Web Scraping Tool

A tool to scrape thorough top repositories of github/topic

Pick a website and describe your objective

- Browse through different sites and pick on to scrape. Check the "Project Ideas" section for inspiration.
- Identify the information you'd like to scrape from the site. Decide the format of the output CSV file.
- Summarize your project idea and outline your strategy in a Juptyer notebook. Use the "New" button above.

Outline:

- We're going to scrape https://github.com/topics
- We'll get a list of topics. For each topic, we'll get topic title, topic page url and topic description.
- For each topic, we'll get the top 25 repositories in the topic from the topic page
- For each repository, we'll grab the repo name, username, stars and repo url.
- For each topic, we'll create a CSV file in the following format:

  ```
  Repo Name, Username, Stars, Repo URL
  ```

Use the requests library to download web pages

- Inspect the website's HTML source and identify the right URLs to download.
- Download and save web pages locally using the requests library.
- Create a function to automate downloading for different topics/search queries.

```
!pip install requests --upgrade --quiet
```

```
import requests
```

```
topics_url = 'https://github.com/topics'
```

```
response = requests.get(topics_url)
```

```
response.status_code
```

```
200
```

```
len(response.text)
```

```
152970
```

```
page_contents = response.text
```

```
page_contents[:1000]
```

```
'\n\n<!DOCTYPE html>\n<html lang="en" data-color-mode="auto" data-light-theme="light"
data-dark-theme="dark" data-a11y-animated-images="system">\n  <head>\n    <meta
charset="utf-8">\n  <link rel="dns-prefetch" href="https://github.githubassets.com">\n
<link rel="dns-prefetch" href="https://avatars.githubusercontent.com">\n  <link
rel="dns-prefetch" href="https://github-cloud.s3.amazonaws.com">\n  <link rel="dns-
prefetch" href="https://user-images.githubusercontent.com/">\n  <link rel="preconnect"
href="https://github.githubassets.com" crossorigin>\n  <link rel="preconnect"
href="https://avatars.githubusercontent.com">\n\n  <link crossorigin="anonymous"
media="all" rel="stylesheet" href="https://github.githubassets.com/assets/light-
fe3f886b577a.css" /><link crossorigin="anonymous" media="all" rel="stylesheet"
href="https://github.githubassets.com/assets/dark-a1dbeda2886c.css" /><link data-color-
theme="dark_dimmed" crossorigin="anonymous" media="all" rel="stylesheet" data-
href="https://github.github'
```

```
with open('webpage.html', 'w') as f:
    f.write(page_contents)
```

```

```

Use Beautiful Soup to parse and extract information

- Parse and explore the structure of downloaded web pages using Beautiful soup.
- Use the right properties and methods to extract the required information.
- Create functions to extract from the page into lists and dictionaries.
- (Optional) Use a REST API to acquire additional information if required.

```
!pip install beautifulsoup4 --upgrade --quiet
```

```
from bs4 import BeautifulSoup
```

```
doc = BeautifulSoup(page_contents, 'html.parser')
```

```
type(doc)
```

```
bs4.BeautifulSoup
```

```
selection_class = 'f3 lh-condensed mb-0 mt-1 Link--primary'
topic_title_tags = doc.find_all('p', {'class': selection_class})
```

```
len(topic_title_tags)
```

```
30
```

```
topic_title_tags[:5]
```

```
[<p class="f3 lh-condensed mb-0 mt-1 Link--primary">3D</p>,
 <p class="f3 lh-condensed mb-0 mt-1 Link--primary">Ajax</p>,
 <p class="f3 lh-condensed mb-0 mt-1 Link--primary">Algorithm</p>,
 <p class="f3 lh-condensed mb-0 mt-1 Link--primary">Amp</p>,
 <p class="f3 lh-condensed mb-0 mt-1 Link--primary">Android</p>]
```

```
desc_selector = 'f5 color-fg-muted mb-0 mt-1'
topic_desc_tags = doc.find_all('p', {'class': desc_selector})
```

```

```

```
topic_link_tags = doc.find_all('a', {'class': 'no-underline flex-1 d-flex flex-column'}
```

```

