Get started

Open in app

Follow

551K Followers

Scraping Reddit data

How to scrape data from Reddit using the Python Reddit API Wrapper(PRAW)

Gilbert Tanner Jan 5, 2019 · 5 min read

Photo by Fabian Grohs on Unsplash

As its name suggests PRAW is a Python wrapper for the Reddit API, which enables you to scrape data from subreddits, create a bot and much more.

In this article, we will learn how to use PRAW to scrape posts from different subreddits as well as how to get comments from a specific post.

Getting Started

PRAW can be installed using pip or conda:

```
pip install praw
or
conda install -c conda-forge praw
```

install_praw.txt hosted with ♡ by GitHub view raw

Now PRAW can be imported by writting:

```
import praw
```

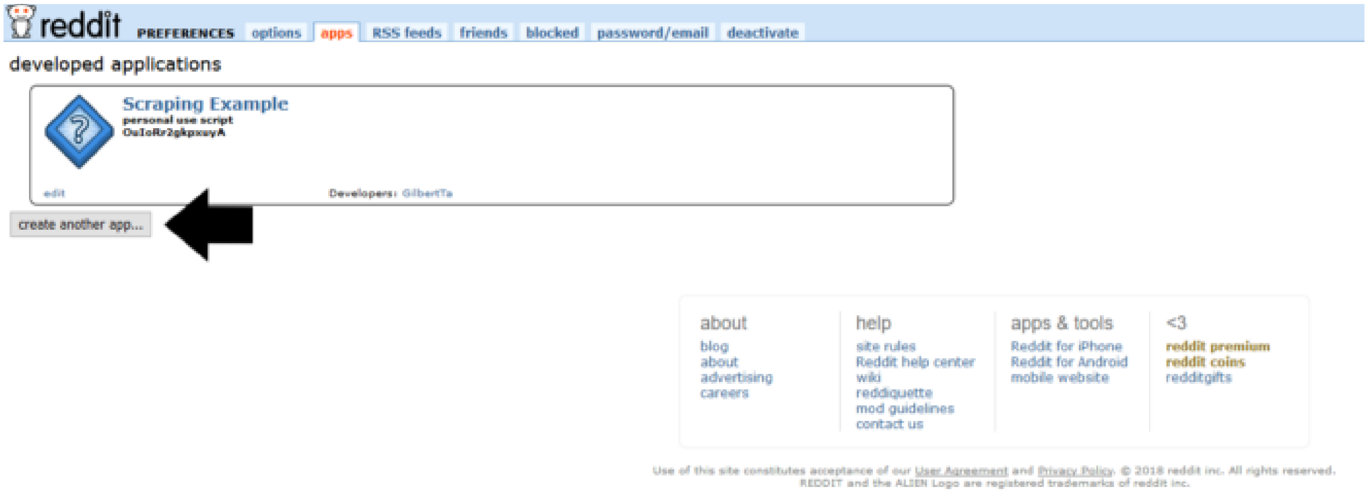
Before it can be used to scrape data we need to authenticate ourselves. For this we need to create a Reddit instance and provide it with a `client_id` , `client_secret` and a `user_agent` .

```
reddit = praw.Reddit(client_id='my_client_id', client_secret='my_client_secret', user_agent='my_u
```

create_reddit_instance.py hosted with ♡ by GitHub view raw

To get the authentication information we need to create a reddit app by navigating to this page and clicking **create app** or **create another app.**

GET NEW REDDIT
MY SUBREDDITS
HOME - POPULAR - ALL - RANDOM - USERS | DATAISBEAUTIFUL - PROGRAMMING - MATH - MACHINELEARNING - DATASCIENCE - LEARNMATH - STATISTICS - ANNOUNCEMENTS - LEARNMACH



developed applications

Scraping Example
personal use script
OuIoRr2gkpxuyA

edit
Developers: GilbertTa

create another app...

