Libraries required for web scraping

As we know, python is a open source programming language. You may find many libraries to perform one function. Hence, it is necessary to find the best to use library. I prefer **BeautifulSoup** (python library), since it is easy and intuitive to work on. Precisely, I’ll use two Python modules for scraping data:

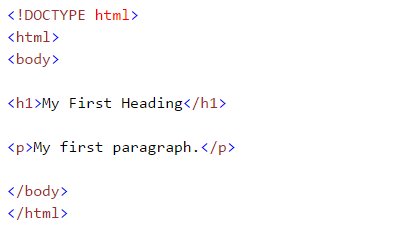
* **Urllib2**: It is a Python module which can be used for fetching URLs. It defines functions and classes to help with URL actions (basic and digest authentication, redirections, cookies, etc). For more detail refer to the [documentation page](https://docs.python.org/2/library/urllib2.html).
* **BeautifulSoup:** It is an incredible 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. In this article, we will use latest version BeautifulSoup 4. You can look at the installation instruction in its [documentation page](http://www.crummy.com/software/BeautifulSoup/bs4/doc/).

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

Python has several other options for HTML scraping in addition to BeatifulSoup. Here are some others:

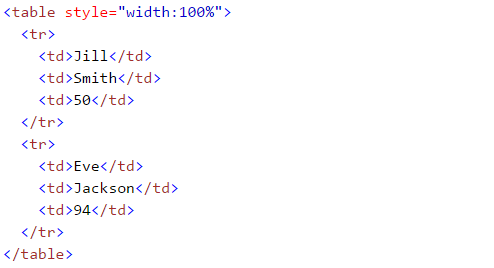
* [mechanize](http://wwwsearch.sourceforge.net/mechanize/)
* [scrapemark](http://arshaw.com/scrapemark/)
* [scrapy](http://scrapy.org/)

Basics – Get familiar with HTML (Tags)

While performing web scarping, we deal with html tags. Thus, we must have good understanding of them. If you already know basics of HTML, you can skip this section. Below is the basic syntax of HTML:[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/HTML.png)This syntax has various tags as elaborated below:

1. **<!DOCTYPE html> :**HTML documents must start with a type declaration
2. HTML document is contained between **<html>** and **</html>**
3. The visible part of the HTML document is between **<body>** and **</body>**
4. HTML headings are defined with the **<h1>** to **<h6>** tags
5. HTML paragraphs are defined with the **<p>** tag

Other useful HTML tags are:

1. HTML links are defined with the **<a>** tag, “<a href=“http://www.test.com”>This is a link for test.com</a>”
2. HTML tables are defined with<Table>, row as <tr> and rows are divided into data as <td>  
   
3. HTML list starts with <ul> (unordered) and <ol> (ordered). Each item of list starts with <li>

If you are new to this HTML tags, I would also recommend you to refer [HTML tutorial from W3schools](http://www.w3schools.com/html/). This will give you a clear understanding about HTML tags.

Scrapping a web Page using BeautifulSoup

Here, I am scraping data from a [Wikipedia page](https://en.wikipedia.org/wiki/List_of_state_and_union_territory_capitals_in_India). Our final goal is to extract list of state, union territory capitals in India. And some basic detail like establishment, former capital and others form this [wikipedia page](https://en.wikipedia.org/wiki/List_of_state_and_union_territory_capitals_in_India" \t "_blank). Let’s learn with doing this project step wise step:

1. **Import necessary libraries:**

#import the library used to query a website

import urllib2

#specify the url

wiki = "https://en.wikipedia.org/wiki/List\_of\_state\_and\_union\_territory\_capitals\_in\_India"

#Query the website and return the html to the variable 'page'