```

```
topic_titles = []
for tag in topic_title_tags:
    topic_titles.append(tag.text)
print(topic_titles)
```

```
['3D', 'Ajax', 'Algorithm', 'Amp', 'Android', 'Angular', 'Ansible', 'API', 'Arduino',
'ASP.NET', 'Atom', 'Awesome Lists', 'Amazon Web Services', 'Azure', 'Babel', 'Bash',
'Bitcoin', 'Bootstrap', 'Bot', 'C', 'Chrome', 'Chrome extension', 'Command line
interface', 'Clojure', 'Code quality', 'Code review', 'Compiler', 'Continuous
integration', 'COVID-19', 'C++']
```

```
topic_desc = []
for desc in topic_desc_tags:
    topic_desc.append(desc.text.strip())
print(topic_desc)
```

```
['3D refers to the use of three-dimensional graphics, modeling, and animation in
various industries.', '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 library 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 subroutines for building
software.', 'Arduino is an open source platform for building electronic devices.',
```

'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 basis.', '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.', 'Chrome extensions enable users to customize the Chrome browsing experience.', 'A CLI, or command-line interface, is a console that helps users issue commands to a program.', '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-level languages (e.g. machine code).', 'Automatically build and test your code as you push it upstream, preventing bugs from being 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.']

```
topic_url = []
base_url = 'https://github.com'
for url in topic_link_tags:
    topic_url.append(base_url + url['href'])
topic_url
```

```
['https://github.com/topics/3d',
 'https://github.com/topics/ajax',
 'https://github.com/topics/algorithm',
 'https://github.com/topics/amphp',
 'https://github.com/topics/android',
 'https://github.com/topics/angular',
 'https://github.com/topics/ansible',
 'https://github.com/topics/api',
 'https://github.com/topics/arduino',
 'https://github.com/topics/aspnet',
 'https://github.com/topics/atom',
 'https://github.com/topics/awesome',
 'https://github.com/topics/aws',
 'https://github.com/topics/azure',
 'https://github.com/topics/babel',
 'https://github.com/topics/bash',
 'https://github.com/topics/bitcoin',
 'https://github.com/topics/bootstrap',
 'https://github.com/topics/bot',
 'https://github.com/topics/c',
```

```
'https://github.com/topics/chrome',
'https://github.com/topics/chrome-extension',
'https://github.com/topics/cli',
'https://github.com/topics/clojure',
'https://github.com/topics/code-quality',
'https://github.com/topics/code-review',
'https://github.com/topics/compiler',
'https://github.com/topics/continuous-integration',
'https://github.com/topics/covid-19',
'https://github.com/topics/cpp']
```

Create CSV file(s) with the extracted information

- Create functions for the end-to-end process of downloading, parsing, and saving CSVs.
- Execute the function with different inputs to create a dataset of CSV files.
- Verify the information in the CSV files by reading them back using Pandas.

```
!pip install pandas --upgrade --quiet
```

```
import pandas as pd
```

```
topic_dict = {
    'title': topic_titles,
    'description': topic_desc,
    'URL': topic_url
}
```

```
topics_df = pd.DataFrame(topic_dict)
topics_df
```

	title	description	URL
0	3D	3D refers to the use of three-dimensional grap...	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 library for ...	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 platform for buildin...	https://github.com/topics/arduino

	title	description	URL
9	ASP.NET	ASP.NET is a web framework for building modern...	https://github.com/topics/aspnet
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	Chrome extensions enable users to customize th...	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

```
topics_df.to_csv('topics.csv', index=None)
```

