**What is Web Scraping?**

Web scraping is a computer software technique of extracting information from websites. This technique mostly focuses on the transformation of unstructured data (HTML format) on the web into structured data (database or spreadsheet).

You can perform web scrapping in various ways, including use of Google Docs to almost every programming language.

Python has a library known as **‘BeautifulSoup’** which assists this task.

**Libraries required for web scraping**

**BeautifulSoup** (python library), since it is easy and intuitive to work on.

* **Urllib2**: It is a Python module which can be used for fetching URLs.
* **BeautifulSoup:** It is an tool for pulling out information from a webpage. You can use it to extract tables, lists, paragraph and you can also put filters to extract information from web pages.

**Note :** BeautifulSoup does not fetch the web page for us. That’s why, I use urllib2 in combination with the BeautifulSoup library.

**Steps to install beautifulsoup4**

Open Command Prompt and type as **pip install beautifulsoup4**.

**Example's**

**Example 1:**

html\_doc = **"""  
<html><head><title> WebScrapping </title></head>  
<body>  
<p class="title"><b>By Naveen Kumar</b></p>  
  
<p class="story">Once upon a time there were three little sisters; and their names were  
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,  
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and  
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;  
and they lived at the bottom of a well.</p>  
  
<p class="story">...</p>  
"""  
  
from** bs4 **import** BeautifulSoup  
soup = BeautifulSoup(html\_doc, **'html.parser'**)  
  
print(soup.prettify())  
  
print(soup.title) *# <title> WebScrapping </title>*print(soup.title.name) *# 'title'*print(soup.title.string) *# 'WebScrapping'*print(soup.title.parent.name) *# 'head'*print(soup.p) *# <p class="title"><b>By Naveen Kumar</b></p>*print(soup.p[**'class'**]) *# ['title']*print(soup.a) *# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>*print(soup.find\_all(**'a'**))  
*# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,  
# <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,  
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]*print(soup.find(id=**"link3"**))  
*# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>***for** link **in** soup.find\_all(**'a'**):  
 print(link.get(**'href'**))  
*# http://example.com/elsie  
# http://example.com/lacie  
# http://example.com/tillie*print(soup.get\_text())  
*# WebScrapping  
#  
# By Naveen Kumar  
# Once upon a time there were three little sisters; and their names were  
# Elsie,  
# Lacie and  
# Tillie;  
# and they lived at the bottom of a well.  
# ...*

Example 2:

**from** bs4 **import** BeautifulSoup  
**import** requests  
  
res = requests.get(**"https://www.sathyatech.com"**)  
  
bs = BeautifulSoup(res.text,**"html.parser"**)  
  
print(bs.title.string)  
  
print(bs.find(**"div"**,{**"class"**:**"fusion-contact-info"**}))  
  
print(bs.find\_all(**"div"**,class\_ =**"fusion-contact-info"**))  
  
**for** x **in** bs.find\_all(**"div"**,class\_ =**"fusion-contact-info"**):  
 print(**"1"**,bs.find(**"b"**))  
 print(**"2"**,bs.find(**"b"**).get\_text())  
 print(**"3"**,bs.find(**"a"**))  
 print(**"4"**,x.get\_text())

Example 3:

**from** bs4 **import** BeautifulSoup  
**import** requests  
  
res = requests.get(**"https://www.flipkart.com/"**)  
  
bs = BeautifulSoup(res.text,**"html.parser"**)  
  
print(bs.title.string)  
  
print(bs.find\_all(**"div"**,{**"class"**:**"iUmrbN"**}))  
  
**for** x **in** bs.find\_all(**"div"**,{**"class"**:**"iUmrbN"**}):  
 print(x.get\_text()) *# or print(x.text)*

**HTTP status codes**

Status codes are issued by a server in response to a client's request made to the server.

This is a list of **Hypertext Transfer Protocol** (HTTP) response status codes. There are five classes defined by the standard:

* **1xx informational response** .
* **2xx successful .**
* **3xx redirection .**
* **4xx client side error .**
* **5xx server side error.**

Note : refer to <https://httpstatuses.com/>