Forward School

Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Introduction of BeautifulSoup - Part 1

Name: Ooi Caaron

IC Number: 990701-07-5837

Date : 3/7/23

Introduction : Learn how to get the data from Web page by accessing the HTML attributes

Conclusion : Still need to practice more and revise again

P13 - Introduction of BeautifulSoup - Part 1

HTML and CSS

Let's start explore how a Web page HTML doc works with CSS (Cascading Style Sheet)

https://codepen.io/buzztracer/pen/yLYJBOd (https://codepen.io/buzztracer/pen/yLYJBOd)

HTML div Tag

Definition and Usage

The div tag defines a division or a section in an HTML document.

The div element is often used as a container for other HTML elements to style them with CSS or to perform certain tasks with JavaScript.

HTML Id Attributes

Definition and Usage

The id attribute is a unique identifier which is used to specify the document.

It is used by CSS and JavaScript to perform a certain task for a unique element.

In CSS, the id attribute is used using # symbol followed by id.

HTML Class Attribute

Definition and Usage

Class in html:

The class is an attribute which specifies one or more class names for an HTML element.

The class attribute can be used on any HTML element.

The class name can be used by CSS and JavaScript to perform certain tasks for elements with the specified class name.

How a web page call another?

By using the < a >: Anchor element

https://developer.mozilla.org/en-US/docs/Web/HTML/Element/a (https://developer.mozilla.org/en-US/docs/Web/HTML/Element/a)

In [1]:

```
# Get all hyperlinks

import requests

from bs4 import BeautifulSoup

html = requests.get('http://en.wikipedia.org/wiki/Malaysia')
bs = BeautifulSoup(html.content, 'html.parser')

for link in bs.find_all('a'):
    if 'href' in link.attrs:
        print(link.attrs['href'])
```

```
#bodyContent
/wiki/Main_Page
/wiki/Wikipedia:Contents
/wiki/Portal:Current_events
/wiki/Special:Random
/wiki/Wikipedia:About
//en.wikipedia.org/wiki/Wikipedia:Contact_us
https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_sour
ce=donate&utm_medium=sidebar&utm_campaign=C13_en.wikipedia.org&uselang=
en (https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_
source=donate&utm_medium=sidebar&utm_campaign=C13_en.wikipedia.org&usel
ang=en)
/wiki/Help:Contents
/wiki/Help:Introduction
/wiki/Wikipedia:Community_portal
/wiki/Special:RecentChanges
/wiki/Wikipedia:File_upload_wizard
/wiki/Main_Page
/wiki/Special:Search
```

In [3]:

```
# retrieve only desired list of articles by using regular expression  ^(/wiki/)((?!:).)*
import requests
import re
from bs4 import BeautifulSoup

html = requests.get('http://en.wikipedia.org/wiki/Malaysia')
bs = BeautifulSoup(html.content, 'html.parser')


for link in bs.find('div', {'id':'bodyContent'}).find_all('a', href=re.compile('^(/wiki/
    if 'href' in link.attrs:
        print(link.attrs['href'])
```

```
/wiki/Geographic_coordinate_system
/wiki/Malesia
/wiki/Flag_of_Malaysia
/wiki/Coat_of_arms_of_Malaysia
/wiki/Jawi_script
/wiki/Unity_makes_strength
/wiki/Negaraku
/wiki/Asia
/wiki/ASEAN
/wiki/Kuala_Lumpur
/wiki/Putrajaya
/wiki/Malaysian_Malay
/wiki/English_language
/wiki/Ethnic_group
/wiki/Bumiputera_(Malaysia)
/wiki/Malaysian_Malays
/wiki/Sabah
/wiki/Sarawak
/wiki/Orang_Asli
```

Saving the results to a CSV file

In [4]:

```
import csv
import requests

from bs4 import BeautifulSoup

html = requests.get("http://en.wikipedia.org/wiki/Comparison_of_text_editors")
bsObj = BeautifulSoup(html.content, 'html.parser')

#The main comparison table is currently the first table on the page
table = bsObj.findAll("table",{"class":"wikitable"})[0]
rows = table.findAll("tr")

csvFile = open("editors.csv", 'w', encoding='utf8')
writer = csv.writer(csvFile)

try:
    for row in rows:
        csvRow = []
        for cell in row.findAll(['td', 'th']):
                csvRow.append(cell.get_text())
                writer.writerow(csvRow)
finally:
    csvFile.close()
```

In [ ]:

```

```