WEB-SCRAPING in Python

Scraping a webpage:

Extracting the unstructured data used website from wikipidia (data is List of Asian countries by area)

In [1]:

```
#import the library to query a website
import requests
#specify the url
wiki_link="https://en.wikipedia.org/wiki/List_of_Asian_countries_by_area"
link=requests.get(wiki_link).text
```

In [2]:

```
print(link)
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-
feature-language-in-main-page-header-disabled vector-feature-language-aler
t-in-sidebar-enabled vector-feature-sticky-header-disabled vector-feature-
page-tools-disabled vector-feature-page-tools-pinned-disabled vector-featu
re-main-menu-pinned-disabled vector-feature-limited-width-enabled vector-f
eature-limited-width-content-enabled" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>List of Asian countries by area - Wikipedia</title>
<script>document.documentElement.className="client-js vector-feature-langu
age-in-header-enabled vector-feature-language-in-main-page-header-disabled
vector-feature-language-alert-in-sidebar-enabled vector-feature-sticky-hea
der-disabled vector-feature-page-tools-disabled vector-feature-page-tools-
pinned-disabled vector-feature-main-menu-pinned-disabled vector-feature-li
mited-width-enabled vector-feature-limited-width-content-enabled";(functio
n(){var cookie=document.cookie.match(/(?:^|; )enwikimwclientprefs=([^;]
+)/);if(cookie){var featureName=cookie[1];document.documentElement.classNa
me=document.documentElement.className.replace(featureName+'-enabled',featu
```

In [3]:

```
from bs4 import BeautifulSoup
soup=BeautifulSoup(link,'lxml')
print(soup)
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-
feature-language-in-main-page-header-disabled vector-feature-language-aler
t-in-sidebar-enabled vector-feature-sticky-header-disabled vector-feature-
page-tools-disabled vector-feature-page-tools-pinned-disabled vector-featu
re-main-menu-pinned-disabled vector-feature-limited-width-enabled vector-f
eature-limited-width-content-enabled" dir="ltr" lang="en">
<head>
<meta charset="utf-8"/>
<title>List of Asian countries by area - Wikipedia</title>
<script>document.documentElement.className="client-js vector-feature-langu
age-in-header-enabled vector-feature-language-in-main-page-header-disabled
vector-feature-language-alert-in-sidebar-enabled vector-feature-sticky-hea
der-disabled vector-feature-page-tools-disabled vector-feature-page-tools-
pinned-disabled vector-feature-main-menu-pinned-disabled vector-feature-li
mited-width-enabled vector-feature-limited-width-content-enabled";(functio
n(){var cookie=document.cookie.match(/(?:^|; )enwikimwclientprefs=([^;]
+)/);if(cookie){var featureName=cookie[1];document.documentElement.classNa
me=document.documentElement.className.replace(featureName+'-enabled',featu
```

In [4]:

```
print(soup.prettify())
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-
feature-language-in-main-page-header-disabled vector-feature-language-aler
t-in-sidebar-enabled vector-feature-sticky-header-disabled vector-feature-
page-tools-disabled vector-feature-page-tools-pinned-disabled vector-featu
re-main-menu-pinned-disabled vector-feature-limited-width-enabled vector-f
eature-limited-width-content-enabled" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   List of Asian countries by area - Wikipedia
  </title>
  <script>
   document.documentElement.className="client-js vector-feature-language-i
n-header-enabled vector-feature-language-in-main-page-header-disabled vect
or-feature-language-alert-in-sidebar-enabled vector-feature-sticky-header-
disabled vector-feature-page-tools-disabled vector-feature-page-tools-pinn
ed-disabled vector-feature-main-menu-pinned-disabled vector-feature-limite
d-width-enabled vector-feature-limited-width-content-enabled";(function()
```

