**Introduction to Web Scraping and REST APIs**

Diagram

Description automatically generated

Web scraping is the process of extracting and parsing data from websites in an automated fashion using a computer program. It's a useful technique for creating datasets for research and learning. While web scraping often involves parsing and processing [HTML documents](https://jovian.ai/outlink?url=https%3A%2F%2Fdeveloper.mozilla.org%2Fen-US%2Fdocs%2FWeb%2FHTML), some platforms also offer [REST APIs](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.smashingmagazine.com%2F2018%2F01%2Funderstanding-using-rest-api%2F) to retrieve information in a machine-readable format like [JSON](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.digitalocean.com%2Fcommunity%2Ftutorials%2Fan-introduction-to-json). In this tutorial, we'll use web scraping and REST APIs to create a real-world dataset.

This tutorial covers the following topics:

* Downloading web pages using the requests library
* Inspecting the HTML source code of a web page
* Parsing parts of a website using Beautiful Soup
* Writing parsed information into CSV files
* Using a REST API to retrieve data as JSON
* Combining data from multiple sources
* Using links on a page to crawl a website

**How to Run the Code**

The best way to learn the material is to execute the code and experiment with it yourself. This tutorial is an executable [Jupyter notebook](https://jovian.ai/outlink?url=https%3A%2F%2Fjupyter.org). You can *run* this tutorial and experiment with the code examples in a couple of ways: *using free online resources* (recommended) or *on your computer*.

**Option 1: Running using free online resources (1-click, recommended)**

The easiest way to start executing the code is to click the **Run** button at the top of this page and select **Run on Binder**. You can also select "Run on Colab" or "Run on Kaggle", but you'll need to create an account on [Google Colab](https://jovian.ai/outlink?url=https%3A%2F%2Fcolab.research.google.com) or [Kaggle](https://jovian.ai/outlink?url=https%3A%2F%2Fkaggle.com) to use these platforms.

**Option 2: Running on your computer locally**

To run the code on your computer locally, you'll need to set up [Python](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.python.org), download the notebook and install the required libraries. We recommend using the [Conda](https://jovian.ai/outlink?url=https%3A%2F%2Fdocs.conda.io%2Fprojects%2Fconda%2Fen%2Flatest%2Fuser-guide%2Finstall%2F) distribution of Python. Click the **Run** button at the top of this page, select the **Run Locally** option, and follow the instructions.

**Problem**

Over the course of this tutorial, we'll solve the following problem to learn the tools and techniques used for web scraping:

**QUESTION**: Write a Python function that creates a CSV file (comma-separated values) containing details about the 25 top GitHub repositories for any given topic. You can view the top repositories for the topic machine-learning on this page: [https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning). The output CSV should contain these details: repository name, owner's username, no. of stars, repository URL.

[Graphical user interface, application

Description automatically generated](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)

How would you go about solving this problem in Python? Explore the web page and take a couple of minutes to come up with an approach before proceeding further. How many lines of code do you think the solution will require?

**Downloading a web page using requests**

When you access a URL like [https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning) using a web browser, it downloads the contents of the web page the URL points to and displays the output on the screen. Before we can extract information from a web page, we need to download the page using Python.

We'll use a library called [requests](https://jovian.ai/outlink?url=https%3A%2F%2Fdocs.python-requests.org%2Fen%2Fmaster%2F) to download web pages from the internet. Let's begin by installing and importing the library.

*# Install the library*

!pip install requests --upgrade --quiet

*# Import the library*

import requests

We can download a web page using the requests.get function.

topic\_url = 'https://github.com/topics/machine-learning'

response = requests.get(topic\_url)

type(response)

requests.models.Response

requests.get returns a response object with the page contents and some information indicating whether the request was successful, using a status code. Learn more about HTTP status codes here: [https://developer.mozilla.org/en-US/docs/Web/HTTP/Status](https://jovian.ai/outlink?url=https%3A%2F%2Fdeveloper.mozilla.org%2Fen-US%2Fdocs%2FWeb%2FHTTP%2FStatus).

If the request was successful, response.status\_code is set to a value between 200 and 299.

response.status\_code

200

The contents of the web page can be accessed using the .text property of the response.

page\_contents = response.text

len(page\_contents)

424155

The page contains over 60,000 characters! Let's view the first 1000 characters of the web page.

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](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com)">\n <link rel="dns-prefetch" href="[https://avatars.githubusercontent.com](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com)">\n <link rel="dns-prefetch" href="[https://github-cloud.s3.amazonaws.com](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub-cloud.s3.amazonaws.com)">\n <link rel="dns-prefetch" href="[https://user-images.githubusercontent.com/](https://jovian.ai/outlink?url=https%3A%2F%2Fuser-images.githubusercontent.com%2F)">\n <link rel="preconnect" href="[https://github.githubassets.com](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com)" crossorigin>\n <link rel="preconnect" href="[https://avatars.githubusercontent.com](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com)">\n\n\n\n <link crossorigin="anonymous" media="all" rel="stylesheet" href="[https://github.githubassets.com/assets/light-5178aee0ee76.css](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com%2Fassets%2Flight-5178aee0ee76.css)" /><link crossorigin="anonymous" media="all" rel="stylesheet" href="[https://github.githubassets.com/assets/dark-217d4f9c8e70.css](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com%2Fassets%2Fdark-217d4f9c8e70.css)" /><link data-color-theme="dark\_dimmed" crossorigin="anonymous" media="all" rel="stylesheet" data-href="[https://github.gith](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.gith)'

What you see above is the *source code* of the web page. It written in a language called [HTML](https://jovian.ai/outlink?url=https%3A%2F%2Fdeveloper.mozilla.org%2Fen-US%2Fdocs%2FWeb%2FHTML). It defines the content and structure of the web page.

Let's save the contents to a file with the .html extension.

with open('machine-learning-topics.html', 'w', encoding="utf-8") as file:

file.write(page\_contents)

You can now view the file using the "File > Open" menu option within Jupyter and clicking on *machine-learning.html* in the list of files displayed. Here's what you'll see when you open the file:

Graphical user interface, application

Description automatically generated

While this looks similar to the original web page, note that it's simply a copy. You will notice that none of the links or buttons work. To view or edit the source code of the file, click "File > Open" within Jupyter, then select the file *machine-learning.html* from the list and click the "Edit" button.

Graphical user interface, text, application, email

Description automatically generated

As you might expect, the source code looks something like this:

Text

Description automatically generated

Try scrolling through the source code. Can you make sense of it? Can you see how the information on the page is organized within the file? We'll learn more about it in the next section.

**EXERCISE**: Download the web page for a different topic, e.g., [https://github.com/topics/data-analysis](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fdata-analysis) using requests and save it to a file, e.g., data-analysis.html. View the page and compare it with the previously downloaded page? How are the two different? Can you spot the differences in the source code?

Let's save our work using jovian before continuing.

!pip install jovian --upgrade --quiet

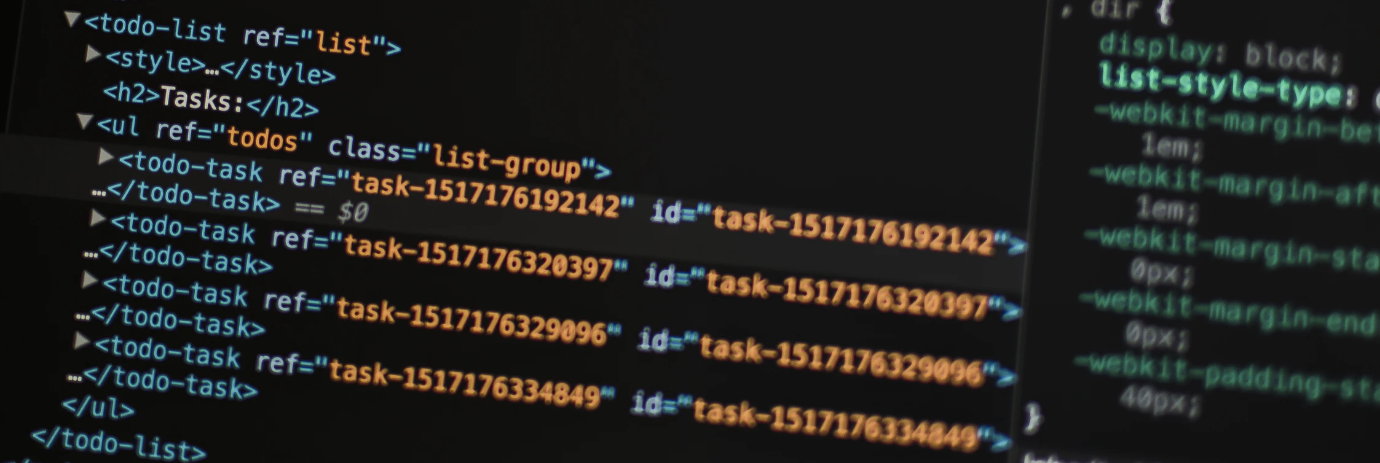
import jovian

jovian.commit(project='python-web-scraping-and-rest-api')

[jovian] Updating notebook "samanvitha/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Committed successfully! <https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>

'<https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>'

**Inspecting the HTML source code of a web page**



As mentioned earlier, web pages are written in a language called HTML (Hyper Text Markup Language). HTML is a fairly simple language comprised of *tags* (also called *nodes* or *elements*) e.g. <a href="https://jovian.ai" target="\_blank">Go to Jovian</a>. An HTML tag has three parts:

1. **Name**: (html, head, body, div, etc.) Indicates what the tag represents and how a browser should interpret the information inside it.
2. **Attributes**: (href, target, class, id, etc.) Properties of tag used by the browser to customize how a tag is displayed and decide what happens on user interactions.
3. **Children**: A tag can contain some text or other tags or both between the opening and closing segments, e.g., <div>Some content</div>.

**Inside an HTML Document**

Here's a simple HTML document that uses many commonly used tags:

<html>

<head>

<title>All About Python</title>

</head>

<body>

<div style="width: 640px; margin: 40px auto">

<h1 style="text-align:center;">Python - A Programming Language</h1>

<img src="https://www.python.org/static/community\_logos/python-logo-master-v3-TM.png" alt="python-logo" style="width:240px;margin:0 auto;display:block;">

<div>

<h2>About Python</h2>

<p>

Python is an <span style="font-style: italic">interpreted, high-level and general-purpose</span> programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Visit the <a href="https://docs.python.org/3/">official documentation</a> to learn more.

</p>

</div>

<div>

<h2>Some Python Libraries</h2>

<ul id="libraries">

<li>Numpy</li>

<li>Pandas</li>

<li>PyTorch</li>

<li>Scikit Learn</li>

</ul>

</div>

<div>

<h2>Recent Python Versions</h2>

<table id="versions-table">

<tr>

<th class="bordered-table">Version</th>

<th class="bordered-table">Released on</th>

</tr>

<tr>

<td class="bordered-table">Python 3.8</td>

<td class="bordered-table">October 2019</td>

</tr>

<tr>

<td class="bordered-table">Python 3.7</td>

<td class="bordered-table">June 2018</td>

</tr>

</table>

<style>

.bordered-table {

border: 1px solid black; padding: 8px;

}

</style>

</div>

</div>

</body>

</html>

**EXERCISE**: Copy the above HTML code and paste it into a new file called webpage.html. To create a new file, select "File > Open" from the menu bar, then select "New > Text" file. View the saved file. Can you see how the different tags are displayed in different ways by the browser?

Text

Description automatically generated

**EXERCISE**: Make some changes to the code inside webpage.html. Save the file and view it again. Do you see your changes reflected? Play with the structure of the file. Try to break things and fix them!

**Common Tags and Attributes**

Following are some of the most commonly used HTML tags:

* html
* head
* title
* body
* div
* span
* h1 to h6
* p
* img
* ul, ol and li
* table, tr, th and td
* style
* ...

Each tag supports several attributes. Following are some common attributes used to modify the behavior of tags:

* id
* style
* class
* href (used with <a>)
* src (used with <img>)

**EXERCISE**: Complete this tutorial on HTML: [https://www.htmldog.com/guides/html/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.htmldog.com%2Fguides%2Fhtml%2F) . Once done, try describing what the above tags and attributes are used for. Try creating a new HTML page using the tags you find most interesting.

To learn how to style HTML tags, check out this tutorial on CSS: [https://www.htmldog.com/guides/css/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.htmldog.com%2Fguides%2Fcss%2F)

**Inspecting HTML in the Browser**

You can view the source code of any webpage right within your browser by right-clicking anywhere on a page and selecting the "Inspect" option. It opens the "Developer Tools" pane, where you can see the source code as a tree. You can expand and collapse various nodes and find the source code for a specific portion of the page.

Here's what it looks like on the Chrome browser:

Graphical user interface, text, application

Description automatically generated

**EXERCISE**: Explore the source code of the web page [https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning) . Try to find the portions in the source code corresponding to the repository name, owner's username, and the number of stars for each repository listed on the page.

Let's save our work before continuing.

jovian.commit()

[jovian] Updating notebook "samanvitha/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Committed successfully! <https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>

'<https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>'

**Extracting information from HTML using Beautiful Soup**

To extract information from the HTML source code of a webpage programmatically, we can use the [Beautiful Soup](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.crummy.com%2Fsoftware%2FBeautifulSoup%2Fbs4%2Fdoc%2F) library. Let's install the library and import the BeautifulSoup class from the bs4 module.

*# Install the library*

!pip install beautifulsoup4 --upgrade --quiet

*# Import the library*

from bs4 import BeautifulSoup

?BeautifulSoup

Next, let's read the contents of the file machine-learning.html and create a BeautifulSoup object to parse the content.

with open('machine-learning-topics.html', 'r') as f:

html\_source = f.read()

html\_source[: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](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com)">\n <link rel="dns-prefetch" href="[https://avatars.githubusercontent.com](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com)">\n <link rel="dns-prefetch" href="[https://github-cloud.s3.amazonaws.com](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub-cloud.s3.amazonaws.com)">\n <link rel="dns-prefetch" href="[https://user-images.githubusercontent.com/](https://jovian.ai/outlink?url=https%3A%2F%2Fuser-images.githubusercontent.com%2F)">\n <link rel="preconnect" href="[https://github.githubassets.com](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com)" crossorigin>\n <link rel="preconnect" href="[https://avatars.githubusercontent.com](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com)">\n\n\n\n <link crossorigin="anonymous" media="all" rel="stylesheet" href="[https://github.githubassets.com/assets/light-5178aee0ee76.css](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com%2Fassets%2Flight-5178aee0ee76.css)" /><link crossorigin="anonymous" media="all" rel="stylesheet" href="[https://github.githubassets.com/assets/dark-217d4f9c8e70.css](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.githubassets.com%2Fassets%2Fdark-217d4f9c8e70.css)" /><link data-color-theme="dark\_dimmed" crossorigin="anonymous" media="all" rel="stylesheet" data-href="[https://github.gith](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.gith)'

doc = BeautifulSoup(html\_source, 'html.parser')

type(doc)

bs4.BeautifulSoup

The doc object contains several properties and methods for extracting information from the HTML document. Let's look at a few examples below.

**NOTE**: You don't need to remember all (or any) of the properties/methods. You can look up [the documentation of BeautifulSoup](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.crummy.com%2Fsoftware%2FBeautifulSoup%2Fbs4%2Fdoc%2F) or [search online](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.google.co.in%2Fsearch%3Fq%3Dbeautifulsoup%2Bhow%2Bto%2Bget%2Bhref%2Bof%2Blink) to find what you need when you need it.