Scraping out of a Topic Page

```
topic_page_url = topic_url[0]
topic_page_url
```

```
'https://github.com/topics/3d'
```

```
response = requests.get(topic_page_url)
response.status_code
```

```
200
```

```
len(response.text)
```

```
458426
```

```
topic_doc = BeautifulSoup(response.text, 'html.parser')
```

```
h3_selection_class = 'f3 color-fg-muted text-normal lh-condensed'
repo_tags = topic_doc.find_all('h3', {'class': h3_selection_class})
repo_tags
```

```
[<h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="4bdbc49d3c05ae7f70b531fbce709a384200b0768554e0172950286a8db30940" data
             mrdoob
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="517d3d5cb9d89752156923904a4238816bc9b51ab7772f3e3644ce897d8dd4e5"
             three.js
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="14658fab6217ec4ba70f16dd98006d4334793fae49cc25ce2e1c0bb5a8950006"
             pmndrs
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="629be4efc1260d27fe29201a1901eb808cbf995e4a51d877282b7164242dbadf"
             react-three-fiber
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="760dcd7b253cb1a27d9b1a8675e86db885295be4e0d8d9fa7397adf923075d36" data
             libgdx
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="ff9d8fbd4b6a268d54aa44ebd06922a789e146ae9d21db01b8ba7839646f5507"
             libgdx
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
```

```
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="35041b8540fc503301f61f50122b6ae6d1b78719943ab6392df86920498edb30" data
            BabylonJS
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="2806ba0b1f7f4081c38662a53b466b7bc022050b5dafe4108bfab142f5214b41"
            Babylon.js
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="35cf14368807d0a0abce48667cf2c0778c4e44ceeb4edde9a860e17a9efe6443" data
            ssloy
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="b27b44faaea7b496d3a92fd58b20e39b4306098a16dc535f3a74969bd25fc472"
            tinyrenderer
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="b3db1ab47cddd377d61855a33924676044d55c8724ce9233f202e64b2a59e40e"
            aframevr
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="1e97be781c78a538510c9e0a7eb97d3d14666ccab0fce09440cf2ef4f543317a"
            aframe
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="92e006e158e11505867ec48dd3b1f9f2e0a12e03556d650ac6163f635b6018db" data
            lettier
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="632d2fdc55d44af08fe1e943134b255567a5fc78d44f267f13687602afc3c8f4"
            3d-game-shaders-for-beginners
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="a27e82740ebd440eb8aec51759f134d74f147b9f07c9e1b2a8e960ff36c0e0dd" data
            FreeCAD
 </a>          /
```

```
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="f409ec71fa689bb04d504c411bd9676bb200cf6639b2678f72b9af816768dbbf"
            FreeCAD
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="61f9f002cf1a4e74bf16f253674df30f6d7e65ee4900647326a173ccb8f31afe"
            CesiumGS
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="415ccd1ea052027e4073ace6132657d6ac12d43cea98b66453bbcdbed555faf5"
            cesium
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="f5b3a8fc92d3f30b4468ca6255d50ef1ef41af4a277e99a44368380d05695b49" data
            metafizzy
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="6ef1f8dc07d99135ad796148e15edb48576348e33a218fa22dbb0a3954450801"
            zdog
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="0077f42015645d900471e0cf0ce8a3d643d649d4fa3ff4d0b3fca48249c1b695"
            timzhang642
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="5ac29541d56b0a4f137b3adb34578840be52d1f547359a52b3f4ac007241de29"
            3D-Machine-Learning
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="2690981e9e9eeb03ddf9f0f49f0c3167881395f2850f1e2b8012d64f5db4088d"
            isl-org
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="3376b3eb05aaf87cba8d79a982e9ed5940507b1d0c89a16fb29986e1bdd5e302"
            Open3D
 </a> </h3>,
```

```
