two numbers and display the result, all n
   First number: <span id= "firstNum">2</span> <br>
   Second number: <span id= "secondNum">7</span> 
   Therefore, the sum of the two of those numbers is: <span id= "answer">(place)
   <input type="button" id="sumButton" value="Click to add!">
</body>
</html>
```

This is a first level heading in HTML. With CSS, I will turn this into red color

This is a second level heading in HTML. With CSS, I will turn this into blue color

This is a third level heading in HTML. With CSS, I will turn this into green color

This is a paragraph As you can see, I placed an empahsis on the word "paragraph". Now, I will change also the background color of the word "paragraph" to black, and its text color to green, all with just CSS.

The main essence of this tutorial is to:

- . Show you how to format a web document with HTML
- Show you how to design a web page with CSS
 Show you how to program a web document with JavaScript

Next, I am going to add two numbers and display the result, all with JavaScript

First number:2

Second number

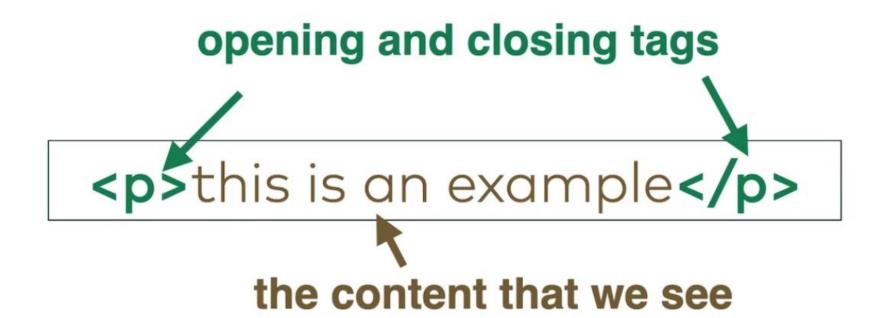
Therefore, the sum of the two of those is (placeholder for the answer)

Get the Sum

HTML Elements

Elements (links, paragraphs, headings, blocks) are wrapped between tags.

Content is between opening and closing tags. Example: sentence in a paragraph () element.



HTML Elements and Attributes

HTML elements may also have attributes that contain additional information about the element.

Attributes are defined in the opening tags with the following syntax: attribute name="attribute value".



CSS

With CSS you can set the colour and background of your elements, as well as the typeface, margins, spacing, padding and so much more.