about
blog
about
advertising
careers

help
site rules
Reddit help center
wiki
reddiquette
mod guidelines
contact us

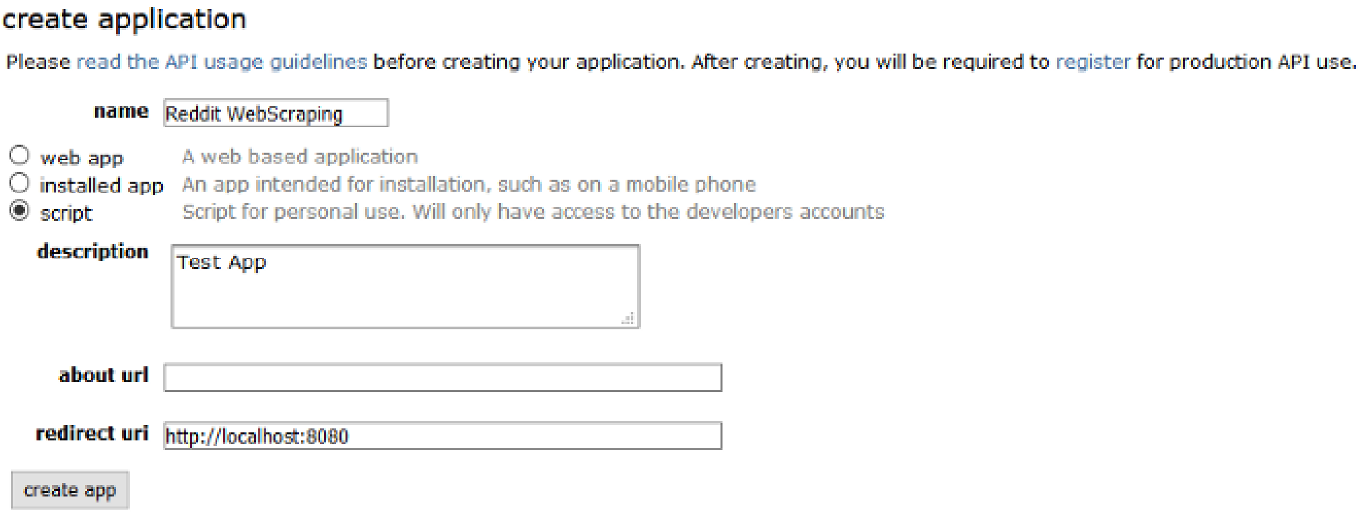
apps & tools
Reddit for iPhone
Reddit for Android
mobile website

<3
reddit premium
reddit coins
redditgifts

Use of this site constitutes acceptance of our User Agreement and Privacy Policy. © 2018 reddit inc. All rights reserved.
REDDIT and the ALIEN Logo are registered trademarks of reddit inc.

Figure 1: Reddit Application

This will open a form where you need to fill in a name, description and redirect uri. For the redirect uri you should choose `http://localhost:8080` as described in the excellent PRAW documentation.



create application

Please read the API usage guidelines before creating your application. After creating, you will be required to register for production API use.

name: Reddit WebScraping

○ web app — A web based application
○ installed app — An app intended for installation, such as on a mobile phone
● script — Script for personal use. Will only have access to the developers accounts

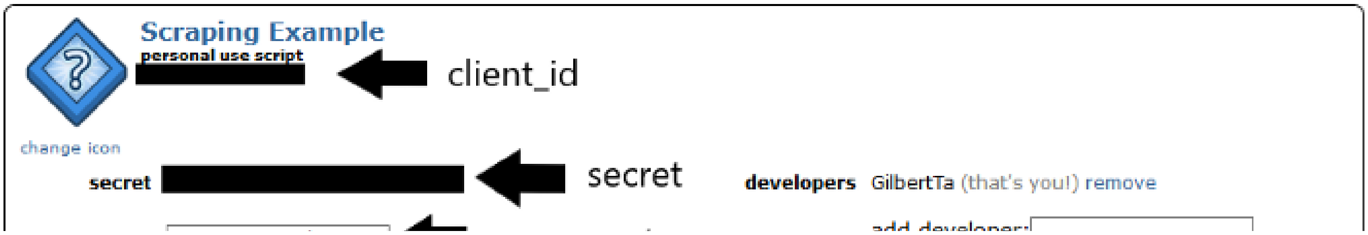
description: Test App

about url:

redirect uri: http://localhost:8080

Figure 2: Create new Reddit Application

After pressing **create app** a new application will appear. Here you can find the authentication information needed to create the `praw.Reddit` instance.

