Web Scraping Wikipedia Tables using BeautifulSoup and Python

Stewyn Chaudhary Follow
May 1, 2018 · 3 min read

Source: SixFeetUp

'Data is the new oil'

As an aspiring data scientist, I do a lot of projects which involve scraping data from various websites. Some companies like Twitter do provide APIs to get their information in a more organized way while we have to scrape other websites to get data in a structured format.

The general idea behind web scraping is to retrieve data that exists on a website and convert it into a format that is usable for analysis. In this tutorial, I will be going through a detail but simple explanation of how to scrape data in Python using BeautifulSoup. I will be scraping Wikipedia to find out all the countries in Asia.

Rank	Country	Area (km²)	Notes
1	Russia*	13,100,000	17,125,200 including European part
2	China	9,596,961	excludes Hong Kong, Macau, Taiwan and disputed areas/islands
3	India[1]	3,287,263	
4	Kazakhstan*	2,455,034	2,724,902 km² including European part
5	Saudi Arabia	2,149,690	
6	Iran	1,648,195	
7	Mongolia	1,564,110	
8	Indonesia*	1,472,639	1,904,569 km² including Oceanian part
9	Pakistan	796,095	882,363 km² including Gilgit-Baltistan and AJK
10	Turkey*	747,272	783,562 km² including European part
11	Myanmar	676,578	
12	Afghanistan	652,230	
13	Yemen	527,968	
14	Thailand	513,120	

Table with names of Asian countries on Wiki

Firstly we are going to **import requests** library. Requests allows you to send *organic, grass-fed* HTTP/1.1 requests, without the need for manual labor.

import requests

Now we assign the link of the website through which we are going to scrape the data and assign it to variable named **website_url**.

requests.get(url).text will ping a website and return you HTML of the website.

```
website_url =
requests.get('https://en.wikipedia.org/wiki/List_of_Asian_countries_by_area').text
```

We begin by reading the source code for a given web page and creating a BeautifulSoup (soup)object with the BeautifulSoup function. Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be

used to extract data from HTML, which is useful for web scraping. Prettify() function in BeautifulSoup will enable us to view how the tags are nested in the document.

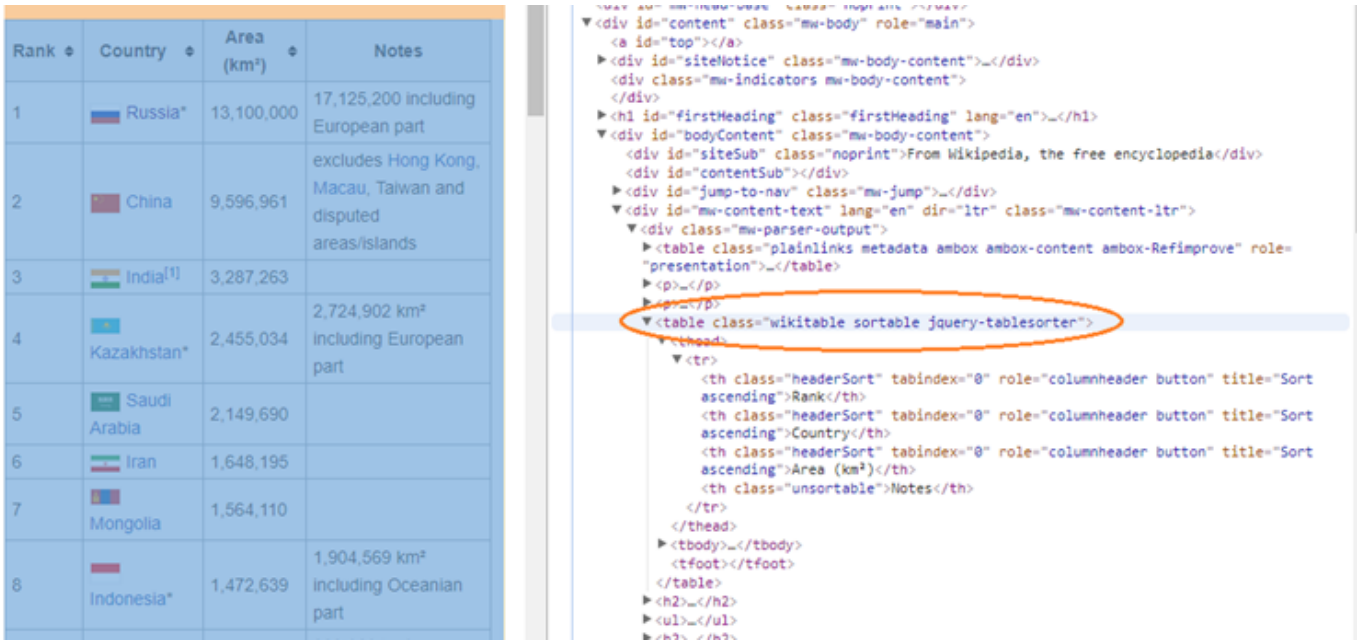
> *from bs4 import BeautifulSoup*
> *soup = BeautifulSoup(website_url,'lxml')*
> *print(soup.prettify())*

```
import requests
website_url = requests.get('https://en.wikipedia.org/wiki/List_of_Asian_countries_by_area').text

from bs4 import BeautifulSoup
soup = BeautifulSoup(website_url,'lxml')
print(soup.prettify())
```

```
<!DOCTYPE html>
<html class="client-nojs" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   List of Asian countries by area - Wikipedia
  </title>
  <script>
   document.documentElement.className = document.documentElement.className.replace( /(^|\s)client-nojs(\s|$)/, "$1client-js
$2" );
```

If you carefully inspect the HTML script all the table contents i.e. names of the countries which we intend to extract is under class Wikitable Sortable.



