Ryerson University | TED ROGERS SCHOOL OF MANAGEMENT | SocialMediaLab.ca "Making sense of a networked world"

Session 4:

Getting Started with Communalytic

Toxicity Analysis of Twitter Data using Perspective API

June 24, 2021

SocialMediaLab.ca

CSS Bootcamp Schedule Summer 2021

Session #1	Getting Started with Communalytic: Data Collection from Reddit	May 13, 2021, 10:00-11:30am (EDT)
Session #2	Toxicity Analysis with Reddit Data using Perspective API	May 27, 2021, 10:00-11:30am (EDT)
Session #3	Getting Started with Communalytic: Data Collection from Twitter (Twitter Thread via API v2.0 and Twitter Academic Track)	June 10, 2021, 10:00-11:30am (EDT)
Session #4	Toxicity Analysis of Twitter data using Perspective API	June 24, 2021, 10:00-11:30am (EDT)
Session #5	Social Network Analysis of Signed Networks with Reddit and Twitter data	July 8, 2021, 10:00-11:30am (EDT)
Session #6	Getting Started with Communalytic: Data Collection from Facebook & Instagram (via CrowdTangle API) + Social Network Analysis of Two-mode Semantic Networks with CrowdTangle data	July 22, 2021, 10:00-11:30am (EDT)

Today’s agenda

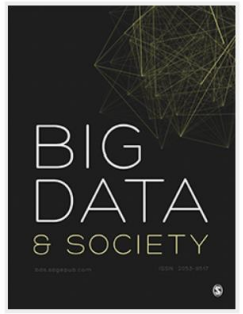
Guest talk by Dr. Paola Pascual-Ferra

Brief recap of how to collect Twitter data and run the Toxicity Analysis in Communalytic

Analyzing Twitter data using Jupyter Notebook

Today's agenda

Guest talk by Dr. Paola Pascual-Ferra



Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic

Paola Pascual-Ferrá, Neil Alperstein, Daniel J Barnett, Rajiv N Rimal

Show less

First Published June 10, 2021 | Research Article | Check for updates

https://doi.org/10.1177/20539517211023533

Article information

Altmetric 10

https://doi.org/10.1177/20539517211023533

Today's agenda

Guest talk by Dr. Paola Pascual-Ferra

Brief recap of how to collect Twitter data and run the Toxicity Analysis in Communalytic

Analyzing Twitter data using Jupyter Notebook

Brief Recap of Sessions 2 & 3

Data Collection from Twitter

How to request Twitter's Bearer authentication token

Toxicity Analysis & How to request Google's Perspective API key

Communalytic is a research tool for studying online communities and online discourse.

Communalytic can collect and analyze public data from social media platforms. It uses advanced text and social network analysis techniques to automatically pinpoint toxic and anti-social interactions, identify influencers, map shared interests and the spread of misinformation, and detect signs of possible coordination among seemingly disparate actors.

collect

extract

analyze

visualize

https://communalytic.com/

How to choose between Communalytic Edu and Pro.	communalytic EDU	communalytic PRO
Account Type	Free	$349/6-mo. to support site infrastructure (server-side data collection, storage, processing, analysis and visualization)
Designed For	Students and is ideal for teaching and learning about social media analytics	Academic researchers and is ideal for large scale academic research projects
Account Caps	≤ 30K records shared across 3 datasets	≤ 10M records shared across 50 datasets
Reddit	Live-collection* of public posts from any public subreddit for ≤ 7 consecutive days (Limit: ≤ 30K posts)	Live* & historical collection of public posts from any public subreddit for ≤ 31 consecutive days (Limit: Account Caps)
Twitter Threads (API ver.2) req. Twitter developer's account	Public replies to any public tweet posted within the previous 7 days (Limit: **≤ 30K tweets**)	Public replies to any public tweet posted within the previous 7 days (Limit: **≤500K tweets/month**)
Twitter Academic Track (API ver.2) req. Application to Twitter	Not supported	Full-archive historical-search of tweets back to 2006 (Limit: **≤ 10M tweets/month**)
CrowdTangle (FB/IG) URL Search req. a CrowdTangle account	Public Facebook or Instagram posts that shared the same URL (Limit: ≤ 30K posts)	Public Facebook or Instagram posts that shared the same URL (Limit: Account Caps)

* Live-collection = the collection of posts/tweets posted on or after the date when you initiated the data collection.

communalytic (EDU)

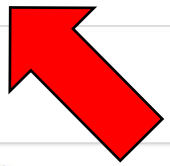
FAQ Tutorials Publications My Datasets

My Datasets

Server Time: June 10, 2021 12:19 UTC

Collect data from

Reddit (LIVE) Twitter Thread CrowdTangle CSV File



Search dataset list...

My Datasets Shared With Me

Donald J. Trump @realDonaldTrump · Sep 10

Sending out 80 MILLION BALLOTS to people who aren't even asking for a Ballot is unfair and a total fraud in the making. Look at what's going on right now!

20.4K 30.5K 98.6K

David Former Trump Supporter Weissman @davidm... · Sep 10

Replying to @realDonaldTrump

Tweet reported for misinformation on the elections.

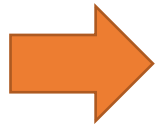
104 146 2.3K

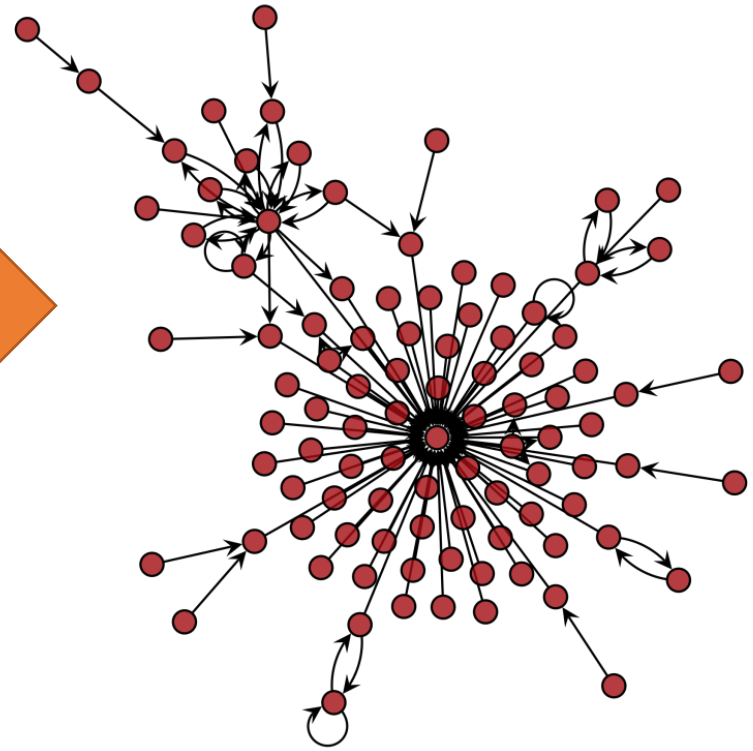
Osha Davidson @OshaDavidson · Sep 10

Me too.

5 2 181

Show replies





A network representation of replies to a given tweet

communalytic (EDU)

FAQ Tutorials Publications My Datasets

Server Time: May 11, 2021 14:04 UTC

Collect data from

Reddit (LIVE) Twitter Thread CrowdTangle CSV File

Search dataset list...

Back to My Datasets

Twitter Thread Data Collection

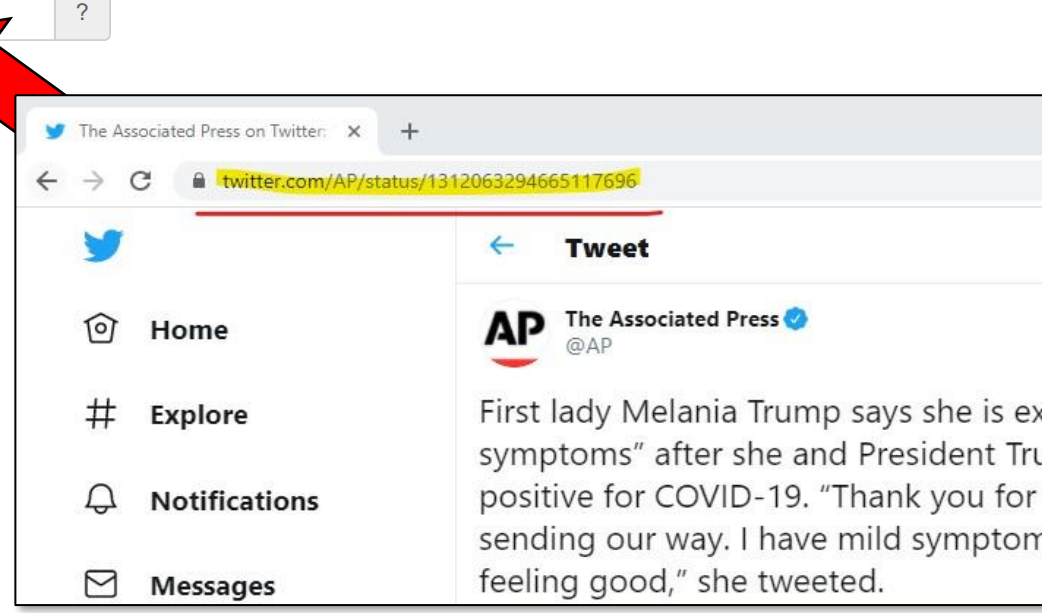
Please enter the following details to begin collection

Dataset name

Enter the URL of a popular tweet (with many replies) ?

Email me once job completes

Start Data Collection



Reddit (LIVE) Twitter Thread CrowdTangle CSV File

Search dataset list...

