Building a Reddit Web Scraper in Python Programming

Reddit is home to countless communities, interminable discussions, and genuine human connections. Reddit has a community for every interest, including breaking news, sports, TV fan theories, and an endless stream of the internet's prettiest animals.

Using Python's PRAW (Python Reddit API Wrapper) package, this tutorial will demonstrate how to scrape data from Reddit. PRAW is a Python wrapper for the Reddit API, allowing you to scrape data from subreddits, develop bots, and much more.

By the end of this tutorial, we will attempt to scrape as much Python-related data as possible from the subreddit and gain access to what Reddit users are truly saying about Python. Let's start having fun!

## Introduction

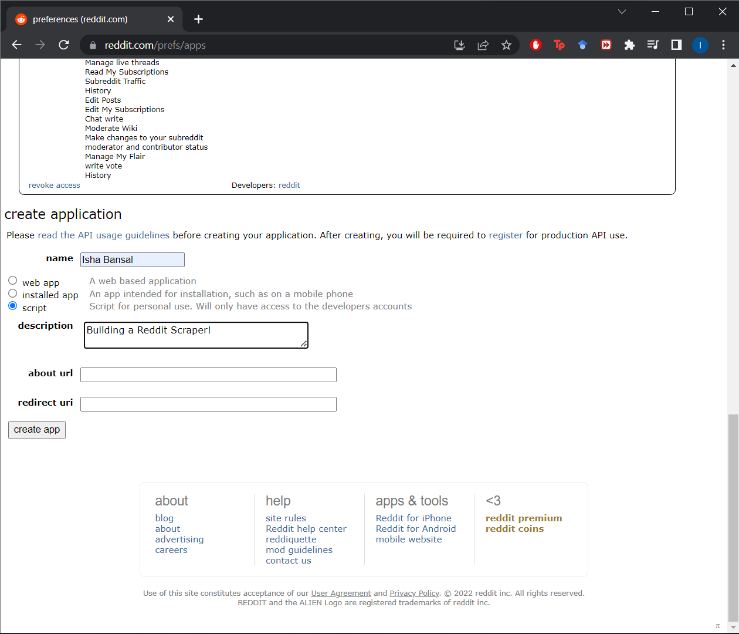
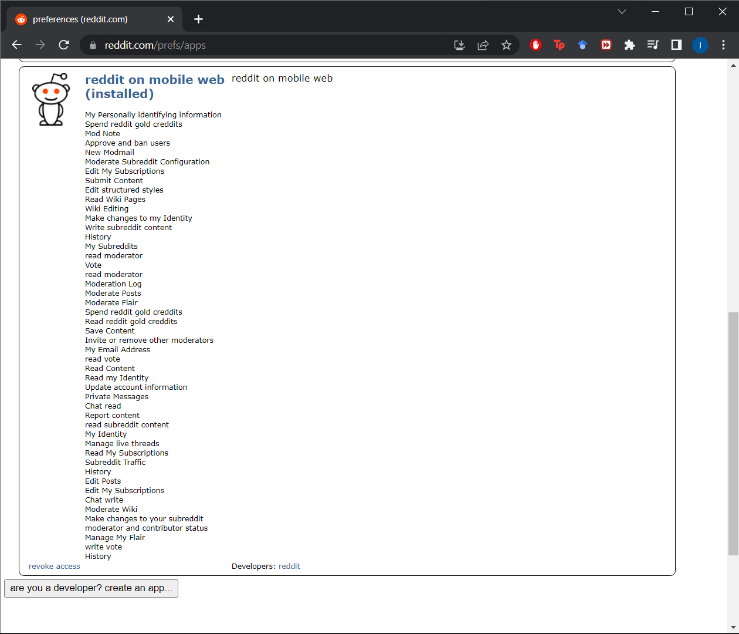
As the name suggests, it is a technique for "scraping" or extracting data from online pages. Everything that can be seen on the Internet using a web browser, including this guide, can be scraped onto a local hard disc. There are numerous applications for web scraping. Data capture is the first phase of any data analysis. The internet is a massive repository of all human history and knowledge, and you have the power to extract any information you desire and use it as you see fit.

Although there are various techniques to scrape data from Reddit, PRAW simplifies the process. It adheres to all Reddit API requirements and eliminates the need for sleep calls in the developer's code. Before installing the scraper, authentication for the Reddit scraper must be set up. The respective steps are listed below.

## Authentication steps for Reddit Scraper

Working with PRAW requires authentication. To accomplish this, we will take the following steps:

1. Follow this link to access the Reddit developer account.
2. Scroll to the bottom of the page to locate the "are you a developer?" button to develop an app.
3. The next step is to build an application, fill out the form, and develop the app.
4. This will take you to a page containing all of the information required for the scraper.



Now that the authentication phase is complete, we will be moving on to the implementation of the Reddit scraper in the next step.

