Assignment 1: Web Scrapping

Topic: Scraping Top Repositories for Topics on GitHub

Name: Shukla Shivam Ashokbhai

Roll No: 19301

Department: EECS

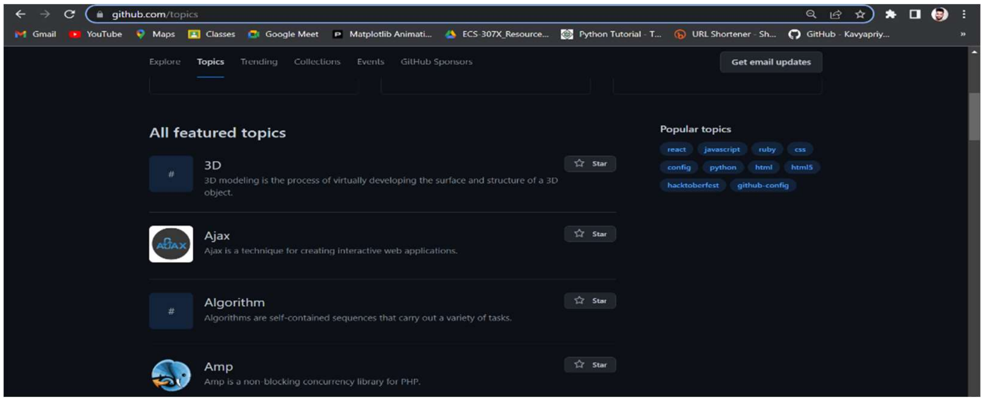
Aim: To get the information about top 20 repositories on any topic on GitHub.

Motivation: For a data scientist, software engineers, and researchers GitHub is a one of the most important platforms to find already done work in their domain. For any subject or any topic there are thousands of repositories on GitHub. If we can get the information about the top repositories of any topic available on GitHub than it will be very useful for users of GitHub. With motivation to get information like user, repository name, link, and most important number of stars about the top 20 repositories of any topic I am build this web scrapping script.

Python Module Used: request, beautifulsoup, pandas

Work Description and Results:

In this script I have scraped the website URL=https://github.com/topics



All featured topics

3D

3D modeling is the process of virtually developing the surface and structure of a 3D object.

Ajax

Ajax is a technique for creating interactive web applications.

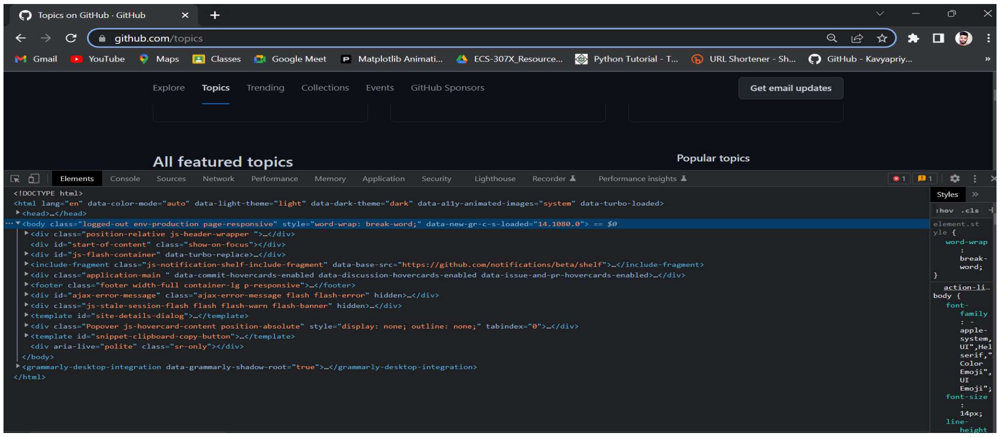
Algorithm

Algorithms are self-contained sequences that carry out a variety of tasks.

Amp

Amp is a non-blocking concurrency library for PHP.