Rank	Country	Area (km²)	Notes
1	Russia*	13,100,000	17,125,200 including European part
2	China	9,596,961	excludes Hong Kong, Macau, Taiwan and disputed areas/islands
3	India[1]	3,287,263	
4	Kazakhstan*	2,455,034	2,724,902 km² including European part
5	Saudi Arabia	2,149,690	
6	Iran	1,648,195	
7	Mongolia	1,564,110	
8	Indonesia*	1,472,639	1,904,569 km² including Oceanian part

So our first task is to find class 'wikitable sortable' in the HTML script.

> *My_table = soup.find('table',{'class':'wikitable sortable'})*

Under table class 'wikitable sortable' we have links with country name as title.

```
My_table = soup.find('table',{'class':'wikitable sortable'})
My_table
```

```
<table class="wikitable sortable">
<tr>
<th>Rank</th>
<th>Country</th>
<th>Area (km²)</th>
<th class="unsortable">Notes</th>
</tr>
<tr>
<td>1</td>
<td><span class="flagicon" style="display:inline-block;width:25px;"><img alt="" class="thumbborder" data-file-height="600" da
ta-file-width="900" height="15" src="//upload.wikimedia.org/wikipedia/en/thumb/f/f3/Flag_of_Russia.svg/23px-Flag_of_Russia.sv
g.png" srcset="//upload.wikimedia.org/wikipedia/en/thumb/f/f3/Flag_of_Russia.svg/35px-Flag_of_Russia.svg.png 1.5x, //upload.w
ikimedia.org/wikipedia/en/thumb/f/f3/Flag_of_Russia.svg/45px-Flag_of_Russia.svg.png 2x" width="23"/></span> <a href="/wiki/Ru
ssia" title="Russia">Russia</a>*</td>
<td>13,100,000</td>
<td>17,125,200 including European part</td>
</tr>
<tr>
<td>2</td>
```

Now to extract all the links within <a>, we will use **find_all().**

```
links = My_table.findAll('a')
links
```

```
[<a href="/wiki/Russia" title="Russia">Russia</a>,
 <a href="/wiki/China" title="China">China</a>,
 <a href="/wiki/Hong_Kong" title="Hong Kong">Hong Kong</a>,
 <a href="/wiki/Macau" title="Macau">Macau</a>,
 <a href="/wiki/India" title="India">India</a>,
 <a href="#cite_note-1">[1]</a>,
 <a href="/wiki/Kazakhstan" title="Kazakhstan">Kazakhstan</a>,
 <a href="/wiki/Saudi_Arabia" title="Saudi Arabia">Saudi Arabia</a>,
 <a href="/wiki/Iran" title="Iran">Iran</a>,
 <a href="/wiki/Mongolia" title="Mongolia">Mongolia</a>,
 <a href="/wiki/Indonesia" title="Indonesia">Indonesia</a>,
 <a href="/wiki/Pakistan" title="Pakistan">Pakistan</a>,
 <a href="/wiki/Gilgit-Baltistan" title="Gilgit-Baltistan">Gilgit-Baltistan</a>,
 <a href="/wiki/Azad_Kashmir" title="Azad Kashmir">AJK</a>,
 <a href="/wiki/Turkey" title="Turkey">Turkey</a>,
```

From the links, we have to extract the title which is the name of countries.

To do that we create a list **Countries** so that we can extract the name of countries from the link and append it to the list countries.

```
Countries = []
for link in links:
    Countries.append(link.get('title'))

print(Countries)
```

```
['Russia', 'China', 'Hong Kong', 'Macau', 'India', None, 'Kazakhstan', 'Saudi Arabia', 'Iran', 'Mongolia', 'Indonesia', 'Pakist
```

```
an', 'Gilgit-Baltistan', 'Azad Kashmir', 'Turkey', 'Myanmar', 'Afghanistan', 'Yemen', 'Thailand', 'Turkmenistan', 'Uzbekistan',
'Iraq', 'Japan', 'Vietnam', 'Malaysia', 'Oman', 'Philippines', 'Laos', 'Kyrgyzstan', 'Syria', 'Golan Heights', 'Cambodia', 'Ban
gladesh', 'Nepal', 'Tajikistan', 'North Korea', 'South Korea', 'Jordan', 'Azerbaijan', 'United Arab Emirates', 'Georgia (countr
y)', 'Sri Lanka', 'Egypt', 'Bhutan', 'Taiwan', 'Armenia', 'Israel', 'Kuwait', 'East Timor', 'Qatar', 'Lebanon', 'Cyprus', 'Nort
hern Cyprus', 'State of Palestine', 'Brunei', 'Bahrain', 'Singapore', 'Maldives']
```

Convert the list countries into Pandas DataFrame to work in python.

Thank you for reading my first article on Medium. I will make it a point to write regularly about my journey towards Data Science. Thanks again for choosing to spend your time here — means the world.

You can find my code on Github.

Beautifulsoup Wikipedia Website Scraping Scrapy Programming

About Help Legal