**Accessing a tag**

**QUESTION**: Find the title of the page represented by doc.

The title of the page is contained within the <title> tag. We can access the title tag using doc.title.

title\_tag = doc.title

title\_tag

<title>machine-learning · GitHub Topics · GitHub</title>

type(title\_tag)

bs4.element.Tag

We can access a tag's name using the .name property.

title\_tag.name

'title'

The text within a tag can be accessed using .text.

title\_tag.text

'machine-learning · GitHub Topics · GitHub'

**EXERCISE**: Explore the html, body, and head tags of doc. Do you see what you expect to see?

If a tag occurs more than once in a document e.g. <a> (which represents links), then doc.a finds the first <a> tag.

first\_link = doc.a

first\_link

<a class="px-2 py-4 color-bg-accent-emphasis color-fg-on-emphasis show-on-focus js-skip-to-content" href="#start-of-content">Skip to content</a>

first\_link.text

'Skip to content'

**EXERCISE**: Find the first occurrence of each of these tags in doc: div, img, span, p, etc.

**Finding all tags of the same type**

To find all the occurrences of a tag, use the find\_all method.

**QUESTION**: Find all the link tags on the page. How many links does the page contain?

all\_link\_tags = doc.find\_all('a')

len(all\_link\_tags)

417

all\_link\_tags[:3]

[<a class="px-2 py-4 color-bg-accent-emphasis color-fg-on-emphasis show-on-focus js-skip-to-content" href="#start-of-content">Skip to content</a>,

<a aria-label="Homepage" class="mr-4 color-fg-inherit" data-ga-click="(Logged out) Header, go to homepage, icon:logo-wordmark" href="[https://github.com/](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2F)">

<svg aria-hidden="true" class="octicon octicon-mark-github" data-view-component="true" height="32" version="1.1" viewbox="0 0 16 16" width="32">

<path d="M8 0C3.58 0 0 3.58 0 8c0 3.54 2.29 6.53 5.47 7.59.4.07.55-.17.55-.38 0-.19-.01-.82-.01-1.49-2.01.37-2.53-.49-2.69-.94-.09-.23-.48-.94-.82-1.13-.28-.15-.68-.52-.01-.53.63-.01 1.08.58 1.23.82.72 1.21 1.87.87 2.33.66.07-.52.28-.87.51-1.07-1.78-.2-3.64-.89-3.64-3.95 0-.87.31-1.59.82-2.15-.08-.2-.36-1.02.08-2.12 0 0 .67-.21 2.2.82.64-.18 1.32-.27 2-.27.68 0 1.36.09 2 .27 1.53-1.04 2.2-.82 2.2-.82.44 1.1.16 1.92.08 2.12.51.56.82 1.27.82 2.15 0 3.07-1.87 3.75-3.65 3.95.29.25.54.73.54 1.48 0 1.07-.01 1.93-.01 2.2 0 .21.15.46.55.38A8.013 8.013 0 0016 8c0-4.42-3.58-8-8-8z" fill-rule="evenodd"></path>

</svg>

</a>,

<a class="d-inline-block d-lg-none f5 no-underline border color-border-default rounded-2 px-2 py-1 mr-3 mr-sm-5 color-fg-inherit" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="e755e218b5b5c8ea0dcf7c1afea104ce5b0e7bb40929d477942f46cf5b206533" href="/signup?ref\_cta=Sign+up&amp;ref\_loc=header+logged+out&amp;ref\_page=%2Ftopics%2Fmachine-learning&amp;source=header">

Sign up

</a>]

**EXERCISE**: Get a list of all the img tags on the page. How many images does the page contain?

**Accessing attributes**

The attributes of a tag can be accessed using the indexing notation, e.g., first\_link['href']

first\_link

<a class="px-2 py-4 color-bg-accent-emphasis color-fg-on-emphasis show-on-focus js-skip-to-content" href="#start-of-content">Skip to content</a>

first\_link['href']

'#start-of-content'

first\_link['class']

['px-2',

'py-4',

'color-bg-accent-emphasis',

'color-fg-on-emphasis',

'show-on-focus',

'js-skip-to-content']

Note that the class attribute is automatically split into a list of classes (this isn't done for any other attribute). This is because it's common practice to check for a specific class within a tag.

You can use the .attrs property to view all the attributes as a dictionary.

first\_link.attrs

{'href': '#start-of-content',

'class': ['px-2',

'py-4',

'color-bg-accent-emphasis',

'color-fg-on-emphasis',

'show-on-focus',

'js-skip-to-content']}

**EXERCISE**: Find the 5th image tag on the page (counting from 0). Which attributes does the tag contain? Find the values of the src and alt attributes of the tag.

**Searching by Attribute Value**

**QUESTION**: Find the img tag(s) on the page with the alt attribute set to transformers.

We can provide a dictionary of attributes as the second argument to find\_all

doc.find\_all('img', { 'alt': 'transformers'})

[<img alt="transformers" class="d-block width-full" loading="lazy" src="[https://repository-images.githubusercontent.com/155220641/a16c4880-a501-11ea-9e8f-646cf611702e](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F155220641%2Fa16c4880-a501-11ea-9e8f-646cf611702e)"/>]

If we're just interested in the first element, we can use the find method. Keep in mind that find returns None if no matching tag is found.

doc.find('img', { 'alt': 'transformers'})

<img alt="transformers" class="d-block width-full" loading="lazy" src="[https://repository-images.githubusercontent.com/155220641/a16c4880-a501-11ea-9e8f-646cf611702e](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F155220641%2Fa16c4880-a501-11ea-9e8f-646cf611702e)"/>

**EXERCISE**: Find the src attribute of the first img tag with the alt attribute set to julia. Visit the link and check what the image represents.

**Searching by Class**

The class attribute is one of the most frequently used attributes on HTML tags (used for layout and styling). We can search for tags containing a class using the class\_ argument in find\_all (note that class is a reserved keyword in Python, hence the underscore in the argument name).

**QUESTION**: Find all the tags containing the class HeaderMenu-link.

matching\_tags = doc.find\_all(class\_='HeaderMenu-link')

matching\_tags

[<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Product

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>,

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Team","label":"ref\_page:/topics/machine-learning;ref\_cta:Team;"}' href="/team">Team</a>,

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Enterprise","label":"ref\_page:/topics/machine-learning;ref\_cta:Enterprise;"}' href="/enterprise">Enterprise</a>,

<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Explore

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>,

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Marketplace","label":"ref\_page:/topics/machine-learning;ref\_cta:Marketplace;"}' href="/marketplace">Marketplace</a>,

<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Pricing

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>,

<a class="HeaderMenu-link flex-shrink-0 no-underline" data-ga-click="(Logged out) Header, clicked Sign in, text:sign-in" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/login?return\_to=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning">

Sign in

</a>,

<a class="HeaderMenu-link flex-shrink-0 d-inline-block no-underline border color-border-default rounded px-2 py-1" data-analytics-event='{"category":"Sign up","action":"click to sign up for account","label":"ref\_page:/topics/machine-learning;ref\_cta:Sign up;ref\_loc:header logged out"}' data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/signup?ref\_cta=Sign+up&amp;ref\_loc=header+logged+out&amp;ref\_page=%2Ftopics%2Fmachine-learning&amp;source=header">

Sign up

</a>]

We can also for a specific type of tag e.g. <a> matching the given class.

header\_link\_tags = doc.find\_all('a', class\_='HeaderMenu-link')

header\_link\_tags

[<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Team","label":"ref\_page:/topics/machine-learning;ref\_cta:Team;"}' href="/team">Team</a>,

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Enterprise","label":"ref\_page:/topics/machine-learning;ref\_cta:Enterprise;"}' href="/enterprise">Enterprise</a>,

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Marketplace","label":"ref\_page:/topics/machine-learning;ref\_cta:Marketplace;"}' href="/marketplace">Marketplace</a>,

<a class="HeaderMenu-link flex-shrink-0 no-underline" data-ga-click="(Logged out) Header, clicked Sign in, text:sign-in" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/login?return\_to=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning">

Sign in

</a>,

<a class="HeaderMenu-link flex-shrink-0 d-inline-block no-underline border color-border-default rounded px-2 py-1" data-analytics-event='{"category":"Sign up","action":"click to sign up for account","label":"ref\_page:/topics/machine-learning;ref\_cta:Sign up;ref\_loc:header logged out"}' data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/signup?ref\_cta=Sign+up&amp;ref\_loc=header+logged+out&amp;ref\_page=%2Ftopics%2Fmachine-learning&amp;source=header">

Sign up

</a>]

**Parsing Information from Tags**

Once we have a list of tags matching some criteria, it's easy to extract information and convert it to a more convenient format.

**QUESTION**: Find the link text and URL of all the links withing the page header on [https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning) .

We'll create a list of dictionaries containing the required information. We'll add the base URL [https://github.com](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com) as a prefix because the href attribute only contains the relative path e.g. /explore.

header\_link\_tags[0]['href']

'/team'

header\_links = []

base\_url = 'https://github.com'

for tag in header\_link\_tags:

header\_links.append({ 'title': tag.text.strip(), 'url': base\_url + tag['href']})

header\_links

[{'title': 'Team', 'url': '[https://github.com/team](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fteam)'},

{'title': 'Enterprise', 'url': '[https://github.com/enterprise](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fenterprise)'},

{'title': 'Marketplace', 'url': '[https://github.com/marketplace](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fmarketplace)'},

{'title': 'Sign in',

'url': '[https://github.com/login?return\_to=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Flogin%3Freturn_to%3Dhttps%253A%252F%252Fgithub.com%252Ftopics%252Fmachine-learning)'},

{'title': 'Sign up',

'url': '[https://github.com/signup?ref\_cta=Sign+up&ref\_loc=header+logged+out&ref\_page=%2Ftopics%2Fmachine-learning&source=header](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fsignup%3Fref_cta%3DSign%2Bup%26ref_loc%3Dheader%2Blogged%2Bout%26ref_page%3D%252Ftopics%252Fmachine-learning%26source%3Dheader)'}]

We have successfully extracted the required information about links in the page header. This is precisely what web scraping is: downloading a webpage, parsing the HTML, and extracting useful information.

**EXERCISE**: Find the list of all the images matching the class d-block width-full. Each list element should be a dictionary containing two keys, "username" and "url". You can obtain the username using the alt attribute of a tag and the URL using the src attribute.

**Elements inside a tag**

**QUESTION**: Find the li tags that are direct children of ul tag with the class top-list in the sample HTML document below.

sample\_html = """

<html>

<body>

<ul class="top-list">

<li>Item 1</li>

<li>Item 2</li>

<li>

<ul>

<li>Item 3.1</li>

<li>Item 3.2</li>

<li>Item 3.3</li>

</ul>

</li>

</ul>

</body>

</html>"""

sample\_doc = BeautifulSoup(sample\_html)

list\_tag = sample\_doc.find('ul', class\_='top-list')

We can use the find\_all method on the tag, and set recursive=False to find just the direct children.

list\_item\_tags = list\_tag.find\_all('li', recursive=False)

list\_item\_tags

[<li>Item 1</li>,

<li>Item 2</li>,

<li>

<ul>

<li>Item 3.1</li>

<li>Item 3.2</li>

<li>Item 3.3</li>

</ul>

</li>]

Without recursive=False, the inner list items are also included in the result.

list\_tag.find\_all('li')

[<li>Item 1</li>,

<li>Item 2</li>,

<li>

<ul>

<li>Item 3.1</li>

<li>Item 3.2</li>

<li>Item 3.3</li>

</ul>

</li>,

<li>Item 3.1</li>,

<li>Item 3.2</li>,

<li>Item 3.3</li>]

Keep in mind that you don't need to remember all (or any) of the methods or properties offered by Beautiful Soup documents and tags. You should be able to figure out what you need to do, when you need to do it. Here's how:

* Look up the documentation: [https://www.crummy.com/software/BeautifulSoup/bs4/doc/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.crummy.com%2Fsoftware%2FBeautifulSoup%2Fbs4%2Fdoc%2F)
* Google what you're trying to do: [https://www.google.co.in/search?q=beautiful+soup+get+href](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.google.co.in%2Fsearch%3Fq%3Dbeautiful%2Bsoup%2Bget%2Bhref)
* Ask a question on StackOverflow: [https://stackoverflow.com/questions/tagged/beautifulsoup](https://jovian.ai/outlink?url=https%3A%2F%2Fstackoverflow.com%2Fquestions%2Ftagged%2Fbeautifulsoup)

Let's save our work before continuing.

jovian.commit()

**Top Repositories for a Topic**

Let's return to our original problem statement of finding the top repositories for a given topic. Before we parse a page and find the top repositories, let's define a helper function to get the web page for any topic.

**QUESTION**: Define a function get\_topic\_page that downloads the GitHub web page for a given topic and returns a beautiful soup document representing the page.

def **get\_topic\_page**(topic):

*# Construct the URL*

topic\_repos\_url = 'https://github.com/topics/' + topic

*# Get the HTML page content using requests*

response = requests.get(topic\_repos\_url)

*# Ensure that the reponse is valid*

if response.status\_code != 200:

print('Status code:', response.status\_code)

raise Exception('Failed to fetch web page ' + topic\_repos\_url)

*# Construct a beautiful soup document*

doc = BeautifulSoup(response.text)

return doc

doc = get\_topic\_page('machine-learning')

doc.title.text

'machine-learning · GitHub Topics · GitHub'

Getting the topic page for another topic is now as simple as invoking the function with a different argument.

doc2 = get\_topic\_page('data-analysis')

doc2.title.text

'data-analysis · GitHub Topics · GitHub'

**QUESTION**: Develop an approach to find the repository name, owner's username, no. of stars, and repository link for the repositories listed on a topic page.

Graphical user interface, text

Description automatically generated

Upon inspecting the box containing the information for a repository, you will find an article tag for each repository, with class attribute set to border rounded color-shadow-small color-bg-secondary my-4.

Let's find all the article tags matching this class.

article\_tags = doc.find\_all('article', class\_='border rounded color-shadow-small color-bg-subtle my-4')

len(article\_tags)

30

There are 30 repositories listed on the page, and our query resulted in 30 article tags. It looks like we've found the enclosing tag for each repository.

