In [1]:

```
# Name: - L Prathyusha
```

In [2]:

```
# WEB SCRAPING
```

In [3]:

```
import requests

res = requests.get('https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_pop

print(res.text)
print(res.status_code)
```

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>List of countries and dependencies by population - Wikipedia</title
>
<script>document.documentElement.className="client-js";RLCONF={"wgBreakFra
mes":!1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":
["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","Februar
y","March","April","May","June","July","August","September","October","Nov
ember","December"],"wgRequestId":"c7590c4c-57a0-4a05-b6e4-66e06843645a","w
gCSPNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,"w
gNamespaceNumber":0,"wgPageName":"List_of_countries_and_dependencies_by_po
pulation","wgTitle":"List of countries and dependencies by population","wg
CurRevisionId":1036683771,"wgRevisionId":1036683771,"wgArticleId":69058,"w
gIsArticle":!0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":null,"wgUs
erGroups":["*"],"wgCategories":["Pages with non-numeric formatnum argument
s","CS1 Indonesian-language sources (id)","CS1 Arabic-language sources (a
r)","CS1 Romanian-language sources (ro)","CS1 errors: URL","CS1 German-lan
```

In [4]:

```
#title
from bs4 import BeautifulSoup

page = requests.get("https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_po
soup = BeautifulSoup(page.content, 'html.parser')
page_title = soup.title.text

# print the result
print(page_title)
```

```
List of countries and dependencies by population - Wikipedia
```

In [5]:

```
#body and head
import requests
from bs4 import BeautifulSoup

# Make a request
page = requests.get("https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_po
soup = BeautifulSoup(page.content, 'html.parser')

# Extract title of page
page_title = soup.title.text

# Extract body of page
page_body = soup.body

# Extract head of page
page_head = soup.head

# print the result
print(page_body, page_head)
```

```
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject m
w-editable page-List_of_countries_and_dependencies_by_population rootpage-
List_of_countries_and_dependencies_by_population skin-vector action-view s
kin-vector-legacy"><div class="noprint" id="mw-page-base"></div>
<div class="noprint" id="mw-head-base"></div>
<div class="mw-body" id="content" role="main">
<a id="top"></a>
<div id="siteNotice"><!-- CentralNotice --></div>
<div class="mw-indicators">
</div>
<h1 class="firstHeading" id="firstHeading">List of countries and dependenc
ies by population</h1>
<div class="vector-body" id="bodyContent">
<div class="noprint" id="siteSub">From Wikipedia, the free encyclopedia</d
iv>
<div id="contentSub"></div>
<div id="contentSub2"></div>
<div id="jump-to-nav"></div>
<a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
```

In [6]:

```
import requests
from bs4 import BeautifulSoup
# Make a request
page = requests.get("https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_po
soup = BeautifulSoup(page.content, 'html.parser')

# Create top_items as empty list
image_data = []

# Extract and store in top_items according to instructions on the left
images = soup.select('img')
for image in images:
    src = image.get('src')
    alt = image.get('alt')
    image_data.append({"src": src, "alt": alt})