My Datasets | Shared With Me

Total records in the account: 186 / 30000

1%

Dataset Name	Number of Records	Toxicity Analysis	Network Analysis	Download Dataset	Collaborators	Delete
Test Reddit reddit : CanadaPolitics Click to check progress	Extracting until 2021-05-12 100					
Test Twitter twitter : https://twitter.com/NBA/status/1392123045968498692	86					

Back to My Datasets

Overview

Dataset Name: **Test Twitter**

Platform: **twitter**

Search Criteria: **https://twitter.com/NBA/status/1392123045968496892**

Collection started: **05/11/2021**

Collection ended: **05/11/2021**

Records: **86**

Posts Per Day

Word Cloud

Emoji Cloud

Top Ten Posters

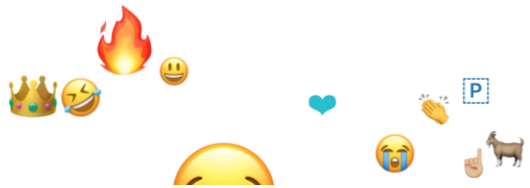
Word Cloud

CSV EXCEL PNG

This chart shows up to 100 most frequently used words in your dataset. (Excluding numbers, URLs, and stop words in 15 different languages.)

Emoji Cloud

CSV EXCEL PNG



Working with Twitter's Academic Product Track in **Communalytic Pro**

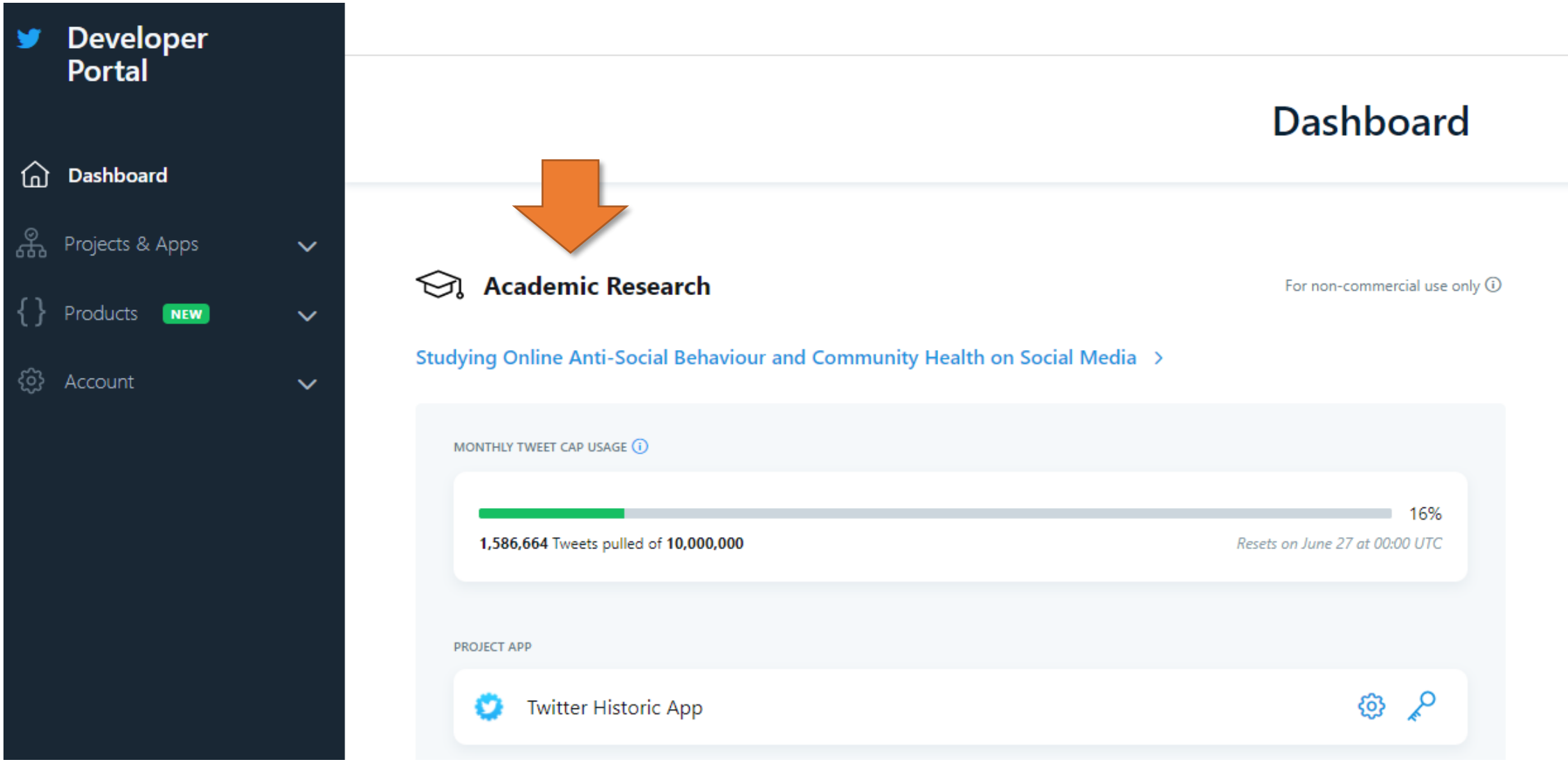
https://pro.communalytic.com

Eligibility requirements:

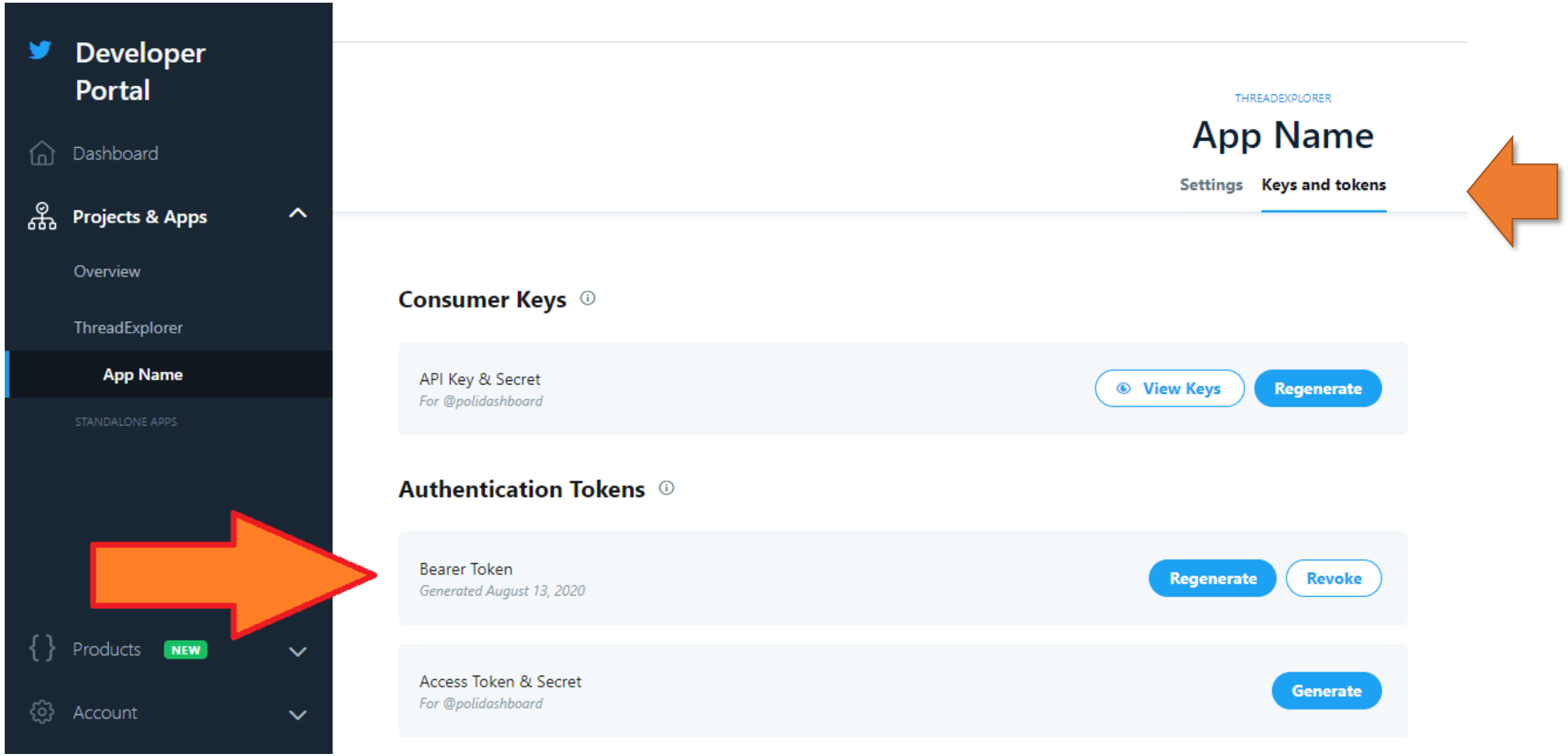
- Master's student, doctoral candidate, postdoc, faculty, or research-focused employee at an academic institution
- A clearly defined research project (i.e., you have a specific plan for Twitter data use, analysis, and sharing)
- Non-commercial use

- **Apply at https://developer.twitter.com/en/portal/petition/academic/is-it-right-for-you**