In [5]:

```
soup.title
```

Out[5]:

```
<title>List of Asian countries by area - Wikipedia</title>
```

In [6]:

```
soup.title.string
```

Out[6]:

```
'List of Asian countries by area - Wikipedia'
```

In [7]:

```
soup.a
```

Out[7]:

```
<a class="mw-jump-link" href="#bodyContent">Jump to content</a>
```

In [8]:

```
soup.find_all("a")
```

Out[8]:

```
[<a class="mw-jump-link" href="#bodyContent">Jump to content</a>,
 <a class="mw-logo" href="/wiki/Main_Page">
 <img alt="" aria-hidden="true" class="mw-logo-icon" height="50" src="/sta
tic/images/icons/wikipedia.png" width="50"/>
 <span class="mw-logo-container">
 <img alt="Wikipedia" class="mw-logo-wordmark" src="/static/images/mobile/
copyright/wikipedia-wordmark-en.svg" style="width: 7.5em; height: 1.125e
m;"/>
 <img alt="The Free Encyclopedia" class="mw-logo-tagline" height="13" src
="/static/images/mobile/copyright/wikipedia-tagline-en.svg" style="width:
7.3125em; height: 0.8125em;" width="117"/>
 </span>
 </a>,
 <a accesskey="f" class="mw-ui-button mw-ui-quiet mw-ui-icon mw-ui-icon-el
ement mw-ui-icon-wikimedia-search search-toggle" href="/wiki/Special:Searc
h" title="Search Wikipedia [f]">
 <span>Search</span>
 </a>,
```

In [9]:

```
all_link=soup.find_all("a")
for link in all_link:
    print(link.get("href"))
```

```
#bodyContent
/wiki/Main_Page
/wiki/Special:Search
/w/index.php?title=Special:CreateAccount&returnto=List+of+Asian+countries+
by+area
/w/index.php?title=Special:CreateAccount&returnto=List+of+Asian+countries+
by+area
/w/index.php?title=Special:UserLogin&returnto=List+of+Asian+countries+by+a
rea
/wiki/Help:Introduction
/wiki/Special:MyTalk
/wiki/Special:MyContributions
/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_source=
```

In [11]:

```
right_table=soup.find('table',class_='wikitable sortable')
right_table
```

Out[11]:

```
<table class="wikitable sortable">
<tbody><tr>
<th rowspan="2">Rank
</th>
<th rowspan="2">Country
</th>
<th colspan="2">Area
</th>
<th class="unsortable" rowspan="2">Notes
</th>
<th rowspan="2">Facts
</th></tr>
<tr>
<th>km²
</th>
<th>sq mi
</th></tr>
<tr>
```

In [10]:

```
all_tables=soup.find_all('table')
print(all_tables)
```

```
[<table class="box-More_citations_needed plainlinks metadata ambox ambox-c
ontent ambox-Refimprove" role="presentation"><tbody><tr><td class="mbox-im
age"><div class="mbox-image-div"><a class="image" href="/wiki/File:Questio
n_book-new.svg"><img alt="" data-file-height="399" data-file-width="512" d
ecoding="async" height="39" src="//upload.wikimedia.org/wikipedia/en/thum
b/9/99/Question_book-new.svg/50px-Question_book-new.svg.png" srcset="//upl
oad.wikimedia.org/wikipedia/en/thumb/9/99/Question_book-new.svg/75px-Quest
ion_book-new.svg.png 1.5x, //upload.wikimedia.org/wikipedia/en/thumb/9/99/
Question_book-new.svg/100px-Question_book-new.svg.png 2x" width="50"/></a>
</div></td><td class="mbox-text"><div class="mbox-text-span">This article
<b>needs additional citations for <a href="/wiki/Wikipedia:Verifiability"
title="Wikipedia:Verifiability">verification</a></b>.<span class="hide-whe
n-compact"> Please help <a class="external text" href="https://en.wikipedi
a.org/w/index.php?title=List_of_Asian_countries_by_area&amp;action=edit">i
mprove this article</a> by <a href="/wiki/Help:Referencing_for_beginners"
title="Help:Referencing for beginners">adding citations to reliable source
s</a>. Unsourced material may be challenged and removed.<br/><small><span
class="plainlinks"><i>Find sources:</i> <a class="external text" href="//w
ww.google.com/search?as_eq=wikipedia&amp;q=%22List+of+Asian+countries+by+a
```