We need to extract the following information from each tag:

1. Repository name
2. Owner's username
3. Number of stars
4. Repository link

Look at the source of any of the article tags. You will notice that the repository name, owner's username, and the repository link are all part of an h1 tag.

article\_tag = article\_tags[4]

*# Uncomment to view*

*# article\_tag*

Let's retrieve the first h1 inside an article.

h3\_tag = article\_tag.find('h3')

h3\_tag

<h3 class="f3 color-fg-muted text-normal lh-condensed">

<a data-ga-click="Explore, go to repository owner, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"OWNER","click\_visual\_representation":"REPOSITORY\_OWNER\_HEADING","actor\_id":null,"record\_id":8401422,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="c14e22ad4cc65812caacdf8b76899e4e4979ea7d56890715930cef3abddde9c5" data-view-component="true" href="/tesseract-ocr">

tesseract-ocr

</a> /

<a class="text-bold wb-break-word" data-ga-click="Explore, go to repository, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"REPOSITORY","click\_visual\_representation":"REPOSITORY\_NAME\_HEADING","actor\_id":null,"record\_id":22887094,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="0d1578eafced610d5100a5880e0d084554b47f4b1fa6ce1de3ce9d413687d2a2" data-view-component="true" href="/tesseract-ocr/tesseract">

tesseract

</a> </h3>

The h1 has a tags inside it, one containing the owner's username and the second containing the repository title. The href of the second tag also includes the relative path of the repository. Let's extract this information from the a tags.

a\_tags = h3\_tag.find\_all('a', recursive=False)

username = a\_tags[0].text

username

'\n tesseract-ocr\n'

Looks like the username contains some leading and trailing whitespace. We can get rid of it using strip.

username = a\_tags[0].text.strip()

username

'tesseract-ocr'

We can get the repository name and repository path in the same fashion.

repo\_name = a\_tags[1].text.strip()

repo\_name

'tesseract'

repo\_path = a\_tags[1]['href'].strip()

repo\_path

'/tesseract-ocr/tesseract'

To get the full URL to the repository, we can append the base URL https://github.com at the beginning of the path.

base\_url = 'https://github.com'

repo\_url = base\_url + repo\_path

repo\_url

'[https://github.com/tesseract-ocr/tesseract](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftesseract-ocr%2Ftesseract)'

Next, to get the number of starts, we notice that it is contained within an span tag which has the count Counter js-social-count.

article\_tags[4]

<article class="border rounded color-shadow-small color-bg-subtle my-4">

<div class="px-3">

<div class="d-flex flex-justify-between my-3">

<div class="d-flex flex-auto">

<span style="margin-top:2px">

<svg aria-hidden="true" class="octicon octicon-repo color-fg-muted mr-2" data-view-component="true" height="16" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M2 2.5A2.5 2.5 0 014.5 0h8.75a.75.75 0 01.75.75v12.5a.75.75 0 01-.75.75h-2.5a.75.75 0 110-1.5h1.75v-2h-8a1 1 0 00-.714 1.7.75.75 0 01-1.072 1.05A2.495 2.495 0 012 11.5v-9zm10.5-1V9h-8c-.356 0-.694.074-1 .208V2.5a1 1 0 011-1h8zM5 12.25v3.25a.25.25 0 00.4.2l1.45-1.087a.25.25 0 01.3 0L8.6 15.7a.25.25 0 00.4-.2v-3.25a.25.25 0 00-.25-.25h-3.5a.25.25 0 00-.25.25z" fill-rule="evenodd"></path>

</svg>

</span>

<h3 class="f3 color-fg-muted text-normal lh-condensed">

<a data-ga-click="Explore, go to repository owner, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"OWNER","click\_visual\_representation":"REPOSITORY\_OWNER\_HEADING","actor\_id":null,"record\_id":8401422,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="c14e22ad4cc65812caacdf8b76899e4e4979ea7d56890715930cef3abddde9c5" data-view-component="true" href="/tesseract-ocr">

tesseract-ocr

</a> /

<a class="text-bold wb-break-word" data-ga-click="Explore, go to repository, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"REPOSITORY","click\_visual\_representation":"REPOSITORY\_NAME\_HEADING","actor\_id":null,"record\_id":22887094,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="0d1578eafced610d5100a5880e0d084554b47f4b1fa6ce1de3ce9d413687d2a2" data-view-component="true" href="/tesseract-ocr/tesseract">

tesseract

</a> </h3>

</div>

<div class="d-flex flex-items-center ml-3">

<div class="BtnGroup d-flex" data-view-component="true">

<a aria-label="You must be signed in to star a repository" class="tooltipped tooltipped-s btn-sm btn BtnGroup-item color-bg-default" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"star button","repository\_id":22887094,"auth\_type":"LOG\_IN","originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="b65dfec4fdcf928d6b7828d6fa3736091509ebc16a861dc9ebca17b980101a9a" data-view-component="true" href="/login?return\_to=%2Ftesseract-ocr%2Ftesseract" rel="nofollow"> <svg aria-hidden="true" class="octicon octicon-star v-align-text-bottom d-inline-block mr-2" data-view-component="true" height="16" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M8 .25a.75.75 0 01.673.418l1.882 3.815 4.21.612a.75.75 0 01.416 1.279l-3.046 2.97.719 4.192a.75.75 0 01-1.088.791L8 12.347l-3.766 1.98a.75.75 0 01-1.088-.79l.72-4.194L.818 6.374a.75.75 0 01.416-1.28l4.21-.611L7.327.668A.75.75 0 018 .25zm0 2.445L6.615 5.5a.75.75 0 01-.564.41l-3.097.45 2.24 2.184a.75.75 0 01.216.664l-.528 3.084 2.769-1.456a.75.75 0 01.698 0l2.77 1.456-.53-3.084a.75.75 0 01.216-.664l2.24-2.183-3.096-.45a.75.75 0 01-.564-.41L8 2.694v.001z" fill-rule="evenodd"></path>

</svg><span class="d-inline" data-view-component="true">

Star

</span> <span aria-label="43557 users starred this repository" class="Counter js-social-count" data-pjax-replace="true" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-view-component="true" id="repo-stars-counter-star" title="43,557">43.6k</span>

</a> <button aria-label="You must be signed in to add this repository to a list" class="btn-sm btn BtnGroup-item px-2" data-view-component="true" disabled="disabled" type="button"> <svg aria-hidden="true" class="octicon octicon-triangle-down" data-view-component="true" height="16" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M4.427 7.427l3.396 3.396a.25.25 0 00.354 0l3.396-3.396A.25.25 0 0011.396 7H4.604a.25.25 0 00-.177.427z"></path>

</svg>

</button></div>

</div>

</div>

</div>

<nav aria-label="Repository menu" class="tabnav px-3 mb-0" data-view-component="true">

<ul class="tabnav-tabs" data-view-component="true">

<li class="d-inline-flex" data-view-component="true">

<a aria-current="true" class="tabnav-tab f6 px-2 py-1" data-ga-click="Explore, go to repository, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"REPOSITORY","click\_visual\_representation":"CODE\_TAB","actor\_id":null,"record\_id":22887094,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="80ac6cc31bd00478a8bd80b601193c7b4bbb8b5139a01fb5066bedba45a7bf5f" data-view-component="true" href="/tesseract-ocr/tesseract" id="code-tab-22887094">

<svg aria-hidden="true" class="octicon octicon-code color-fg-default" data-view-component="true" height="16" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M4.72 3.22a.75.75 0 011.06 1.06L2.06 8l3.72 3.72a.75.75 0 11-1.06 1.06L.47 8.53a.75.75 0 010-1.06l4.25-4.25zm6.56 0a.75.75 0 10-1.06 1.06L13.94 8l-3.72 3.72a.75.75 0 101.06 1.06l4.25-4.25a.75.75 0 000-1.06l-4.25-4.25z" fill-rule="evenodd"></path>

</svg> Code

</a></li>

<li class="d-inline-flex" data-view-component="true">

<a class="tabnav-tab f6 px-2 py-1" data-ga-click="Explore, go to repository issues, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"REPOSITORY\_ISSUES","click\_visual\_representation":"ISSUES\_TAB","actor\_id":null,"record\_id":22887094,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="416a2c38c2576549fd811ae282b109cda1ea66ae176969689676e10c19f0ad85" data-view-component="true" href="/tesseract-ocr/tesseract/issues" id="issues-tab-22887094">

<svg aria-hidden="true" class="octicon octicon-issue-opened color-gray-light" height="16" version="1.1" viewbox="0 0 16 16" width="16"><path d="M8 9.5a1.5 1.5 0 100-3 1.5 1.5 0 000 3z"></path><path d="M8 0a8 8 0 100 16A8 8 0 008 0zM1.5 8a6.5 6.5 0 1113 0 6.5 6.5 0 01-13 0z" fill-rule="evenodd"></path></svg> Issues

</a></li>

<li class="d-inline-flex" data-view-component="true">

<a class="tabnav-tab f6 px-2 py-1" data-ga-click="Explore, go to repository pulls, location:explore feed" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"REPOSITORY\_PULL\_REQUESTS","click\_visual\_representation":"PULL\_REQUESTS\_TAB","actor\_id":null,"record\_id":22887094,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="7e2018d96533d65ff3b7423c9a301beaaf3b7ab780e49502dafbad36cf41b846" data-view-component="true" href="/tesseract-ocr/tesseract/pulls" id="pull-requests-tab-22887094">

<svg aria-hidden="true" class="octicon octicon-git-pull-request color-fg-muted" data-view-component="true" height="16" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M7.177 3.073L9.573.677A.25.25 0 0110 .854v4.792a.25.25 0 01-.427.177L7.177 3.427a.25.25 0 010-.354zM3.75 2.5a.75.75 0 100 1.5.75.75 0 000-1.5zm-2.25.75a2.25 2.25 0 113 2.122v5.256a2.251 2.251 0 11-1.5 0V5.372A2.25 2.25 0 011.5 3.25zM11 2.5h-1V4h1a1 1 0 011 1v5.628a2.251 2.251 0 101.5 0V5A2.5 2.5 0 0011 2.5zm1 10.25a.75.75 0 111.5 0 .75.75 0 01-1.5 0zM3.75 12a.75.75 0 100 1.5.75.75 0 000-1.5z" fill-rule="evenodd"></path>

</svg> Pull requests

</a></li>

</ul>

</nav>

<div class="color-bg-default rounded-bottom-1">

<div class="px-3 pt-3">

<p class="color-fg-muted mb-0"></p><div>Tesseract Open Source OCR Engine (main repository)</div>

</div>

<div class="d-flex flex-wrap border-bottom color-border-muted px-3 pt-2 pb-2">

<a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":2169,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="2ae010ad1be3e78b1c6ad952e345c55d08969a9561b27d50e493f46d399ebfcb" data-view-component="true" href="/topics/machine-learning" title="Topic: machine-learning">

machine-learning

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":7573,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="0433e82aedb8be1925854ebb504f1cec679aa18581ca9304cea2aeeee5953bcc" data-view-component="true" href="/topics/ocr" title="Topic: ocr">

ocr

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":21692,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="5cf5e16da1d566fe587136d0ff7902f72a9db771edfc208250681c493b738469" data-view-component="true" href="/topics/tesseract" title="Topic: tesseract">

tesseract

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":39557,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="bea05b25027d09323812acf59f0e7c866c1af7745fc18ad0ac284f30832e7235" data-view-component="true" href="/topics/lstm" title="Topic: lstm">

lstm

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":60505,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="432db8dc97bd307285fcc806716c00e6aa0201edf6d823774945c9742ec8b440" data-view-component="true" href="/topics/tesseract-ocr" title="Topic: tesseract-ocr">

tesseract-ocr

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":196454,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="519eda089dceb603999dc257be730819fdc911dbcbe18c4764edbf59a221744b" data-view-component="true" href="/topics/hacktoberfest" title="Topic: hacktoberfest">

hacktoberfest

</a> <a class="topic-tag topic-tag-link f6 mb-2" data-ga-click="Explore, go to topic, location:explore feed repository" data-hydro-click='{"event\_type":"explore.click","payload":{"click\_context":"REPOSITORY\_CARD","click\_target":"TOPIC","click\_visual\_representation":"TOPIC\_TAG","actor\_id":null,"record\_id":297861,"originating\_url":"[https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning)","user\_id":null}}' data-hydro-click-hmac="d593f14eb50d76e87962e7b017bd3a827d64cf663cd6640258d0c4e570750eba" data-view-component="true" href="/topics/ocr-engine" title="Topic: ocr-engine">

ocr-engine

</a> </div>

<div class="p-3">

<ul class="d-flex f6 list-style-none color-fg-muted">

<li class="mr-4">

Updated

<relative-time class="no-wrap" datetime="2022-01-12T18:22:52Z">Jan 12, 2022</relative-time>

</li>

<li class="mr-4">

<span class="f6 my-1 ml-0">

<span class="repo-language-color" style="background-color: #f34b7d"></span>

<span itemprop="programmingLanguage">C++</span>

</span>

</li>

</ul>

</div>

</div>

</article>

a\_star\_tag = article\_tags[4].find('span', class\_='Counter js-social-count')

a\_star\_tag

<span aria-label="43557 users starred this repository" class="Counter js-social-count" data-pjax-replace="true" data-plural-suffix="users starred this repository" data-singular-suffix="user starred this repository" data-view-component="true" id="repo-stars-counter-star" title="43,557">43.6k</span>

Let's extract the star count from the a tag.

a\_star\_tag.text.strip()

'43.6k'

The k at the end indicates 1000. Let's write a helper function which can convert strings like 40.3k into the number 40,300.

def **parse\_star\_count**(stars\_str):

stars\_str = stars\_str.strip()

if stars\_str[-1] == 'k':

return int(float(stars\_str[:-1]) \* 1000)

else:

return int(stars\_str)

parse\_star\_count('40.3k')

40300

parse\_star\_count('991')

991

We can now determine the star count as a number.

star\_count = parse\_star\_count(a\_star\_tag.text.strip())

star\_count

43600

Perfect, we've extracted all the information we were interested in.

print('Repository name:', repo\_name)

print("Owner's username:", username)

print('Stars:', star\_count)

print('Repository URL:', repo\_url)

Repository name: tesseract Owner's username: tesseract-ocr Stars: 43600 Repository URL: [https://github.com/tesseract-ocr/tesseract](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftesseract-ocr%2Ftesseract)

Let's extract the logic for parsing the required information from an article tag into a function.

**QUESTION**: Write a function parse\_repostory that returns a dictionary containing the repository name, owner's username, number of stars, and repository URL by parsing a given article tag representing a repository.

def **parse\_repository**(article\_tag):

*# <a> tags containing username, repository name and URL*

a\_tags = article\_tag.h3.find\_all('a')

*# Owner's username*

username = a\_tags[0].text.strip()

*# Repository name*

repo\_name = a\_tags[1].text.strip()

*# Repository URL*

repo\_url = base\_url + a\_tags[1]['href'].strip()

*# Star count*

stars\_tag = article\_tag.find('span', class\_='Counter js-social-count')

star\_count = parse\_star\_count(stars\_tag.text.strip())

*# Return a dictionary*

return {

'repository\_name': repo\_name,

'owner\_username': username,

'stars': star\_count,

'repository\_url': repo\_url

}

We can now use the function to parse any article tag.

parse\_repository(article\_tags[0])

{'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'}

parse\_repository(article\_tags[10])

{'repository\_name': '100-Days-Of-ML-Code',

'owner\_username': 'Avik-Jain',

'stars': 34200,

'repository\_url': '[https://github.com/Avik-Jain/100-Days-Of-ML-Code](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2FAvik-Jain%2F100-Days-Of-ML-Code)'}

We can use a list comprehension to parse all the article tags in one go.

top\_repositories = [parse\_repository(tag) for tag in article\_tags]

len(top\_repositories)

30

top\_repositories[:5]

[{'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'},

{'repository\_name': 'keras',

'owner\_username': 'keras-team',

'stars': 53700,

'repository\_url': '[https://github.com/keras-team/keras](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fkeras-team%2Fkeras)'},

{'repository\_name': 'pytorch',

'owner\_username': 'pytorch',

'stars': 53300,

'repository\_url': '[https://github.com/pytorch/pytorch](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fpytorch%2Fpytorch)'},

{'repository\_name': 'scikit-learn',

'owner\_username': 'scikit-learn',

'stars': 48500,

'repository\_url': '[https://github.com/scikit-learn/scikit-learn](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fscikit-learn%2Fscikit-learn)'},

{'repository\_name': 'tesseract',

'owner\_username': 'tesseract-ocr',

'stars': 43600,

'repository\_url': '[https://github.com/tesseract-ocr/tesseract](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftesseract-ocr%2Ftesseract)'}]

**QUESTION**: Write a function that takes a BeautifulSoup object representing a topic page and returns a list of dictionaries containing information about the top repositories for the topic.

def **get\_top\_repositories**(doc):

article\_tags = doc.find\_all('article', class\_='border rounded color-shadow-small color-bg-subtle my-4')

topic\_repos = [parse\_repository(tag) for tag in article\_tags]

return topic\_repos

We can now use the functions we've defined to get the top repositories for any topic.

topic\_page\_ml = get\_topic\_page('machine-learning')

top\_repos\_ml = get\_top\_repositories(topic\_page\_ml)

top\_repos\_ml[:5]

