Web Scraping Mini Project

In [75]:
```
from bs4 import BeautifulSoup
import requests
import pandas as pd
```

In [2]:
```
url = 'https://github.com/topics'
response = requests.get(url)
```

In [3]:
```
response.status_code
```

Out[3]:
```
200
```

In [4]:
```
page_contents = response.text
```

In [5]:
```
page_contents[:1000]
```

Out[5]:
```
'\n\n<!DOCTYPE html>\n<html lang="en" data-color-mode="auto" data-light-theme="light" da
ta-dark-theme="dark">\n  <head>\n    <meta charset="utf-8">\n  <link rel="dns-prefetch"
href="https://github.githubassets.com">\n  <link rel="dns-prefetch" href="https://avatar
s.githubusercontent.com">\n  <link rel="dns-prefetch" href="https://github-cloud.s3.amaz
onaws.com">\n  <link rel="dns-prefetch" href="https://user-images.githubusercontent.co
m/">\n  <link rel="preconnect" href="https://github.githubassets.com" crossorigin>\n  <l
ink rel="preconnect" href="https://avatars.githubusercontent.com">\n\n\n\n  <link crosso
rigin="anonymous" media="all" integrity="sha512-E9wnWjoxQmh5A1jiWVYDPKOvA8VPf0iKQYoc+9yc
MJvtAi9gOSlaUci+W2smxFIlWkV8hkX+O27S8NIB59iIDw==" rel="stylesheet" href="https://github.
githubassets.com/assets/light-13dc275a3a314268790358e25956033c.css" /><link crossorigin
="anonymous" media="all" integrity="sha512-nYSv3KrFhMlGUpjkFQBLMEN6HvHhijcoubQLjV3DWlcAB
Ei2yDYf6KGUjRubJ5R+dJnKXR7jA4wu5Dg200SApA==" rel="s'
```

In [7]:
```
with open('webpage.html','w',encoding='utf-8') as f:
    f.write(page_contents)
```

Beautiful Soup starts

In [12]:
```
doc = BeautifulSoup(page_contents,'html.parser')
```

Extract the titles of all topics

In [29]:
```
p_tag = doc.find_all('p',class_='f3 lh-condensed mb-0 mt-1 Link--primary')
topic_title = []
for title in p_tag:
    topic_title.append(title.text)
print(topic_title)
```

```
['3D', 'Ajax', 'Algorithm', 'Amp', 'Android', 'Angular', 'Ansible', 'API', 'Arduino', 'A
SP.NET', 'Atom', 'Awesome Lists', 'Amazon Web Services', 'Azure', 'Babel', 'Bash', 'Bitc
oin', 'Bootstrap', 'Bot', 'C', 'Chrome', 'Chrome extension', 'Command line interface',
'Clojure', 'Code quality', 'Code review', 'Compiler', 'Continuous integration', 'COVID-1
9', 'C++']
```

Extract Description of Topics

In [35]:

```
p_tag_desc = doc.find_all('p',class_='f5 color-fg-muted mb-0 mt-1')
topic_desc = []
for desc in p_tag_desc:
    desc_text = desc.text
    topic_desc.append(desc_text)
topic_desc
```