Twitter's Developer Dashboard: Confirm Access to the Academic API

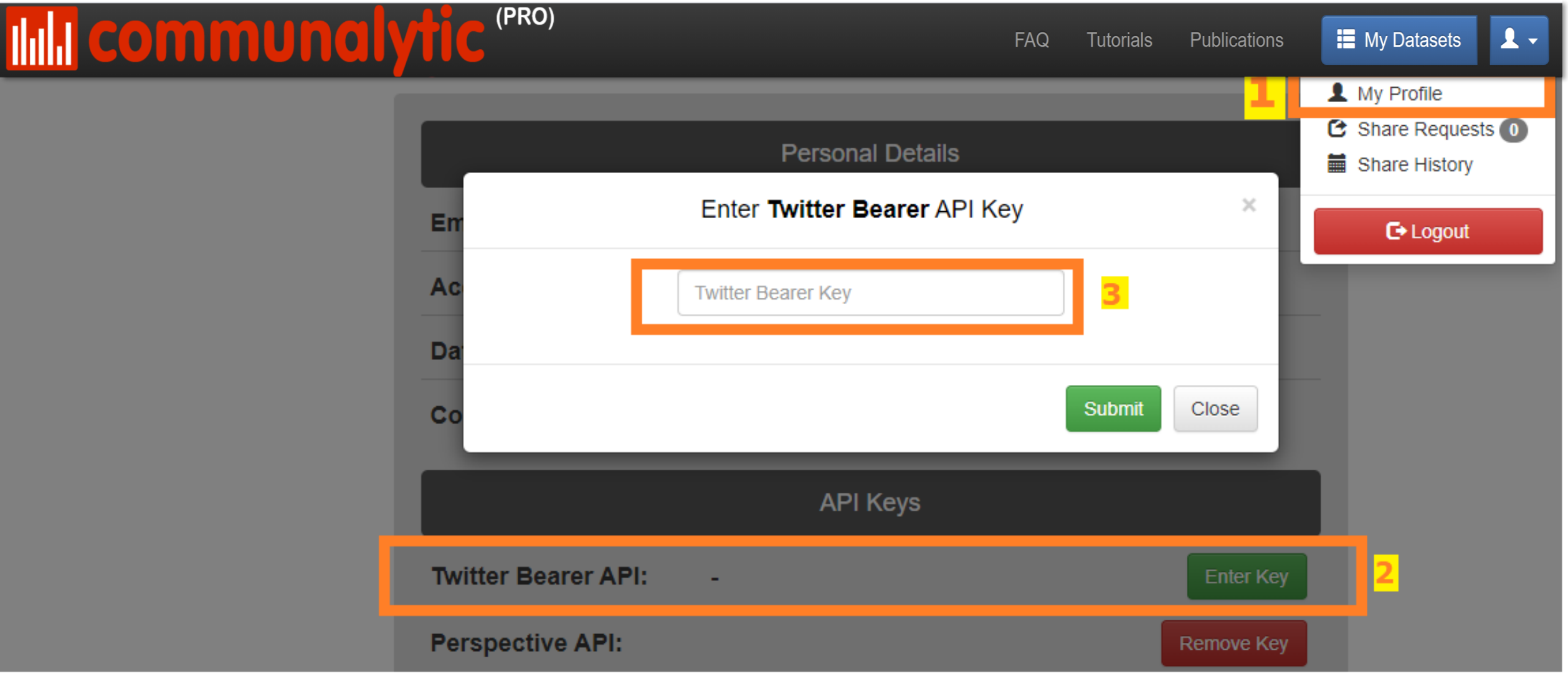


Twitter's Developer Dashboard: Retrieve your "Bearer" access token



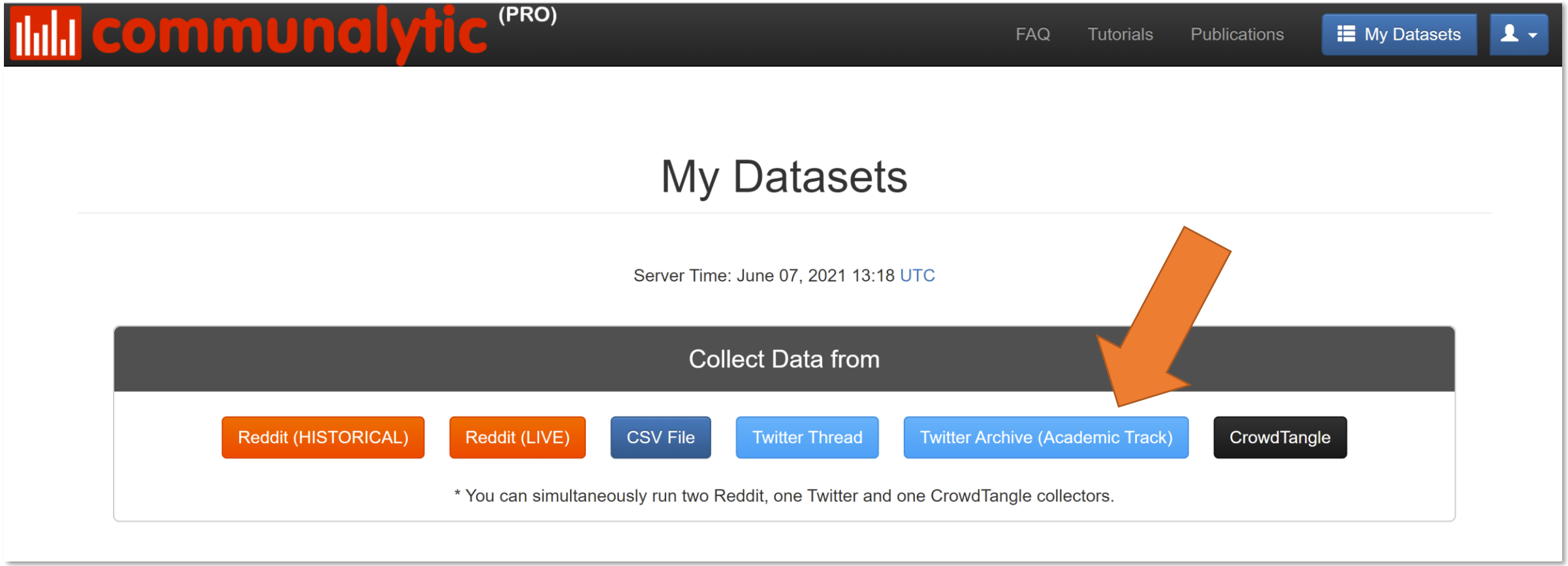
1. Twitter API → 2. Collection → 3. Analysis → 4. Export → 5. Sharing

Adding your “Bearer” access token to your Communalytic **PRO** account



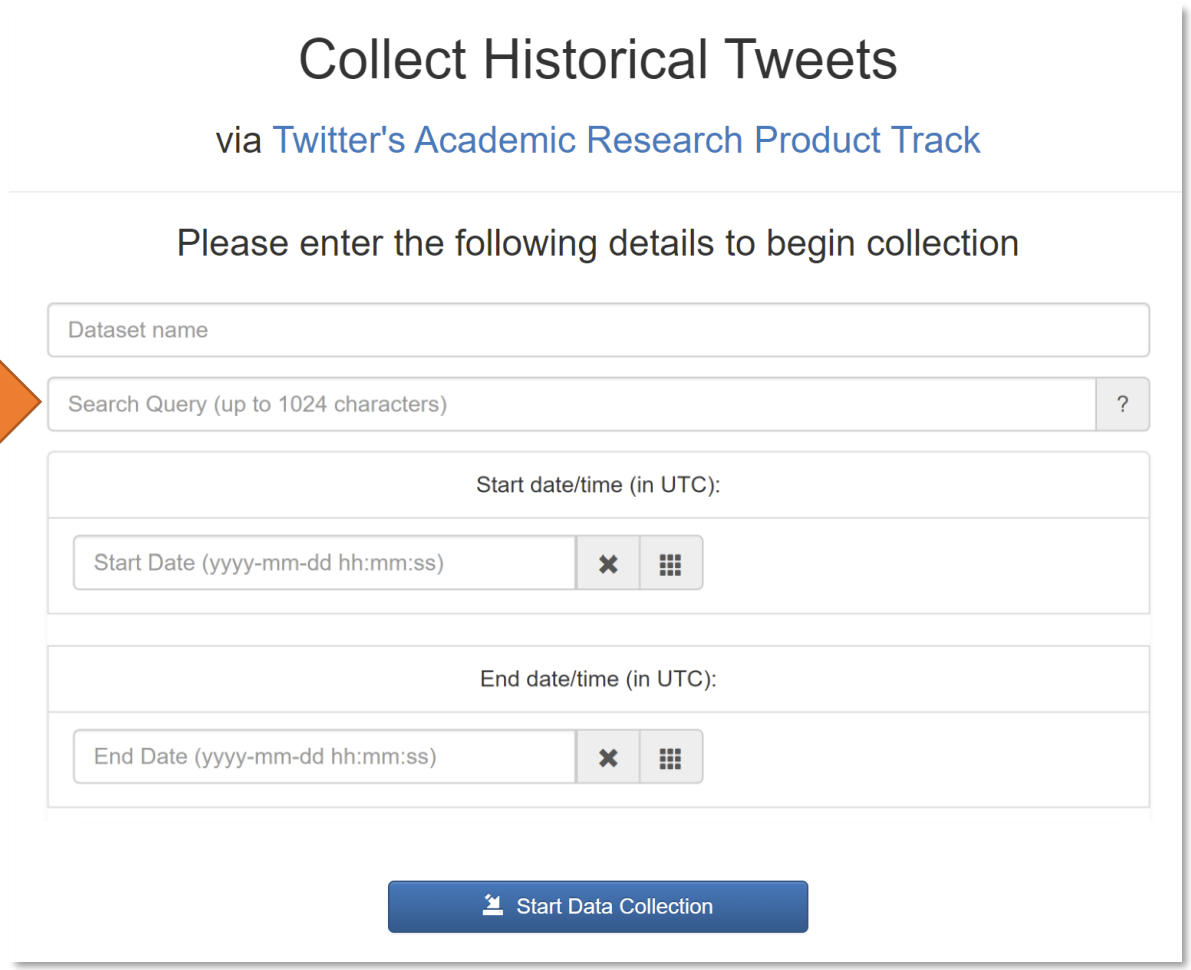
1. Twitter API → 2. Collection → 3. Analysis → 4. Export → 5. Sharing

Full-archive Search API: Create a new search (1)



Full-archive Search API: Create a new search (2)

Search Query
(up to 1024 chars)



Toxicity Analysis with Perspective API

Perspective | Developers

About the API Docs Contact Us

- Overview
- Key Concepts
- Attributes & Languages
- Best Practices & Risks
- Methods
- Limits & Errors
- FAQs

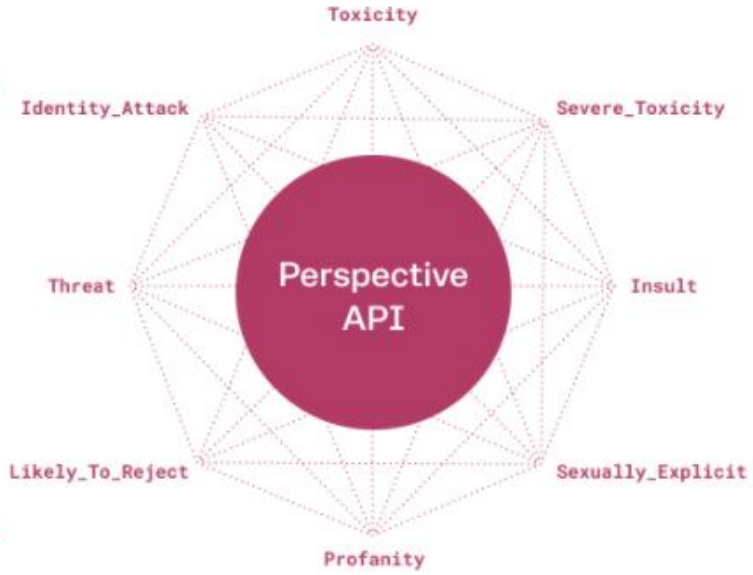
To access the API, review the Get Started documentation.