[{'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'},

{'repository\_name': 'keras',

'owner\_username': 'keras-team',

'stars': 53700,

'repository\_url': '[https://github.com/keras-team/keras](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fkeras-team%2Fkeras)'},

{'repository\_name': 'pytorch',

'owner\_username': 'pytorch',

'stars': 53300,

'repository\_url': '[https://github.com/pytorch/pytorch](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fpytorch%2Fpytorch)'},

{'repository\_name': 'scikit-learn',

'owner\_username': 'scikit-learn',

'stars': 48500,

'repository\_url': '[https://github.com/scikit-learn/scikit-learn](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fscikit-learn%2Fscikit-learn)'},

{'repository\_name': 'tesseract',

'owner\_username': 'tesseract-ocr',

'stars': 43600,

'repository\_url': '[https://github.com/tesseract-ocr/tesseract](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftesseract-ocr%2Ftesseract)'}]

Here are the top repositories for the keyword data-analysis.

topic\_page\_da = get\_topic\_page('data-analysis')

top\_repos\_da = get\_top\_repositories(topic\_page\_da)

top\_repos\_da[:5]

[{'repository\_name': 'scikit-learn',

'owner\_username': 'scikit-learn',

'stars': 48500,

'repository\_url': '[https://github.com/scikit-learn/scikit-learn](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fscikit-learn%2Fscikit-learn)'},

{'repository\_name': 'superset',

'owner\_username': 'apache',

'stars': 43500,

'repository\_url': '[https://github.com/apache/superset](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fapache%2Fsuperset)'},

{'repository\_name': 'pandas',

'owner\_username': 'pandas-dev',

'stars': 32299,

'repository\_url': '[https://github.com/pandas-dev/pandas](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fpandas-dev%2Fpandas)'},

{'repository\_name': 'metabase',

'owner\_username': 'metabase',

'stars': 27100,

'repository\_url': '[https://github.com/metabase/metabase](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fmetabase%2Fmetabase)'},

{'repository\_name': 'streamlit',

'owner\_username': 'streamlit',

'stars': 17300,

'repository\_url': '[https://github.com/streamlit/streamlit](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fstreamlit%2Fstreamlit)'}]

Here are the top repositories for the keyword python

get\_top\_repositories(get\_topic\_page('python'))[:5]

[{'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'},

{'repository\_name': 'system-design-primer',

'owner\_username': 'donnemartin',

'stars': 158000,

'repository\_url': '[https://github.com/donnemartin/system-design-primer](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fdonnemartin%2Fsystem-design-primer)'},

{'repository\_name': 'CS-Notes',

'owner\_username': 'CyC2018',

'stars': 145000,

'repository\_url': '[https://github.com/CyC2018/CS-Notes](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2FCyC2018%2FCS-Notes)'},

{'repository\_name': 'Python',

'owner\_username': 'TheAlgorithms',

'stars': 127000,

'repository\_url': '[https://github.com/TheAlgorithms/Python](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2FTheAlgorithms%2FPython)'},

{'repository\_name': 'awesome-python',

'owner\_username': 'vinta',

'stars': 113000,

'repository\_url': '[https://github.com/vinta/awesome-python](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fvinta%2Fawesome-python)'}]

Do you see the power of defining functions and using libraries? With just one line of code, we can scrape GitHub and find the top repositories for any topic.

Let's save our work before continuing.

jovian.commit()

[jovian] Updating notebook "aakashns/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Committed successfully! <https://jovian.ai/aakashns/python-web-scraping-and-rest-api>

'<https://jovian.ai/aakashns/python-web-scraping-and-rest-api>'

**Writing information to CSV files**

Let's create a helper function which takes a list of dictionaries and writes them to a CSV file.

The input to our function will be a list of dictionary of the form:

[

{'key1': 'abc', 'key2': 'def', 'key3': 'ghi'},

{'key1': 'jkl', 'key2': 'mno', 'key3': 'pqr'},

{'key1': 'stu', 'key2': 'vwx', 'key3': 'yza'}

...

]

The function will create a file with a given name containing the following data:

key1,key2,key3

abc,def,ghi

jkl,mno,pqr

stu,vwx,yza

def **write\_csv**(items, path):

*# Open the file in write mode*

with open(path, 'w') as f:

*# Return if there's nothing to write*

if len(items) == 0:

return

*# Write the headers in the first line*

headers = list(items[0].keys())

f.write(','.join(headers) + '\n')

*# Write one item per line*

for item in items:

values = []

for header in headers:

values.append(str(item.get(header, "")))

f.write(','.join(values) + "\n")

Let's write the data stored in top\_repos\_ml into a CSV file.

len(top\_repos\_ml)

30

top\_repos\_ml[:3]

[{'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'},

{'repository\_name': 'keras',

'owner\_username': 'keras-team',

'stars': 53700,

'repository\_url': '[https://github.com/keras-team/keras](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fkeras-team%2Fkeras)'},

{'repository\_name': 'pytorch',

'owner\_username': 'pytorch',

'stars': 53300,

'repository\_url': '[https://github.com/pytorch/pytorch](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fpytorch%2Fpytorch)'}]

write\_csv(top\_repositories, 'machine-learning.csv')

We can now read the file and inspect its contents. The contents of the file can also be inspected using the "File > Open" menu option within Jupyter.

with open('machine-learning.csv', 'r') as f:

print(f.read())

repository\_name,owner\_username,stars,repository\_url tensorflow,tensorflow,162000,https://github.com/tensorflow/tensorflow keras,keras-team,53700,https://github.com/keras-team/keras pytorch,pytorch,53300,https://github.com/pytorch/pytorch scikit-learn,scikit-learn,48500,https://github.com/scikit-learn/scikit-learn tesseract,tesseract-ocr,43600,https://github.com/tesseract-ocr/tesseract face\_recognition,ageitgey,42800,https://github.com/ageitgey/face\_recognition TensorFlow-Examples,aymericdamien,41600,https://github.com/aymericdamien/TensorFlow-Examples faceswap,deepfakes,40100,https://github.com/deepfakes/faceswap julia,JuliaLang,37800,https://github.com/JuliaLang/julia awesome-scalability,binhnguyennus,37000,https://github.com/binhnguyennus/awesome-scalability 100-Days-Of-ML-Code,Avik-Jain,34200,https://github.com/Avik-Jain/100-Days-Of-ML-Code caffe,BVLC,32200,https://github.com/BVLC/caffe DeepFaceLab,iperov,30700,https://github.com/iperov/DeepFaceLab d2l-zh,d2l-ai,29700,https://github.com/d2l-ai/d2l-zh MadeWithML,GokuMohandas,29400,https://github.com/GokuMohandas/MadeWithML ML-For-Beginners,microsoft,28600,https://github.com/microsoft/ML-For-Beginners machine-learning-for-software-engineers,ZuzooVn,25600,https://github.com/ZuzooVn/machine-learning-for-software-engineers cs-video-courses,Developer-Y,25500,https://github.com/Developer-Y/cs-video-courses handson-ml,ageron,23900,https://github.com/ageron/handson-ml Coursera-ML-AndrewNg-Notes,fengdu78,23800,https://github.com/fengdu78/Coursera-ML-AndrewNg-Notes awesome-deep-learning-papers,terryum,23500,https://github.com/terryum/awesome-deep-learning-papers openpose,CMU-Perceptual-Computing-Lab,23100,https://github.com/CMU-Perceptual-Computing-Lab/openpose data-science-ipython-notebooks,donnemartin,22300,https://github.com/donnemartin/data-science-ipython-notebooks spaCy,explosion,22200,https://github.com/explosion/spaCy xgboost,dmlc,22100,https://github.com/dmlc/xgboost fastai,fastai,21800,https://github.com/fastai/fastai google-research,google-research,21400,https://github.com/google-research/google-research yolov5,ultralytics,20900,https://github.com/ultralytics/yolov5 ML-From-Scratch,eriklindernoren,20700,https://github.com/eriklindernoren/ML-From-Scratch NLP-progress,sebastianruder,19600,https://github.com/sebastianruder/NLP-progress

Perfect! We've created a CSV containing the information about the top GitHub repositories for the topic machine-learning. We can now put together everything we've done so far to solve the original problem.

**QUESTION**: Write a Python function that creates a CSV file (comma-separated values) containing details about the 25 top GitHub repositories for any given topic. The top repositories for the topic machine-learning can be found on this page: [https://github.com/topics/machine-learning](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning). The output CSV should contain these details: repository name, owner's username, no. of stars, repository URL.

import requests

from bs4 import BeautifulSoup

base\_url = 'https://github.com'

def **scrape\_topic\_repositories**(topic, path=None):

"""Get the top repositories for a topic and write them to a CSV file"""

if path is None:

path = topic + '.csv'

topic\_page\_doc = get\_topic\_page(topic)

topic\_repositories = get\_top\_repositories(topic\_page\_doc)

write\_csv(topic\_repositories, path)

print('Top repositories for topic "{}" written to file "{}"'.format(topic, path))

return path

def **get\_top\_repositories**(doc):

"""Parse the top repositories for a topic given a Beautiful Soup document"""

article\_tags = doc.find\_all('article', class\_='border rounded color-shadow-small color-bg-subtle my-4')

topic\_repos = [parse\_repository(tag) for tag in article\_tags]

return topic\_repos

def **get\_topic\_page**(topic):

"""Get the web page containing the top repositories for a topic as a Beautiful Soup document"""

topic\_repos\_url = 'https://github.com/topics/' + topic

response = requests.get(topic\_repos\_url)

if response.status\_code != 200:

print('Status code:', response.status\_code)

raise Exception('Failed to fetch web page ' + topic\_repos\_url)

return BeautifulSoup(response.text)

def **parse\_repository**(article\_tag):

"""Parse information about a repository from an <article> tag"""

a\_tags = article\_tag.h3.find\_all('a')

username = a\_tags[0].text.strip()

repo\_name = a\_tags[1].text.strip()

repo\_url = base\_url + a\_tags[1]['href'].strip()

stars\_tag = article\_tag.find('span', class\_='Counter js-social-count')

star\_count = parse\_star\_count(stars\_tag.text.strip())

return {'repository\_name': repo\_name, 'owner\_username': username, 'stars': star\_count, 'repository\_url': repo\_url}

def **parse\_star\_count**(stars\_str):

"""Parse strings like 40.3k and get the no. of stars as a number"""

stars\_str = stars\_str.strip()

return int(float(stars\_str[:-1]) \* 1000) if stars\_str[-1] == 'k' else int(stars\_str)

def **write\_csv**(items, path):

"""Write a list of dictionaries to a CSV file"""

with open(path, 'w') as f:

if len(items) == 0:

return

headers = list(items[0].keys())

f.write(','.join(headers) + '\n')

for item in items:

values = []

for header in headers:

values.append(str(item.get(header, "")))

f.write(','.join(values) + "\n")

The entire code of this problem is only about 50 lines long. Isn't that neat?

Put another way, if you understand these 50 lines of code, you know pretty much all there is to know about web scraping. Use the interactive nature of Jupyter to experiment with each function and add print statements wherever required to display intermediate output. Reading and understanding code is an essential skill for programmers.

scrape\_topic\_repositories('machine-learning')

Top repositories for topic "machine-learning" written to file "machine-learning.csv"

'machine-learning.csv'

Now that we have a CSV file, we can use the pandas library to view its contents.

import pandas as pd

pd.read\_csv('machine-learning.csv')

scrape\_topic\_repositories('data-analysis')

Top repositories for topic "data-analysis" written to file "data-analysis.csv"

'data-analysis.csv'

pd.read\_csv('data-analysis.csv')

scrape\_topic\_repositories('python')

Top repositories for topic "python" written to file "python.csv"

'python.csv'

pd.read\_csv('python.csv')

Of course, we can go even further and write a function that scrapes top repositories for several topics.

**EXERCISE**: Write a function scrape\_topics which takes a list of topics and creates CSV files containing top repositories for a list of topics. Test it out using the empty cells below.

Let's save our work before continuing.

jovian.commit(files=['machine-learning.csv', 'python.csv', 'data-analysis.csv'])

[jovian] Updating notebook "aakashns/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Uploading additional files... [jovian] Committed successfully! <https://jovian.ai/aakashns/python-web-scraping-and-rest-api>

'<https://jovian.ai/aakashns/python-web-scraping-and-rest-api>'

**Using a REST API to retrieve data as JSON**

Not all URLs point to an HTML page. Consider this URL for example: [https://api.github.com/repos/octocat/hello-world](https://jovian.ai/outlink?url=https%3A%2F%2Fapi.github.com%2Frepos%2Foctocat%2Fhello-world) . It points to a JSON document, which has a structure like this:

{

"name": "Hello-World",

"full\_name": "octocat/Hello-World",

"private": false,

"owner": {

"login": "octocat",

"id": 583231,

},

"html\_url": "https://github.com/octocat/Hello-World",

}

It's quite similar to a Python dictionary. In fact, you can use the json module from python to convert a JSON document into a Python dictionary.

response = requests.get('https://api.github.com/repos/octocat/hello-world')

import json

data\_dict = json.loads(response.text)

data\_dict

{'id': 1296269,

'node\_id': 'MDEwOlJlcG9zaXRvcnkxMjk2MjY5',

'name': 'Hello-World',

'full\_name': 'octocat/Hello-World',

'private': False,

'owner': {'login': 'octocat',

'id': 583231,

'node\_id': 'MDQ6VXNlcjU4MzIzMQ==',

'avatar\_url': 'https://avatars.githubusercontent.com/u/583231?v=4',

'gravatar\_id': '',

'url': 'https://api.github.com/users/octocat',

'html\_url': 'https://github.com/octocat',

'followers\_url': 'https://api.github.com/users/octocat/followers',

'following\_url': 'https://api.github.com/users/octocat/following{/other\_user}',

'gists\_url': 'https://api.github.com/users/octocat/gists{/gist\_id}',

'starred\_url': 'https://api.github.com/users/octocat/starred{/owner}{/repo}',

'subscriptions\_url': 'https://api.github.com/users/octocat/subscriptions',

'organizations\_url': 'https://api.github.com/users/octocat/orgs',

'repos\_url': 'https://api.github.com/users/octocat/repos',

'events\_url': 'https://api.github.com/users/octocat/events{/privacy}',

'received\_events\_url': 'https://api.github.com/users/octocat/received\_events',

'type': 'User',

'site\_admin': False},

'html\_url': 'https://github.com/octocat/Hello-World',

'description': 'My first repository on GitHub!',

'fork': False,

'url': 'https://api.github.com/repos/octocat/Hello-World',

'forks\_url': 'https://api.github.com/repos/octocat/Hello-World/forks',

'keys\_url': 'https://api.github.com/repos/octocat/Hello-World/keys{/key\_id}',

'collaborators\_url': 'https://api.github.com/repos/octocat/Hello-World/collaborators{/collaborator}',

'teams\_url': 'https://api.github.com/repos/octocat/Hello-World/teams',

'hooks\_url': 'https://api.github.com/repos/octocat/Hello-World/hooks',

'issue\_events\_url': 'https://api.github.com/repos/octocat/Hello-World/issues/events{/number}',

'events\_url': 'https://api.github.com/repos/octocat/Hello-World/events',

'assignees\_url': 'https://api.github.com/repos/octocat/Hello-World/assignees{/user}',

'branches\_url': 'https://api.github.com/repos/octocat/Hello-World/branches{/branch}',

'tags\_url': 'https://api.github.com/repos/octocat/Hello-World/tags',

'blobs\_url': 'https://api.github.com/repos/octocat/Hello-World/git/blobs{/sha}',

'git\_tags\_url': 'https://api.github.com/repos/octocat/Hello-World/git/tags{/sha}',

