

# Capturing Twitter data with the Twitter API, v2 (Academic Track) and 4CAT

<https://tinyurl.com/nmrw-twitter-v2-4cat>

Department of Media Studies  
University of Amsterdam

## New Media Reference Worksheet

Version: April 2022

Department of Media Studies  
University of Amsterdam

<https://www.uva.nl/en/disciplines/media-studies>  
<http://www.mediastudies.nl/>



Created by Stijn Peeters

Content is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International licence ([CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/))

### In this worksheet you learn

- What the process is of getting access to the 'Academic Track' of Twitter's API
- How you can use this access to create Twitter datasets with the 4CAT Capture and Analysis Toolkit
- How to design a query that will get you the tweets you are looking for
- What further analyses you can think of after collecting the data

**⚠ NOTE: Following Elon Musk's acquisition of Twitter, API access to the platform (now named X) has been greatly restricted. As of mid 2024, 4CAT still worked with the Twitter API, if you manage to get an access token, but development of the 4CAT Twitter module has mostly been halted and may cease to work at some point in the future. ⚠**

## Preface

Twitter has long been a darling of social media research, as a platform that offers a view on public debates through small, easy-to-analyse bits of data - tweets. As short bits of text that are easy to

group together via hashtags, users or reply threads, tweets are compelling objects of research and can be used to e.g. map debates on a particular topic structurally, or to find out who the important actors in a certain debate are.

In contrast to many other social media platforms, Twitter actively facilitates such data-driven research with an API, an interface you can programmatically connect to to request data within certain parameters. If you have access to this API, you can use it to 'ask' Twitter for tweets with e.g. a certain hashtag or keyword, which it will then send to you, as 'raw' data objects in batches.

You can interface with this API manually, or program your computer to do it, but this can be complicated particularly if you are less familiar with programming. There are tools that handle this part - requesting and receiving data from Twitter - for you. One of these tools is 4CAT, a general-purpose social media capture and analysis toolkit.

The goal of this worksheet is to explain how you can use 4CAT to create a dataset of tweets matching your query. At the end, you will know what needs to be done to get access to Twitter's API, how to use that access to request tweets with 4CAT, and what you can do next.

## 1. Requesting access

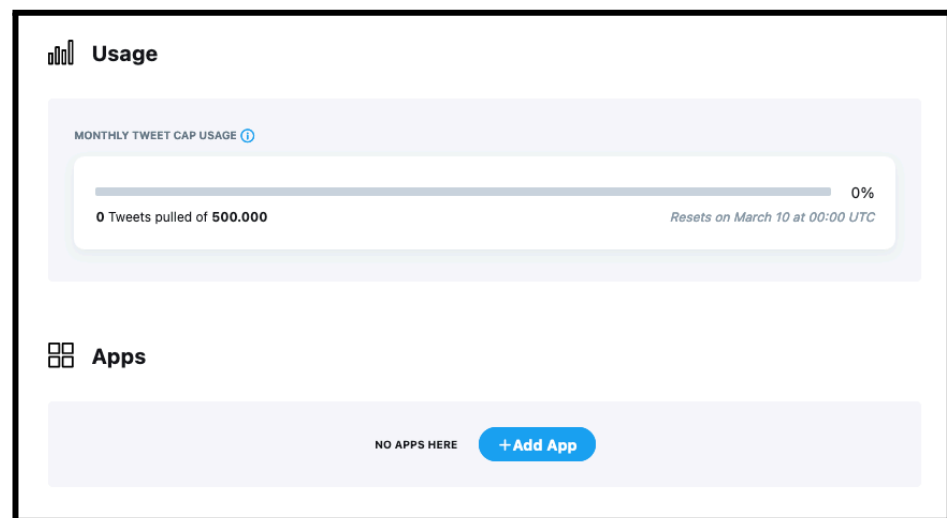
Access to the [academic track of the Twitter API](#), which allows 'full-archive search' – searching the full archive of all tweets posted since the platform started – is only available by request. You can read more about the process [here](#). To request access, you can follow these steps:

1. Start the process by going to the [relevant page](#) in the developer portal. You need to be logged in to Twitter to start the process.
2. If you match the criteria listed on the page, click 'Start Academic Research Application'.
3. You will be asked to fill in a series of questions about how you plan to use the API. It is recommended that you keep a copy of your answers to these questions in a separate document, since you will not be able to see them after submitting!
4. If you are a student requesting access for your MA thesis, ask your supervisor to add a statement to their university profile page confirming that you are their thesis student; a link to this page can then serve as proof that you indeed qualify for access.
5. After filling in the form, Twitter will manually vet your request. This process takes a few days, or sometimes up to one week. They may ask you to clarify some of your answers before granting access.
6. If you have been granted access, you will receive an e-mail saying so at the address you provided.