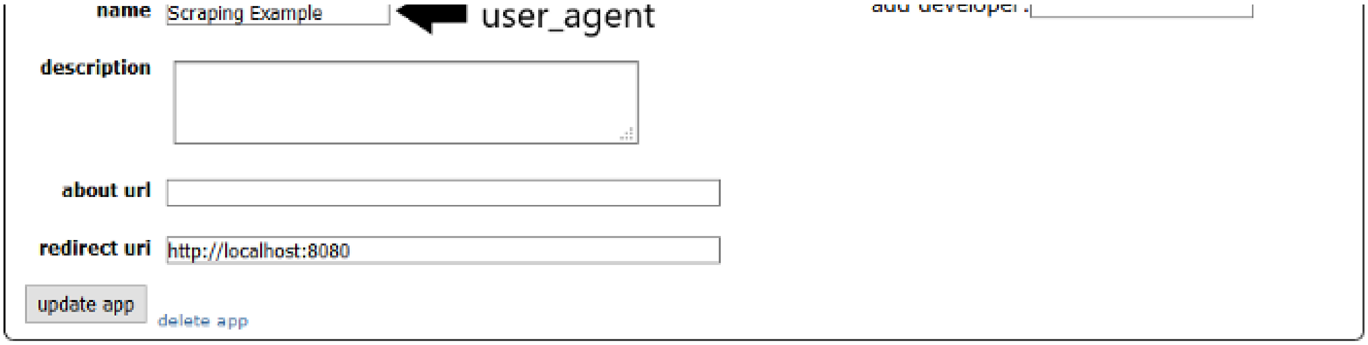


Scraping Example
personal use script
[redacted]

secret [redacted]

developers GilbertTa (that's you!) remove

add developer:


name Scraping Example

add developer:

description

about url

redirect uri http://localhost:8080

Figure 3: Authentication information

Get subreddit data

Now that we have a `praw.Reddit` instance we can access all available functions and use it, to for example get the 10 "hottest" posts from the Machine Learning subreddit.

```python
# get 10 hot posts from the MachineLearning subreddit
hot_posts = reddit.subreddit('MachineLearning').hot(limit=10)
for post in hot_posts:
    print(post.title)
```

get_hottest_ml_reddit_posts.py hosted with ♡ by GitHub view raw

Output:

```
[D] What is the best ML paper you read in 2018 and why?
[D] Machine Learning - WAYR (What Are You Reading) - Week 53
[R] A Geometric Theory of Higher-Order Automatic Differentiation
UC Berkeley and Berkeley AI Research published all materials of CS
188: Introduction to Artificial Intelligence, Fall 2018
[Research] Accurate, Data-Efficient, Unconstrained Text Recognition
with Convolutional Neural Networks
...
```

We can also get the 10 "hottest" posts of all subreddits combined by specifying "all" as the subreddit name.

```python
# get hottest posts from all subreddits
hot_posts = reddit.subreddit('all').hot(limit=10)
```

```python
for post in hot_posts:
    print(post.title)
```