Out[35]:

```
['\n          3D modeling is the process of virtually developing the surface and structu
re of a 3D object.\n        ',
 '\n          Ajax is a technique for creating interactive web applications.\n        ',
 '\n          Algorithms are self-contained sequences that carry out a variety of task
s.\n        ',
 '\n          Amp is a non-blocking concurrency framework for PHP.\n        ',
 '\n          Android is an operating system built by Google designed for mobile device
s.\n        ',
 '\n          Angular is an open source web application platform.\n        ',
 '\n          Ansible is a simple and powerful automation engine.\n        ',
 '\n          An API (Application Programming Interface) is a collection of protocols an
d subroutines for building software.\n        ',
 '\n          Arduino is an open source hardware and software company and maker communit
y.\n        ',
 '\n          ASP.NET is a web framework for building modern web apps and services.\n
',
 '\n          Atom is a open source text editor built with web technologies.\n        ',
 '\n          An awesome list is a list of awesome things curated by the community.\n
',
 '\n          Amazon Web Services provides on-demand cloud computing platforms on a subs
cription basis.\n        ',
 '\n          Azure is a cloud computing service created by Microsoft.\n        ',
 '\n          Babel is a compiler for writing next generation JavaScript, today.\n
',
 '\n          Bash is a shell and command language interpreter for the GNU operating sys
tem.\n        ',
 '\n          Bitcoin is a cryptocurrency developed by Satoshi Nakamoto.\n        ',
 '\n          Bootstrap is an HTML, CSS, and JavaScript framework.\n        ',
 '\n          A bot is an application that runs automated tasks over the Internet.\n
',
 '\n          C is a general purpose programming language that first appeared in 1972.\n
',
 '\n          Chrome is a web browser from the tech company Google.\n        ',
 '\n          Google Chrome Extensions are add-ons that allow users to customize their C
hrome web browser.\n        ',
 '\n          A CLI, or command-line interface, is a console that helps users issue comm
ands to a program.\n        ',
 '\n          Clojure is a dynamic, general-purpose programming language.\n        ',
 '\n          Automate your code review with style, quality, security, and test-coverage
checks when you need them.\n        ',
 '\n          Ensure your code meets quality standards and ship with confidence.\n
',
 '\n          Compilers are software that translate higher-level programming languages t
o lower-level languages (e.g. machine code).\n        ',
 '\n          Automatically build and test your code as you push it upstream, preventing
```

```
bugs from being deployed to production.\n        ',
 '\n          The coronavirus disease 2019 (COVID-19) is an infectious disease caused by
SARS-CoV-2.\n        ',
 '\n          C++ is a general purpose and object-oriented programming language.\n
']
```

Extract urls of all topics title

In [74]:
```
topic_link_tag = doc.find_all('a',class_='no-underline flex-1 d-flex flex-column',href=
topic_url = []
for i in topic_link_tag:
    ok_url = 'https://github.com'+i['href']
    topic_url.append(ok_url)
#     print(i['href'])
print(topic_url)
```

```
['https://github.com/topics/3d', 'https://github.com/topics/ajax', 'https://github.com/t
opics/algorithm', 'https://github.com/topics/amphp', 'https://github.com/topics/androi
d', 'https://github.com/topics/angular', 'https://github.com/topics/ansible', 'https://g
ithub.com/topics/api', 'https://github.com/topics/arduino', 'https://github.com/topics/a
spnet', 'https://github.com/topics/atom', 'https://github.com/topics/awesome', 'https://
github.com/topics/aws', 'https://github.com/topics/azure', 'https://github.com/topics/ba
bel', 'https://github.com/topics/bash', 'https://github.com/topics/bitcoin', 'https://gi
thub.com/topics/bootstrap', 'https://github.com/topics/bot', 'https://github.com/topics/
c', 'https://github.com/topics/chrome', 'https://github.com/topics/chrome-extension', 'h
ttps://github.com/topics/cli', 'https://github.com/topics/clojure', 'https://github.com/
topics/code-quality', 'https://github.com/topics/code-review', 'https://github.com/topic
s/compiler', 'https://github.com/topics/continuous-integration', 'https://github.com/top
ics/covid-19', 'https://github.com/topics/cpp']
```

In [83]:
```
final_desc = []
for i in topic_desc:
    i = i.strip()
    final_desc.append(i)
#     print(i)
final_desc
```