## 2. Creating a bearer token

You've been granted access to the Twitter API for Academic Research. Great! To actually use the API, you first need to create a 'bearer token', a string of text to use while interacting with the API to verify that you actually have access, similar to a password. To do this, follow these steps:

1. Go to the Twitter Developer Portal at <https://developer.twitter.com/en/portal/> and log in.
2. Navigate to the page of the project you were granted access for, in the 'Projects & Apps' list on the left side of the screen.
3. You will see a dashboard page with an indication of how much of your monthly quota you have used (it should be 10 million), and a list of apps (which should be empty). Click the '+ Add App' button. An 'app' here is simply an entry point to the API, which you can use later to get data from it.



4. You will be asked to enter a name for your app. Enter anything here - it doesn't really matter. After submitting, you will be shown a screen with an 'API Key', an 'API Secret key' and a 'Bearer token'. These are used to allow Twitter to verify that you really have access, while interacting with the API. **Copy the 'bearer token'** and save it somewhere; you will not be able to see it again (though you can generate a new one if you lose it).
5. Click 'No thanks, go to the dashboard' at the bottom of the screen.

## 3. Creating a dataset with tweets in 4CAT

Now you have access to the academic track of the Twitter API, you can use it to (among other things) retrieve historical tweets from Twitter for a given query. To do this you can use [4CAT](#), the capture & analysis toolkit, which can interface with the Twitter API and retrieve tweets according to parameters you specify.

1. Now you have a bearer token, you can use 4CAT to create a dataset of historical tweets matching a given query.
2. In 4CAT, go to 'Create dataset'. On the 'Create new dataset' page, choose the 'Twitter API (v2) Search' data source.
3. You will now be able to set the parameters of your query:

This data source uses the full-archive search endpoint of the Twitter API, v2. To use this endpoint, you must have access to the Academic Research track of the Twitter API, and provide a valid **bearer token**. The bearer token **will be sent to the 4CAT server**, where it will be deleted after the scrape has started.

Please refer to the **Twitter API documentation** for more information about this API endpoint and the syntax you can use in your search query. Note that any tweets retrieved with 4CAT will count towards your monthly Tweet retrieval cap. Also note that results are saved as **NDJSON**, not CSV.

Query:

Tweets to retrieve:  (0 = unlimited)

By default, Twitter returns tweets up til 30 days ago. If you want to go back further, you need to explicitly set a date range.

Date range:  to

API Bearer token

4CAT can replace author names with their hash value. Other personal information may persist; it is your responsibility to further anonymise data where appropriate.

Pseudonymise: ☒ Replace author names with hash values

4. The syntax you can use for your query is the same syntax you can use to search for tweets on Twitter itself. A full reference is available [on Twitter's developer site](#) - this page also explains how you can combine various queries. You can combine various operators and keywords to make your query specific. For example, the following query would return all tweets from [o i l a b](#) containing the word 'qanon', but ignores retweets and replies:

`qanon from:o_i_l_a_b -is:retweet -is:reply`

5. Note that if you want to search for multiple keywords/hashtags, you need to explicitly indicate this in the query. By default, if the query contains multiple keywords, only tweets

containing *all* of these keywords would be returned; the query matches every tweet that matches all of the query's components. You can use the **OR** keyword and (brackets) to combine various parts of the queries in a more complex way. If you would want to modify the previous query to return all tweets containing *either* 'qanon' or 'dmi', this would be:

```
(qanon OR dmi) from:o_i_l_a_b -is:retweet -is:reply
```

There are many operators you can add to your query to make it more specific. They can all be negated by prepending a hyphen -, for example (as in the query above) to select tweets that are *not* retweets. Here are some commonly useful ones (but the Twitter documentation has [a full list](#)):

Syntax	Selects
from:jack	Tweets by @jack
to:jack	Tweets from any user that are replies to tweets by @jack
is:retweet	Tweets that are retweets
is:reply	Tweets that are replies
is:quote	Tweets that are quote tweets
url:"nytimes.com"	Tweets containing a link to nytimes.com
has:media / has:images / has:videos	Tweets containing images, videos, or media (i.e. videos and/or images)
lang:nl	Tweets classified as Dutch by Twitter's language recognition algorithm. See Twitter's documentation for a list of valid language codes.
place_country:nl	Tweets geo-tagged with locations in The Netherlands. See Twitter's documentation for a list of valid country codes. Note that most tweets are <b>not</b> geo-tagged!

- By default, 10 tweets are retrieved. You can set this amount as high or low as you want, but note that you will not be able to retrieve more than 10 million tweets per month (something to keep in mind especially when retrieving '0', i.e. unlimited tweets). You can see how many tweets you can retrieve on your Twitter dashboard.

The more tweets there are to collect, the longer your query will take! It is often good to




start small, with short date ranges and specific keywords, and then only broaden your scope once you have a decent idea of how much data that returns. You can collect approximately 115,000 tweets per hour in ideal circumstances.

