Creating a dataset using an API with Python

Karan Bhanot Sep 22, 2018 · 5 min read

"person using laptop" by rawpixel on Unsplash

Whenever we begin a Machine Learning project, the first thing that we need is a dataset. While there are many datasets that you can find online with varied information, sometimes you wish to extract data on your own and begin your own investigation. This is when an API provided by a website can come to the rescue.

> *An application program interface (API) is code that allows two software programs to communicate with each other. The API defines the correct way for a developer to write a program that requests services from an operating system (OS) or other application. — TechTarget*

API is actually a very simple tool that allows anyone to access information from a given website. You might require the use of certain headers but some APIs require just the URL. In this particular article, I'll use the Twitch API provided by Free Code Camp Twitch API Pass-through. This API route does not require any client id to run, thus making it very simple to access Twitch Data. The whole project is available as a Notebook in the Create-dataset-using-API repository.

Import Libraries

As part of accessing the API content and getting the data into a .CSV file, we'll have to import a number of Python Libraries.

1. **requests** library helps us get the content from the API by using the `get()` method. The `json()` method converts the API response to JSON format for easy handling.
2. **json** library is needed so that we can work with the JSON content we get from the API. In this case, we get a dictionary for each Channel's information such as name, id, views and other information.
3. **pandas** library helps to create a dataframe which we can export to a .CSV file in correct format with proper headings and indexing.

Understand the API

We first need to understand what all information can be accessed from the API. For that we use the example of the channel *Free Code Camp* to make the API call and check the information we get.

```
import numpy as np
import pandas as pd
import requests
import json

url = "https://wind-bow.glitch.me/twitch-api/channels/freecodecamp"
JSONContent = requests.get(url).json()
content = json.dumps(JSONContent, indent = 4, sort_keys=True)
print(content)
```

API.py hosted with ♡ by GitHub view raw

This prints the response of the API

To access the API response, we use the function call `requests.get(url).json()` which not only gets the response from the API for the `url` but also gets the JSON format for it. We then dump the data using `dump()` method into `content` so that we can view it in a more presentable view. The output of the code is as follows:

```
{
    "_id": 79776140,
    "_links": {
        "chat": "https://api.twitch.tv/kraken/chat/freecodecamp",
        "commercial": "https://api.twitch.tv/kraken/channels/freecodecamp/commercial",
        "editors": "https://api.twitch.tv/kraken/channels/freecodecamp/editors",
        "features": "https://api.twitch.tv/kraken/channels/freecodecamp/features",
        "follows": "https://api.twitch.tv/kraken/channels/freecodecamp/follows",
        "self": "https://api.twitch.tv/kraken/channels/freecodecamp",
        "stream_key": "https://api.twitch.tv/kraken/channels/freecodecamp/stream_key",
        "subscriptions": "https://api.twitch.tv/kraken/channels/freecodecamp/subscriptions",
```

```
        "teams": "https://api.twitch.tv/kraken/channels/freecodecamp/teams",
        "videos": "https://api.twitch.tv/kraken/channels/freecodecamp/videos"
    },
    "background": null,
    "banner": null,
    "broadcaster_language": "en",
    "created_at": "2015-01-14T03:36:47Z",
    "delay": null,
    "display_name": "FreeCodeCamp",
    "followers": 11770,
    "game": "Creative",
    "language": "en",
    "logo": "https://static-cdn.jtvnw.net/jtv_user_pictures/freecodecamp-profile_image-d9514f2df
    "mature": false,
    "name": "freecodecamp",
    "partner": false,
    "profile_banner": "https://static-cdn.jtvnw.net/jtv_user_pictures/freecodecamp-profile_banne
    "profile_banner_background_color": null,
    "status": "Some GoLang Today #go #golang #youtube",
    "updated_at": "2018-09-19T23:01:33Z",
    "url": "https://www.twitch.tv/freecodecamp",
    "video_banner": "https://static-cdn.jtvnw.net/jtv_user_pictures/freecodecamp-channel_offline_
    "views": 216340
}
```


response.json hosted with ♡ by GitHub view raw

Response for API

If we look closely at the output, we can see that there is a lot of information that we have received. We get the id, links to various other sections, followers, name, language, status, url, views and much more. Now, we can loop through a list of channels, get information for each channel and compile it into a dataset. I will be using a few properties from this list including *_id*, *display_name*, *status*, *followers* and *views*.

Create the dataset

Now that we are aware of what to expect from the API response, let's start with compiling the data together and creating our dataset. For this blog, we'll consider a list of channels that I collected online.

We will first start by defining out list of channels in an array. Then for each channel, we'll use the API to get its information and store each channel's information inside another array `channels_list` using the `append()` method till we get all information collected together in one place. The request response is in JSON format, so to access any key value pair we simply write the key's name within square brackets after the `JSONContent` variable. Now, we use the pandas library to convert this array into a pandas Dataframe using the method `DataFrame()` provided in pandas library. A dataframe is a representation of the data in a tabular form similar to a table, where data is expressed in terms of rows and columns. This dataframe allows fast manipulation of data using various methods.

```
# List of channels we want to access
channels = ["ESL_SC2", "OgamingSC2", "cretetion", "freecodecamp", "storbeck", "habathcx", "Robot
            "ninja", "shroud", "Dakotaz", "esltv_cs", "pokimane", "tsm_bjergsen", "boxbox", "wtc
           "kinggothalion", "amazhs", "jahrein", "thenadeshot", "sivhd", "kingrichard"]

channels_list = []
# For each channel, we access its information through its API
for channel in channels:
    JSONContent = requests.get("https://wind-bow.glitch.me/twitch-api/channels/" + channel).json
    if 'error' not in JSONContent:
        channels_list.append([JSONContent['_id'], JSONContent['display_name'], JSONContent['stat
```

```
                                JSONContent['followers'], JSONContent['views']])

dataset = pd.DataFrame(channels_list)
dataset.sample(5)
```