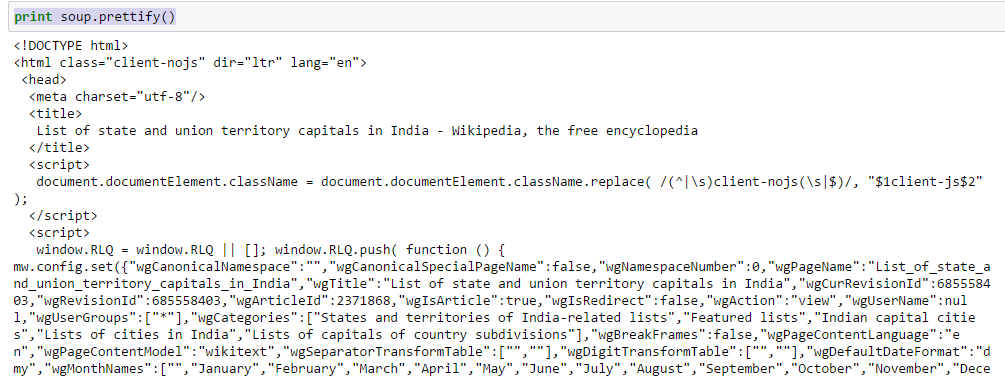
page = urllib2.urlopen(wiki)

#import the Beautiful soup functions to parse the data returned from the website

from bs4 import BeautifulSoup

#Parse the html in the 'page' variable, and store it in Beautiful Soup format

soup = BeautifulSoup(page)

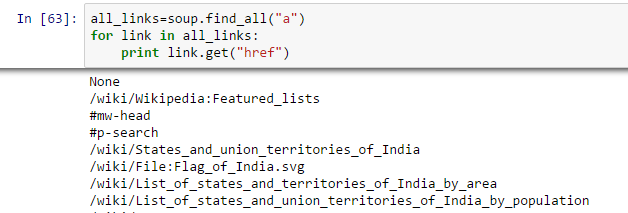
1. **Use function “prettify” to look at nested structure of HTML page**[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/FirstP.png)Above, you can see that structure of the HTML tags. This will help you to know about different available tags and how can you play with these to extract information.

1. **Work with HTML tags**
2. **soup.<tag>:**Return content between opening and closing tag including tag.
3. In[30]:soup.title

Out[30]:<title>List of state and union territory capitals in India - Wikipedia, the free encyclopedia</title>

1. **soup.<tag>.string:**Return string within given tag
2. In [38]:soup.title.string
3. Out[38]:u'List of state and union territory capitals in India - Wikipedia, the free encyclopedia'
4. **Find all the links within page’s <a> tags::** We know that, we can tag a link using tag “<a>”. So, we should go with option **soup.a** and it should return the links available in the web page. Let’s do it.
5. In [40]:soup.a

Out[40]:<a id="top"></a>

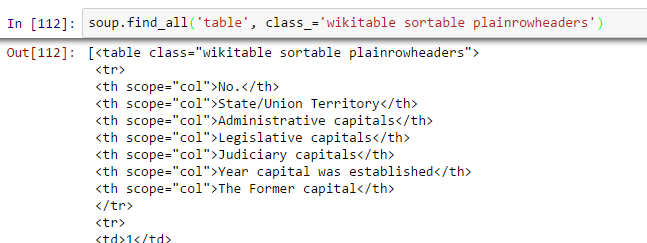
Above, you can see that, we have only one output. Now to extract all the links within <a>, we will use “**find\_all().  
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/All_Links.png)**Above, it is showing all links including titles, links and other information.  Now to show only links, we need to iterate over each a tag and then return the link using attribute “href” with **get**.  
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/All_Links2.png)

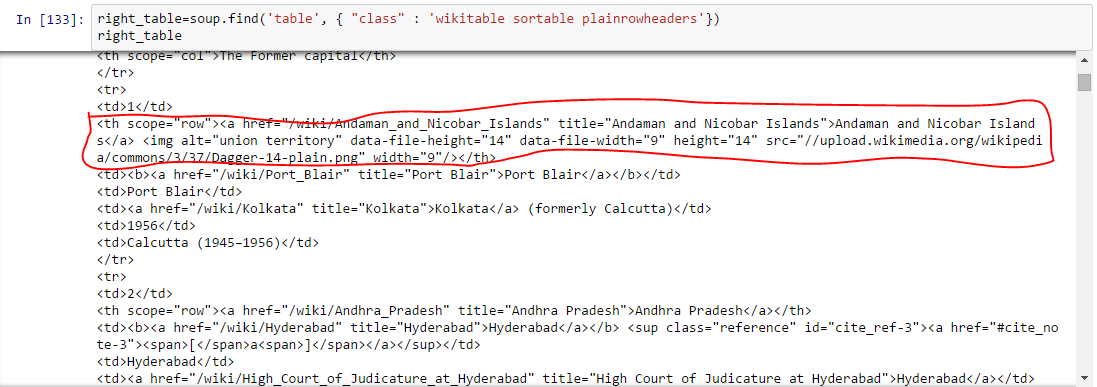
1. **Find the right table:**As we are seeking a table to extract information about state capitals, we should identify the right table first. Let’s write the command to extract information within all **table** tags.
2. all\_tables=soup.find\_all('table')