7. If you want to search for tweets older than 30 days, you additionally need to set a date range. Date ranges need to end in the past.
8. Fill in the bearer token you generated earlier in the 'API Bearer token' field.
9. By default, 4CAT pseudonymises tweet author information. This is good practice; you can disable this if author information is relevant to your research. Note that you still bear the responsibility of collecting and using the data ethically.
10. After clicking 'Create dataset', tweets will be retrieved from Twitter. You can keep track of the progress at the right side of the screen and you will receive a notice when the dataset is complete, after which a link to it will appear in the panel on the right. You can also find your dataset (and others you have created) on the 'Past results' page in 4CAT.

## 4. Working with the dataset in 4CAT

4CAT has many 'processors' you can use to further process and analyse the dataset you have created. You can find them below the dataset details on the result page:

The screenshot shows the 'Query' result page in 4CAT. At the top, there are buttons for 'Add to favourites', 'Permalink', 'Delete dataset', and 'Re-run dataset'. Below these, the dataset details are listed: 'Data source' is 'twittrv2', 'Queued at' is '04 Mar 2021, 19:36', 'Queued by' is 'stijn.peeters@uva.nl', 'Parameters' is 'query: from:thierrybaudet OR from:marjovrij amount: 1000', and 'Result file' is 'Download ndjson (1,000 items, 888.51kB)'. Below the details, there is a section titled 'Analytical processors' with instructions to start analysis by choosing a module. It also mentions the '4CAT Cookbook' for new users. At the bottom, there is a 'Presets' section with two options: 'Annotate images with Google Vision API' and 'Create image wall'.

Query	
<a href="#">★ Add to favourites</a>	<a href="#">Permalink</a>
<a href="#">Delete dataset</a>	<a href="#">Re-run dataset</a>
Data source	twittrv2
Queued at	04 Mar 2021, 19:36
Queued by	stijn.peeters@uva.nl
Parameters	query: from:thierrybaudet OR from:marjovrij amount: 1000
Result file	<a href="#">Download ndjson (1,000 items, 888.51kB)</a>
Analytical processors	
Start your analysis of the retrieved data by choosing one of the analysis modules below. Note that some may take a while to complete, so carefully consider which one you want to run before queueing it.	
If you are new to 4CAT, we recommend you consult the <a href="#">4CAT Cookbook</a> , which explains this interface and includes examples of several analyses you can run.	
Presets	
 Options	<b>Annotate images with Google Vision API</b>  Use the Google Vision API to annotate images linked to in the dataset the most often. Note that the Google Vision API is a paid service and using this processor will count towards your Google API credit!
 Options	<b>Create image wall</b> Use a sample of the images (up to 125) linked to the most in the dataset and put them in a single image, side by side.

Discussing them all is beyond the scope of this document, but here are a few things you can for example use to draw insights from your dataset:

- If you want a quick impression of the activity within the dataset, you can run the 'Monthly histogram' processor under 'Presets', which will generate a histogram graph showing the amount of tweets per month.
- Use the 'Count values' processor under 'Post metrics' to find out what the most prolific author in the dataset is, or the most popular hashtag, or the most linked-to site. You can configure the processor to provide results per day, week, month, year or overall. The result can then e.g. be visualised in a RankFlow diagram.
- Use the 'Co-tag network' processor to create a co-hashtag network that can be opened with [Gephi](#). This allows you to see what hashtags are used together often in the tweets in your tweet. Often, there are clusters of related hashtags, which correspond to particular interests or to specific positions in the discourse.
- Use the 'Download images' processor under 'Visual analysis' to download images used in Twitter, from the 'images' column of the dataset. These can then be downloaded for off-line analysis or visualised together with e.g. the 'Image Wall' processor.
- A somewhat more advanced analysis: if you want to know what words are used most often in the dataset...
  - First run the 'Tokenise' processor, under 'Text analysis', filtering stopwords for the appropriate language. This will split up each tweet into separate tokens (words), grouping all words by the chosen time interval (e.g. day, month or overall).
  - Then, run the 'Vectorise tokens' processor on the result of that, which will count how often each token (word) is used.
  - Then, run the 'Top vectors' processor on the result of *that*, which will generate a file with the amount of occurrences for the most used words, sorting by occurrences, and optionally grouping the results per time interval
  - The result of this is a csv file, which you can open and browse to see what the most popular word was per (for example) month.

## Further notes & reading

There are some short tutorial videos about 4CAT if you want to know more about how the tool works:

- [4CAT Tutorials on YouTube](#)

For a refresher on how to work with spreadsheets and networks, you can take a look at the relevant other Media Studies worksheets:

- [Reference Worksheet I: Data Management](#)
- [Reference Worksheet III: Data Visualisation](#), section 5, 'Gephi'