

**Beautiful Soup**

|  |
| --- |
| from bs4 import BeautifulSoup |
|  |

|  |
| --- |
| import requests |
|  |

|  |
| --- |
| import csv |
|  |

|  |
| --- |
|  |
|  |

source = requests.get('http://coreyms.com').text

soup = BeautifulSoup(source, 'lxml'**)//It will have the entire HTML structure saved to the soup object**

Print(soup**)// it will print the html file, in left allighned manner**

Print(soup.prettify())**// will automatically convert the fill and indend it according to the tags**

**//if we want to access any tag then we will access it as a attribute**

Match= soup.title //**This will give the first title of the page we can add, DIV also it will still select the first div**

Print(match) //Structure



//to access the text part

Then,

Match=soup.title.text

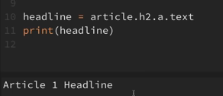
Print(match) 

//**In order to find a specific part of html**

**Suppose I want to get the div of class Xiomi from smartphones,**

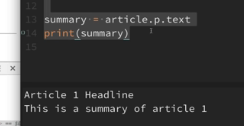
**Then,**

Match= soup.find(‘div’,class\_=’Xiomi’)**// will give me the div of the class Xiomi**



matched elements

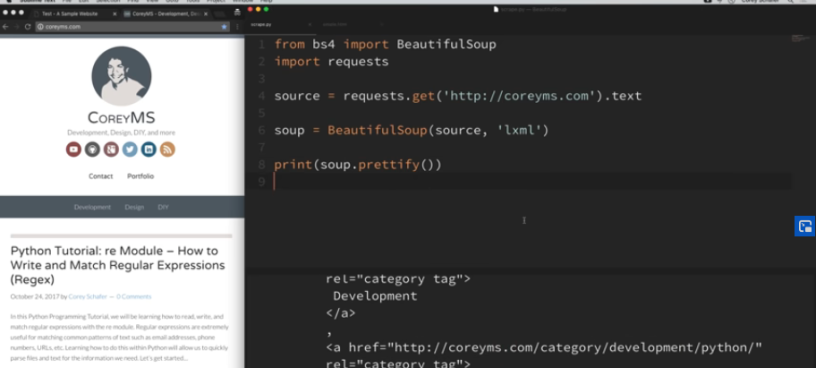
To access the separate text in the ‘a’ tag



It will give the entire div.



Gives List of all the matched elements

 New Website:

Tip: Try to get One Element and then you can loop over it and find rest of the elements.

from bs4 import BeautifulSoup

import requests

import csv

source = requests.get('http://coreyms.com').text

soup = BeautifulSoup(source, 'lxml')

csv\_file = open('cms\_scrape.csv', 'w')

csv\_writer = csv.writer(csv\_file)

csv\_writer.writerow(['headline', 'summary', 'video\_link'])

for article in soup.find\_all('article'):

headline = article.h2.a.text

print(headline)

Splitted using forward slash, took the 4th element because it contains the video code generally. Then again splitted using ? and pickedup the 0th element and we have our video id

summary = article.find('div', class\_='entry-content').p.text

print(summary)

try:

vid\_src = article.find('iframe', class\_='youtube-player')['src']

vid\_id = vid\_src.split('/')[4]

vid\_id = vid\_id.split('?')[0]

yt\_link = f'https://youtube.com/watch?v={vid\_id}'

except Exception as e:

yt\_link = None

print(yt\_link)

print()

csv\_writer.writerow([headline, summary, yt\_link])

csv\_file.close()