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="3ef621e0683e557bb6ec46fbf192cfbfdd0fa2e15d1ed98932b3546e43be6655"
             blender
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="ac5fee340bdeebc5fd06dbf09aca394adad6d0a397aa20f51ddee7ea76ed61f0"
             blender
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="7220683d54951178816028424a94bb30c0e0deb6ed51a32331c4c5a4f1344a62" data
             a1studmuffin
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="02354cc3f2ad7480715b6a482b5f7f035bd5abedba1fc0faead1894e17a61820"
             SpaceshipGenerator
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="2b46ab94ba783e748930d230cdb4e412a8f4935af6bb89136ae4f1c32f1b5ccb" data
             domlysz
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="7c97f5ccba5ac252cf18348fde7349654b1bcc5a2849dff71c6c86aa5659f495"
             BlenderGIS
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="253650db0f2b6f83bde31f3af2cb1342ac44f41987dd296e94a08abb4a8b0298"
             FyroxEngine
 </a>           /
           <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="be159b4cbb1fd8013a482b98f601fd932ffba3b6f0fe97d9f36a1dc468b54236"
             Fyrox
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="6e27233282417634bd67b72ffd9eb417d80e8683b1df666e55551cc9b8be4532" data
             openscad
```

```
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="ce459e10b38eff918a7732ee23229e2547b096565812cfa6a9fb3a60b47ed1bd'
            openscad
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="fefb66c769603a36c83f99d8fcb711851a0c516cf07be3a8a80997d82d0f4ef4" data
            google
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="0e361312deb28484dffa632561cf40a92e83db0f1472a6f00d4d1f335c18ccbd'
            model-viewer
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="f73e50528259296676fd44b4cdcf910acd91b7ec5540ff270bb045e86b1c38d8'
            spritejs
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="a56e16d0a40310d1b96aaa49cca27395eceed1a101cf1a0bacaf5355cd7d0b54'
            spritejs
 </a> </h3>,
 <h3 class="f3 color-fg-muted text-normal lh-condensed">
 <a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
hydro-click-hmac="c3bf27c338bea41a20ec22010a10d077bdc14c5edf8a9d3c14f464c1a18d4c24" data
            jagenjo
 </a>          /
          <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.c
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="f35c57e030fa0d745719a2f7ebb0d63d8025f9adf5beb0f674db3a4fc159026a'
            webglstudio.js
 </a> </h3>]
```

```
len(repo_tags)
```

```
20
```

```
a_tags = repo_tags[0].find_all('a')
a_tags[0]
```

```
<a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="4bdbc49d3c05ae7f70b531fbce709a384200b0768554e0172950286a8db30940'
```

```
            mrdoob
</a>
```

```
a_tags[0].text.strip()
```

```
'mrdoob'
```

```
a_tags[1].text.strip()
```

```
'three.js'
```

```
repo_url = base_url+a_tags[1]['href']
repo_url
```

```
'https://github.com/mrdoob/three.js'
```

```
star_selection_class = 'Counter js-social-count'
star_tags = topic_doc.find_all('span', {'class' : star_selection_class})
star_tags
```

```
[<span aria-label="89718 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="89,718">89.7k</span>,
 <span aria-label="21733 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="21,733">21.7k</span>,
 <span aria-label="21208 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="21,208">21.2k</span>,
 <span aria-label="19540 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="19,540">19.5k</span>,
 <span aria-label="16260 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="16,260">16.3k</span>,
 <span aria-label="15122 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="15,122">15.1k</span>,
 <span aria-label="14688 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-
counter-star" title="14,688">14.7k</span>,
 <span aria-label="13434 users starred this repository" class="Counter js-social-count"
data-plural-suffix="users starred this repository" data-singular-suffix="user starred
```

this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="13,434">13.4k,
 <span aria-label="10026 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="10,026">10k,
 <span aria-label="9630 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="9,630">9.6k,
 <span aria-label="8731 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="8,731">8.7k,
 <span aria-label="8217 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="8,217">8.2k,
 <span aria-label="7866 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="7,866">7.9k,
 <span aria-label="7351 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="7,351">7.4k,
 <span aria-label="6093 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="6,093">6.1k,
 <span aria-label="5961 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="5,961">6k,
 <span aria-label="5398 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="5,398">5.4k,
 <span aria-label="5386 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="5,386">5.4k,
 <span aria-label="5142 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="5,142">5.1k,
 <span aria-label="4895 users starred this repository" class="Counter js-social-count" data-plural-suffix="users starred this repository" data-singular-suffix="user starred

```
this repository" data-turbo-replace="true" data-view-component="true" id="repo-stars-counter-star" title="4,895">4.9k</span>]
```

```
star_tags[0].text
```

```
'89.7k'
```

```
def parse_star_count(star_str):
    if star_str[-1] == 'k':
        return int(float(star_str[:-1])*1000)
    return int(star_str)
```

```
parse_star_count(star_tags[0].text)
```

```
89700
```

```
a_tags = repo_tags[1].find_all('a')
a_tags
```

```
[<a data-hydro-click='{"event_type":"explore.click","payload":
{"click_context":"REPOSITORY_CARD","click_target":"OWNER","click_visual_representation":
data-hydro-click-hmac="14658fab6217ec4ba70f16dd98006d4334793fae49cc25ce2e1c0bb5a8950006'
             pmndrs
 </a>,
 <a class="text-bold wb-break-word" data-hydro-click='{"event_type":"explore.click","pay
{"click_context":"REPOSITORY_CARD","click_target":"REPOSITORY","click_visual_representat
data-hydro-click-hmac="629be4efc1260d27fe29201a1901eb808cbf995e4a51d877282b7164242dbadf'
             react-three-fiber
 </a>]
```

```
def get_repo_info(h3_tag, star_tag):
    #returns all the required info
    a_tags = h3_tag.find_all('a')
    username = a_tags[0].text.strip()
    repo_name = a_tags[1].text.strip()
    repo_url = base_url + a_tags[1]['href']
    stars = parse_star_count(star_tag.text)
    return username, repo_name, repo_url, stars
```

```
get_repo_info(repo_tags[0], star_tags[0])
```

```
('mrdoob', 'three.js', 'https://github.com/mrdoob/three.js', 89700)
```

```
topics_repos_dict = {
    'username' : [],
    'repo_name': [],
    'repo_url': [],
    'stars': []
}
```

```
for i in range(len(repo_tags)):
    repo_info = get_repo_info(repo_tags[i], star_tags[i])
    topics_repos_dict['username'].append(repo_info[0])
    topics_repos_dict['repo_name'].append(repo_info[1])
    topics_repos_dict['repo_url'].append(repo_info[2])
    topics_repos_dict['stars'].append(repo_info[3])
```

```
topics_repos_dict
```

```
{'username': ['mrdoob',
  'pmndrs',
  'libgdx',
  'BabylonJS',
  'ssloy',
  'aframevr',
  'lettier',
  'FreeCAD',
  'CesiumGS',