'git\_refs\_url': 'https://api.github.com/repos/octocat/Hello-World/git/refs{/sha}',

'trees\_url': 'https://api.github.com/repos/octocat/Hello-World/git/trees{/sha}',

'statuses\_url': 'https://api.github.com/repos/octocat/Hello-World/statuses/{sha}',

'languages\_url': 'https://api.github.com/repos/octocat/Hello-World/languages',

'stargazers\_url': 'https://api.github.com/repos/octocat/Hello-World/stargazers',

'contributors\_url': 'https://api.github.com/repos/octocat/Hello-World/contributors',

'subscribers\_url': 'https://api.github.com/repos/octocat/Hello-World/subscribers',

'subscription\_url': 'https://api.github.com/repos/octocat/Hello-World/subscription',

'commits\_url': 'https://api.github.com/repos/octocat/Hello-World/commits{/sha}',

'git\_commits\_url': 'https://api.github.com/repos/octocat/Hello-World/git/commits{/sha}',

'comments\_url': 'https://api.github.com/repos/octocat/Hello-World/comments{/number}',

'issue\_comment\_url': 'https://api.github.com/repos/octocat/Hello-World/issues/comments{/number}',

'contents\_url': 'https://api.github.com/repos/octocat/Hello-World/contents/{+path}',

'compare\_url': 'https://api.github.com/repos/octocat/Hello-World/compare/{base}...{head}',

'merges\_url': 'https://api.github.com/repos/octocat/Hello-World/merges',

'archive\_url': 'https://api.github.com/repos/octocat/Hello-World/{archive\_format}{/ref}',

'downloads\_url': 'https://api.github.com/repos/octocat/Hello-World/downloads',

'issues\_url': 'https://api.github.com/repos/octocat/Hello-World/issues{/number}',

'pulls\_url': 'https://api.github.com/repos/octocat/Hello-World/pulls{/number}',

'milestones\_url': 'https://api.github.com/repos/octocat/Hello-World/milestones{/number}',

'notifications\_url': 'https://api.github.com/repos/octocat/Hello-World/notifications{?since,all,participating}',

'labels\_url': 'https://api.github.com/repos/octocat/Hello-World/labels{/name}',

'releases\_url': 'https://api.github.com/repos/octocat/Hello-World/releases{/id}',

'deployments\_url': 'https://api.github.com/repos/octocat/Hello-World/deployments',

'created\_at': '2011-01-26T19:01:12Z',

'updated\_at': '2022-01-13T04:09:25Z',

'pushed\_at': '2022-01-13T03:39:27Z',

'git\_url': 'git://github.com/octocat/Hello-World.git',

'ssh\_url': 'git@github.com:octocat/Hello-World.git',

'clone\_url': 'https://github.com/octocat/Hello-World.git',

'svn\_url': 'https://github.com/octocat/Hello-World',

'homepage': '',

'size': 1,

'stargazers\_count': 1748,

'watchers\_count': 1748,

'language': None,

'has\_issues': True,

'has\_projects': True,

'has\_downloads': True,

'has\_wiki': True,

'has\_pages': False,

'forks\_count': 1680,

'mirror\_url': None,

'archived': False,

'disabled': False,

'open\_issues\_count': 752,

'license': None,

'allow\_forking': True,

'is\_template': False,

'topics': [],

'visibility': 'public',

'forks': 1680,

'open\_issues': 752,

'watchers': 1748,

'default\_branch': 'master',

'temp\_clone\_token': None,

'network\_count': 1680,

'subscribers\_count': 1746}

Unlike HTML, it's really easy to work with JSON using Python, simply fetch the contents of the URL and convert it to a dictionary. Such URLs are often called **REST APIs** or REST API endpoints. Many websites offer well-documented REST APIs to access data from the site in JSON format:

* GitHub: [https://docs.github.com/en/rest/reference/repos](https://jovian.ai/outlink?url=https%3A%2F%2Fdocs.github.com%2Fen%2Frest%2Freference%2Frepos)
* Facebook: [https://developers.facebook.com/docs/groups-api/reference](https://jovian.ai/outlink?url=https%3A%2F%2Fdevelopers.facebook.com%2Fdocs%2Fgroups-api%2Freference)
* Twitter: [https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user\_timeline](https://jovian.ai/outlink?url=https%3A%2F%2Fdeveloper.twitter.com%2Fen%2Fdocs%2Ftwitter-api%2Fv1%2Ftweets%2Ftimelines%2Fapi-reference%2Fget-statuses-user_timeline)
* Reddit: [https://www.reddit.com/dev/api/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.reddit.com%2Fdev%2Fapi%2F)

Using an API is the *officially supported* way of extracting information from a website. To use an API, you will often need to register as a developer on the platform and generate an API key, which you'll need to send with every request to authenticate yourself.

Since GitHub offers a public API, we can use it without any restrictions to fetch information about public repositories.

**QUESTION**: Write a function get\_repo\_details to find the following information about a repository: description, watcher count, fork count, open issues count, created at time and updated at time.

def **get\_repo\_details**(username, repo\_name):

print('Fetching information for {}/{}'.format(username, repo\_name))

repo\_details\_url = "https://api.github.com/repos/" + username + "/" + repo\_name

response = requests.get(repo\_details\_url)

if not response.ok:

print("Failed to fetch!")

return {}

repo\_data = json.loads(response.text)

return {

'description': repo\_data['description'],

'watchers': repo\_data['watchers\_count'],

'open\_issues': repo\_data['open\_issues\_count'],

'created\_at': repo\_data['created\_at'],

'updated\_at': repo\_data['updated\_at']

}

get\_repo\_details('octocat', 'hello-world')

Fetching information for octocat/hello-world

{'description': 'My first repository on GitHub!',

'watchers': 1748,

'open\_issues': 752,

'created\_at': '2011-01-26T19:01:12Z',

'updated\_at': '2022-01-13T04:09:25Z'}

get\_repo\_details('tensorflow', 'tensorflow')

Fetching information for tensorflow/tensorflow

{'description': 'An Open Source Machine Learning Framework for Everyone',

'watchers': 162000,

'open\_issues': 2544,

'created\_at': '2015-11-07T01:19:20Z',

'updated\_at': '2022-01-13T09:01:41Z'}

**QUESTION**: Augment the list of top repositories for a topic with the repository description, watcher count, fork count, open issues count, created at time and updated at time.

def **add\_repo\_details**(repos):

return [dict(\*\*get\_repo\_details(repo['owner\_username'], repo['repository\_name']), \*\*repo) for repo in repos]

add\_repo\_details(top\_repositories[:5])

Fetching information for tensorflow/tensorflow Fetching information for keras-team/keras Fetching information for pytorch/pytorch Fetching information for scikit-learn/scikit-learn Fetching information for tesseract-ocr/tesseract

[{'description': 'An Open Source Machine Learning Framework for Everyone',

'watchers': 162000,

'open\_issues': 2544,

'created\_at': '2015-11-07T01:19:20Z',

'updated\_at': '2022-01-13T09:01:41Z',

'repository\_name': 'tensorflow',

'owner\_username': 'tensorflow',

'stars': 162000,

'repository\_url': '[https://github.com/tensorflow/tensorflow](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftensorflow%2Ftensorflow)'},

{'description': 'Deep Learning for humans',

'watchers': 53666,

'open\_issues': 253,

'created\_at': '2015-03-28T00:35:42Z',

'updated\_at': '2022-01-13T05:22:16Z',

'repository\_name': 'keras',

'owner\_username': 'keras-team',

'stars': 53700,

'repository\_url': '[https://github.com/keras-team/keras](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fkeras-team%2Fkeras)'},

{'description': 'Tensors and Dynamic neural networks in Python with strong GPU acceleration',

'watchers': 53268,

'open\_issues': 11047,

'created\_at': '2016-08-13T05:26:41Z',

'updated\_at': '2022-01-13T09:03:50Z',

'repository\_name': 'pytorch',

'owner\_username': 'pytorch',

'stars': 53300,

'repository\_url': '[https://github.com/pytorch/pytorch](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fpytorch%2Fpytorch)'},

{'description': 'scikit-learn: machine learning in Python',

'watchers': 48530,

'open\_issues': 2506,

'created\_at': '2010-08-17T09:43:38Z',

'updated\_at': '2022-01-13T07:09:05Z',

'repository\_name': 'scikit-learn',

'owner\_username': 'scikit-learn',

'stars': 48500,

'repository\_url': '[https://github.com/scikit-learn/scikit-learn](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fscikit-learn%2Fscikit-learn)'},

{'description': 'Tesseract Open Source OCR Engine (main repository)',

'watchers': 43557,

'open\_issues': 348,

'created\_at': '2014-08-12T18:04:59Z',

'updated\_at': '2022-01-13T09:01:46Z',

'repository\_name': 'tesseract',

'owner\_username': 'tesseract-ocr',

'stars': 43600,

'repository\_url': '[https://github.com/tesseract-ocr/tesseract](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftesseract-ocr%2Ftesseract)'}]

You may get rate limited if you attempt to make more than 60 requests per hour. To overcome the rate limit, use the Github OAuth token as described here: [https://towardsdatascience.com/all-the-things-you-can-do-with-github-api-and-python-f01790fca131](https://jovian.ai/outlink?url=https%3A%2F%2Ftowardsdatascience.com%2Fall-the-things-you-can-do-with-github-api-and-python-f01790fca131)

Note: Never publish your Github API token publicly, as it can be used to access your Github account. To store your API token without displaying it on the screen, use getpass.

from getpass import getpass

token = getpass()

········

**EXERCISE**: Augment the list of top repositories for a topic with some additional information about the user/organization the repository belong to: name, description, Github URL, no. of repositories, type (user or organization) etc.

**Acronyms**

In case you're feeling overwhelmed by all the acronyms, here are their expansions:

* **REST**: Represetational State Transfer
* **API**: Application Programming Interface
* **JSON**: JavaScript Object Notation
* **URL**: Universal Resource Locator

Don't worry, you needn't remember any of them!

Let's save our work before continuing.

jovian.commit()

[jovian] Updating notebook "aakashns/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Committed successfully! <https://jovian.ai/aakashns/python-web-scraping-and-rest-api>

'<https://jovian.ai/aakashns/python-web-scraping-and-rest-api>'

**Crawling Websites by Parsing Links on a Page**

When you scrape you a web page, you are likely to find several links on the page. For, example, on the page [https://github.com/topics](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics), you will find links to several topic pages. You can parse all the topic page links from this page, and scrape those pages to get the top repositories for each topic. Further, you can parse all the repository links from a topic page and scrape individual repository pages, and so on.

The process of scraping a page, parsing links and then using the links to parsing other pages on the same site is called **web crawling**. It's how search engines like Google are able to index and search data from millions of websites on the internet. Python offer libraries like [Scrapy](https://jovian.ai/outlink?url=https%3A%2F%2Fscrapy.org) for crawling websites easily.

You can do some basic crawling with requests, Beautiful soup, and few simple for loops in Python. Here's an exercise to get you started

**EXERCISE**: Get the top 100 repositories for the all the featured topics on GitHub. You might find these URLs useful:

* Eighth page of featured topics: [https://github.com/topics/?page=8](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2F%3Fpage%3D8)
* Second page of top repositories for a topic: [https://github.com/topics/machine-learning?page=2](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning%3Fpage%3D2)

jovian.commit()

[jovian] Updating notebook "samanvitha/python-web-scraping-and-rest-api" on <https://jovian.ai> [jovian] Committed successfully! <https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>

'<https://jovian.ai/samanvitha/python-web-scraping-and-rest-api>'

**Summary and Further Reading**

We've covered the following topics in this tutorial:

* Downloading web pages using the requests library
* Inspecting the HTML source code of a web page
* Parsing parts of a website using Beautiful Soup
* Writing parsed information into CSV files
* Using a REST API to retrieve data as JSON
* Combining data from multiple sources
* Using links on a page to crawl a website

Here are some things to keep in mind w.r.t. web scraping:

* Most websites disallow web scraping for commercial purposes
* Prefer using web scraping only for learning and research purposes
* Some websites may block your IP or stop sending valid information if you send too many requests
* Review the terms and conditions of a website before scraping data from it
* Remove sensitive and personally identifiable information before publishing a dataset online
* Use official REST APIs wherever available, with proper API keys
* Scraping data that you see after logging in is harder (it requires special cookies and headers)
* Websites change their HTML layout frequently, which may cause your scarping scripts to break
* Websites with dynamic content cannot be scraped using BeautifulSoup. One way to scrape dynamic website is by using Selenium

Here are some more examples of scraping:

* [https://medium.com/@msalmon00/web-scraping-job-postings-from-indeed-96bd588dcb4b](https://jovian.ai/outlink?url=https%3A%2F%2Fmedium.com%2F%40msalmon00%2Fweb-scraping-job-postings-from-indeed-96bd588dcb4b)
* [https://medium.com/the-innovation/scraping-medium-with-python-beautiful-soup-3314f898bbf5](https://jovian.ai/outlink?url=https%3A%2F%2Fmedium.com%2Fthe-innovation%2Fscraping-medium-with-python-beautiful-soup-3314f898bbf5)
* [https://medium.com/brainstation23/how-to-become-a-pro-with-scraping-youtube-videos-in-3-minutes-a6ac56021961](https://jovian.ai/outlink?url=https%3A%2F%2Fmedium.com%2Fbrainstation23%2Fhow-to-become-a-pro-with-scraping-youtube-videos-in-3-minutes-a6ac56021961)
* [https://www.freecodecamp.org/news/web-scraping-python-tutorial-how-to-scrape-data-from-a-website/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.freecodecamp.org%2Fnews%2Fweb-scraping-python-tutorial-how-to-scrape-data-from-a-website%2F)
* [https://www.freecodecamp.org/news/scraping-wikipedia-articles-with-python/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.freecodecamp.org%2Fnews%2Fscraping-wikipedia-articles-with-python%2F)
* [https://towardsdatascience.com/web-scraping-yahoo-finance-477fe3daa852](https://jovian.ai/outlink?url=https%3A%2F%2Ftowardsdatascience.com%2Fweb-scraping-yahoo-finance-477fe3daa852)
* [https://www.analyticsvidhya.com/blog/2020/10/web-scraping-selenium-in-python/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.analyticsvidhya.com%2Fblog%2F2020%2F10%2Fweb-scraping-selenium-in-python%2F)
* [https://medium.com/ml-book/web-scraping-using-selenium-python-3be7b8762747](https://jovian.ai/outlink?url=https%3A%2F%2Fmedium.com%2Fml-book%2Fweb-scraping-using-selenium-python-3be7b8762747)

**Project Ideas**

Here are some project ideas if you're looking to work on a web scraping project. You can work of one of these ideas, or pick something entirely different.