get_hottest_reddit_posts.py hosted with ♡ by GitHub view raw

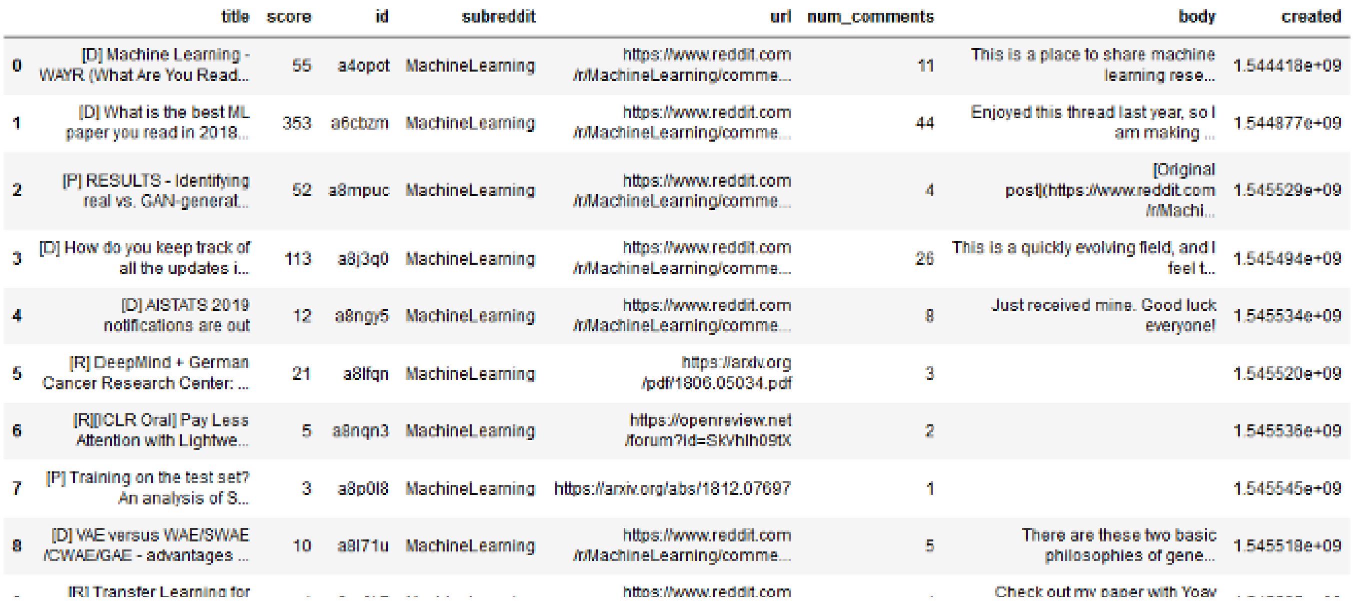
Output:

```
I've been lying to my wife about film plots for years.
I don’t care if this gets downvoted into oblivion! I DID IT REDDIT!!
I’ve had enough of your shit, Karen
Stranger Things 3: Coming July 4th, 2019
...
```

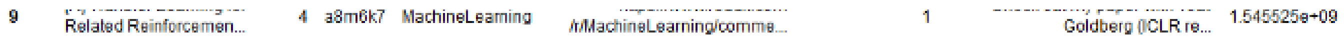
This variable can be iterated over and features including the post title, id and url can be extracted and saved into an `.csv` file.

```python
import pandas as pd
posts = []
ml_subreddit = reddit.subreddit('MachineLearning')
for post in ml_subreddit.hot(limit=10):
    posts.append([post.title, post.score, post.id, post.subreddit, post.url, post.num_comments, p
posts = pd.DataFrame(posts,columns=['title', 'score', 'id', 'subreddit', 'url', 'num_comments', '
print(posts)
```

get_and_save_hottest_ml_posts.py hosted with ♡ by GitHub view raw



	title	score	id	subreddit	url	num_comments	body	created
0	[D] Machine Learning - WAYR (What Are You Read...	55	a4opot	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	11	This is a place to share machine learning rese...	1.544418e+09
1	[D] What is the best ML paper you read in 2018...	353	a6cbzm	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	44	Enjoyed this thread last year, so I am making ...	1.544877e+09
2	[P] RESULTS - Identifying real vs. GAN-generat...	52	a8mpuc	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	4	[Original post](https://www.reddit.com /r/Machi...	1.545529e+09
3	[D] How do you keep track of all the updates i...	113	a8j3q0	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	26	This is a quickly evolving field, and I feel t...	1.545494e+09
4	[D] AISTATS 2019 notifications are out	12	a8ngy5	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	8	Just received mine. Good luck everyone!	1.545534e+09
5	[R] DeepMind + German Cancer Research Center: ...	21	a8lfqn	MachineLearning	https://arxiv.org /pdf/1806.05034.pdf	3		1.545520e+09
6	[R][ICLR Oral] Pay Less Attention with Lightwe...	5	a8nqn3	MachineLearning	https://openreview.net /forum?id=SkVhlh09tX	2		1.545536e+09
7	[P] Training on the test set? An analysis of S...	3	a8p0l8	MachineLearning	https://arxiv.org/abs/1812.07697	1		1.545545e+09
8	[D] VAE versus WAE/SWAE /CWAE/GAE - advantages ...	10	a8l71u	MachineLearning	https://www.reddit.com /r/MachineLearning/comme...	5	There are these two basic philosophies of gene...	1.545518e+09
[illegible]	[R] Transfer Learning for	[illegible]	[illegible]	[illegible]	https://www.reddit.com	[illegible]	Check out my paper with Yoav	[illegible]