In [13]:

```
table_links=right_table.findAll('a')
table_links
```

Out[13]:

```
[<a href="/wiki/Russia" title="Russia">Russia</a>,
 <a href="/wiki/European_Russia" title="European Russia">European Russia</a
>,
 <a href="#cite_note-russiaTotalAreaByCIA-1">[1]</a>,
 <a href="/wiki/China" title="China">China</a>,
 <a href="/wiki/Taiwan" title="Taiwan">Taiwan</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="/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/Western_New_Guinea" title="Western New Guinea">Indonesian Pa
pua</a>,
 <a href="/wiki/Oceania" title="Oceania">Oceania</a>,
 <a href="/wiki/Pakistan" title="Pakistan">Pakistan</a>,
 <a href="/wiki/Turkey" title="Turkey">Turkey</a>,
 <a href="/wiki/East_Thrace" title="East Thrace">European Turkey</a>,
 <a href="/wiki/Myanmar" title="Myanmar">Myanmar</a>,
 <a href="/wiki/Afghanistan" title="Afghanistan">Afghanistan</a>,
 <a href="/wiki/Yemen" title="Yemen">Yemen</a>,
 <a href="/wiki/Thailand" title="Thailand">Thailand</a>,
 <a href="/wiki/Turkmenistan" title="Turkmenistan">Turkmenistan</a>,
 <a href="/wiki/Uzbekistan" title="Uzbekistan">Uzbekistan</a>,
 <a href="/wiki/Iraq" title="Iraq">Iraq</a>,
 <a href="/wiki/Japan" title="Japan">Japan</a>,
 <a href="/wiki/Vietnam" title="Vietnam">Vietnam</a>,
 <a href="/wiki/Malaysia" title="Malaysia">Malaysia</a>,
 <a href="/wiki/Oman" title="Oman">Oman</a>,
 <a href="/wiki/Philippines" title="Philippines">Philippines</a>,
 <a href="/wiki/Laos" title="Laos">Laos</a>,
 <a href="/wiki/Kyrgyzstan" title="Kyrgyzstan">Kyrgyzstan</a>,
 <a href="/wiki/Syria" title="Syria">Syria</a>,
 <a href="/wiki/Golan_Heights" title="Golan Heights">Golan Heights</a>,
 <a href="/wiki/Cambodia" title="Cambodia">Cambodia</a>,
 <a href="/wiki/Bangladesh" title="Bangladesh">Bangladesh</a>,
 <a href="/wiki/Nepal" title="Nepal">Nepal</a>,
 <a href="/wiki/Tajikistan" title="Tajikistan">Tajikistan</a>,
 <a href="/wiki/North_Korea" title="North Korea">North Korea</a>,
 <a href="/wiki/South_Korea" title="South Korea">South Korea</a>,
 <a href="/wiki/Jordan" title="Jordan">Jordan</a>,
 <a href="/wiki/United_Arab_Emirates" title="United Arab Emirates">United Ar
ab Emirates</a>,
 <a href="/wiki/Azerbaijan" title="Azerbaijan">Azerbaijan</a>,
 <a href="/wiki/Caucasus" title="Caucasus">Caucasus</a>,
 <a href="/wiki/Europe" title="Europe">Europe</a>,
 <a href="/wiki/Asia" title="Asia">Asia</a>,
 <a href="/wiki/Georgia_(country)" title="Georgia (country)">Georgia</a>,
 <a href="/wiki/Caucasus" title="Caucasus">Caucasus</a>,
 <a href="/wiki/Europe" title="Europe">Europe</a>,
 <a href="/wiki/Asia" title="Asia">Asia</a>,
 <a href="/wiki/Sri_Lanka" title="Sri Lanka">Sri Lanka</a>,
```

```
 <a href="/wiki/Egypt" title="Egypt">Egypt</a>,
 <a href="/wiki/Bhutan" title="Bhutan">Bhutan</a>,
 <a href="/wiki/Taiwan" title="Taiwan">Taiwan</a>,
 <a href="/wiki/Free_area_of_the_Republic_of_China" title="Free area of the
Republic of China">free area of the Republic of China</a>,
 <a href="/wiki/Armenia" title="Armenia">Armenia</a>,
 <a href="/wiki/Armenian_highlands" title="Armenian highlands">Armenian high
lands</a>,
 <a href="/wiki/Caucasus" title="Caucasus">Caucasus</a>,
 <a href="/wiki/Europe" title="Europe">Europe</a>,
 <a href="/wiki/Asia" title="Asia">Asia</a>,
 <a href="/wiki/Israel" title="Israel">Israel</a>,
 <a href="/wiki/West_Bank" title="West Bank">West Bank</a>,
 <a href="/wiki/Gaza_Strip" title="Gaza Strip">Gaza Strip</a>,
 <a href="/wiki/Golan_Heights" title="Golan Heights">Golan Heights</a>,
 <a href="/wiki/Kuwait" title="Kuwait">Kuwait</a>,
 <a href="/wiki/East_Timor" title="East Timor">Timor-Leste</a>,
 <a href="/wiki/Qatar" title="Qatar">Qatar</a>,
 <a href="/wiki/Lebanon" title="Lebanon">Lebanon</a>,
 <a href="/wiki/Cyprus" title="Cyprus">Cyprus</a>,
 <a href="/wiki/Northern_Cyprus" title="Northern Cyprus">Northern Cyprus</a
>,
 <a href="/wiki/State_of_Palestine" title="State of Palestine">Palestine</a
>,
 <a href="/wiki/West_Bank" title="West Bank">West Bank</a>,
 <a href="/wiki/Gaza_Strip" title="Gaza Strip">Gaza Strip</a>,
 <a href="/wiki/Brunei" title="Brunei">Brunei</a>,
 <a href="/wiki/Bahrain" title="Bahrain">Bahrain</a>,
 <a href="/wiki/Singapore" title="Singapore">Singapore</a>,
 <a href="/wiki/Maldives" title="Maldives">Maldives</a>]
```

In [14]:

```
country=[]
for links in table_links:
    country.append(links.get('title'))
print(country)
```

```
['Russia', 'European Russia', None, 'China', 'Taiwan', 'Hong Kong', 'Macau',
'India', 'Kazakhstan', 'Saudi Arabia', 'Iran', 'Mongolia', 'Indonesia', 'Wes
tern New Guinea', 'Oceania', 'Pakistan', 'Turkey', 'East Thrace', 'Myanmar',
'Afghanistan', 'Yemen', 'Thailand', 'Turkmenistan', 'Uzbekistan', 'Iraq', 'J
apan', 'Vietnam', 'Malaysia', 'Oman', 'Philippines', 'Laos', 'Kyrgyzstan',
'Syria', 'Golan Heights', 'Cambodia', 'Bangladesh', 'Nepal', 'Tajikistan',
'North Korea', 'South Korea', 'Jordan', 'United Arab Emirates', 'Azerbaija
n', 'Caucasus', 'Europe', 'Asia', 'Georgia (country)', 'Caucasus', 'Europe',
'Asia', 'Sri Lanka', 'Egypt', 'Bhutan', 'Taiwan', 'Free area of the Republic
of China', 'Armenia', 'Armenian highlands', 'Caucasus', 'Europe', 'Asia', 'I
srael', 'West Bank', 'Gaza Strip', 'Golan Heights', 'Kuwait', 'East Timor',
'Qatar', 'Lebanon', 'Cyprus', 'Northern Cyprus', 'State of Palestine', 'West
Bank', 'Gaza Strip', 'Brunei', 'Bahrain', 'Singapore', 'Maldives']
```

In [15]:

```
#represent of DataFrame
import pandas as pd
df=pd.DataFrame()
df['Country']=country
df
```

Out[15]:

	Country
0	Russia
1	European Russia
2	None
3	China
4	Taiwan
...	...
72	Gaza Strip
73	Brunei
74	Bahrain
75	Singapore
76	Maldives

77 rows × 1 columns

In [ ]:

```

```