1. **Dataset of Books (Amazon)**: Create a dataset of popular books in different genres by scraping the site: [https://www.amazon.in/gp/bestsellers/books/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.amazon.in%2Fgp%2Fbestsellers%2Fbooks%2F)
2. **Dataset of Quotes (BrainyQuote)**: Create a dataset of quotes for different tags/topics by scraping the site :[https://www.brainyquote.com/topics](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.brainyquote.com%2Ftopics" \t "_blank)
3. **Dataset of Movies (TMDb)**: The Movie Database (TMDb) contains information about thousands of movies from around the world: [https://www.themoviedb.org/movie](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.themoviedb.org%2Fmovie) . Can you scape the site to create a dataset of movies containing information like title, release date, cast, etc. ? You can also create datasets of movie actors/actresses/directors using this site.
4. **Dataset of TV Shows (TMDb)**: The Movie Database (TMDb) contains information about thousands of TV shows from around the world: [https://www.themoviedb.org/tv](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.themoviedb.org%2Ftv) . Can you scrape the site to create a dataset of TV shows containing information like title, release date, cast, crew, etc. ? You can also create datasets of TV actors/actresses/directors using this site.
5. **Collections of Popular Repositories (GitHub)**: Scape GitHub collections ( [https://github.com/collections](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Fcollections) ) to create a dataset of popular repositories organized by different use cases.
6. **Dataset of Books (BooksToScrape)**: Create a dataset of popular books in different genres by scraping the site *Books To Scrape*: [http://books.toscrape.com](https://jovian.ai/outlink?url=http%3A%2F%2Fbooks.toscrape.com)
7. **Dataset of Quotes (QuotesToScrape)**: Create a dataset of popular quotes for different tags by scraping the site *Quotes To Scrape*: [http://quotes.toscrape.com](https://jovian.ai/outlink?url=http%3A%2F%2Fquotes.toscrape.com)
8. **Scrape a User's Repositories (GitHub)**: Given someone's GitHub username, can you scrape their GitHub profile to create a list of their repositories with information like repository name, no. of stars, no. of forks, etc.?
9. **Scrape User's Reviews (ConsumerAffairs)**: Consumeraffairs contains reviews about thousands of brands: [https://www.consumeraffairs.com/](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.consumeraffairs.com%2F). Can you scrape any category from the site to create a dataset of Reviews containing information like Title, Rating, Reviews and toll-free number etc.?.
10. **Songs Dataset (AZLyrics)**: Create a dataset of songs by scraping AZLyrics: [https://www.azlyrics.com/f.html](https://jovian.ai/outlink?url=https%3A%2F%2Fwww.azlyrics.com%2Ff.html) . Capture information like song title, artist name, year of release and lyrics URL.
11. **Scrape a Popular Blog**: Create a dataset of blog posts on a popular blog e.g. [https://m.signalvnoise.com/search/](https://jovian.ai/outlink?url=https%3A%2F%2Fm.signalvnoise.com%2Fsearch%2F) . The dataset can contain information like the blog title, published date, tags, author, link to blog post, etc.
12. **Weekly Top Songs (Top 40 Weekly)**: Create a dataset of the top 40 songs of each week in a given year by scraping the site [https://top40weekly.com](https://jovian.ai/outlink?url=https%3A%2F%2Ftop40weekly.com) . Capture information like song title, artist, weekly rank, etc.

**Questions for Revision**

1. Why do we need to scrape websites?
2. What different tools can we use to scrape websites?
3. What are the applications of web-scraping?
4. What are the steps involved in web-scraping?
5. What are the techniques to get data from websites?
6. What technique is used to retrieve data in a machine-readable format in python?
7. How can one download a webpage from the internet using python?
8. What library do we need for downloading the webpage in python?
9. What function from the library do we need for downloading the webpage?
10. How do we make sure that the webpage is downloaded successfully?
11. How can we access the content of the downloaded webpage?
12. What function do we need to find out the total number of characters in the downloaded webpage?
13. What defines the content and structure of the downloaded webpage?
14. What is a source code? In what language is it usually written in?
15. How different are the original webpage and scraped webpage?
16. How many parts does HTML tag have? What are they?
17. Is it possible to be blocked by website when you scrape more pages? If yes, how can one avoid this?
18. How do we get the information we need from the downloaded website?
19. What library do we need to install to extract information from HTML source code?
20. What is doc object?
21. How can we access attributes of a tag?
22. How do we find the direct children of the tag?
23. What is the purpose of strip()?
24. How can we write the extracted information into CSV files?
25. What are REST APIs? How are they different from usual URLs?
26. What is the official way to extract information from a website? What do we need for that? How does it help one in extracting information?
27. What websites offer public APIs?
28. Can we extract data from all the websites on web? If not, why?
29. What is getpass()?
30. What is web crawling and how is it different from web scraping?
31. What are the applications of web crawling?
32. What does python offer for crawling websites?
33. How do we extract data from dynamic websites?

**Solutions for Exercises**

**EXERCISE**: Find the first occurrence of each of these tags in doc: div, img, span, p, etc.

first\_div=doc.find\_all('div')

*#first=doc.find('div')*

*#first=doc.div*

first\_div[0]

<div class="position-relative js-header-wrapper">

<a class="px-2 py-4 color-bg-accent-emphasis color-fg-on-emphasis show-on-focus js-skip-to-content" href="#start-of-content">Skip to content</a>

<span class="progress-pjax-loader js-pjax-loader-bar Progress position-fixed width-full" data-view-component="true">

<span class="Progress-item progress-pjax-loader-bar left-0 top-0 color-bg-accent-emphasis" data-view-component="true" style="width: 0%;"></span>

</span>

<header class="Header-old header-logged-out js-details-container Details position-relative f4 py-2" role="banner">

<div class="container-lg d-lg-flex flex-items-center p-responsive">

<div class="d-flex flex-justify-between flex-items-center">

<a aria-label="Homepage" class="mr-4 color-fg-inherit" data-ga-click="(Logged out) Header, go to homepage, icon:logo-wordmark" href="https://github.com/">

<svg aria-hidden="true" class="octicon octicon-mark-github" data-view-component="true" height="32" version="1.1" viewbox="0 0 16 16" width="32">

<path d="M8 0C3.58 0 0 3.58 0 8c0 3.54 2.29 6.53 5.47 7.59.4.07.55-.17.55-.38 0-.19-.01-.82-.01-1.49-2.01.37-2.53-.49-2.69-.94-.09-.23-.48-.94-.82-1.13-.28-.15-.68-.52-.01-.53.63-.01 1.08.58 1.23.82.72 1.21 1.87.87 2.33.66.07-.52.28-.87.51-1.07-1.78-.2-3.64-.89-3.64-3.95 0-.87.31-1.59.82-2.15-.08-.2-.36-1.02.08-2.12 0 0 .67-.21 2.2.82.64-.18 1.32-.27 2-.27.68 0 1.36.09 2 .27 1.53-1.04 2.2-.82 2.2-.82.44 1.1.16 1.92.08 2.12.51.56.82 1.27.82 2.15 0 3.07-1.87 3.75-3.65 3.95.29.25.54.73.54 1.48 0 1.07-.01 1.93-.01 2.2 0 .21.15.46.55.38A8.013 8.013 0 0016 8c0-4.42-3.58-8-8-8z" fill-rule="evenodd"></path>

</svg>

</a>

<div class="d-lg-none css-truncate css-truncate-target width-fit p-2">

</div>

<div class="d-flex flex-items-center">

<a class="d-inline-block d-lg-none f5 no-underline border color-border-default rounded-2 px-2 py-1 mr-3 mr-sm-5 color-fg-inherit" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="e755e218b5b5c8ea0dcf7c1afea104ce5b0e7bb40929d477942f46cf5b206533" href="/signup?ref\_cta=Sign+up&amp;ref\_loc=header+logged+out&amp;ref\_page=%2Ftopics%2Fmachine-learning&amp;source=header">

Sign up

</a>

<button aria-expanded="false" aria-label="Toggle navigation" class="js-details-target btn-link d-lg-none mt-1 color-fg-inherit" data-view-component="true" type="button"> <svg aria-hidden="true" class="octicon octicon-three-bars" data-view-component="true" height="24" version="1.1" viewbox="0 0 16 16" width="24">

<path d="M1 2.75A.75.75 0 011.75 2h12.5a.75.75 0 110 1.5H1.75A.75.75 0 011 2.75zm0 5A.75.75 0 011.75 7h12.5a.75.75 0 110 1.5H1.75A.75.75 0 011 7.75zM1.75 12a.75.75 0 100 1.5h12.5a.75.75 0 100-1.5H1.75z" fill-rule="evenodd"></path>

</svg>

</button> </div>

</div>

<div class="HeaderMenu HeaderMenu--logged-out position-fixed top-0 right-0 bottom-0 height-fit position-lg-relative d-lg-flex flex-justify-between flex-items-center flex-auto">

<div class="d-flex d-lg-none flex-justify-end border-bottom color-bg-subtle p-3">

<button aria-expanded="false" aria-label="Toggle navigation" class="js-details-target btn-link" data-view-component="true" type="button"> <svg aria-hidden="true" class="octicon octicon-x color-fg-muted" data-view-component="true" height="24" version="1.1" viewbox="0 0 24 24" width="24">

<path d="M5.72 5.72a.75.75 0 011.06 0L12 10.94l5.22-5.22a.75.75 0 111.06 1.06L13.06 12l5.22 5.22a.75.75 0 11-1.06 1.06L12 13.06l-5.22 5.22a.75.75 0 01-1.06-1.06L10.94 12 5.72 6.78a.75.75 0 010-1.06z" fill-rule="evenodd"></path>

</svg>

</button> </div>

<nav aria-label="Global" class="mt-0 px-3 px-lg-0 mb-5 mb-lg-0">

<ul class="d-lg-flex list-style-none">

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<details class="HeaderMenu-details details-overlay details-reset width-full">

<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Why GitHub?

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>

<div class="dropdown-menu flex-auto rounded px-0 mt-0 pb-4 p-lg-4 position-relative position-lg-absolute left-0 left-lg-n4">

<ul class="list-style-none f5 pb-1">

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Features","label":"ref\_page:/topics/machine-learning;ref\_cta:Features;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Features","label":"ref\_page:/topics/machine-learning;ref\_cta:Features;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="cf5c2c8c8b8a91b082eb88f4c5f7e6db197d52190cae4068b05ccb057f34430c" href="/features">

Features

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Mobile","label":"ref\_page:/topics/machine-learning;ref\_cta:Mobile;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Mobile","label":"ref\_page:/topics/machine-learning;ref\_cta:Mobile;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="44cd73fe2c2087b060b1838489a10f261c51a8d7f2e984eff3621f5f4f60f0aa" href="/mobile">

Mobile

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Actions","label":"ref\_page:/topics/machine-learning;ref\_cta:Actions;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Actions","label":"ref\_page:/topics/machine-learning;ref\_cta:Actions;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="038474db69a707a3c57298c7afccecea912284f0278fc1c1b5f7b4ea50f9b3e7" href="/features/actions">

Actions

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Codespaces","label":"ref\_page:/topics/machine-learning;ref\_cta:Codespaces;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Codespaces","label":"ref\_page:/topics/machine-learning;ref\_cta:Codespaces;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="11553057316654b641be126483cbf89667d3e1c56f969b476af40b523b30cae5" href="/features/codespaces">

Codespaces

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Packages","label":"ref\_page:/topics/machine-learning;ref\_cta:Packages;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Packages","label":"ref\_page:/topics/machine-learning;ref\_cta:Packages;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="4d77db26992345a36eefc83da3fb1c7856bda3bcd86ef0b5fb9468838d8ad916" href="/features/packages">

Packages

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Security","label":"ref\_page:/topics/machine-learning;ref\_cta:Security;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Security","label":"ref\_page:/topics/machine-learning;ref\_cta:Security;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="c1f7a6d533773a2ea93146061c99e4a947747108826c56adedbc0347aa713831" href="/features/security">

Security

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Code review","label":"ref\_page:/topics/machine-learning;ref\_cta:Code review;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Code review","label":"ref\_page:/topics/machine-learning;ref\_cta:Code review;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="c62385ca60fb597914072a4dd997f46b8f4f496cfbaaf926894add62196e5d5b" href="/features/code-review">

Code review

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Issues","label":"ref\_page:/topics/machine-learning;ref\_cta:Issues;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Issues","label":"ref\_page:/topics/machine-learning;ref\_cta:Issues;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="6a1bbd35852767701cbf7e9ef17c4613cfed470160e9e7b4ce36d6de5939727d" href="/features/issues">

Issues

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Integrations","label":"ref\_page:/topics/machine-learning;ref\_cta:Integrations;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Integrations","label":"ref\_page:/topics/machine-learning;ref\_cta:Integrations;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="cd307afc175c539ab670058c8b5463745cd10bd590f6bd0d9d231546d68ef8fb" href="/features/integrations">

Integrations

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold border-top pt-4 pb-2 mt-3" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to GitHub Sponsors","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Sponsors;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to GitHub Sponsors","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Sponsors;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="ec5e04dea2323ef28019dc347ec4616a671d21018dafcd27525b4afe0c91ea55" href="/sponsors">

GitHub Sponsors

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold py-2" data-analytics-event='{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Customer stories","label":"ref\_page:/topics/machine-learning;ref\_cta:Customer stories;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Why GitHub?","action":"click to go to Customer stories","label":"ref\_page:/topics/machine-learning;ref\_cta:Customer stories;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="81f83e7334fbd3ef17561698dc7fffe4b09f718ba14946e55cb0d8fb76e4dd18" href="/customer-stories">

Customer stories

</a> </li>

</ul>

</div>

</details>

</li>

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Team","label":"ref\_page:/topics/machine-learning;ref\_cta:Team;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header menu top item (logged out)","action":"click to go to Team","label":"ref\_page:/topics/machine-learning;ref\_cta:Team;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="b19e66b512aa3d1ec4f913987dac03b09dc9a8dbd5b73af28c5146126299681c" href="/team">Team</a>

</li>

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Enterprise","label":"ref\_page:/topics/machine-learning;ref\_cta:Enterprise;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header menu top item (logged out)","action":"click to go to Enterprise","label":"ref\_page:/topics/machine-learning;ref\_cta:Enterprise;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="a6426118a4bcbc78ac48ffce3035372c339f82be4e9ee10f083874cc79b54ff8" href="/enterprise">Enterprise</a>

</li>

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<details class="HeaderMenu-details details-overlay details-reset width-full">

<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Explore

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>

<div class="dropdown-menu flex-auto rounded px-0 mt-0 pb-4 p-lg-4 position-relative position-lg-absolute left-0 left-lg-n4">

<ul class="list-style-none f5 pb-1">

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Explore GitHub","label":"ref\_page:/topics/machine-learning;ref\_cta:Explore GitHub;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Explore GitHub","label":"ref\_page:/topics/machine-learning;ref\_cta:Explore GitHub;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="fb0ab8746b8916954b3b307ff150aa080672926024b10e230ca6e99bebb6d4be" href="/explore">

Explore GitHub

</a> </li>

<li class="color-fg-muted text-normal f6 text-mono mb-1 border-top pt-3 mt-3 mb-1">Learn and contribute</li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Topics","label":"ref\_page:/topics/machine-learning;ref\_cta:Topics;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Topics","label":"ref\_page:/topics/machine-learning;ref\_cta:Topics;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="1e1504dfae3a5635c4dfdd0c41789f47212109384ee801aed6c421bd500e9a58" href="/topics">

Topics

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Collections","label":"ref\_page:/topics/machine-learning;ref\_cta:Collections;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Collections","label":"ref\_page:/topics/machine-learning;ref\_cta:Collections;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="1c3439b2d817ffe0d32c8bbf39b5b8c6e85a28017aaf8c442bea3eacd227043c" href="/collections">

Collections

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Trending","label":"ref\_page:/topics/machine-learning;ref\_cta:Trending;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Trending","label":"ref\_page:/topics/machine-learning;ref\_cta:Trending;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="c0c89f7b0a9e0f1d57601f468e644d19407c74b7f56c75912722960e793b6d16" href="/trending">

Trending

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Learning Lab","label":"ref\_page:/topics/machine-learning;ref\_cta:Learning Lab;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Learning Lab","label":"ref\_page:/topics/machine-learning;ref\_cta:Learning Lab;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="aa0849018e91da405dbde865cf0b1c28ba73d6fc9efaf861c08113ebbe3d615a" href="https://lab.github.com/">

Learning Lab

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Open source guides","label":"ref\_page:/topics/machine-learning;ref\_cta:Open source guides;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Open source guides","label":"ref\_page:/topics/machine-learning;ref\_cta:Open source guides;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="3d3821f51aae1922404c49447f5a483c5eac17ce3c7058f7782376adc3a26ad0" href="https://opensource.guide">