INPUT: TEXT

"Shut up. You're an idiot!"

OUTPUT: SCORE

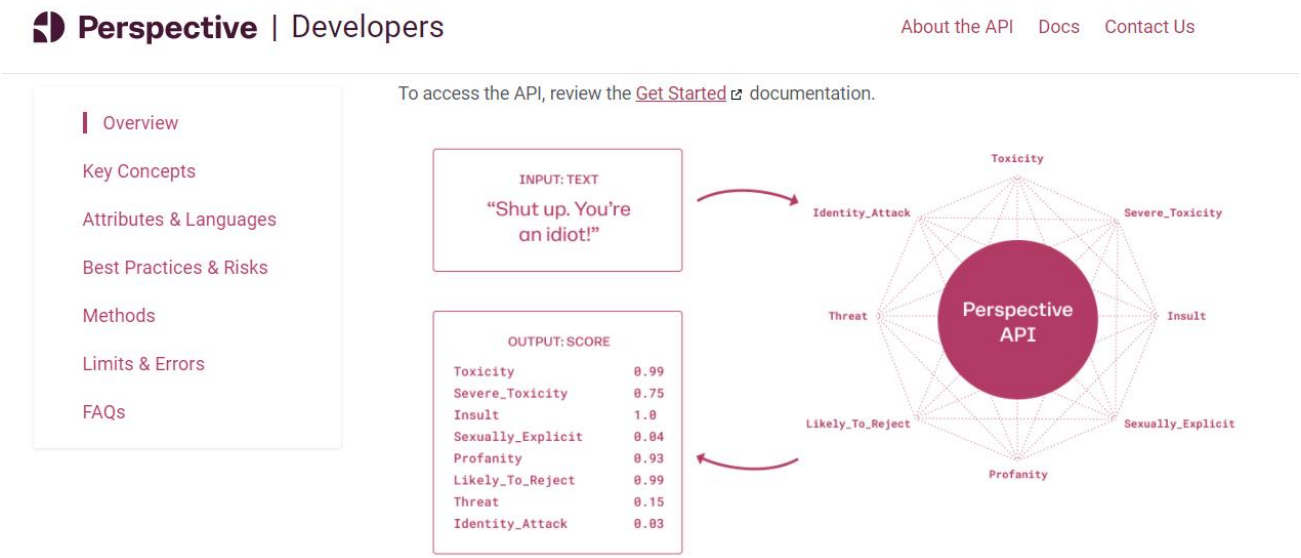
Toxicity	0.99
Severe_Toxicity	0.75
Insult	1.0
Sexually_Explicit	0.04
Profanity	0.93
Likely_To_Reject	0.99
Threat	0.15
Identity_Attack	0.03



https://support.perspectiveapi.com/s/docs-get-started

Toxicity Analysis with Perspective API

- **Training data:** online forums including Wikipedia & New York Times comments, with crowdsourced labels (e.g., a comment is “toxic” or not)
- **Machine Learning:** Convolutional Neural Network (CNN) trained with GloVe word embeddings

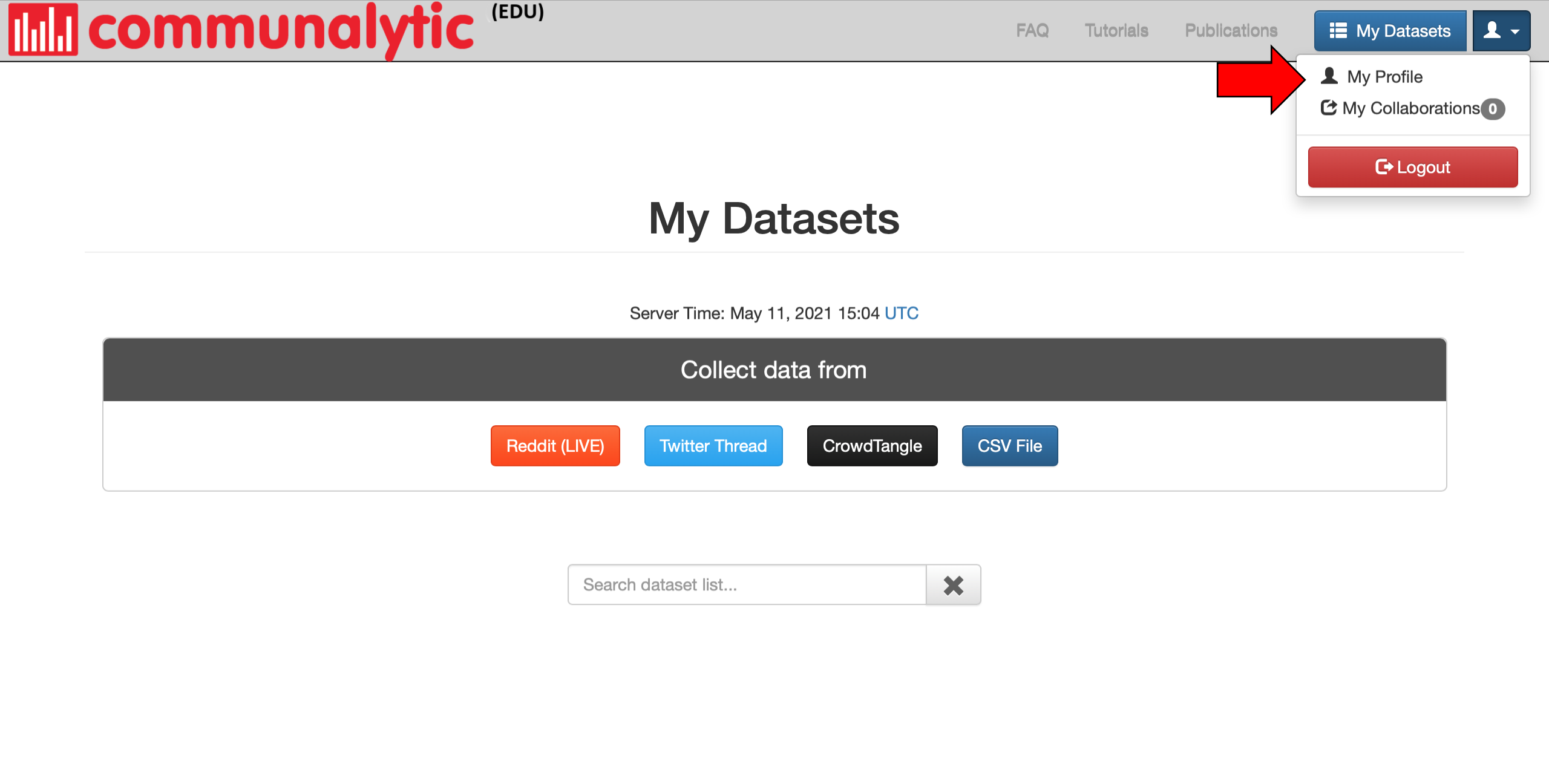


https://support.perspectiveapi.com/s/docs-get-started

Toxicity Scores from Perspective API

	Definition	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	“This is one of stupidest things I've read but fuck me I laughed at the second line”
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	“Fuck off pathetic loser, no one cares about your worthless opinion”
Insult	Insulting, inflammatory, or negative post toward an individual or a group	“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake having him say racist things as a "social experiment"?
Identity attack	Negative post attacking someone because of their identity (including race, gender, sexual orientation, ideology, religion, nationality, etc.)	“You people are a bunch of fags. And I voted for [Political Party Name]”
Profanity	Post with swear words or other obscene language	“Why vote for the [Political Party Name] when you know they won’t win shit.”
Threat	Post with an intention to inflict pain, injury, or violence against an individual or group	“Shoot all yellow vests! We have to kill all Nazis!”

https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages


communalytic (EDU)
FAQ
Tutorials
Publications
My Datasets
My Profile
My Collaborations 0
Logout
My Datasets
Server Time: May 11, 2021 15:04 UTC
Collect data from
Reddit (LIVE)
Twitter Thread
CrowdTangle
CSV File
Search dataset list...