Now to identify the right table, we will use attribute “class” of table and use it to filter the right table. In chrome, you can check the class name by right click on the required table of web page –> Inspect element –> Copy the class name OR go through the output of above command find the class name of right table.

right\_table=soup.find('table', class\_='wikitable sortable plainrowheaders')

right\_table

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/righttable.png)Above, we are able to identify right table.

1. **Extract the information to DataFrame:**Here, we need to iterate through each row (tr) and then assign each element of tr (td) to a variable and append it to a list. Let’s first look at the HTML structure of the table (I am not going to extract information for table heading <th>)  
   [](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/Structure.png)Above, you can notice that second element of <tr> is within tag <th> not <td> so we need to take care for this. Now to access value of each element, we will use “find(text=True)” option with each element.  Let’s look at the code:

#Generate lists

A=[]

B=[]

C=[]

D=[]

E=[]

F=[]

G=[]

for row in right\_table.findAll("tr"):

cells = row.findAll('td')

states=row.findAll('th') #To store second column data

if len(cells)==6: #Only extract table body not heading

A.append(cells[0].find(text=True))

B.append(states[0].find(text=True))

C.append(cells[1].find(text=True))

D.append(cells[2].find(text=True))

E.append(cells[3].find(text=True))

F.append(cells[4].find(text=True))

G.append(cells[5].find(text=True))

#import pandas to convert list to data frame

import pandas as pd

df=pd.DataFrame(A,columns=['Number'])

df['State/UT']=B

df['Admin\_Capital']=C

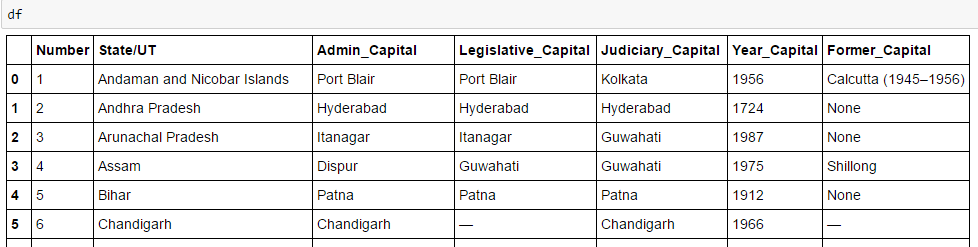
df['Legislative\_Capital']=D

df['Judiciary\_Capital']=E

df['Year\_Capital']=F

df['Former\_Capital']=G

df

Finally, we have data in dataframe:  
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/Output.png)  
Similarly, you can perform various other types of web scraping using “**BeautifulSoup**“. This will reduce your manual efforts to collect data from web pages. You can also look at the other attributes like .parent, .contents, .descendants and .next\_sibling, .prev\_sibling and various attributes to navigate using tag name. These will help you to scrap the web pages effectively.-

But, why can’t I just use Regular Expressions?

Now, if you know [regular expressions](https://www.analyticsvidhya.com/blog/2015/06/regular-expression-python/), you might be thinking that you can write code using regular expression which can do the same thing for you. I definitely had this question. In my experience with BeautifulSoup and Regular expressions to do same thing I found out:

* Code written in BeautifulSoup is usually more robust than the one written using regular expressions. Codes written with regular expressions need to be altered with any changes in pages. Even BeautifulSoup needs that in some cases, it is just that BeautifulSoup is relatively better.
* Regular expressions are much faster than BeautifulSoup, usually by a factor of 100 in giving the same outcome.

So, it boils down to speed vs. robustness of the code and there is no universal winner here. If the information you are looking for can be extracted with simple regex statements, you should go ahead and use them. For almost any complex work, I usually recommend BeautifulSoup more than regex.