Open source guides

</a> </li>

<li class="color-fg-muted text-normal f6 text-mono mb-1 border-top pt-3 mt-3 mb-1">Connect with others</li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to The ReadME Project","label":"ref\_page:/topics/machine-learning;ref\_cta:The ReadME Project;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to The ReadME Project","label":"ref\_page:/topics/machine-learning;ref\_cta:The ReadME Project;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="da3b598e0e73c3697babebfbf9efea0ed03dfe07a47cbcaf95c6cefc62d11bb8" href="/readme">

The ReadME Project

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Events","label":"ref\_page:/topics/machine-learning;ref\_cta:Events;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Events","label":"ref\_page:/topics/machine-learning;ref\_cta:Events;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="1eaa32e7b7e71867a012cfad3bdc9ba346d4b4a5f8afa500c439477b08222fae" href="/events">

Events

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to Community forum","label":"ref\_page:/topics/machine-learning;ref\_cta:Community forum;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to Community forum","label":"ref\_page:/topics/machine-learning;ref\_cta:Community forum;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="3f473c6a4ab5ae1d89aac2d3c63af28c158b908f7eccfa76f616d4ca97d94613" href="https://github.community">

Community forum

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to GitHub Education","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Education;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to GitHub Education","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Education;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="f5ba5879e42f8f98d17eaab25c0866223eee0f7d0ce61089a779f618aba6b593" href="https://education.github.com">

GitHub Education

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Explore","action":"click to go to GitHub Stars program","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Stars program;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Explore","action":"click to go to GitHub Stars program","label":"ref\_page:/topics/machine-learning;ref\_cta:GitHub Stars program;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="63ddf172c17130dfc6ccaa991964fdd638ad698c9348499a173256fb4cbc1be4" href="https://stars.github.com">

GitHub Stars program

</a> </li>

</ul>

</div>

</details>

</li>

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<a class="HeaderMenu-link no-underline py-3 d-block d-lg-inline-block" data-analytics-event='{"category":"Header menu top item (logged out)","action":"click to go to Marketplace","label":"ref\_page:/topics/machine-learning;ref\_cta:Marketplace;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header menu top item (logged out)","action":"click to go to Marketplace","label":"ref\_page:/topics/machine-learning;ref\_cta:Marketplace;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="ba2e84dbf5c036bad823765ef50c9bd49f3b97b7ba4107508aebd051524557e4" href="/marketplace">Marketplace</a>

</li>

<li class="mr-0 mr-lg-3 position-relative flex-wrap flex-justify-between flex-items-center border-bottom border-lg-bottom-0 d-block d-lg-flex flex-lg-nowrap flex-lg-items-center">

<details class="HeaderMenu-details details-overlay details-reset width-full">

<summary class="HeaderMenu-summary HeaderMenu-link px-0 py-3 border-0 no-wrap d-block d-lg-inline-block">

Pricing

<svg class="icon-chevon-down-mktg position-absolute position-lg-relative" fill="none" viewbox="0 0 14 8" x="0" xml:space="preserve" y="0"><path d="M1,1l6.2,6L13,1"></path></svg>

</summary>

<div class="dropdown-menu flex-auto rounded px-0 mt-0 pb-4 p-lg-4 position-relative position-lg-absolute left-0 left-lg-n4">

<ul class="list-style-none f5 pb-1">

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold py-2" data-analytics-event='{"category":"Header dropdown (logged out), Pricing","action":"click to go to Plans","label":"ref\_page:/topics/machine-learning;ref\_cta:Plans;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Pricing","action":"click to go to Plans","label":"ref\_page:/topics/machine-learning;ref\_cta:Plans;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="456f9d2ed8e16b95eb0440a3973a811f723580a53ff9cd574455ccb79679b461" href="/pricing">

Plans

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Pricing","action":"click to go to Compare plans","label":"ref\_page:/topics/machine-learning;ref\_cta:Compare plans;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Pricing","action":"click to go to Compare plans","label":"ref\_page:/topics/machine-learning;ref\_cta:Compare plans;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="34885664fef940ead2dd0db12e8f08fec42b7cbf566f9c685106db208c85d1a8" href="/pricing#compare-features">

Compare plans

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--secondary py-2" data-analytics-event='{"category":"Header dropdown (logged out), Pricing","action":"click to go to Contact Sales","label":"ref\_page:/topics/machine-learning;ref\_cta:Contact Sales;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Pricing","action":"click to go to Contact Sales","label":"ref\_page:/topics/machine-learning;ref\_cta:Contact Sales;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="42a052f23c6fe0d3f9200dff4d26337e348987469935aad0d9f3092360d2548e" href="https://github.com/enterprise/contact">

Contact Sales

</a> </li>

<li>

<a class="lh-condensed-ultra d-block no-underline position-relative Link--primary text-bold border-top pt-4 pb-2 mt-3" data-analytics-event='{"category":"Header dropdown (logged out), Pricing","action":"click to go to Education","label":"ref\_page:/topics/machine-learning;ref\_cta:Education;"}' data-hydro-click='{"event\_type":"analytics.event","payload":{"category":"Header dropdown (logged out), Pricing","action":"click to go to Education","label":"ref\_page:/topics/machine-learning;ref\_cta:Education;","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="899e4c578bb977baa75a67718da705edff5132bbef3270a13de61fe7bb9b750f" href="https://education.github.com">

Education

</a> </li>

</ul>

</div>

</details>

</li>

</ul>

</nav>

<div class="d-lg-flex flex-items-center px-3 px-lg-0 text-center text-lg-left">

<div class="d-lg-flex min-width-0 mb-3 mb-lg-0">

<div class="header-search flex-auto js-site-search position-relative flex-self-stretch flex-md-self-auto mb-3 mb-md-0 mr-0 mr-md-3 scoped-search site-scoped-search js-jump-to">

<div class="position-relative">

<!-- '"` --><!-- </textarea></xmp> --><form accept-charset="UTF-8" action="/topics/machine-learning" aria-label="Site" class="js-site-search-form" data-scope-id="2169" data-scope-type="Topic" data-scoped-search-url="/topics/machine-learning" data-unscoped-search-url="/search" method="get" role="search">

<label class="form-control input-sm header-search-wrapper p-0 js-chromeless-input-container header-search-wrapper-jump-to position-relative d-flex flex-justify-between flex-items-center">

<input aria-autocomplete="list" aria-controls="jump-to-results" aria-expanded="false" aria-haspopup="listbox" aria-label="Search" autocapitalize="off" autocomplete="off" class="form-control input-sm header-search-input jump-to-field js-jump-to-field js-site-search-focus js-site-search-field is-clearable" data-hotkey="s,/" data-jump-to-suggestions-path="/\_graphql/GetSuggestedNavigationDestinations" data-scoped-placeholder="Search" data-test-selector="nav-search-input" data-unscoped-placeholder="Search GitHub" name="q" placeholder="Search" role="combobox" spellcheck="false" type="text" value=""/>

<input class="js-data-jump-to-suggestions-path-csrf" data-csrf="true" type="hidden" value="Eq0FRjNPLIWgw0BEQ1KaRHblIoB35S5qlknXXkls78U0YOSh7/d1GxhNj5RS+KWo3Rn+Pe/TEXnyov4TPRakpg=="/>

<input class="js-site-search-type-field" name="type" type="hidden"/>

<svg aria-hidden="true" class="mr-1 header-search-key-slash" height="20" width="22" xmlns="http://www.w3.org/2000/svg"><path d="M3.5.5h12c1.7 0 3 1.3 3 3v13c0 1.7-1.3 3-3 3h-12c-1.7 0-3-1.3-3-3v-13c0-1.7 1.3-3 3-3z" fill="none" opacity=".4" stroke="#979A9C"></path><path d="M11.8 6L8 15.1h-.9L10.8 6h1z" fill="#979A9C"></path></svg>

<div class="Box position-absolute overflow-hidden d-none jump-to-suggestions js-jump-to-suggestions-container">

<ul class="d-none js-jump-to-suggestions-template-container">

<li class="d-flex flex-justify-start flex-items-center p-0 f5 navigation-item js-navigation-item js-jump-to-suggestion" role="option">

<a class="no-underline d-flex flex-auto flex-items-center jump-to-suggestions-path js-jump-to-suggestion-path js-navigation-open p-2" data-item-type="suggestion" href="" tabindex="-1">

<div class="jump-to-octicon js-jump-to-octicon flex-shrink-0 mr-2 text-center d-none">

<svg aria-label="Repository" class="octicon octicon-repo js-jump-to-octicon-repo d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Repository" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M2 2.5A2.5 2.5 0 014.5 0h8.75a.75.75 0 01.75.75v12.5a.75.75 0 01-.75.75h-2.5a.75.75 0 110-1.5h1.75v-2h-8a1 1 0 00-.714 1.7.75.75 0 01-1.072 1.05A2.495 2.495 0 012 11.5v-9zm10.5-1V9h-8c-.356 0-.694.074-1 .208V2.5a1 1 0 011-1h8zM5 12.25v3.25a.25.25 0 00.4.2l1.45-1.087a.25.25 0 01.3 0L8.6 15.7a.25.25 0 00.4-.2v-3.25a.25.25 0 00-.25-.25h-3.5a.25.25 0 00-.25.25z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Project" class="octicon octicon-project js-jump-to-octicon-project d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Project" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M1.75 0A1.75 1.75 0 000 1.75v12.5C0 15.216.784 16 1.75 16h12.5A1.75 1.75 0 0016 14.25V1.75A1.75 1.75 0 0014.25 0H1.75zM1.5 1.75a.25.25 0 01.25-.25h12.5a.25.25 0 01.25.25v12.5a.25.25 0 01-.25.25H1.75a.25.25 0 01-.25-.25V1.75zM11.75 3a.75.75 0 00-.75.75v7.5a.75.75 0 001.5 0v-7.5a.75.75 0 00-.75-.75zm-8.25.75a.75.75 0 011.5 0v5.5a.75.75 0 01-1.5 0v-5.5zM8 3a.75.75 0 00-.75.75v3.5a.75.75 0 001.5 0v-3.5A.75.75 0 008 3z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Search" class="octicon octicon-search js-jump-to-octicon-search d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Search" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M11.5 7a4.499 4.499 0 11-8.998 0A4.499 4.499 0 0111.5 7zm-.82 4.74a6 6 0 111.06-1.06l3.04 3.04a.75.75 0 11-1.06 1.06l-3.04-3.04z" fill-rule="evenodd"></path>

</svg>

</div>

<img alt="" aria-label="Team" class="avatar mr-2 flex-shrink-0 js-jump-to-suggestion-avatar d-none" height="28" src="" width="28"/>

<div class="jump-to-suggestion-name js-jump-to-suggestion-name flex-auto overflow-hidden text-left no-wrap css-truncate css-truncate-target">

</div>

<div class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none js-jump-to-badge-search">

<span aria-label="in this topic" class="js-jump-to-badge-search-text-default d-none">

In this topic

</span>

<span aria-label="in all of GitHub" class="js-jump-to-badge-search-text-global d-none">

All GitHub

</span>

<span aria-hidden="true" class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

<div aria-hidden="true" class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none d-on-nav-focus js-jump-to-badge-jump">

Jump to

<span class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

</a>

</li>

</ul>

<ul class="d-none js-jump-to-no-results-template-container">

<li class="d-flex flex-justify-center flex-items-center f5 d-none js-jump-to-suggestion p-2">

<span class="color-fg-muted">No suggested jump to results</span>

</li>

</ul>

<ul class="p-0 m-0 js-navigation-container jump-to-suggestions-results-container js-jump-to-suggestions-results-container" id="jump-to-results" role="listbox">

<li class="d-flex flex-justify-start flex-items-center p-0 f5 navigation-item js-navigation-item js-jump-to-scoped-search d-none" role="option">

<a class="no-underline d-flex flex-auto flex-items-center jump-to-suggestions-path js-jump-to-suggestion-path js-navigation-open p-2" data-item-type="scoped\_search" href="" tabindex="-1">

<div class="jump-to-octicon js-jump-to-octicon flex-shrink-0 mr-2 text-center d-none">

<svg aria-label="Repository" class="octicon octicon-repo js-jump-to-octicon-repo d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Repository" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M2 2.5A2.5 2.5 0 014.5 0h8.75a.75.75 0 01.75.75v12.5a.75.75 0 01-.75.75h-2.5a.75.75 0 110-1.5h1.75v-2h-8a1 1 0 00-.714 1.7.75.75 0 01-1.072 1.05A2.495 2.495 0 012 11.5v-9zm10.5-1V9h-8c-.356 0-.694.074-1 .208V2.5a1 1 0 011-1h8zM5 12.25v3.25a.25.25 0 00.4.2l1.45-1.087a.25.25 0 01.3 0L8.6 15.7a.25.25 0 00.4-.2v-3.25a.25.25 0 00-.25-.25h-3.5a.25.25 0 00-.25.25z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Project" class="octicon octicon-project js-jump-to-octicon-project d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Project" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M1.75 0A1.75 1.75 0 000 1.75v12.5C0 15.216.784 16 1.75 16h12.5A1.75 1.75 0 0016 14.25V1.75A1.75 1.75 0 0014.25 0H1.75zM1.5 1.75a.25.25 0 01.25-.25h12.5a.25.25 0 01.25.25v12.5a.25.25 0 01-.25.25H1.75a.25.25 0 01-.25-.25V1.75zM11.75 3a.75.75 0 00-.75.75v7.5a.75.75 0 001.5 0v-7.5a.75.75 0 00-.75-.75zm-8.25.75a.75.75 0 011.5 0v5.5a.75.75 0 01-1.5 0v-5.5zM8 3a.75.75 0 00-.75.75v3.5a.75.75 0 001.5 0v-3.5A.75.75 0 008 3z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Search" class="octicon octicon-search js-jump-to-octicon-search d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Search" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M11.5 7a4.499 4.499 0 11-8.998 0A4.499 4.499 0 0111.5 7zm-.82 4.74a6 6 0 111.06-1.06l3.04 3.04a.75.75 0 11-1.06 1.06l-3.04-3.04z" fill-rule="evenodd"></path>

</svg>

</div>

<img alt="" aria-label="Team" class="avatar mr-2 flex-shrink-0 js-jump-to-suggestion-avatar d-none" height="28" src="" width="28"/>

<div class="jump-to-suggestion-name js-jump-to-suggestion-name flex-auto overflow-hidden text-left no-wrap css-truncate css-truncate-target">

</div>

<div class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none js-jump-to-badge-search">

<span aria-label="in this topic" class="js-jump-to-badge-search-text-default d-none">

In this topic

</span>

<span aria-label="in all of GitHub" class="js-jump-to-badge-search-text-global d-none">

All GitHub

</span>

<span aria-hidden="true" class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

<div aria-hidden="true" class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none d-on-nav-focus js-jump-to-badge-jump">

Jump to

<span class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

</a>

</li>

<li class="d-flex flex-justify-start flex-items-center p-0 f5 navigation-item js-navigation-item js-jump-to-owner-scoped-search d-none" role="option">

<a class="no-underline d-flex flex-auto flex-items-center jump-to-suggestions-path js-jump-to-suggestion-path js-navigation-open p-2" data-item-type="owner\_scoped\_search" href="" tabindex="-1">

<div class="jump-to-octicon js-jump-to-octicon flex-shrink-0 mr-2 text-center d-none">