9	Related Reinforcemen...	4	a8m6k7	MachineLearning	/r/MachineLearning/comme...	1	Goldberg (ICLR re...	1.545525e+09

Figure 4: Hottest ML posts

General information about the subreddit can be obtained using the `.description` function on the subreddit object.

```python
# get MachineLearning subreddit data
ml_subreddit = reddit.subreddit('MachineLearning')

print(ml_subreddit.description)
```

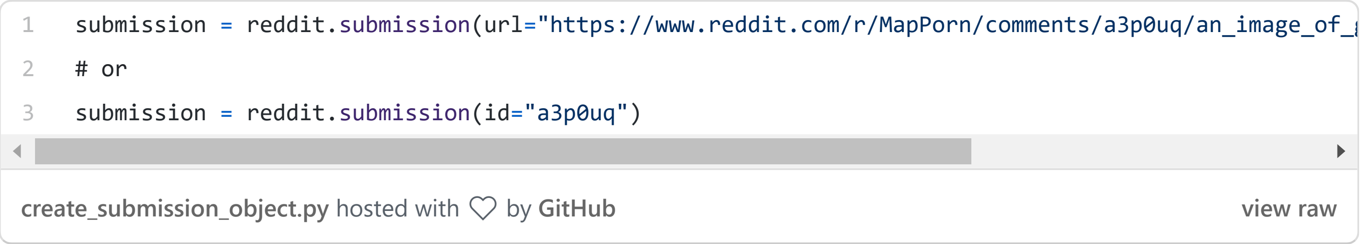
get_subreddit_information.py hosted with ♡ by GitHub view raw

Output:

```
**[Rules For Posts]
(https://www.reddit.com/r/MachineLearning/about/rules/)**
--------
+[Research](https://www.reddit.com/r/MachineLearning/search?
sort=new&restrict_sr=on&q=flair%3AResearch)
--------
+[Discussion](https://www.reddit.com/r/MachineLearning/search?
sort=new&restrict_sr=on&q=flair%3ADiscussion)
--------
+[Project](https://www.reddit.com/r/MachineLearning/search?
sort=new&restrict_sr=on&q=flair%3AProject)
--------
+[News](https://www.reddit.com/r/MachineLearning/search?
sort=new&restrict_sr=on&q=flair%3ANews)
--------
...
```

Get comments from a specific post

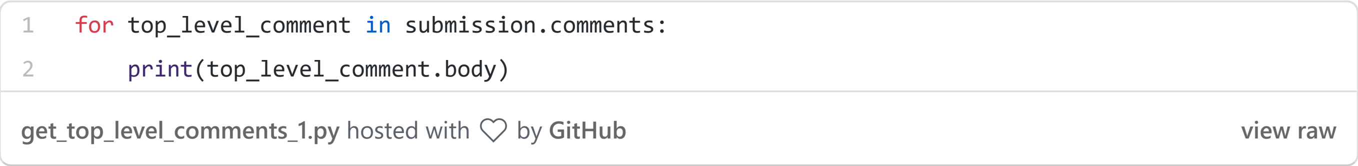
You can get the comments for a post/submission by creating/obtaining a `Submission` object and looping through the `comments` attribute. To get a post/submission we can either iterate through the submissions of a subreddit or specify a specific submission using `reddit.submission` and passing it the submission url or id.



```
submission = reddit.submission(url="https://www.reddit.com/r/MapPorn/comments/a3p0uq/an_image_of_
# or
submission = reddit.submission(id="a3p0uq")
```

create_submission_object.py hosted with ♡ by GitHub view raw

To get the **top-level** comments we only need to iterate over `submission.comments` .



```
for top_level_comment in submission.comments:
    print(top_level_comment.body)
```