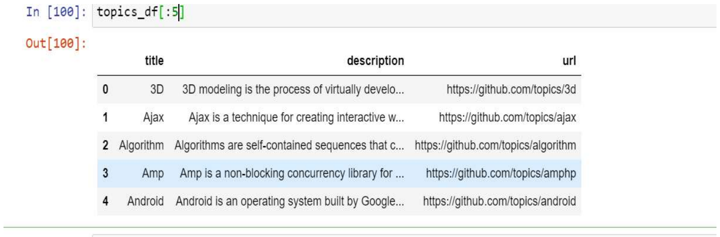
Fig1: webpage for topic



```
<!DOCTYPE html>
<html lang="en" data-color-mode="auto" data-light-theme="light" data-dark-theme="dark" data-a11y-animated-images="system" data-turbo-loaded>
 ▶<head>…</head>
…▼<body class="logged-out env-production page-responsive" style="word-wrap: break-word;" data-new-gr-c-s-loaded="14.1080.0"> == $0
  ▶<div class="position-relative js-header-wrapper ">…</div>
   <div id="start-of-content" class="show-on-focus"></div>
  ▶<div id="js-flash-container" data-turbo-replace>…</div>
  ▶<include-fragment class="js-notification-shelf-include-fragment" data-base-src="https://github.com/notifications/beta/shelf">…</include-fragment>
  ▶<div class="application-main " data-commit-hovercards-enabled data-discussion-hovercards-enabled data-issue-and-pr-hovercards-enabled>…</div>
  ▶<footer class="footer width-full container-lg p-responsive">…</footer>
  ▶<div id="ajax-error-message" class="ajax-error-message flash flash-error" hidden>…</div>
  ▶<div class="js-stale-session-flash flash flash-warn flash-banner" hidden>…</div>
  ▶<template id="site-details-dialog">…</template>
  ▶<div class="Popover js-hovercard-content position-absolute" style="display: none; outline: none;" tabindex="0">…</div>
  ▶<template id="snippet-clipboard-copy-button">…</template>
   <div aria-live="polite" class="sr-only"></div>
  </body>
 ▶<grammarly-desktop-integration data-grammarly-shadow-root="true">…</grammarly-desktop-integration>
</html>
```

Fig2:

For the given GitHub webpage, I have scrapped for all the featured topics on the topic page of GitHub and for all available topics I have extracted the information about top 20 repositories.



```
In [100]: topics_df[:5]
```

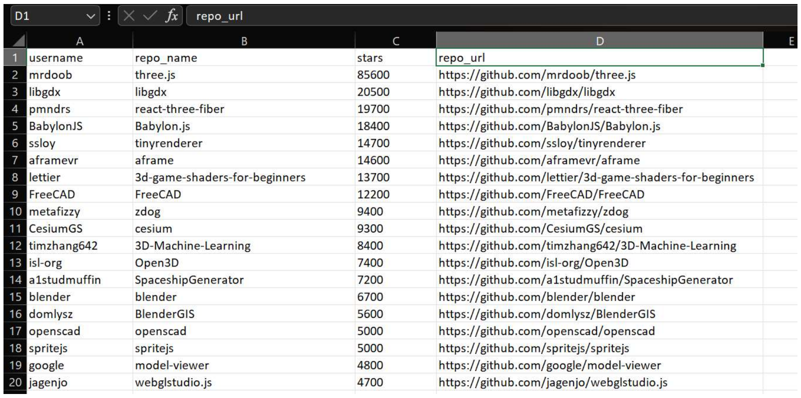
Out[100]:

	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 library for ...	https://github.com/topics/amphp
4	Android	Android is an operating system built by Google...	https://github.com/topics/android

Fig3: top 5 feature topics

Data is stored in csv file format in *“.../data”* folder. For every topic top 20 repositories have been scrapped and for each repository four information have been collected, which are:

1.) Username: Owner/creator of the repository
2.) Repo_name: Name of the repository
3.) Stars: Total number of starts given by GitHub users to the repo.
4.) Repo_Url: link of the repository.