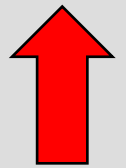
API Keys

Twitter Bearer Token: Remove Key

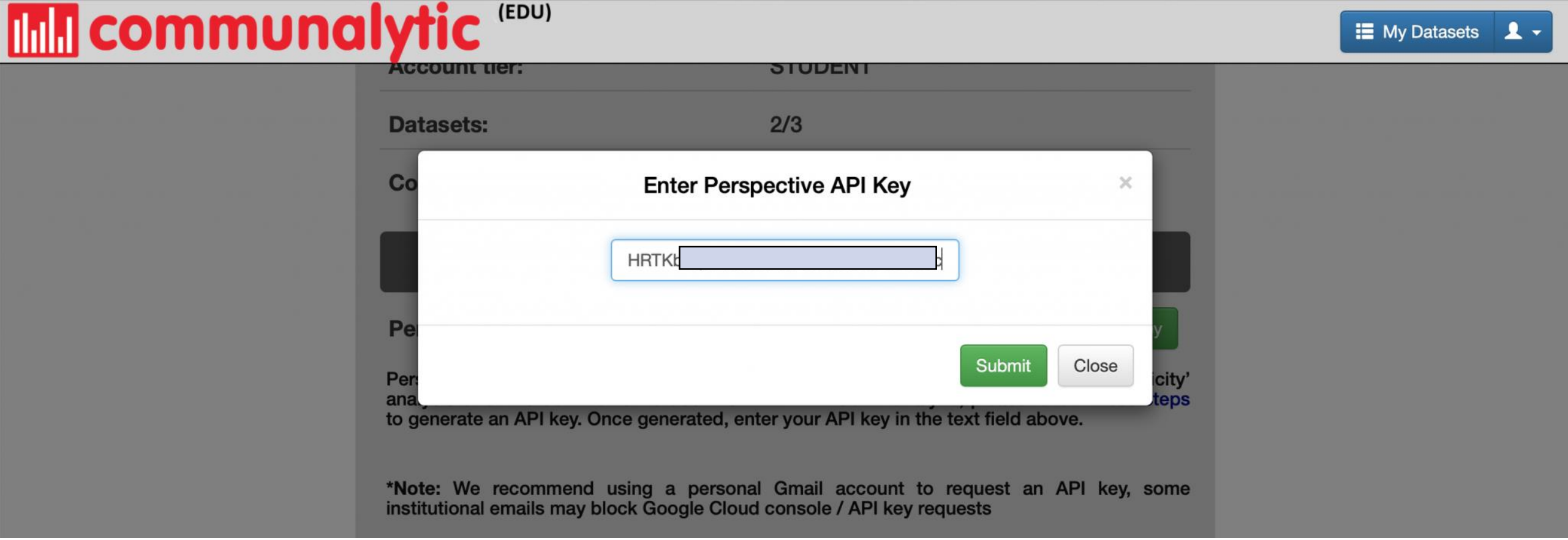
CrowdTangle API: Apply for academic/research access to Facebook's CrowdTangle here — Enter Key

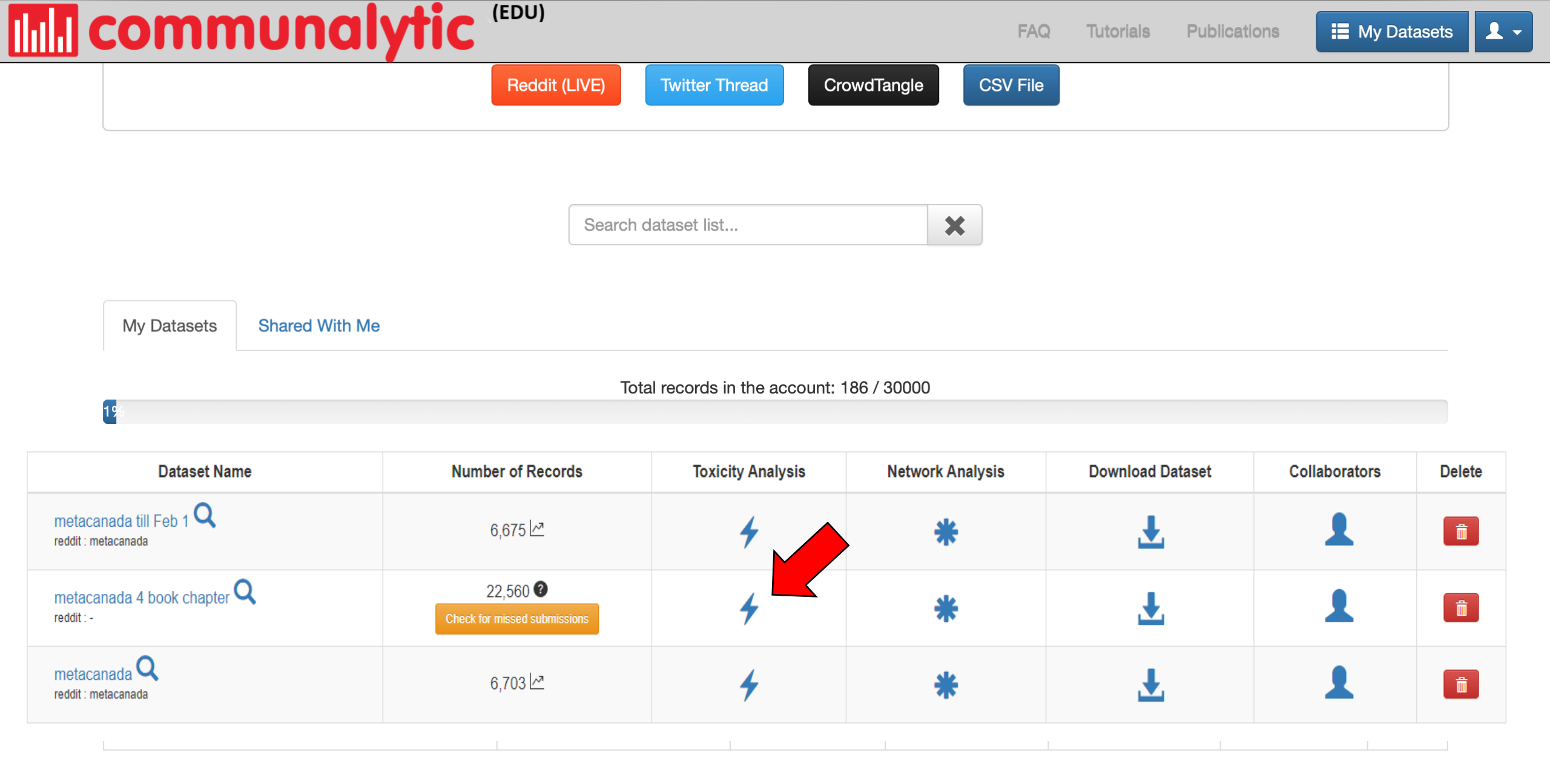
Perspective API: - Enter Key

Perspective is a machine learning API by Jigsaw and Google designed to conduct a 'toxicity' analysis of online comments. To use this API within Communalytic, please follow these steps to generate an API key. Once generated, enter your API key in the text field above.



***Note:** We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests**


communalytic
(EDU)
My Datasets
Account tier:
STUDENT
Datasets:
2/3
Enter Perspective API Key
HRTKb
Submit
Close
to generate an API key. Once generated, enter your API key in the text field above.
*Note: We recommend using a personal Gmail account to request an API key, some institutional emails may block Google Cloud console / API key requests



1. Twitter API → 2. Collection → 3. Analysis → 4. Export → 5. Sharing

Overview

Dataset Name:	**metacanada 4 book chapter**
Platform:	**reddit**
Subreddit:	**-**
Collection started:	**2020-02-11 02:11**
Collection before:	**2020-02-11 02:11**
Records:	**22,560**

- Posts Per Day
- Word Cloud
- Top Ten Posters
- Toxicity Analysis

Toxicity Analysis

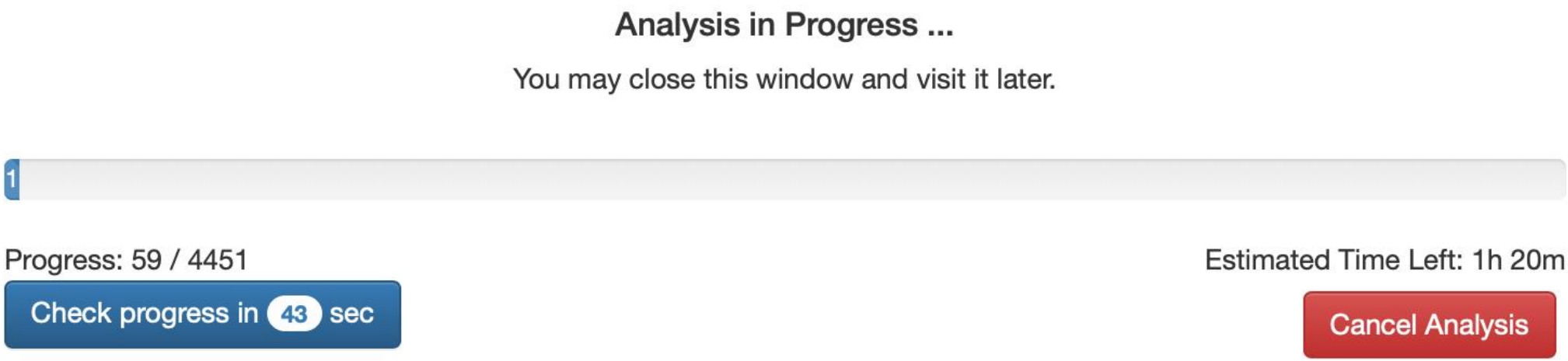
The Toxicity analysis can only run with one language mode at a time.
Please select the primary language of your dataset using the drop-down menu below.

English

Start Analysis

- Toxicity Analysis is based on a machine learning API called Perspective API by Google. If you already have an API key for Perspective, it can be added to your account under My Profile; otherwise, see this guide on how to generate a new Perspective API.
- Due to the post length restriction imposed by the API, Communalytic can only analyze the first 3000 characters of each post. Posts longer than 3000 characters will be automatically truncated.
- Link's and URLs will be removed before text is sent to API for analysis.
- The resulting toxicity scores will be added to the export files and available for download via the "Export Posts" and "Export Network" options.