```
h1 {
  background-color: #ff0000;
}

h2 {
  background-color: #0000FF;
}

h3 {
  background-color: #00FF00;
}

em {
  background-color: #0000000;
  color: #fffffff;
}
```

This is a first level heading in HTML. With CSS, I will turn this into blue color.

This is a second level heading in HTML. With CSS, I will turn this into blue color.

This is a SAMON As you can see, I placed as expelsive on the word "paragraph". Nov. I will change sho the background order of the word "paragraph" to black, and its next order to green, all with just CSS. The main counter of this next seed obscured with HTML.

• Share you have to femous to web document with HTML.

• Share you have to femous a web document with HTML.

• Share you have to graps and those suchers and departed. CSS.

• Share you have to graps and document with I restrictly.

* Seed as an agent and from suchers and departed of the seed, all with JesuScript

* First anather: I

* Second market: 7

Therefore, the sees of the two of those is: (placeholder for the answer)

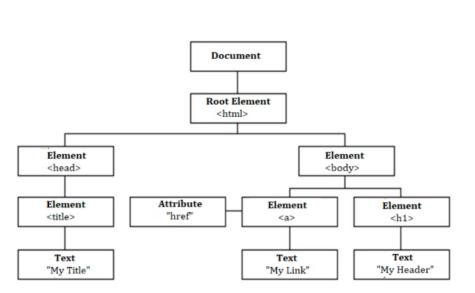
Offitting Sours

Javascript

The DOM is a tree-like representation of the web page that gets loaded into the browser.

Thanks to the DOM, we can use methods like getElementById() to access elements from our web page.

JavaScript allows you to make your webpage "think and act", which is what programming is all about.



Each element on the web page is represented on the DOM

This is a first level heading in HTML. With CSS, I will turn this into red color

This is a second level heading in HTML. With CSS, I will turn this into blue color

This is a third level heading in HTML. With CSS, I will turn this into preca color

This is a second level heading in HTML. With CSS, I will turn this into preca color

This is a second level heading in HTML. With CSS, I will turn this into preca color

This is a second level heading in HTML. With CSS, I will turn this into preca color

This is a second level heading in HTML. With CSS, I will turn this into preca color

This is a fair level heading in HTML. With CSS, I will turn this into preca color

This is a third level heading in HTML. With CSS, I will turn this into preca color

This is a third level heading in HTML. With CSS, I will turn this into preca color

This is a fair level heading in HTML. With CSS, I will turn this into preca color

This is a third level heading in HTML. With CSS, I will turn this into preca color

This is a third level heading in HTML. With CSS, I will turn this into preca color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading in HTML. With CSS, I will turn this into place color

This is a third level heading

Clicking the "Get the sum" button will display the sum of 2 and 7

How to Obtain HTML using Requests

Requests makes it easy to make HTTP requests and access the response data.

To obtain HTML from a website, you need to make a GET request using the requests library. This will return the HTML of the website in the response.

html = response.content
print(html)

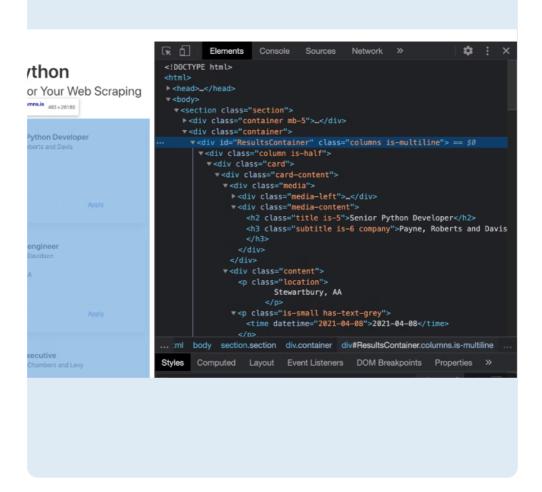
b'<!DOCTYPE html>\n<html lang="en">\n<head>\n\n k rel="preconnect" href="https://d3ogvdx946i4sr.cloudfront.r et"/>\nhr<="dns-prefetch" href="https://d3ogvdx946i4sr.cloudfront.net"/>\n<script type="text/javascript">\n function csmWidgetStart(widgetName) {\n if (typeof uet -- \'function\') {\n uet(\'bb\', w idgetName, {wb: 1});\n function csmWidgetEnd(widgetName) {\n if (typeof uet = \'function\') (\n </script>\ uex(\'ld\', widgetName, {wb: 1});\n }\n <style>\n .hide-when-no-is {\n display: none !important;\n .show-when-no-js {\n display: block !important; \n </noscript>\n\n \mathrm{meta http-equiv="content-type" content="text/html; charset=UTF-8" />\n<meta name="copyright" co ntent="© 2020 Book Depository Ltd." />\n<meta name="author" content="Book Depository" />\n<meta name="viewport" content="width-device-width, initial-scale=1, maximum-scale=2, user-scalable=1">\n<meta name="google-site-verificat ion" content="ogOme2fCVYq3eJxsHzQiMuXsUIRVwYrtI2Xn-U1nNpc" />\n<meta name="msvalidate.01" content="D45E907CC9A963F7 8BD3129AAAAFE4F0" />\n<meta http-equiv="X-UA-Compatible" content="IE=edge; charset=UTF-8"> <meta name="description" content="See our top 1000 bestselling books, charts and future bestsellers. Free delivery worldwide on over 20 mill ion books at Book Depository." />\n <meta name="keywords" content="top 1000, bestseller, book, featured, chart, futur e bestsellers, free delivery worldwide, books" />\n\t<meta name="Revisit-After" content="30 days" />\n\t<link rel="ca nonical" href="https://www.bookdepository.com/bestsellers" />\nlink rel="alternate" hreflang="x-default" href="htt ps://www.bookdepository.com/bestsellers" />\n\t<link rel="alternate" hreflang="en" href="https://www.bookdepositor y.com/bestsellers" />\n\t<link rel="alternate" hreflang="es" href="https://www.bookdepository.com/es/bestsellers" / >\n\t<link rel="next" href="/bestsellers7page=2" />\n\t<title>\n\tBestselling books online with free delivery at Bo

Parsing Data with BeautifulSoup

BeautifulSoup allows to navigate and search the HTML.

It can be used to find specific elements in the HTML, such as tags, attributes, and text.

Once the data is extracted, it can be used for further analysis or manipulation.



Conclusion

Web scraping is a process of extracting data from websites.

Popular libraries: Requests and BeautifulSoup.

Requests is used to obtain HTML from a website.

BeautifulSoup is used to parse the HTML and extract the data.

HTML elements have attributes such as class or id which are used for styling with CSS and adding interactivity with Javascript.

Ethics of Web Scraping

- Be mindful of copyright
- If data is clearly private, then <u>do not</u> scrape
- **Do not** overload other people's server
- Use API when available

Examples of API

- https://developers.google.com/books/docs/overview
- https://wiki.harvard.edu/confluence/display/LibraryStaffDoc/LibraryCloud+APIs
- https://github.com/HKBULIB/TVWeek API