Pandas Dataframe creation.py hosted with ♡ by GitHub view raw

Create dataframe using Pandas

The pandas `sample()` method displays randomly selected rows of the dataframe. In this method, we pass the number of rows we wish to show. Here, let's display 5 rows.

	0	1	2	3	4
2	90401618	cretetion	Logging some Miles With My Pack	2812	37083
15	51950404	wtcN	Bıktım başlık yazmaktan | sümüklü | instagram....	834344	35332694
5	6726509	habathcx	Massively Effective	24	2622
21	27686136	SivHD	Siv HD messing around in WoW. the original FAM...	993961	37815131
7	82534701	noobs2ninjas	Doing some work and felt like streaming. Progr...	974	55127

dataset.sample(5)

On close inspection, we see that the dataset has two minor problems. Let's address them one by one.

1. **Headings:** Presently, the headings are numbers and unreflective of the data each column represents. It might seem less important with this dataset because it has only a few columns. However, when you'll explore datasets with 100s of columns, this step will become really important. Here, we define the columns using the `columns()` method provided by pandas. In this case, we explicitly defined the headings but in certain cases, you can pick up the keys as headings directly.
2. **None/Null/Blank Values:** Some of the rows will have missing values. In such cases, we'll have two options. We can either remove the complete row where any value is blank or we can input some carefully selected value in the blank spaces. Here, the `status` column will have `None` in some cases. We'll remove these rows by using the method `dropna(axis = 0, how = 'any', inplace = True)` which drops rows with blank values in the dataset itself. Then, we change the index of the numbers from 0 to the length of the dataset using the method `RangeIndex(len(dataset.index))`.

```
dataset.columns = ['Id', 'Name', 'Status', 'Followers', 'Views']
dataset.dropna(axis = 0, how = 'any', inplace = True)
dataset.index = pd.RangeIndex(len(dataset.index))
```

Headings-index.py hosted with ♡ by GitHub view raw

Add column headings and update index

Export Dataset

Our dataset is now ready, and can be exported to an external file. We use the `to_csv()` method. We define two paramteres. The first parameter refers to the name of the file. The second parameter is a boolean that represents if the first column in the exported file will have the index or not. We now have a .CSV file with the dataset we created.

	Id	Name	Status	Followers	Views
0	30220059	ESL_SC2	RERUN: ShoWTimE vs. GuMiho [PvT] - Group D - I...	233265	75380804
1	71852806	OgamingSC2	Olimoleague Weekly #134	79265	38973783
2	90401618	cretetion	Logging some Miles With My Pack	2812	37083
3	79776140	FreeCodeCamp	Some GoLang Today #go #golang #youtube	11771	216347
4	6726509	habathcx	Massively Effective	24	2622
5	54925078	RobotCaleb	Sabering	35	6684
6	82534701	noobs2ninjas	Doing some work and felt like streaming. Progr...	974	55127
7	19571641	Ninja	Duos with Travis Scott! See how you could squ...	11343822	320594181
8	37402112	shroud	gaminxS @Shroud for updates!	4352191	206944462
9	39298218	dakotaz	TSM Dakotaz - instagram/twitter: @dakotaz	2625586	59728959
10	31239503	ESL_CSGO	RERUN: IEM Shanghai 2018	2475125	296624596
11	44445592	pokimane	See you guys soon! Follow my twitter & insta f...	2170913	43709758
12	38421618	TSM_Bjergsen	Monday stream as per usualllllll!	1440737	91181335
13	38881685	boxbox	Mr. Challenger Climb I don't FEel so good	1389852	83498885
14	51950404	wtcN	Bıktım başlık yazmaktan | sümüklü | instagram....	834344	35332694
15	19070311	A_Seagull	cozy	864185	24793514
16	43830727	KingGothalion	Raid. Day. Today.	875741	34100062
17	43356746	AmazHS	✧AMAZ✧ It's me! =D | !MTGA #sponsored , Ravnic...	894551	93073898
18	6768122	Jahrein	60 Parsecs uzay versiyonu | instagram.com/jahr...	935101	36168882
19	21130533	Nadeshot	The Return to Fortnite | 100 Thieves	948418	42710352
20	27686136	SivHD	Siv HD messing around in WoW. the original FAM...	993961	37815131
21	66691674	KingRichard	Fall Skirmish Practice #bushbandits @KingRicha...	1075328	17037876

Dataset.csv

In this article, we discussed an approach to create our own dataset using the Twitch API and Python. You can follow a similar approach to access information through any other API. Give it a try using the GitHub API.

. . .

Please feel free to share your thoughts and hit me up with any questions you might have.

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

Get this newsletter

Emails will be sent to mcanovasdre@gmail.com.
Not you?

Data Science Python Machine Learning Web Scraping API

Medium About Help Legal