Toxicity Analysis



Toxicity Analysis

There were **22560** comments analyzed in **English**

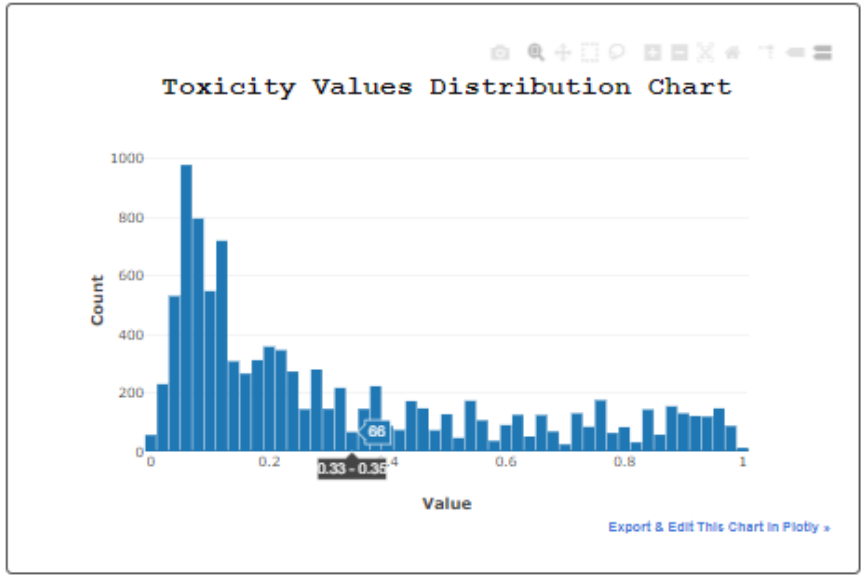
Reset Analysis

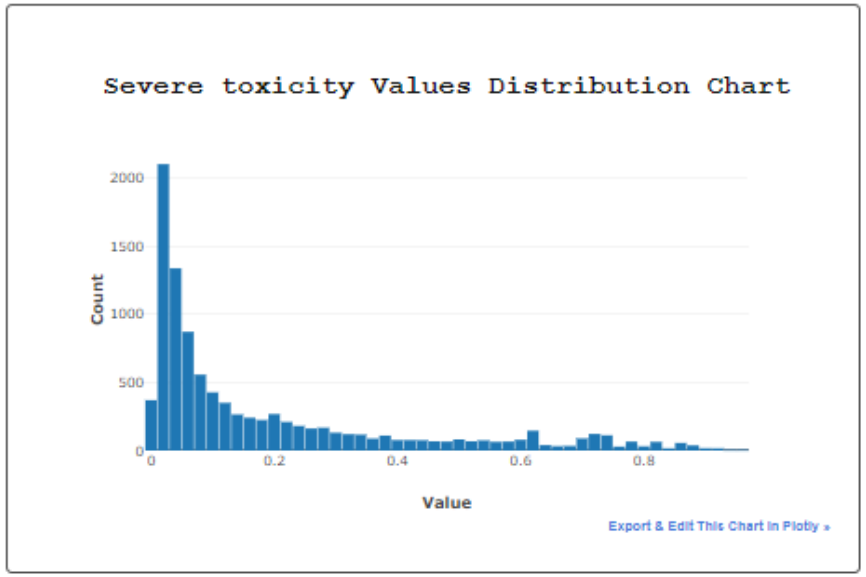
⚠ The Toxicity analysis can only run with one language mode at a time. To change the primary language selection for this dataset, click on the Reset Analysis button and rerun the analysis as needed.

Click on the highest/lowest values to see the top posts for each category

Download the Toxicity analysis results as a CSV file.

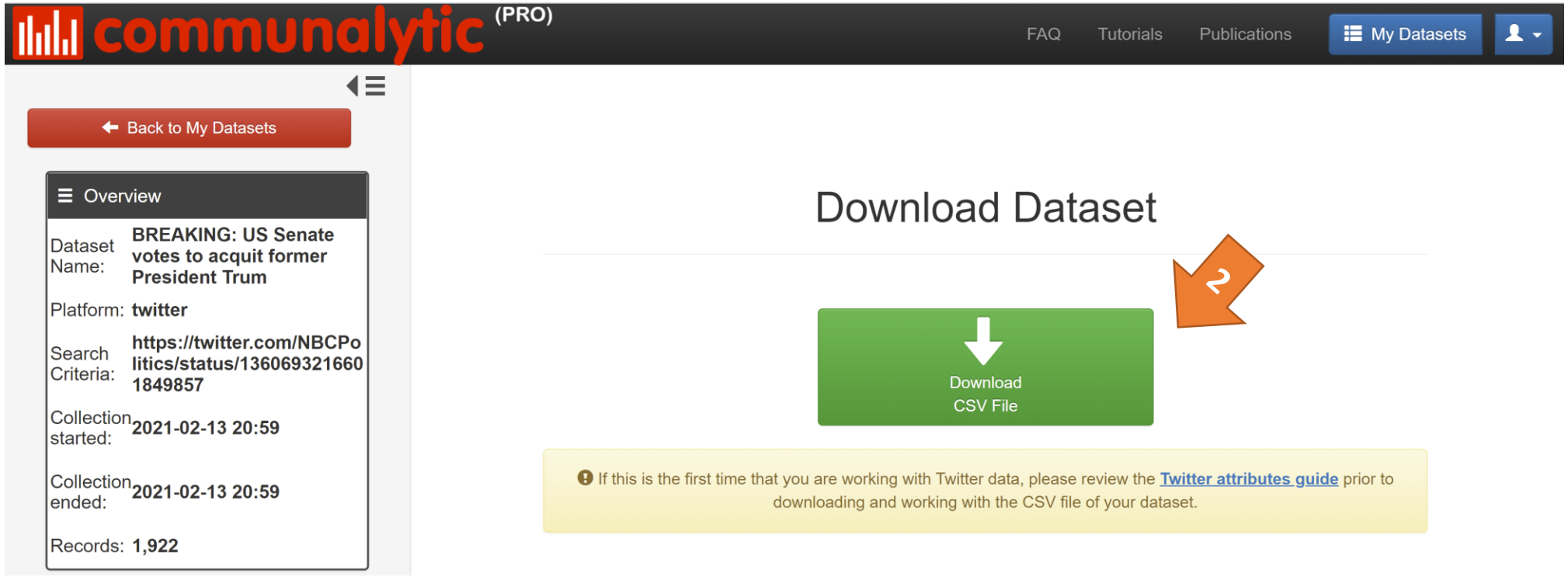
	Average for dataset	Highest value	Lowest value
Toxicity	0.32	1.00	0.00
Severe toxicity	0.19	0.95	0.00
Identity attack	0.25	0.98	0.00
Insult	0.28	0.99	0.00
Profanity	0.26	0.99	0.00
Threat	0.22	0.99	0.01
Sexually Explicit	0.17	1.00	0.00
Flirtation	0.30	0.98	0.02





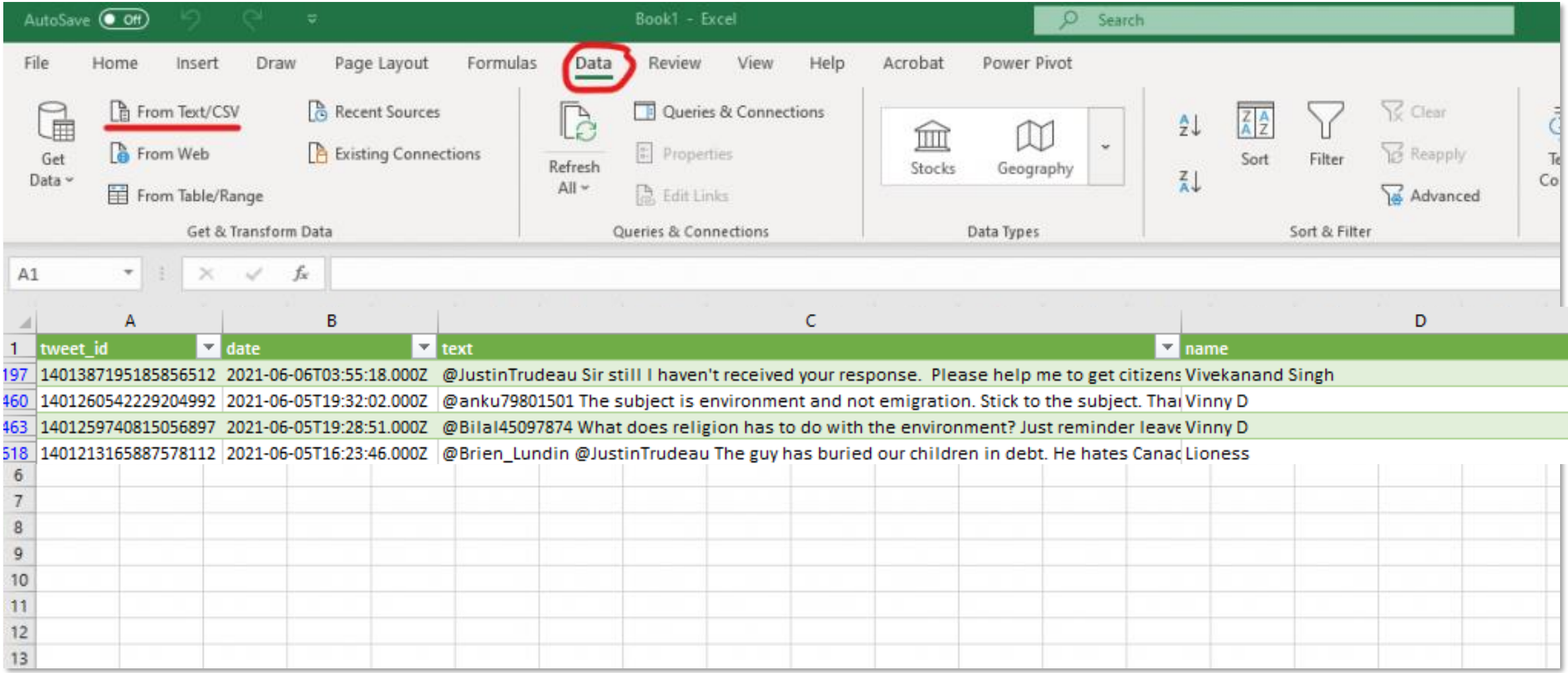
Export Twitter Dataset from Communalytic

BREAKING: US Senate votes to acquit former President Trum twitter : https://twitter.com/NBCPolitics/status/1360693216601849857	1,922					