## Implementation of the Scraper

This part will explain everything you must do to obtain the data that this tutorial aims to obtain. We will begin by importing all required modules and libraries into the program file. Before importing the PRAW library, we must install PRAW by executing the following line at the command prompt:



Now that PRAW has been successfully installed, the following code snippet may be used to import PRAW together with other required libraries.



The authentication procedure we just completed will be useful immediately. Before utilizing PRAW to scrape any data, we must authenticate in the software. This can be accomplished by creating either a Reddit instance or an Authorized Instance.

In this guide, we will create an authorized instance that will allow us to perform any action we desire with our Reddit account. You only need to provide a client id, client secret, user agent, username, and password to the instance. Examine the code fragment below (fill in your keys instead of the blank strings).



We aim to find what redditors talk about Python on the platform but in case you change your mind and want to know about something else, we will take the input on the topic from the user itself. Look at the code snippet below.



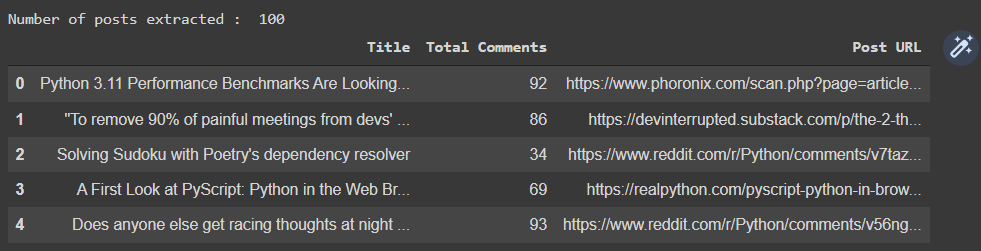
Using the code below, we will next attempt to gain access to the subreddit using the instance object we generated before. In addition, we will provide some basic information about the subreddit to check that we have access.



We will try to extract the top weekly, monthly and yearly posts on the top in the upcoming code snippets to understand what are the topmost posts on the topic. We will be extracting the title of the post, the number of comments, and the URL of the post with the help of a for loop on the post objects extracted. To make the analysis easier, we will convert the data into a dataframe. The code below will extract the top posts of the week on the topic.



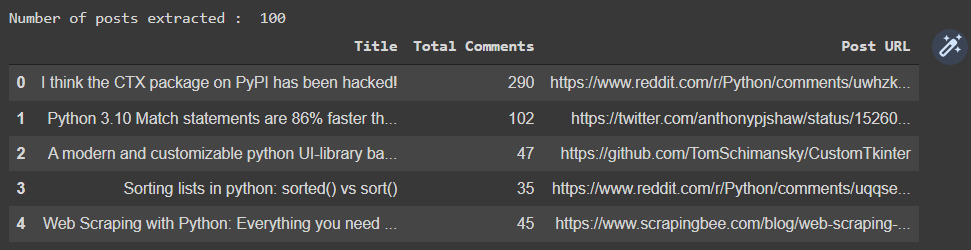
The output looks somewhat like shown below and you can see we were able to extract data for 100 posts.



The next code snippet will get top posts of the month on the topic. All you need to change is the parameter of the `subreddit.top` function.



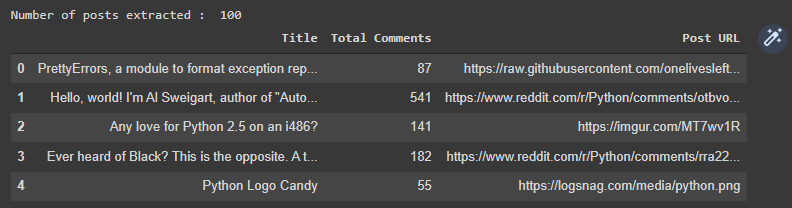
Have a look at the top monthly posts extracted through the code.



Lastly the following code snippet will get top posts of the year on the topic. Again, all you need to change is the parameter of the `subreddit.top` function.



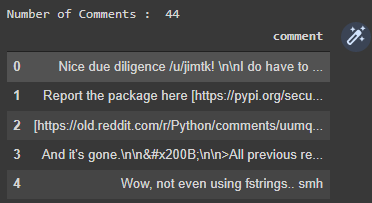
Have a look at the top yearly posts extracted through the code above.



Lastly, let’s also try to extract all the comments of a post with the help of the post URL using the code snippet below. This will help to know how people are reacting to the posts on Python. We will extract the best comments from the initial post of the most popular Python articles published each month. The MoreComments under the praw module will be required to achieve this.



Have a look at all the 44 comments extracted for the post in the following image.



## Conclusion

Praw is a Python wrapper for the Reddit API, allowing us to use the Reddit API with a straightforward Python interface. The API can be used for web scraping, bot creation, and other purposes. This tutorial addressed authentication, retrieving the most popular weekly, monthly, and yearly posts from a subreddit, as well as extracting the post's comments.

I hope you enjoyed the article, and if you did, I recommend that you also check out the following tutorials: