



Published in Towards AI

You have **1** free member-only story left this month. [Sign up for Medium and get an extra one](#)



Zoumana Keita

[Follow](#)Dec 9, 2022 · 5 min read · ✨ · [Listen](#)

Save



Extract Tweets Without Limitations in a Few Lines of Code Using Python

Data quantity should not be a limitation.

Scrape Tweets Without Limitations Using Python



Like — Follow — Subscribe — Share

Introduction

If you are familiar with the Tweepy library, you might also be familiar with the fact that you can not go beyond a certain number of tweets, which is I believe a huge drawback for experimentations requiring a significant amount of data.

There is a new player in town: [snsrape](#) a python library specifically built for social networking services (SNS for short).

It can scrape information like user profiles, hashtags, and specific user posts from a variety of platforms such as Facebook, Instagram, Mastodon, Reddit, Telegram, Twitter, VKontakte, and Weibo.

As just a little extra icing on the cake 🍰, you don't need to apply for any API credentials 🎉.

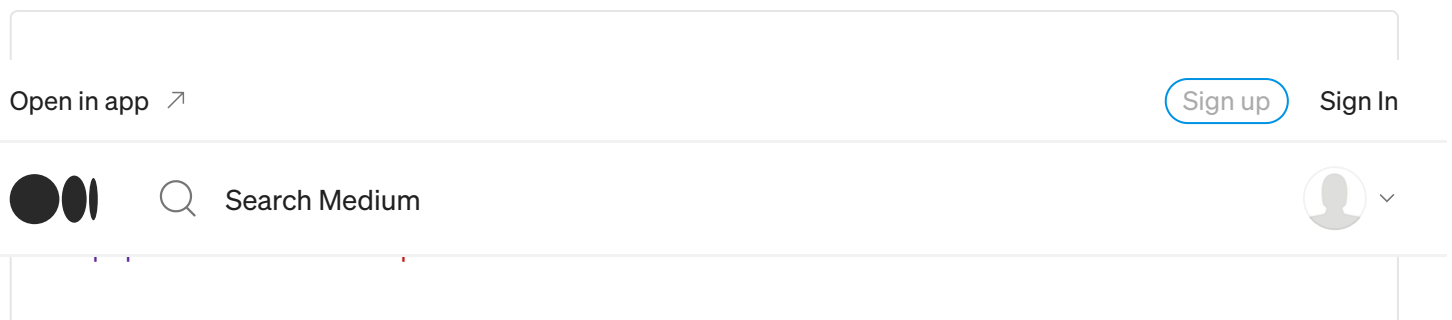
Let's get started

In this conceptual tutorial, we will use [snsrape](#) to pull thousands of tweets and save them locally as a DataFrame. You can get more than that. The sky is the limit 🚀!

Prerequisites

First of all, you must install the `snsrape` library. But to do this, you need to have Python installed on your computer, which can be done by downloading and installing it from the [official website](#).

Once you have Python installed, you can finally install `snsrape` with the following `pip` command from your terminal or within Jupyter notebook:



Once successfully, you should be able to import it for further use using the following statement:

```
import snsrape.modules.twitter as snt
```

This way, you will be able to use `snt` wherever you want to use the `snsrape.modules.twitter` module.

Scrape Tweets

Once the library is properly set up, you can start scraping the tweets you need. In this section, we will be scraping tweets related to WorldCup 2022.

Pulling Tweets requires the `TwitterSearchScrapper` module. Before diving deep, let's understand the output format of `extract_tweet` information by getting a sample of tweets with a hashtag `#worldcup2022`

```
# Get Tweets with the #worldcup
world_cup_scraper = snt.TwitterSearchScrapper("#worldcup")

print(type(world_cup_scraper))
```

The print statement shows:

`<class 'snsrape.modumes.twitter.TwitterSearchScrapper'>` . This simply means that the application of the `TwitterSearchScrapper` module also returns a `TwitterSearchScrapper` object.

To be able to grab the actual tweets data, you need to apply the `get_items()` function as shown below.

 104 |  2 |

```
# Let's get the first tweet from the world_cup_scraper search
for world_cup_tweet in world_cup_scraper.get_items():
    break
```

The previous for loop only grabs the first tweet's data. We can show the raw data of the `world_cup_tweet` by typing the code below.

```
world_cup_tweet
```

```
Tweet(url='https://twitter.com/rodmur/status/1600301040917676032', date=datetime.datetime(2022, 12, 7, 1, 27, 54, tzinfo=datetime.timezone.utc), content='I just earned the \'World Pint (2022)\' badge on @untappd! https://t.co/TkgkQ4edqo #football #fifaworldcup #worldcup', renderedContent='I just earned the \'World Pint (2022)\' badge on @untappd! untp.beer/s/b1050646581 #football #fifaworldcup #worldcup', id=1600301040917676032, user=User(username='rodmur', id=9553672, displayname='Dale', description='Geek, photographer, jack of all trades, master of none. Amateur thespian and professional heckler. I\'m a bad, bad person. I\'ll bake you cookies.\n(he/him)', rawDescription='Geek, photographer, jack of all trades, master of none. Amateur thespian and professional heckler. I\'m a bad, bad person. I\'ll bake you cookies.\n(he/him)', descriptionUrls=None, verified=False, created=datetime.datetime(2007, 10, 20, 0, 38, 55, tzinfo=datetime.timezone.utc), followersCount=187, friendsCount=1104, statusesCount=37113, favouritesCount=14697, listedCount=5, mediaCount=1045, location='Brooklyn, NY', protected=False, linkUrl=None, linkTcourl=None, profileImageUrl='https://pbs.twimg.com/profile_images/1557442945669898240/rhR5oF9S_normal.jpg', profileBannerUrl='https://pbs.twimg.com/profile_banners/9553672/1631131936', label=None), replyCount=0, retweetCount=0, likeCount=0, quoteCount=0, conversationId=1600301040917676032, lang='en', source='<a href="https://untappd.com" rel="nofollow">Untappd</a>', sourceUrl='https://untappd.com', sourceLabel='Untappd', outlinks=['https://untp.beer/s/b1050646581'], tcooutlinks=['https://t.co/TkgkQ4edqo'], media=None, retweetedTweet=None, quotedTweet=None, inReplyToTweetId=None, inReplyToUser=None, mentionedUsers=[User(username='untappd', id=147845476, displayname='untappd', description=None, rawDescription=None, descriptionUrls=None, verified=None, created=None, followersCount=None, friendsCount=None, statusesCount=None, favouritesCount=None, listedCount=None, mediaCount=None, location=None, protected=None, linkUrl=None, linkTcourl=None, profileImageUrl=None, profileBannerUrl=None, label=None)], coordinates=None, place=None, hashtags=['football', 'fifaworldcup', 'worldcup'], cashtags=None)
```

First tweet's raw data (Image by Author)

As you can see the raw data is in the format of `key=value` and some of them are underlined in green. Not all the columns are useful. So let's consider only those specified in the `column_name` list.

```
column_names = ['url', 'date', 'content', 'username', 'displayname',
                'description', 'followersCount', 'friendsCount',
                'likeCount', 'world_cup_tweet']
```

Keep in mind that if you do not specify the number of tweets to be collected, this scraping process might run forever, trying to grab all the tweets. For simplicity's sake, let's say we want a maximum of 200000 tweets.

```
total_tweet = 20000
```

Putting all this together, we get the following helper function that grabs the required number of tweets.

```
# Putting all together
def grab_tweets(total_number):

    final_tweets = []

    for index, world_cup_tweet in enumerate(world_cup_scraper.get_items()):

        user = world_cup_tweet.user

        tweet_data = [world_cup_tweet.url,
                      world_cup_tweet.date,
                      world_cup_tweet.content,
                      user.username,
                      user.displayname,
                      user.description,
                      user.followersCount,
                      user.friendsCount,
                      world_cup_tweet.likeCount,
                      world_cup_tweet.retweetCount
                      ]

        final_tweets.append(tweet_data)

        if(index == total_number):
            break

    # Create the dataframe
    final_tweets_df = pd.DataFrame(final_tweets, columns = column_names)

    return final_tweets_df
```

- The `break` statement is important because it allows the program to not continue once we reach the `total_number`.

Finally, we can call the function specifying the `total_number` parameter, then we show the shape of the data with the `.shape` attribute, and the first five rows with the `.head()` function.

```
# Call the grab_tweets() function
final_tweets_data = grab_tweets(20000)

# Show the shape
print(final_tweets_data.shape)

# Show the first 5 rows
final_tweets_data.head()
```

- The shape is (20000, 10) → 20000 rows and 10 columns.

Below are the first five rows.

	url	date	content	username	displayname	description	followersCount	friendsCount	likeCount	world_cup_tweet
0	https://twitter.com/pakunnnnn/status/16003158...	2022-12-07 02:26:43+00:00	グループ1位通過で負けたん日本だ けか。\\n今までで1番ベスト8いけ る可能性あったしぜん勝...	pakunnnnn		101"osaka 誰とでも気 軽に スプラ 3 エンジョ イ第一 ち エンソーマ ン3周目	30	31	0	0
1	https://twitter.com/jobsalution/status/1600315...	2022-12-07 02:26:40+00:00	Join now Coco's Carnival: https://t.co/HzZQBKcS...	jobsalution	rashid mehmood		17	10	0	0
2	https://twitter.com/shankarrkn/status/16003158...	2022-12-07 02:26:38+00:00	நேத்து போர்க்கல் வெளையாடனதை பாத்தாக்க அவங்க ப...	shankarrkn	Dr Shankar	Consciously try to think different, against mi...	952	3830	0	0
3	https://twitter.com/browncito/status/160031577...	2022-12-07 02:26:28+00:00	@alimo_philip I want to see #Portugal @Cristia...	browncito	Dunnya	Marketista de profesion, amo la cocina cuando ...	2232	1590	0	0
4	https://twitter.com/sun8866557/status/16003157...	2022-12-07 02:26:13+00:00	#世界杯 #worldcup 足坛烽烟又 起，斗志薪火相传。 https://t.co/0u15P	sun8866557	soon (互fo)	互fo关注必 回	1160	1156	0	0

First 5 rows of the tweets (Image by Author)

What about specific language tweets?

Previously tweets are collected no matter the language, which is not ideal if we are interested in language-specific tweets. For instance, let's say we are only interested in French Tweets, this can be done by specifying the `lang` parameter in the `TwitterSearchScrapper` module as follows:

- `TwitterSearchScrapper("topic lang:language")`

To do that, we will create a new function and also slightly modify the previous one:

```
def get_language_specific_tweets(topic, total_number, lang="fr"):
```

```
# Get the topic from using the scraper and the language
topic_scraper = snt.TwitterSearchScraper(f"{topic} lang:{lang}")

# Grab the tweets
final_tweets_as_df = grab_tweets(topic_scraper, total_number)

return final_tweets_as_df
```

This new function takes as parameters the topic of interest (e.g. #worldcup), the total number of tweets, and finally the language of interest which is french by default.

In addition to that, we slightly modify the original function to meet the previous function's requirement, because this time we have a new parameter: `topic_scraper`.

```
# Putting all together
def grab_tweets(scraper, total_number):

    final_tweets = []

    for index, world_cup_tweet in enumerate(scraper.get_items()):

        user = world_cup_tweet.user

        tweet_data = [world_cup_tweet.url,
                       world_cup_tweet.date,
                       world_cup_tweet.content,
                       user.username,
                       user.displayname,
                       user.description,
                       user.followersCount,
                       user.friendsCount,
                       world_cup_tweet.likeCount,
                       world_cup_tweet.retweetCount
                      ]

        final_tweets.append(tweet_data)

        if(index == total_number):
            break

    # Create the dataframe
    final_tweets_df = pd.DataFrame(final_tweets, columns = column_names)
```

```
return final_tweets_df
```

Here are finally some examples of grabbing French Tweets and English Tweets.




```
# French Tweets
topic = "#worldcup"
lang = "fr"
fr_df = get_language_specific_tweets(topic, 200, lang)

fr_df.head()
```

	url	date	content	username	displayname	description	followersCount	friendsCount	likeCount	retweetCount
0	https://twitter.com/ActuSport_EDF/status/16019...	2022-12-11 11:54:43+00:00	Biathlon : #WorldCup\nVictoire du relais franç...	ActuSport_EDF	Actu Sport France 🇫🇷	Toute l'actualité du sport français avec @Actu...	305	270	0	0
1	https://twitter.com/SarahBelmir/status/1601907...	2022-12-11 11:50:02+00:00	Du "blanchiment" de joueur. Tout un réseau de ...	SarahBelmir	Sarah Belmir	Frenchie living in Dakar and tweeting in Engli...	198	713	0	0
2	https://twitter.com/joetke/status/160190708006...	2022-12-11 11:49:44+00:00	#WorldCup #WorldcupQatar2022 France-#Maroc, du...	joetke	Joetke • bedeef1 • Joetke's Hub	Madagascar 🇲🇵 ❤️ & Asia/Amerindia- centric Activi...	2212	4687	0	0
3	https://twitter.com/STaskadi/status/1601906926...	2022-12-11 11:49:07+00:00	le but de Youssef En- Nesyri 🇲🇵 au rythme de la ...	STaskadi	Sofia Taskadi 💎	lfrane	1433	1254	0	0
4	https://twitter.com/CaptainPalestin/status/160...	2022-12-11 11:45:30+00:00	Un pays arabe va gagner la coupe du monde...\nÇa...	CaptainPalestin	Super Captain Jerusalem	meilleur tueur à gages de Poitou Charentes, dé...	27	135	0	0

French Tweets (Image by Author)

```
# English Tweets about worldcup
topic = "#worldcup"
lang = "en"
en_df = get_language_specific_tweets(topic, 200, lang)
```


	url	date	content	username	displayname	description	followersCount	friendsCount	likeCount	retweetCount
0	https://twitter.com/se7enmadseason/status/1601...	2022-12-11 12:06:47+00:00	That one #KanyeWest fan who still defends him....	se7enmadseason		a Mad Season	6736	6738	0	0
1	https://twitter.com/QQarramis/status/160191135...	2022-12-11 12:06:42+00:00	Congrats to #Morocco on being the first Africa...	QQarramis		Proud #Afar #Self- determination Views on #...	3028	633	1	0
2	https://twitter.com/healthyuke/status/16019112...	2022-12-11 12:06:26+00:00	Picked up your @VitaCoco Footie Bundle? Why n...	healthyuke	@healthyuke	We wake up every morning to inspire our custom...	4725	574	0	0
3	https://twitter.com/abdallah_rm11/status/16019...	2022-12-11 12:06:25+00:00	1 available ticket for the Semi-Final game bet...	abdallah_rm11		Real Madrid blogger Contenido de fútbol ...	9772	766	0	0
4	https://twitter.com/thelaymanshaman/status/160...	2022-12-11 12:06:23+00:00	Kane fully using that penalty in the #WorldCup...	thelaymanshaman	Charles Bown	Spare thoughts // All things Psych // Hassenbr...	4201	107	0	0

English Tweets (Image by Author)

Conclusion

In this blog, we have explained how to scrape tweets using the `snsrape` library. We have also demonstrated how to customize the scraping process to meet your need. `snsrape` is definitely a must-go-for library to efficiently scrape tweets for multiple purposes.

Also, If you like reading my stories and wish to support my writing, consider [becoming a Medium member](#). With a \$ 5-a-month commitment, you unlock unlimited access to stories on Medium.

Feel free to follow me on [Medium](#), [Twitter](#), and [YouTube](#), or say Hi on [LinkedIn](#). It is always a pleasure to discuss AI, ML, Data Science, NLP, and MLOps stuff!

Data Science Python

Enjoy the read? Reward the writer. ^{Beta}

Your tip will go to Zoumana Keita through a third-party platform of their choice, letting them know you appreciate their story.

[Give a tip](#)

Sign up for This AI newsletter is all you need

By Towards AI

We have moved our newsletter to: ws.towardsai.net/subscribe [Take a look.](#)

By signing up, you will create a Medium account if you don't already have one. Review our [Privacy Policy](#) for more information about our privacy practices.

[Get this newsletter](#)

[About](#) [Help](#) [Terms](#) [Privacy](#)

Get the Medium app