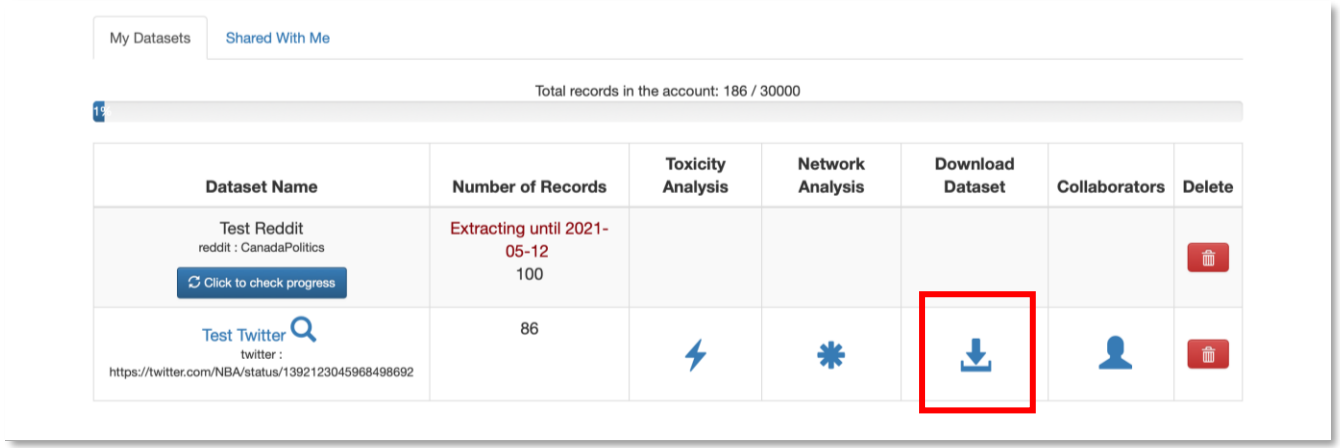
1. Twitter API → 2. Collection → 3. Analysis → 4. Export → 5. Sharing

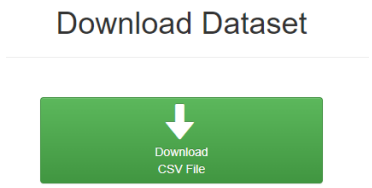
How to open the exported CSV file in Excel

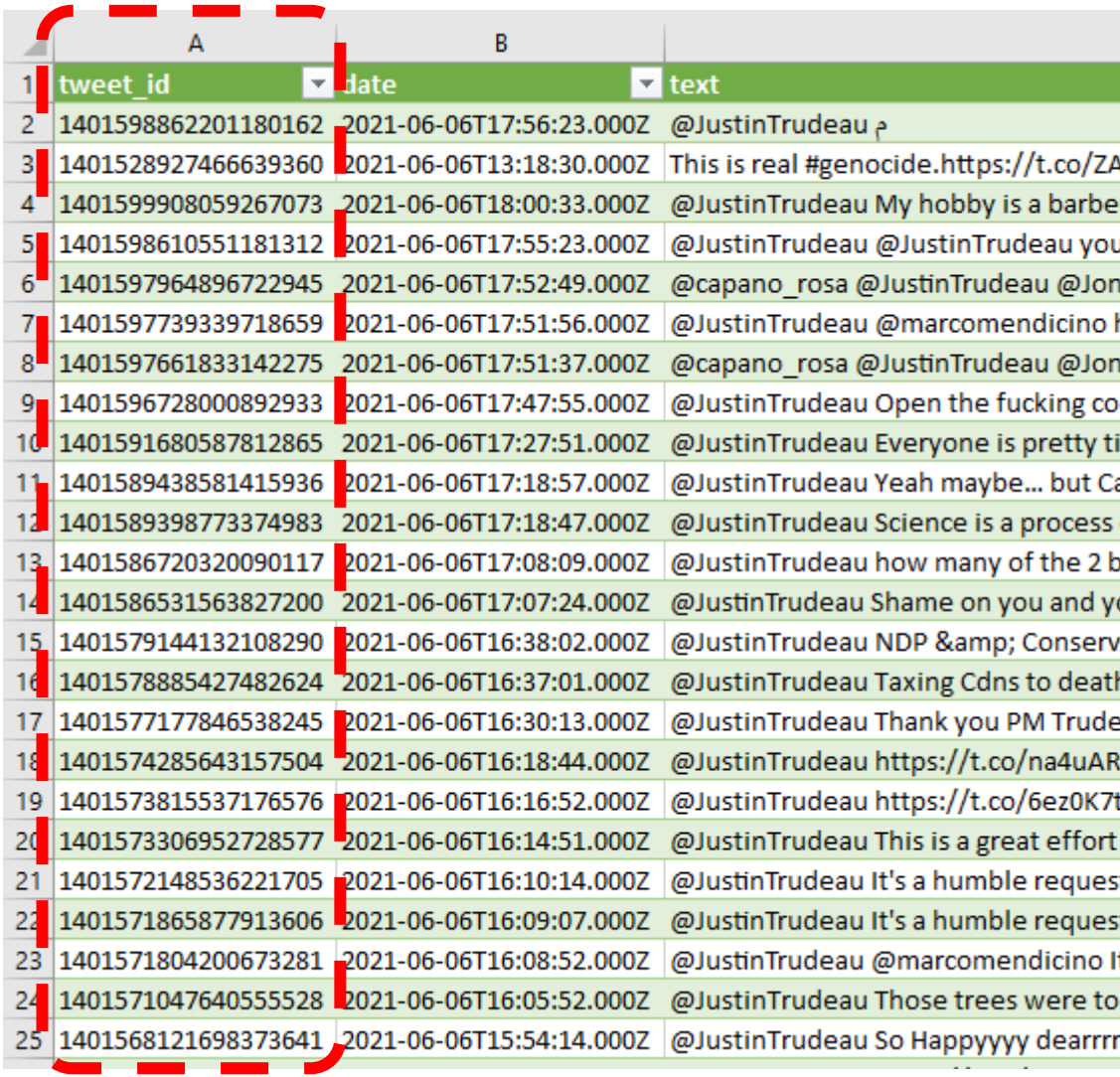


Data Sharing via Tweet Ids

Collect / Export tweets as CSV







	A	B	
1	tweet_id	date	text
2	1401598862201180162	2021-06-06T17:56:23.000Z	@JustinTrudeau
3	1401528927466639360	2021-06-06T13:18:30.000Z	This is real #genocide.https://t.co/ZA
4	1401599908059267073	2021-06-06T18:00:33.000Z	@JustinTrudeau My hobby is a barbe
5	1401598610551181312	2021-06-06T17:55:23.000Z	@JustinTrudeau @JustinTrudeau you
6	1401597964896722945	2021-06-06T17:52:49.000Z	@capano_rosa @JustinTrudeau @Jon
7	1401597739339718659	2021-06-06T17:51:56.000Z	@JustinTrudeau @marcomendicino h
8	1401597661833142275	2021-06-06T17:51:37.000Z	@capano_rosa @JustinTrudeau @Jon
9	1401596728000892933	2021-06-06T17:47:55.000Z	@JustinTrudeau Open the fucking co
10	1401591680587812865	2021-06-06T17:27:51.000Z	@JustinTrudeau Everyone is pretty ti
11	1401589438581415936	2021-06-06T17:18:57.000Z	@JustinTrudeau Yeah maybe... but Ca
12	1401589398773374983	2021-06-06T17:18:47.000Z	@JustinTrudeau Science is a process
13	1401586720320090117	2021-06-06T17:08:09.000Z	@JustinTrudeau how many of the 2 b
14	1401586531563827200	2021-06-06T17:07:24.000Z	@JustinTrudeau Shame on you and yo
15	1401579144132108290	2021-06-06T16:38:02.000Z	@JustinTrudeau NDP & Conserv
16	1401578885427482624	2021-06-06T16:37:01.000Z	@JustinTrudeau Taxing Cdns to death
17	1401577177846538245	2021-06-06T16:30:13.000Z	@JustinTrudeau Thank you PM Trude
18	1401574285643157504	2021-06-06T16:18:44.000Z	@JustinTrudeau https://t.co/na4uAR
19	1401573815537176576	2021-06-06T16:16:52.000Z	@JustinTrudeau https://t.co/6ez0K7t
20	1401573306952728577	2021-06-06T16:14:51.000Z	@JustinTrudeau This is a great effort
21	1401572148536221705	2021-06-06T16:10:14.000Z	@JustinTrudeau It's a humble reques
22	1401571865877913606	2021-06-06T16:09:07.000Z	@JustinTrudeau It's a humble reques
23	1401571804200673281	2021-06-06T16:08:52.000Z	@JustinTrudeau @marcomendicino I
24	1401571047640555528	2021-06-06T16:05:52.000Z	@JustinTrudeau Those trees were to
25	1401568121698373641	2021-06-06T15:54:14.000Z	@JustinTrudeau So Happyyyy dearrrr

We only shared **tweet_ids** + **Toxicity Scores**
+ **tweet statuses** (as checked using the **twarc** library)

1. Twitter API → **2. Collection** → **3. Metadata** → **4. Export** → **5. Sharing**

Today's agenda

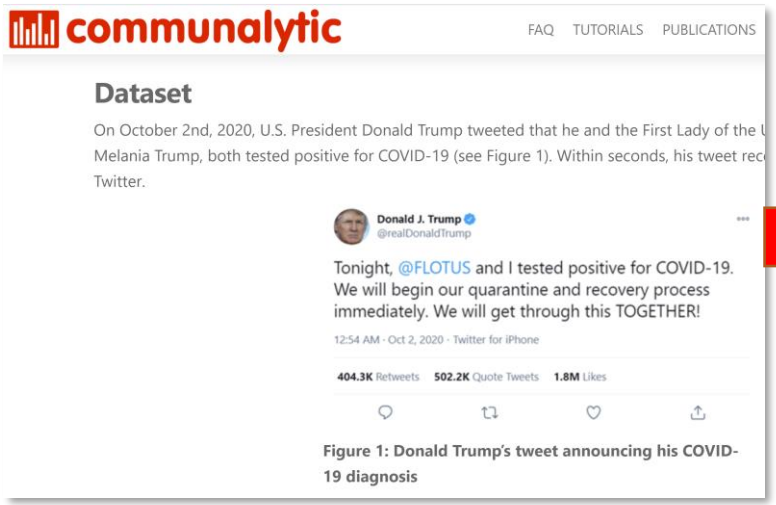
Guest talk by Dr. Paola Pascual-Ferra

Brief recap of how to collect Twitter data and run the Toxicity Analysis in Communalytic