print(image_data)
```

```
[{'src': '//upload.wikimedia.org/wikipedia/commons/thumb/a/a5/World_Popula
tion.svg/280px-World_Population.svg.png', 'alt': ''}, {'src': '//upload.wi
kimedia.org/wikipedia/commons/thumb/f/fa/Flag_of_the_People%27s_Republic_o
f_China.svg/23px-Flag_of_the_People%27s_Republic_of_China.svg.png', 'alt':
''}, {'src': '//upload.wikimedia.org/wikipedia/en/thumb/4/41/Flag_of_Indi
a.svg/23px-Flag_of_India.svg.png', 'alt': ''}, {'src': '//upload.wikimedi
a.org/wikipedia/en/thumb/a/a4/Flag_of_the_United_States.svg/23px-Flag_of_t
he_United_States.svg.png', 'alt': ''}, {'src': '//upload.wikimedia.org/wik
ipedia/commons/thumb/9/9f/Flag_of_Indonesia.svg/23px-Flag_of_Indonesia.sv
g.png', 'alt': ''}, {'src': '//upload.wikimedia.org/wikipedia/commons/thum
b/3/32/Flag_of_Pakistan.svg/23px-Flag_of_Pakistan.svg.png', 'alt': ''},
{'src': '//upload.wikimedia.org/wikipedia/en/thumb/0/05/Flag_of_Brazil.sv
g/22px-Flag_of_Brazil.svg.png', 'alt': ''}, {'src': '//upload.wikimedia.or
g/wikipedia/commons/thumb/7/79/Flag_of_Nigeria.svg/23px-Flag_of_Nigeria.sv
g.png', 'alt': ''}, {'src': '//upload.wikimedia.org/wikipedia/commons/thum
b/f/f9/Flag_of_Bangladesh.svg/23px-Flag_of_Bangladesh.svg.png', 'alt':
''}, {'src': '//upload.wikimedia.org/wikipedia/en/thumb/f/f3/Flag_of_Russi
a.svg/23px-Flag_of_Russia.svg.png', 'alt': ''}, {'src': '//upload.wikimedi
a.org/wikipedia/en/thumb/9/9e/Flag_of_Japan.svg/23px-Flag_of_Japan.svg.pn
```

In [7]:

```
all_links = []
links = soup.select('a')
for ahref in links:
    text = ahref.text
    text = text.strip() if text is not None else ''

    href = ahref.get('href')
    href = href.strip() if href is not None else ''
    all_links.append({"href": href, "text": text})

print(all_links)
```

```
[{'href': '', 'text': ''}, {'href': '#mw-head', 'text': 'Jump to navigatio
n'}, {'href': '#searchInput', 'text': 'Jump to search'}, {'href': '/wiki/F
ile:World_Population.svg', 'text': ''}, {'href': '/wiki/File:World_Populat
ion.svg', 'text': ''}, {'href': '/wiki/Sovereign_state', 'text': 'sovereig
n states'}, {'href': '/wiki/Dependent_territory', 'text': 'dependent terri
tories'}, {'href': '/wiki/Country#Sovereignty', 'text': 'constituent count
ries'}, {'href': '/wiki/ISO', 'text': 'ISO'}, {'href': '/wiki/ISO_3166-1',
'text': 'ISO 3166-1'}, {'href': '/wiki/United_Kingdom', 'text': 'United Ki
ngdom'}, {'href': '/wiki/Kingdom_of_the_Netherlands', 'text': 'Kingdom of
the Netherlands'}, {'href': '/wiki/List_of_states_with_limited_recognitio
n', 'text': 'states with limited recognition'}, {'href': '/wiki/World_popu
lation', 'text': 'world population'}, {'href': '/wiki/United_Nations', 'te
xt': 'United Nations'}, {'href': '#Method', 'text': '1 Method'}, {'href':
'#Sovereign_states_and_dependencies_by_population', 'text': '2 Sovereign s
tates and dependencies by population'}, {'href': '#Notes', 'text': '3 Note
s'}, {'href': '#References', 'text': '4 References'}, {'href': '/w/index.p
hp?title=List_of_countries_and_dependencies_by_population&action=edit&sect
ion=1', 'text': 'edit'}, {'href': '/wiki/List_of_countries_and_dependencie
s_by_population_density', 'text': 'List of countries and dependencies by p
```

In [ ]:

```
import csv
import requests
from bs4 import BeautifulSoup


def scrape_data(url):

    response = requests.get(url, timeout=10)
    soup = BeautifulSoup(response.content, 'html.parser')

    table = soup.find_all('table')[1]

    rows = table.select('tbody > tr')

    header = [th.text.rstrip() for th in rows[0].find_all('th')]

    with open('output.csv', 'w') as csv_file:
        writer = csv.writer(csv_file)
        writer.writerow(header)
        for row in rows[1:]:
            data = [th.text.rstrip() for th in row.find_all('td')]
            writer.writerow(data)


if __name__=="__main__":
    url = "https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population"
    scrape_data(url)
```