<svg aria-label="Repository" class="octicon octicon-repo js-jump-to-octicon-repo d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Repository" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M2 2.5A2.5 2.5 0 014.5 0h8.75a.75.75 0 01.75.75v12.5a.75.75 0 01-.75.75h-2.5a.75.75 0 110-1.5h1.75v-2h-8a1 1 0 00-.714 1.7.75.75 0 01-1.072 1.05A2.495 2.495 0 012 11.5v-9zm10.5-1V9h-8c-.356 0-.694.074-1 .208V2.5a1 1 0 011-1h8zM5 12.25v3.25a.25.25 0 00.4.2l1.45-1.087a.25.25 0 01.3 0L8.6 15.7a.25.25 0 00.4-.2v-3.25a.25.25 0 00-.25-.25h-3.5a.25.25 0 00-.25.25z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Project" class="octicon octicon-project js-jump-to-octicon-project d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Project" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M1.75 0A1.75 1.75 0 000 1.75v12.5C0 15.216.784 16 1.75 16h12.5A1.75 1.75 0 0016 14.25V1.75A1.75 1.75 0 0014.25 0H1.75zM1.5 1.75a.25.25 0 01.25-.25h12.5a.25.25 0 01.25.25v12.5a.25.25 0 01-.25.25H1.75a.25.25 0 01-.25-.25V1.75zM11.75 3a.75.75 0 00-.75.75v7.5a.75.75 0 001.5 0v-7.5a.75.75 0 00-.75-.75zm-8.25.75a.75.75 0 011.5 0v5.5a.75.75 0 01-1.5 0v-5.5zM8 3a.75.75 0 00-.75.75v3.5a.75.75 0 001.5 0v-3.5A.75.75 0 008 3z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Search" class="octicon octicon-search js-jump-to-octicon-search d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Search" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M11.5 7a4.499 4.499 0 11-8.998 0A4.499 4.499 0 0111.5 7zm-.82 4.74a6 6 0 111.06-1.06l3.04 3.04a.75.75 0 11-1.06 1.06l-3.04-3.04z" fill-rule="evenodd"></path>

</svg>

</div>

<img alt="" aria-label="Team" class="avatar mr-2 flex-shrink-0 js-jump-to-suggestion-avatar d-none" height="28" src="" width="28"/>

<div class="jump-to-suggestion-name js-jump-to-suggestion-name flex-auto overflow-hidden text-left no-wrap css-truncate css-truncate-target">

</div>

<div class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none js-jump-to-badge-search">

<span aria-label="in all of GitHub" class="js-jump-to-badge-search-text-default d-none">

Search

</span>

<span aria-label="in all of GitHub" class="js-jump-to-badge-search-text-global d-none">

All GitHub

</span>

<span aria-hidden="true" class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

<div aria-hidden="true" class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none d-on-nav-focus js-jump-to-badge-jump">

Jump to

<span class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

</a>

</li>

<li class="d-flex flex-justify-start flex-items-center p-0 f5 navigation-item js-navigation-item js-jump-to-global-search d-none" role="option">

<a class="no-underline d-flex flex-auto flex-items-center jump-to-suggestions-path js-jump-to-suggestion-path js-navigation-open p-2" data-item-type="global\_search" href="" tabindex="-1">

<div class="jump-to-octicon js-jump-to-octicon flex-shrink-0 mr-2 text-center d-none">

<svg aria-label="Repository" class="octicon octicon-repo js-jump-to-octicon-repo d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Repository" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M2 2.5A2.5 2.5 0 014.5 0h8.75a.75.75 0 01.75.75v12.5a.75.75 0 01-.75.75h-2.5a.75.75 0 110-1.5h1.75v-2h-8a1 1 0 00-.714 1.7.75.75 0 01-1.072 1.05A2.495 2.495 0 012 11.5v-9zm10.5-1V9h-8c-.356 0-.694.074-1 .208V2.5a1 1 0 011-1h8zM5 12.25v3.25a.25.25 0 00.4.2l1.45-1.087a.25.25 0 01.3 0L8.6 15.7a.25.25 0 00.4-.2v-3.25a.25.25 0 00-.25-.25h-3.5a.25.25 0 00-.25.25z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Project" class="octicon octicon-project js-jump-to-octicon-project d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Project" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M1.75 0A1.75 1.75 0 000 1.75v12.5C0 15.216.784 16 1.75 16h12.5A1.75 1.75 0 0016 14.25V1.75A1.75 1.75 0 0014.25 0H1.75zM1.5 1.75a.25.25 0 01.25-.25h12.5a.25.25 0 01.25.25v12.5a.25.25 0 01-.25.25H1.75a.25.25 0 01-.25-.25V1.75zM11.75 3a.75.75 0 00-.75.75v7.5a.75.75 0 001.5 0v-7.5a.75.75 0 00-.75-.75zm-8.25.75a.75.75 0 011.5 0v5.5a.75.75 0 01-1.5 0v-5.5zM8 3a.75.75 0 00-.75.75v3.5a.75.75 0 001.5 0v-3.5A.75.75 0 008 3z" fill-rule="evenodd"></path>

</svg>

<svg aria-label="Search" class="octicon octicon-search js-jump-to-octicon-search d-none flex-shrink-0" data-view-component="true" height="16" role="img" title="Search" version="1.1" viewbox="0 0 16 16" width="16">

<path d="M11.5 7a4.499 4.499 0 11-8.998 0A4.499 4.499 0 0111.5 7zm-.82 4.74a6 6 0 111.06-1.06l3.04 3.04a.75.75 0 11-1.06 1.06l-3.04-3.04z" fill-rule="evenodd"></path>

</svg>

</div>

<img alt="" aria-label="Team" class="avatar mr-2 flex-shrink-0 js-jump-to-suggestion-avatar d-none" height="28" src="" width="28"/>

<div class="jump-to-suggestion-name js-jump-to-suggestion-name flex-auto overflow-hidden text-left no-wrap css-truncate css-truncate-target">

</div>

<div class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none js-jump-to-badge-search">

<span aria-label="in this topic" class="js-jump-to-badge-search-text-default d-none">

In this topic

</span>

<span aria-label="in all of GitHub" class="js-jump-to-badge-search-text-global d-none">

All GitHub

</span>

<span aria-hidden="true" class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

<div aria-hidden="true" class="border rounded-1 flex-shrink-0 color-bg-subtle px-1 color-fg-muted ml-1 f6 d-none d-on-nav-focus js-jump-to-badge-jump">

Jump to

<span class="d-inline-block ml-1 v-align-middle">↵</span>

</div>

</a>

</li>

</ul>

</div>

</label>

</form> </div>

</div>

</div>

<div class="position-relative mr-3 mb-4 mb-lg-0 d-inline-block">

<a class="HeaderMenu-link flex-shrink-0 no-underline" data-ga-click="(Logged out) Header, clicked Sign in, text:sign-in" data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/login?return\_to=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning">

Sign in

</a>

</div>

<a class="HeaderMenu-link flex-shrink-0 d-inline-block no-underline border color-border-default rounded px-2 py-1" data-analytics-event='{"category":"Sign up","action":"click to sign up for account","label":"ref\_page:/topics/machine-learning;ref\_cta:Sign up;ref\_loc:header logged out"}' data-hydro-click='{"event\_type":"authentication.click","payload":{"location\_in\_page":"site header menu","repository\_id":null,"auth\_type":"SIGN\_UP","originating\_url":"https://github.com/topics/machine-learning","user\_id":null}}' data-hydro-click-hmac="77fa0805c2ee5e083e5dbe5571a2e37d8661eca3a115806e97d54914b8332ea9" href="/signup?ref\_cta=Sign+up&amp;ref\_loc=header+logged+out&amp;ref\_page=%2Ftopics%2Fmachine-learning&amp;source=header">

Sign up

</a>

</div>

</div>

</div>

</header>

</div>

first\_img=doc.find('img')

*#first\_img=doc.find\_all('img')*

*#first\_img=doc.img*

first\_img

<img alt="" aria-label="Team" class="avatar mr-2 flex-shrink-0 js-jump-to-suggestion-avatar d-none" height="28" src="" width="28"/>

first\_span=doc.span

*#first\_span=doc.find\_all['span']*

*#first\_span=doc.find['span']*

first\_span

<span class="progress-pjax-loader js-pjax-loader-bar Progress position-fixed width-full" data-view-component="true">

<span class="Progress-item progress-pjax-loader-bar left-0 top-0 color-bg-accent-emphasis" data-view-component="true" style="width: 0%;"></span>

</span>

first\_p=doc.find\_all('p')

first\_p[0]

<p>Machine learning is the practice of teaching a computer to learn. The concept uses pattern recognition, as well as other forms of predictive algorithms, to make judgments on incoming data. This field is closely related to artificial intelligence and computational statistics.</p>

**EXERCISE**: Get a list of all the img tags on the page. How many images does the page contain?

all\_images=doc.find\_all('img')

len(all\_images)

16

**EXERCISE**: Find the 5th image tag on the page (counting from 0). Which attributes does the tag contain? Find the values of the src and alt attributes of the tag.

fifth\_image=all\_images[5]

fifth\_image

<img alt="janeyx99" class="avatar avatar-user avatar-small" height="32" src="[https://avatars.githubusercontent.com/u/31798555?v=4](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com%2Fu%2F31798555%3Fv%3D4)" width="32"/>

fifth\_image['src']

'[https://avatars.githubusercontent.com/u/31798555?v=4](https://jovian.ai/outlink?url=https%3A%2F%2Favatars.githubusercontent.com%2Fu%2F31798555%3Fv%3D4)'

fifth\_image['alt']

'janeyx99'

**EXERCISE**: Find the src attribute of the first img tag with the alt attribute set to julia. Visit the link and check what the image represents.

doc.find('img',{'alt':'julia'})['src']

'[https://repository-images.githubusercontent.com/1644196/ddfc1e00-6638-11e9-9b80-0fe7b9aedd72](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F1644196%2Fddfc1e00-6638-11e9-9b80-0fe7b9aedd72)'

**EXERCISE**: Find the list of all the images matching the class d-block width-full. Each list element should be a dictionary containing two keys, "username" and "url". You can obtain the username using the alt attribute of a tag and the URL using the src attribute.

image\_link\_tags = doc.find\_all('img', class\_='d-block width-full')

avatar\_users = []

for tag in image\_link\_tags:

avatar\_users.append({

'username' : tag['alt'],

'url': tag['src']

})

avatar\_users

[{'username': 'transformers',

'url': '[https://repository-images.githubusercontent.com/155220641/a16c4880-a501-11ea-9e8f-646cf611702e](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F155220641%2Fa16c4880-a501-11ea-9e8f-646cf611702e)'},

{'username': 'ML-For-Beginners',

'url': '[https://repository-images.githubusercontent.com/343965132/549b1a80-c897-11eb-9436-918072d2e0f8](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F343965132%2F549b1a80-c897-11eb-9436-918072d2e0f8)'},

{'username': 'awesome-scalability',

'url': '[https://repository-images.githubusercontent.com/115478820/109a8e00-283a-11ea-8891-ad7215b06a4c](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F115478820%2F109a8e00-283a-11ea-8891-ad7215b06a4c)'},

{'username': 'julia',

'url': '[https://repository-images.githubusercontent.com/1644196/ddfc1e00-6638-11e9-9b80-0fe7b9aedd72](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F1644196%2Fddfc1e00-6638-11e9-9b80-0fe7b9aedd72)'},

{'username': 'Made-With-ML',

'url': '[https://repository-images.githubusercontent.com/156157055/5d88d9c5-1030-4da2-a153-d2836b299eac](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F156157055%2F5d88d9c5-1030-4da2-a153-d2836b299eac)'},

{'username': 'yolov5',

'url': '[https://repository-images.githubusercontent.com/264818686/40f8c2c3-7919-4652-b278-ec6a7fb06a53](https://jovian.ai/outlink?url=https%3A%2F%2Frepository-images.githubusercontent.com%2F264818686%2F40f8c2c3-7919-4652-b278-ec6a7fb06a53)'}]

**EXERCISE**: Write a function scrape\_topics which takes a list of topics and creates CSV files containing top repositories for a list of topics. Test it out using the empty cells below.

topics=['data-analysis','python','deep-learning']

def **scrape\_topics**(topics):

for topic in topics:

scrape\_topic\_repositories(topic)

scrape\_topics(topics)

Top repositories for topic "data-analysis" written to file "data-analysis.csv" Top repositories for topic "python" written to file "python.csv" Top repositories for topic "deep-learning" written to file "deep-learning.csv"

pd.read\_csv('data-analysis.csv')

pd.read\_csv('python.csv')

pd.read\_csv('deep-learning.csv')

**EXERCISE**: Get the top 100 repositories for the all the featured topics on GitHub. You might find these URLs useful:

* Eighth page of featured topics: [https://github.com/topics/?page=8](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2F%3Fpage%3D8)
* Second page of top repositories for a topic: [https://github.com/topics/machine-learning?page=2](https://jovian.ai/outlink?url=https%3A%2F%2Fgithub.com%2Ftopics%2Fmachine-learning%3Fpage%3D2)

def **get\_feature\_page**(n):

docs=[]

for i in range(1,n):

topic\_repos\_url = 'https://github.com/topics/?page=' + str(i)

response = requests.get(topic\_repos\_url)

if response.status\_code != 200:

print('Status code:', response.status\_code)

raise Exception('Failed to fetch web page ' + topic\_repos\_url)

doc = BeautifulSoup(response.text)

docs.append(doc)

return docs

def **get\_featured\_topics**(docs):

hrefs=[]

for doc in docs:

href=doc.find\_all('a',class\_='no-underline d-flex flex-column flex-justify-center')

for a in href:

link=a['href']

hrefs.append(link)

return hrefs

def **scrape\_featured\_repositories**(topic, path=None):

"""Get the top repositories for a topic and write them to a CSV file"""

for i in range(1,6):

path = topic.strip('/topics')+ str(i) + '.csv'

topic\_repos\_url = 'https://github.com/' + topic +'?page=' + str(i)

response = requests.get(topic\_repos\_url)

if response.status\_code != 200:

print('Status code:', response.status\_code)

raise Exception('Failed to fetch web page ' + topic\_repos\_url)

topic\_page\_doc = BeautifulSoup(response.text)

topic\_repositories = get\_top\_repositories(topic\_page\_doc)

write\_csv(topic\_repositories, path)

print('Top repositories for topic "{}" written to file "{}"'.format(topic, path))

return path

def **get\_top\_featured\_repositories**(n):

docs=get\_feature\_page(n)

topics=get\_featured\_topics(docs)

dfs=[]

for topic in topics:

scrape\_featured\_repositories(topic)

*#scarping featured topics in first page. You can go ahead and scrape for as many pages as you'd like.*

get\_top\_featured\_repositories(2)

Top repositories for topic "/topics/matlab" written to file "matlab1.csv" Top repositories for topic "/topics/matlab" written to file "matlab2.csv" Top repositories for topic "/topics/matlab" written to file "matlab3.csv" Top repositories for topic "/topics/matlab" written to file "matlab4.csv" Top repositories for topic "/topics/matlab" written to file "matlab5.csv" Top repositories for topic "/topics/covid-19" written to file "vid-191.csv" Top repositories for topic "/topics/covid-19" written to file "vid-192.csv" Top repositories for topic "/topics/covid-19" written to file "vid-193.csv" Top repositories for topic "/topics/covid-19" written to file "vid-194.csv" Top repositories for topic "/topics/covid-19" written to file "vid-195.csv" Top repositories for topic "/topics/r" written to file "r1.csv" Top repositories for topic "/topics/r" written to file "r2.csv" Top repositories for topic "/topics/r" written to file "r3.csv" Top repositories for topic "/topics/r" written to file "r4.csv" Top repositories for topic "/topics/r" written to file "r5.csv"

pd.read\_csv('matlab1.csv')

Try to combine each topic's all pages CSVs into single one ;)