Analyzing Twitter data using Jupyter Notebook

with Shahnawaz Attarwala

Sample Dataset of Tweet IDs



https://communalytic.com/2021/01/13/toxicity-analysis-of-a-twitter-thread/

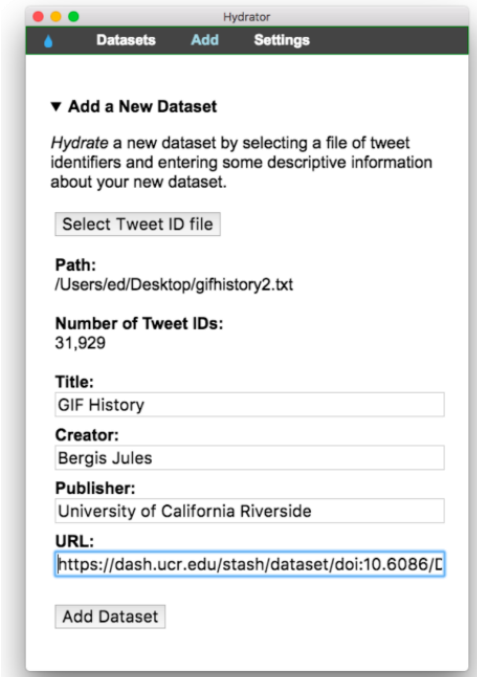
https://doi.org/10.5683/SP2/BMCGI3

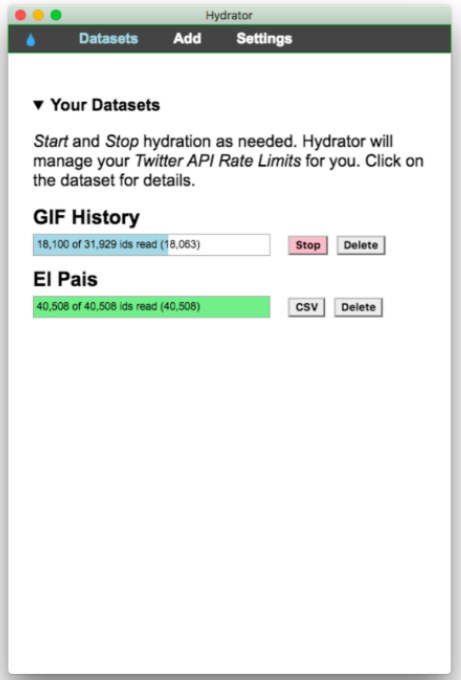
You can recollect tweets using programs like Hydrator

tweet_id
1401598862201180162
1401528927466639360
1401599908059267073
1401598610551181312
1401597964896722945
1401597739339718659
1401597661833142275
1401596728000892933
1401591680587812865
1401589438581415936
1401589398773374983
1401586720320090117
1401586531563827200
1401579144132108290
1401578885427482624

Hydrator

https://github.com/DocNow/hydrator





Why use Python for Data Science projects..

Source: https://www.nobledesktop.com/learn/python/python-vs-excel

Benefits of using Jupyter Notebook

- Jupyter notebook is an Integrated Development Environment (IDE) that acts as an interface between user and python.
- It is an open-source project.
- Supports multiple programming languages including Python and R.
- It allows users to create and share code and documents easily.
- It supports code, text, and interactive data visualization.
- Support single code cell execution, independent of other code cells

Source: Jupyter logo

```
1 #Display brief info about the dataset
2 pdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 298173 entries, 0 to 298172
Data columns (total 11 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   tweet_id           298173 non-null  int64
 1   tweetstatus        298173 non-null  object
 2   language           298173 non-null  object
 3   Toxicity           298173 non-null  float64
 4   Severe Toxicity    298173 non-null  float64
 5   Profane            298173 non-null  float64
 6   Identity Attack    298173 non-null  float64
 7   Insult             298173 non-null  float64
 8   Threat             298173 non-null  float64
 9   Sexually Explicit  298173 non-null  float64
 10  Flirtation         298173 non-null  float64
dtypes: float64(8), int64(1), object(2)
memory usage: 25.0+ MB
```

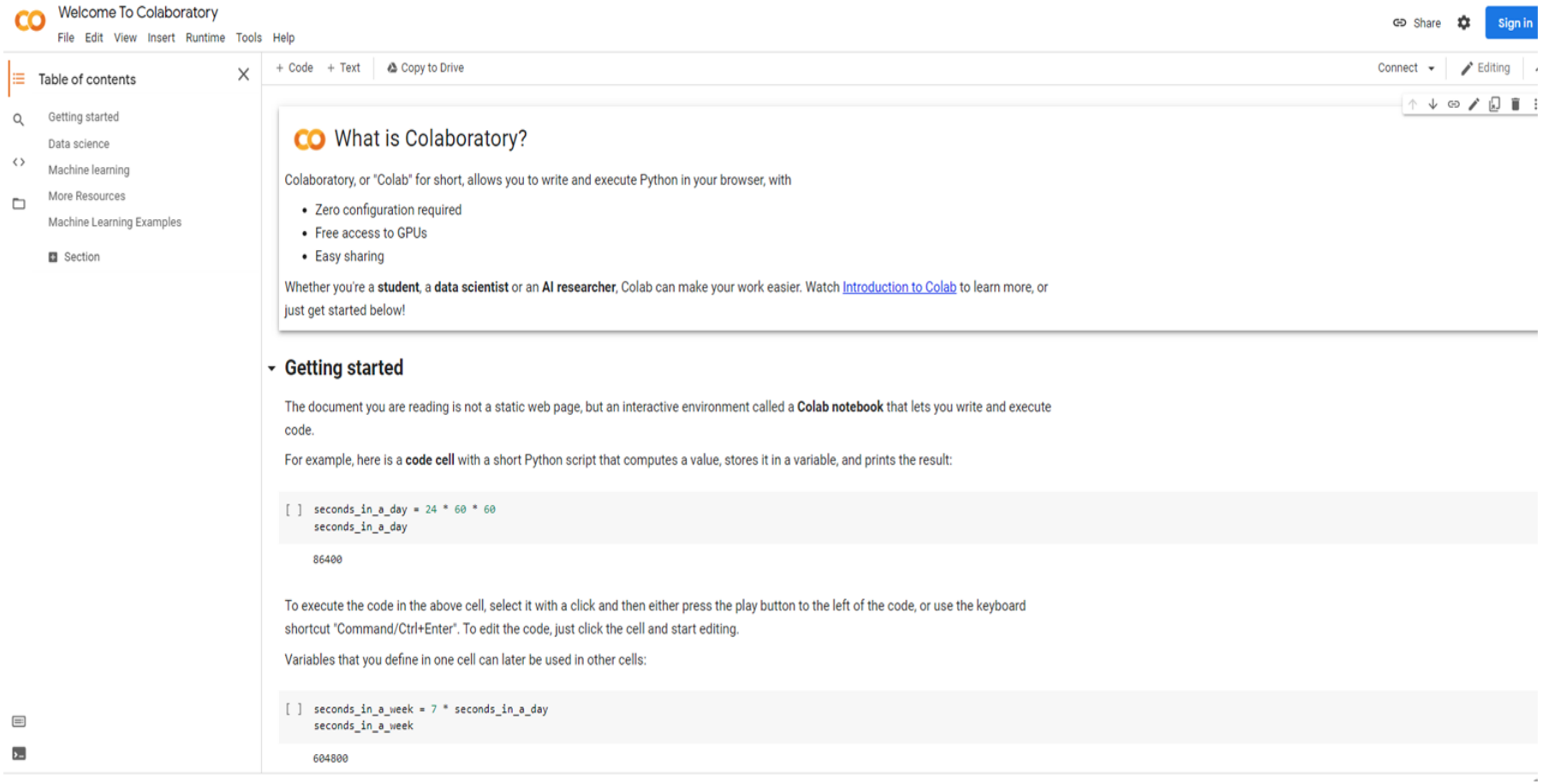
```
[5]  1 #How many different language used in the tweets?
     2 len(pdf["language"].unique())
```

```
62
```

Google’s free version of Jupyter Notebook

Benefits of using Google Colab’s Jupyter Notebook:

- Pre-Installed Libraries
- Saved on the Cloud
- Collaboration
- Free RAM and Storage



Sample Google colab's Jupyter Notebook

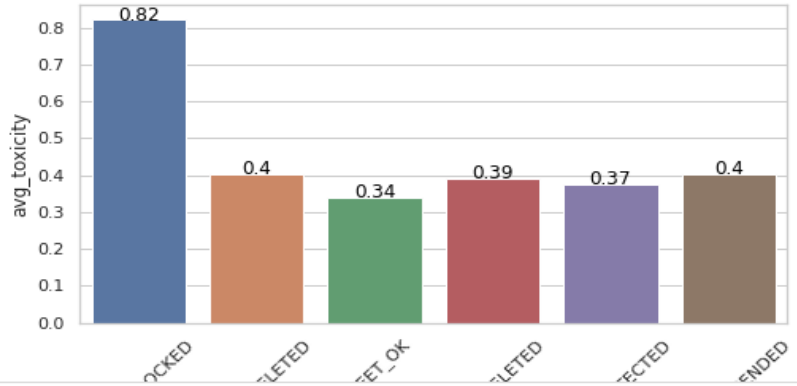
The complete Jupyter Notebook of Toxicity Analysis is at this link.

SMLab-TwitterEDA-June2021.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
1 #Average toxicity based on the tweet status
2 pdf_en_avgToxicity_vs_status = pdf_en.groupby(pdf_en['tweetstatus'])['Toxicity'].agg("mean").reset_index(name="avg_toxicity")
3 pdf_en_avgToxicity_vs_status
```

	tweetstatus	avg_toxicity
0	TWEET_BLOCKED	0.819918
1	TWEET_DELETED	0.403394
2	TWEET_OK	0.340719
3	USER_DELETED	0.388920
4	USER_PROTECTED	0.373923
5	USER_SUSPENDED	0.401506

[16]
```
1 #Bar graph for average toxicity versus tweet status
2 sns.set_theme(style="whitegrid")
3 plt.figure(figsize=(8,4))
4 plt.xticks(rotation=45)
5 ax2 = sns.barplot(x="tweetstatus", y="avg_toxicity", data=pdf_en_avgToxicity_vs_status)
6 for index, row in pdf_en_avgToxicity_vs_status.iterrows():
7     ax2.text(row.name,row.avg_toxicity, round(row.avg_toxicity,2), color='black', ha="center")
```