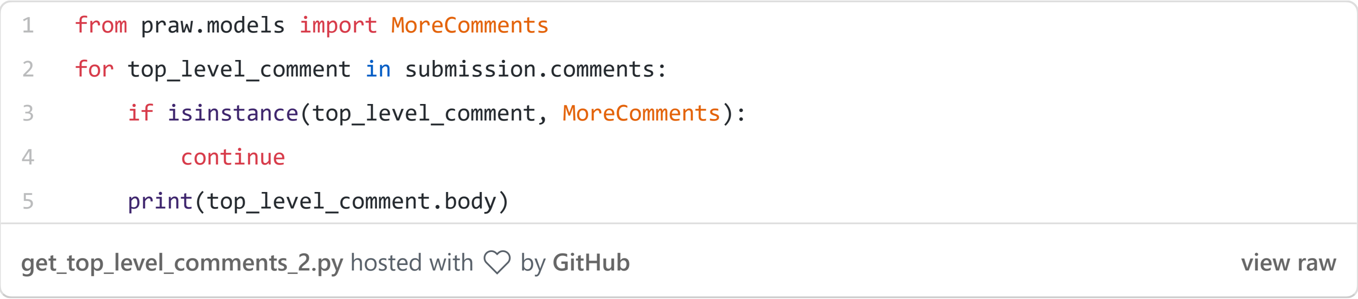
get_top_level_comments_1.py hosted with ♡ by GitHub view raw

This will work for some submission, but for others that have more comments this code will throw an AttributeError saying:

```
AttributeError: 'MoreComments' object has no attribute 'body'
```

These `MoreComments` object represent the "load more comments" and "continue this thread" links encountered on the websites, as described in more detail in the comment documentation.

There get rid of the `MoreComments` objects, we can check the datatype of each comment before printing the body.



```
from praw.models import MoreComments
for top_level_comment in submission.comments:
    if isinstance(top_level_comment, MoreComments):
        continue
    print(top_level_comment.body)
```

get_top_level_comments_2.py hosted with ♡ by GitHub view raw

But Praw already provides a method called `replace_more` , which replaces or removes the `MoreComments` . The method takes an argument called limit, which when set to 0 will remove all `MoreComments` .

```python
submission.comments.replace_more(limit=0)
for top_level_comment in submission.comments:
    print(top_level_comment.body)
```

get_top_level_comments_3.py hosted with ♡ by GitHub view raw

Both of the above code blocks successfully iterate over all the **top-level** comments and print their body. The output can be seen below.

```
Source: [https://www.facebook.com/VoyageursWolfProject/]
(https://www.facebook.com/VoyageursWolfProject/)
I thought this was a shit post made in paint before I read the title
Wow, that’s very cool.  To think how keen their senses must be to
recognize and avoid each other and their territories.  Plus, I like
to think that there’s one from the white colored clan who just goes
way into the other territories because, well, he’s a badass.
That’s really cool. The edges are surprisingly defined.
...
```

However, the comment section can be arbitrarily deep and most of the time we surely also want to get the comments of the comments. `CommentForest` provides the `.list` method, which can be used for getting all comments inside the comment section.

```python
submission.comments.replace_more(limit=0)
for comment in submission.comments.list():
    print(comment.body)
```

get_all_comments.py hosted with ♡ by GitHub view raw

The above code will first of output all the top-level comments, followed by the second-level comments and so on until there are no comments left.

Recommended Reading

Web Scraping using Selenium and BeautifulSoup

How to use Selenium to navigate between pages and use it to scrap

HTML loaded with JavaScript.

towardsdatascience.com

Conclusion

Praw is a Python wrapper for the Reddit API, which enables us to use the Reddit API with a clean Python interface. The API can be used for webscraping, creating a bot as well as many others.

This article covered authentication, getting posts from a subreddit and getting comments. To learn more about the API I suggest to take a look at their excellent documentation.

If you liked this article consider subscribing on my Youtube Channel and following me on social media.

The code covered in this article is available as a Github Repository.

If you have any questions, recommendations or critiques, I can be reached via Twitter or the comment section.

Thanks to Chamin Nalinda.

Reddit Web Scraping API Python Data Science