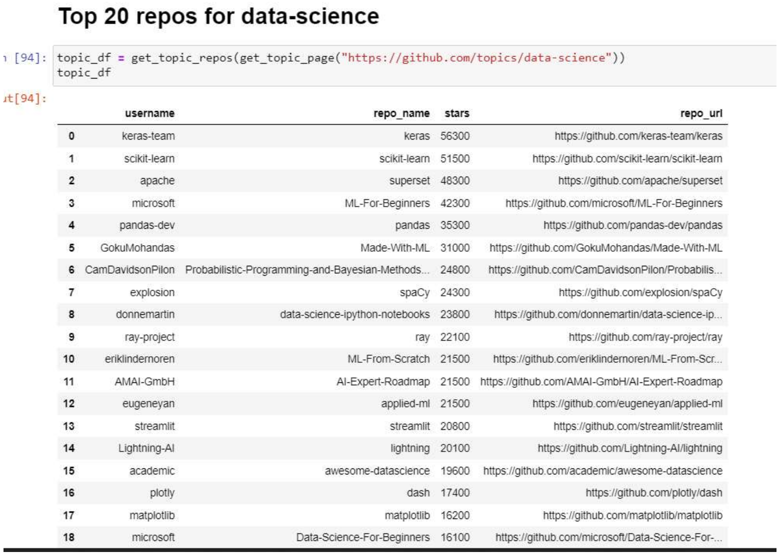


	A	B	C	D	E
1	username	repo_name	stars	repo_url	
2	mrdoob	three.js	85600	https://github.com/mrdoob/three.js	
3	libgdx	libgdx	20500	https://github.com/libgdx/libgdx	
4	pmndrs	react-three-fiber	19700	https://github.com/pmndrs/react-three-fiber	
5	BabylonJS	Babylon.js	18400	https://github.com/BabylonJS/Babylon.js	
6	ssloy	tinyrenderer	14700	https://github.com/ssloy/tinyrenderer	
7	aframevr	aframe	14600	https://github.com/aframevr/aframe	
8	lettier	3d-game-shaders-for-beginners	13700	https://github.com/lettier/3d-game-shaders-for-beginners	
9	FreeCAD	FreeCAD	12200	https://github.com/FreeCAD/FreeCAD	
10	metafizzy	zdog	9400	https://github.com/metafizzy/zdog	
11	CesiumGS	cesium	9300	https://github.com/CesiumGS/cesium	
12	timzhang642	3D-Machine-Learning	8400	https://github.com/timzhang642/3D-Machine-Learning	
13	isl-org	Open3D	7400	https://github.com/isl-org/Open3D	
14	a1studmuffin	SpaceshipGenerator	7200	https://github.com/a1studmuffin/SpaceshipGenerator	
15	blender	blender	6700	https://github.com/blender/blender	
16	domlysz	BlenderGIS	5600	https://github.com/domlysz/BlenderGIS	
17	openscad	openscad	5000	https://github.com/openscad/openscad	
18	spritejs	spritejs	5000	https://github.com/spritejs/spritejs	
19	google	model-viewer	4800	https://github.com/google/model-viewer	
20	jagenjo	webglstudio.js	4700	https://github.com/jagenjo/webglstudio.js	

Fig: information of top 20 repos for topic: 3d

Further we can use this scrip to get same information about any topic just by getting the URL of the topic page.

For example: Top 20 repositories for topic: Data Science



Top 20 repos for data-science

```
[94]: topic_df = get_topic_repos(get_topic_page("https://github.com/topics/data-science"))
      topic_df
```

t[94]:

	username	repo_name	stars	repo_url
0	keras-team	keras	56300	https://github.com/keras-team/keras
1	scikit-learn	scikit-learn	51500	https://github.com/scikit-learn/scikit-learn
2	apache	superset	48300	https://github.com/apache/superset
3	microsoft	ML-For-Beginners	42300	https://github.com/microsoft/ML-For-Beginners
4	pandas-dev	pandas	35300	https://github.com/pandas-dev/pandas
5	GokuMohandas	Made-With-ML	31000	https://github.com/GokuMohandas/Made-With-ML
6	CamDavidsonPilon	Probabilistic-Programming-and-Bayesian-Methods...	24800	https://github.com/CamDavidsonPilon/Probabilis...
7	explosion	spaCy	24300	https://github.com/explosion/spaCy
8	donnemartin	data-science-ipython-notebooks	23800	https://github.com/donnemartin/data-science-ip...
9	ray-project	ray	22100	https://github.com/ray-project/ray
10	eriklindernoren	ML-From-Scratch	21500	https://github.com/eriklindernoren/ML-From-Scr...
11	AMAI-GmbH	AI-Expert-Roadmap	21500	https://github.com/AMAI-GmbH/AI-Expert-Roadmap
12	eugeneyan	applied-ml	21500	https://github.com/eugeneyan/applied-ml
13	streamlit	streamlit	20800	https://github.com/streamlit/streamlit
14	Lightning-AI	lightning	20100	https://github.com/Lightning-AI/lightning
15	academic	awesome-datascience	19600	https://github.com/academic/awesome-datascience
16	plotly	dash	17400	https://github.com/plotly/dash
17	matplotlib	matplotlib	16200	https://github.com/matplotlib/matplotlib
18	microsoft	Data-Science-For-Beginners	16100	https://github.com/microsoft/Data-Science-For-...

Future Work:

1.) We can develop and script which can scrap information of all the repositories with specific username in any topic and list them according to their ranking in terms of stars.