  'metafizzy',
  'timzhang642',
  'isl-org',
  'blender',
  'a1studmuffin',
  'domlysz',
  'FyroxEngine',
  'openscad',
  'google',
  'spritejs',
  'jagenjo'],
 'repo_name': ['three.js',
  'react-three-fiber',
  'libgdx',
  'Babylon.js',
  'tinyrenderer',
  'aframe',
  '3d-game-shaders-for-beginners',
  'FreeCAD',
  'cesium',
  'zdog',
  '3D-Machine-Learning',
  'Open3D',
  'blender',
  'SpaceshipGenerator',
  'BlenderGIS',
  'Fyrox',
  'openscad',
  'model-viewer',
  'spritejs',
  'webglstudio.js'],
```

```
 'repo_url': ['https://github.com/mrdoob/three.js',
  'https://github.com/pmndrs/react-three-fiber',
  'https://github.com/libgdx/libgdx',
  'https://github.com/BabylonJS/Babylon.js',
  'https://github.com/ssloy/tinyrenderer',
  'https://github.com/aframevr/aframe',
  'https://github.com/lettier/3d-game-shaders-for-beginners',
  'https://github.com/FreeCAD/FreeCAD',
  'https://github.com/CesiumGS/cesium',
  'https://github.com/metafizzy/zdog',
  'https://github.com/timzhang642/3D-Machine-Learning',
  'https://github.com/isl-org/Open3D',
  'https://github.com/blender/blender',
  'https://github.com/a1studmuffin/SpaceshipGenerator',
  'https://github.com/domlysz/BlenderGIS',
  'https://github.com/FyroxEngine/Fyrox',
  'https://github.com/openscad/openscad',
  'https://github.com/google/model-viewer',
  'https://github.com/spritejs/spritejs',
  'https://github.com/jagenjo/webglstudio.js'],
 'stars': [89700,
  21700,
  21200,
  19500,
  16300,
  15100,
  14700,
  13400,
  10000,
  9600,
  8700,
  8200,
  7900,
  7400,
  6100,
  6000,
  5400,
  5400,
  5100,
  4900]}
```

```
import os
def get_topic_page(topic_url):
    # Download the page
    response = requests.get(topic_url)

    # Check Successful response
    if response.status_code != 200:
        raise Exception('Failed to load page {}'.format(topic_url))
```

```
    # Parse using Beautiful Soup
    topic_doc = BeautifulSoup(response.text, 'html.parser')
    return topic_doc

def get_repo_info(h3_tag, star_tag):
    #returns all the required info
    a_tags = h3_tag.find_all('a')
    username = a_tags[0].text.strip()
    repo_name = a_tags[1].text.strip()
    repo_url = base_url + a_tags[1]['href']
    stars = parse_star_count(star_tag.text)
    return username, repo_name, repo_url, stars

def get_topic_repos(topic_doc):
    # Get h3 tags containing repo info
    repo_tags = topic_doc.find_all('h3', {'class': h3_selection_class})

    # Get star tags containing star info
    star_tags = topic_doc.find_all('span', {'class' : star_selection_class})

    # Get repo info
    topics_repos_dict = {
        'username' : [],
        'repo_name': [],
        'repo_url': [],
        'stars': []
    }

    for i in range(len(repo_tags)):
        repo_info = get_repo_info(repo_tags[i], star_tags[i])
        topics_repos_dict['username'].append(repo_info[0])
        topics_repos_dict['repo_name'].append(repo_info[1])
        topics_repos_dict['repo_url'].append(repo_info[2])
        topics_repos_dict['stars'].append(repo_info[3])
    return pd.DataFrame(topics_repos_dict)

def scrape_topic(topic_url, path):
    if os.path.exists(path):
        print('The file {} already exists. Skipping....'.format(path))
        return
    topic_df = get_topic_repos(get_topic_page(topic_url))
    topic_df.to_csv(path, index = None)
```

Write a single Function to:

- Get a list of topics from the topic page
- Get the list of top repos from the individual topic pages
- For a each create a CSV of the top repos for the topic

Final Code

```
import os
import requests
import pandas as pd
from bs4 import BeautifulSoup

def get_topic_page(topic_url):
    # Download the page
    response = requests.get(topic_url)

    # Check Successful response
    if response.status_code != 200:
        raise Exception('Failed to load page {}'.format(topic_url))

    # Parse using Beautiful Soup
    topic_doc = BeautifulSoup(response.text, 'html.parser')
    return topic_doc

def get_repo_info(h3_tag, star_tag):
    #returns all the required info
    a_tags = h3_tag.find_all('a')
    username = a_tags[0].text.strip()
    repo_name = a_tags[1].text.strip()
    repo_url = base_url + a_tags[1]['href']
    stars = parse_star_count(star_tag.text)
    return username, repo_name, repo_url, stars

def get_topic_repos(topic_doc):
    # Get h3 tags containing repo info
    repo_tags = topic_doc.find_all('h3', {'class': h3_selection_class})

    # Get star tags containing star info
    star_tags = topic_doc.find_all('span', {'class' : star_selection_class})

    # Get repo info
    topics_repos_dict = {
        'username' : [],
        'repo_name': [],
        'repo_url': [],
        'stars': []
    }

    for i in range(len(repo_tags)):
        repo_info = get_repo_info(repo_tags[i], star_tags[i])
        topics_repos_dict['username'].append(repo_info[0])
        topics_repos_dict['repo_name'].append(repo_info[1])
        topics_repos_dict['repo_url'].append(repo_info[2])
        topics_repos_dict['stars'].append(repo_info[3])
    return pd.DataFrame(topics_repos_dict)
```

```
def scrape_topic(topic_url, path):
    if os.path.exists(path):
        print('The file {} already exists. Skipping....'.format(path))
        return
    topic_df = get_topic_repos(get_topic_page(topic_url))
    topic_df.to_csv(path, index = None)
```

```
def get_topic_titles(doc):
    selection_class = 'f3 lh-condensed mb-0 mt-1 Link--primary'
    topic_title_tags = doc.find_all('p', {'class': selection_class})
    topic_titles = []
    for tag in topic_title_tags:
        topic_titles.append(tag.text)
    return topic_titles

def get_topic_desc(doc):
    desc_selector = 'f5 color-fg-muted mb-0 mt-1'
    topic_desc_tags = doc.find_all('p', {'class': desc_selector})
    topic_desc = []
    for desc in topic_desc_tags:
        topic_desc.append(desc.text.strip())
    return topic_desc

def get_topic_url(doc):
    topic_link_tags = doc.find_all('a', {'class': 'no-underline flex-1 d-flex flex-colu
    topic_url = []
    base_url = 'https://github.com'
    for url in topic_link_tags:
        topic_url.append(base_url + url['href'])
    return topic_url

def scrape_topics():
    topics_url = 'https://github.com/topics'
    response = requests.get(topics_url)
    if response.status_code != 200:
        raise Exception('Failed to load page {}'.format(topic_url))
    topics_dict = {
        'title': get_topic_titles(doc),
        'description': get_topic_desc(doc),
        'URL': get_topic_url(doc)
    }

    return pd.DataFrame(topics_dict)
```

```
def scrape_topic_repos():
    print('Scraping list of topics')
    topics_df = scrape_topics()

    # Create Folder here
    os.makedirs('data', exist_ok=True)
```

```
    for index, row in topics_df.iterrows():
        print('Scraping top repositories for "{}" '.format(row['title']))
        scrape_topic(row['URL'], 'data/{}.csv'.format(row['title']))
```

```
scrape_topic_repos()
```

```
Scraping list of topics
Scraping top repositories for "3D"
The file data/3D.csv already exists. Skipping....
Scraping top repositories for "Ajax"
The file data/Ajax.csv already exists. Skipping....
Scraping top repositories for "Algorithm"
The file data/Algorithm.csv already exists. Skipping....
Scraping top repositories for "Amp"
The file data/Amp.csv already exists. Skipping....
Scraping top repositories for "Android"
The file data/Android.csv already exists. Skipping....
Scraping top repositories for "Angular"
The file data/Angular.csv already exists. Skipping....
Scraping top repositories for "Ansible"
The file data/Ansible.csv already exists. Skipping....
Scraping top repositories for "API"
The file data/API.csv already exists. Skipping....
Scraping top repositories for "Arduino"
The file data/Arduino.csv already exists. Skipping....
Scraping top repositories for "ASP.NET"
The file data/ASP.NET.csv already exists. Skipping....
Scraping top repositories for "Atom"
The file data/Atom.csv already exists. Skipping....
Scraping top repositories for "Awesome Lists"
The file data/Awesome Lists.csv already exists. Skipping....
Scraping top repositories for "Amazon Web Services"
The file data/Amazon Web Services.csv already exists. Skipping....
Scraping top repositories for "Azure"
The file data/Azure.csv already exists. Skipping....
Scraping top repositories for "Babel"
The file data/Babel.csv already exists. Skipping....
Scraping top repositories for "Bash"
The file data/Bash.csv already exists. Skipping....
Scraping top repositories for "Bitcoin"
The file data/Bitcoin.csv already exists. Skipping....
Scraping top repositories for "Bootstrap"
The file data/Bootstrap.csv already exists. Skipping....
```

```
Scraping top repositories for "Bot"
The file data/Bot.csv already exists. Skipping....
Scraping top repositories for "C"
The file data/C.csv already exists. Skipping....
Scraping top repositories for "Chrome"
The file data/Chrome.csv already exists. Skipping....
Scraping top repositories for "Chrome extension"
The file data/Chrome extension.csv already exists. Skipping....
Scraping top repositories for "Command line interface"
The file data/Command line interface.csv already exists. Skipping....
Scraping top repositories for "Clojure"
The file data/Clojure.csv already exists. Skipping....
Scraping top repositories for "Code quality"
The file data/Code quality.csv already exists. Skipping....
Scraping top repositories for "Code review"
The file data/Code review.csv already exists. Skipping....
Scraping top repositories for "Compiler"
The file data/Compiler.csv already exists. Skipping....
Scraping top repositories for "Continuous integration"
The file data/Continuous integration.csv already exists. Skipping....
Scraping top repositories for "COVID-19"
The file data/COVID-19.csv already exists. Skipping....
Scraping top repositories for "C++"
The file data/C++.csv already exists. Skipping....
```

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "rishavdas-0307/web-scraping-tool" on https://jovian.com
[jovian] Attaching records (metrics, hyperparameters, dataset etc.)
[jovian] Committed successfully! https://jovian.com/rishavdas-0307/web-scraping-tool
```

```
'https://jovian.com/rishavdas-0307/web-scraping-tool'
```

```

```