Out[83]:
```
['3D modeling is the process of virtually developing the surface and structure of a 3D o
bject.',
 'Ajax is a technique for creating interactive web applications.',
 'Algorithms are self-contained sequences that carry out a variety of tasks.',
 'Amp is a non-blocking concurrency framework for PHP.',
 'Android is an operating system built by Google designed for mobile devices.',
 'Angular is an open source web application platform.',
 'Ansible is a simple and powerful automation engine.',
 'An API (Application Programming Interface) is a collection of protocols and subroutine
s for building software.',
 'Arduino is an open source hardware and software company and maker community.',
 'ASP.NET is a web framework for building modern web apps and services.',
 'Atom is a open source text editor built with web technologies.',
 'An awesome list is a list of awesome things curated by the community.',
 'Amazon Web Services provides on-demand cloud computing platforms on a subscription bas
is.',
 'Azure is a cloud computing service created by Microsoft.',
 'Babel is a compiler for writing next generation JavaScript, today.',
 'Bash is a shell and command language interpreter for the GNU operating system.',
 'Bitcoin is a cryptocurrency developed by Satoshi Nakamoto.',
 'Bootstrap is an HTML, CSS, and JavaScript framework.',
 'A bot is an application that runs automated tasks over the Internet.',
```

```
 'C is a general purpose programming language that first appeared in 1972.',
 'Chrome is a web browser from the tech company Google.',
 'Google Chrome Extensions are add-ons that allow users to customize their Chrome web br
owser.',
 'A CLI, or command-line interface, is a console that helps users issue commands to a pr
ogram.',
 'Clojure is a dynamic, general-purpose programming language.',
 'Automate your code review with style, quality, security, and test-coverage checks when
you need them.',
 'Ensure your code meets quality standards and ship with confidence.',
 'Compilers are software that translate higher-level programming languages to lower-leve
l languages (e.g. machine code).',
 'Automatically build and test your code as you push it upstream, preventing bugs from b
eing deployed to production.',
 'The coronavirus disease 2019 (COVID-19) is an infectious disease caused by SARS-CoV-
2.',
 'C++ is a general purpose and object-oriented programming language.']
```

Creating dataset from extract info

In [84]:
```
df = pd.DataFrame({
    'title':topic_title,
    'description': final_desc,
    'url': topic_url
})
```

In [85]:
```
df
```

Out[85]:

	title	description	url
0	3D	3D modeling is the process of virtually develo...	https://github.com/topics/3d
1	Ajax	Ajax is a technique for creating interactive w...	https://github.com/topics/ajax
2	Algorithm	Algorithms are self-contained sequences that c...	https://github.com/topics/algorithm
3	Amp	Amp is a non-blocking concurrency framework fo...	https://github.com/topics/amphp
4	Android	Android is an operating system built by Google...	https://github.com/topics/android
5	Angular	Angular is an open source web application plat...	https://github.com/topics/angular
6	Ansible	Ansible is a simple and powerful automation en...	https://github.com/topics/ansible
7	API	An API (Application Programming Interface) is ...	https://github.com/topics/api
8	Arduino	Arduino is an open source hardware and softwar...	https://github.com/topics/arduino
9	ASP.NET	ASP.NET is a web framework for building modern...	https://github.com/topics/aspnet

	title	description	url
10	Atom	Atom is a open source text editor built with w...	https://github.com/topics/atom
11	Awesome Lists	An awesome list is a list of awesome things cu...	https://github.com/topics/awesome
12	Amazon Web Services	Amazon Web Services provides on-demand cloud c...	https://github.com/topics/aws
13	Azure	Azure is a cloud computing service created by ...	https://github.com/topics/azure
14	Babel	Babel is a compiler for writing next generatio...	https://github.com/topics/babel
15	Bash	Bash is a shell and command language interpret...	https://github.com/topics/bash
16	Bitcoin	Bitcoin is a cryptocurrency developed by Satos...	https://github.com/topics/bitcoin
17	Bootstrap	Bootstrap is an HTML, CSS, and JavaScript fram...	https://github.com/topics/bootstrap
18	Bot	A bot is an application that runs automated ta...	https://github.com/topics/bot
19	C	C is a general purpose programming language th...	https://github.com/topics/c
20	Chrome	Chrome is a web browser from the tech company ...	https://github.com/topics/chrome
21	Chrome extension	Google Chrome Extensions are add-ons that allo...	https://github.com/topics/chrome-extension
22	Command line interface	A CLI, or command-line interface, is a console...	https://github.com/topics/cli
23	Clojure	Clojure is a dynamic, general-purpose programm...	https://github.com/topics/clojure
24	Code quality	Automate your code review with style, quality,...	https://github.com/topics/code-quality
25	Code review	Ensure your code meets quality standards and s...	https://github.com/topics/code-review
26	Compiler	Compilers are software that translate higher-l...	https://github.com/topics/compiler
27	Continuous integration	Automatically build and test your code as you ...	https://github.com/topics/continuous-integration
28	COVID-19	The coronavirus disease 2019 (COVID-19) is an ...	https://github.com/topics/covid-19
29	C++	C++ is a general purpose and object-oriented p...	https://github.com/topics/cpp

Convert the data into csv file

In [87]:

```
df.to_csv('topics.csv',index=False)
```

This is a mini web scraping poject

-----------------------*---------------

In [ ]: