DH500 Computational Social Media

Twitter Data Collection Tutorial

Lakmal Meegahapola

idiap RESEARCH INSTITUTE EPFL

Outline

1. Twitter API Basics

- API (Search vs. Realtime)
- Rate Limits and TOS
- Authentication
- Data Access

2. Inside the Data

- User
- Tweets

3. Demo

1. Twitter API Basics

Twitter API

- API: Application Programming Interface
- Search API vs. Realtime API

	Search API	Realtime API
Tweet Results	Recent and archived tweets (e.g., last week)	Realtime tweets (e.g., right now)
Queries	Using multiple filters to fill tweets: keywords (e.g., covid2019), location (e.g., bounding box), etc.	
Categories	Standard (Free), Premium (Pay), Enterprise (Pay)	
Interface Examples	Tweet Lookup, Timelines, Search Tweets	Filtered stream, Sampled Stream

Twitter API Examples

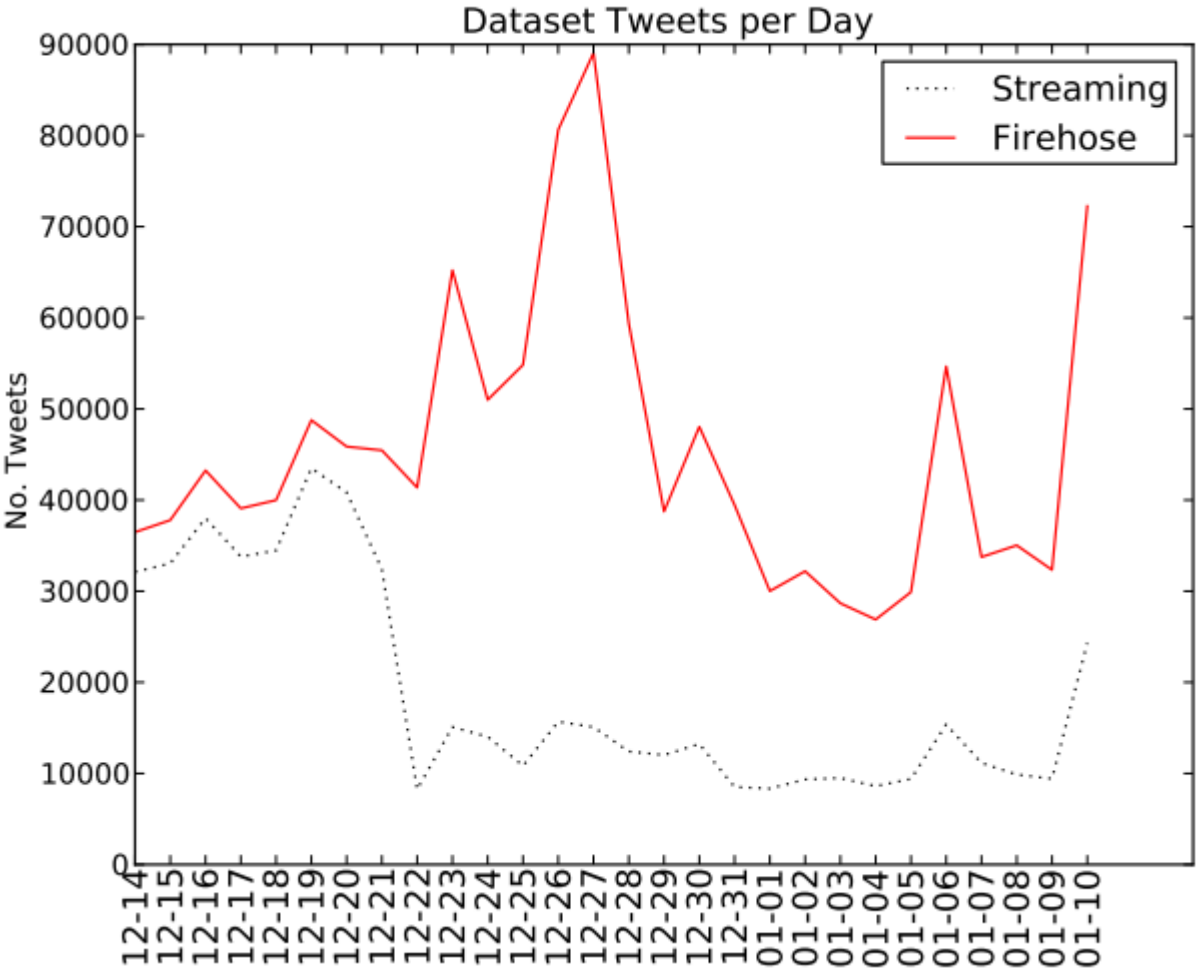
Group	API	Description
Search API	Tweet Lookup	Tweet ID → Tweet data
	Search Tweets	Search recent and archived tweets (e.g., last week)
	Timelines	User ID → Return user tweet timeline
Realtime API	Filtered Stream	Using multiple filters to fill tweets: keywords (e.g., covid19), location (e.g., bounding box), etc.
	Sampled Stream	Return 1% random tweets in real-time

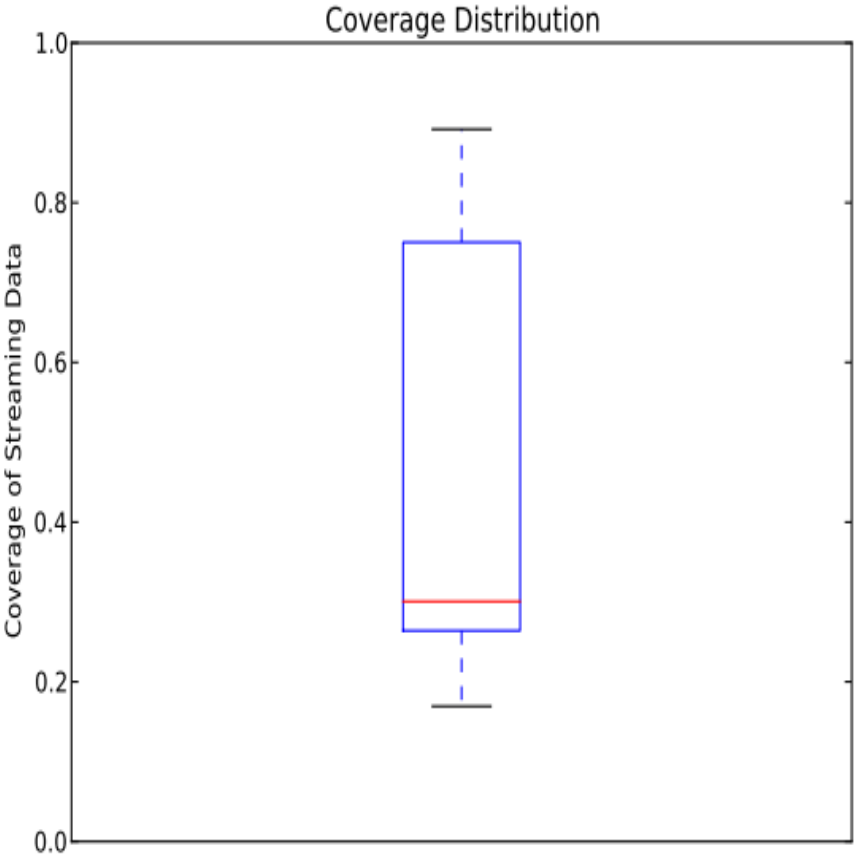
Rate Limits and Bias

- This applies for all categories (Standard, Premium, Enterprise)
- Imposes clear rate limits for both Realtime API and Search API
- Around 1% for Sampled Stream
- Enterprise category
  - 100% of all public tweets
  - High cost and Computing resources

Source: https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction

Sample Bias





Streaming API was previous name for **Sampled Stream**
Firehose was previous name for **Enterprise Category**

F Morstatter,et al. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. International AAAI Conference on Web and Social Media, 2013.

Terms of Service (ToS)
One for Outside and One for Inside
the EU, EFTA States, the UK, and the USA

Redistribution of Twitter content

If you need to share Twitter content you obtained via the Twitter APIs with another party, the best way to do so is by sharing Tweet IDs, Direct Message IDs, and/or User IDs, which the end user of the content can then rehydrate (i.e. request the full Tweet, user, or Direct Message content) using the Twitter APIs. This helps ensure that end users of Twitter content always get the most current information directly from us.

We permit limited redistribution of hydrated Twitter content via non-automated means. If you choose to share hydrated Twitter content with another party in this way, you may only share up to 50,000 hydrated public Tweet Objects and/or User Objects per recipient, per day, and should not make this data publicly available (for example, as an attachment to a blog post or in a public Github repository).

There are a few other points to keep in mind about redistributing Twitter content:

- You may only distribute up to a total of 1,500,000 Tweet IDs to a single entity within a 30 day period unless you've received prior express written permission from Twitter.
- Individuals redistributing Tweet IDs and/or User IDs on behalf of an academic institution for the sole purpose of non-commercial research are permitted to redistribute an unlimited number of Tweet IDs and/or User IDs.
- To request permission to share Twitter content as outlined above, please use the API Policy support form.

To the extent you are permitted to distribute Twitter content to a third party, note that this content remains subject to the Developer Agreement and Policy, and you must ensure such third party has agreed to the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy before receiving Twitter content.

Source: https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Authentication

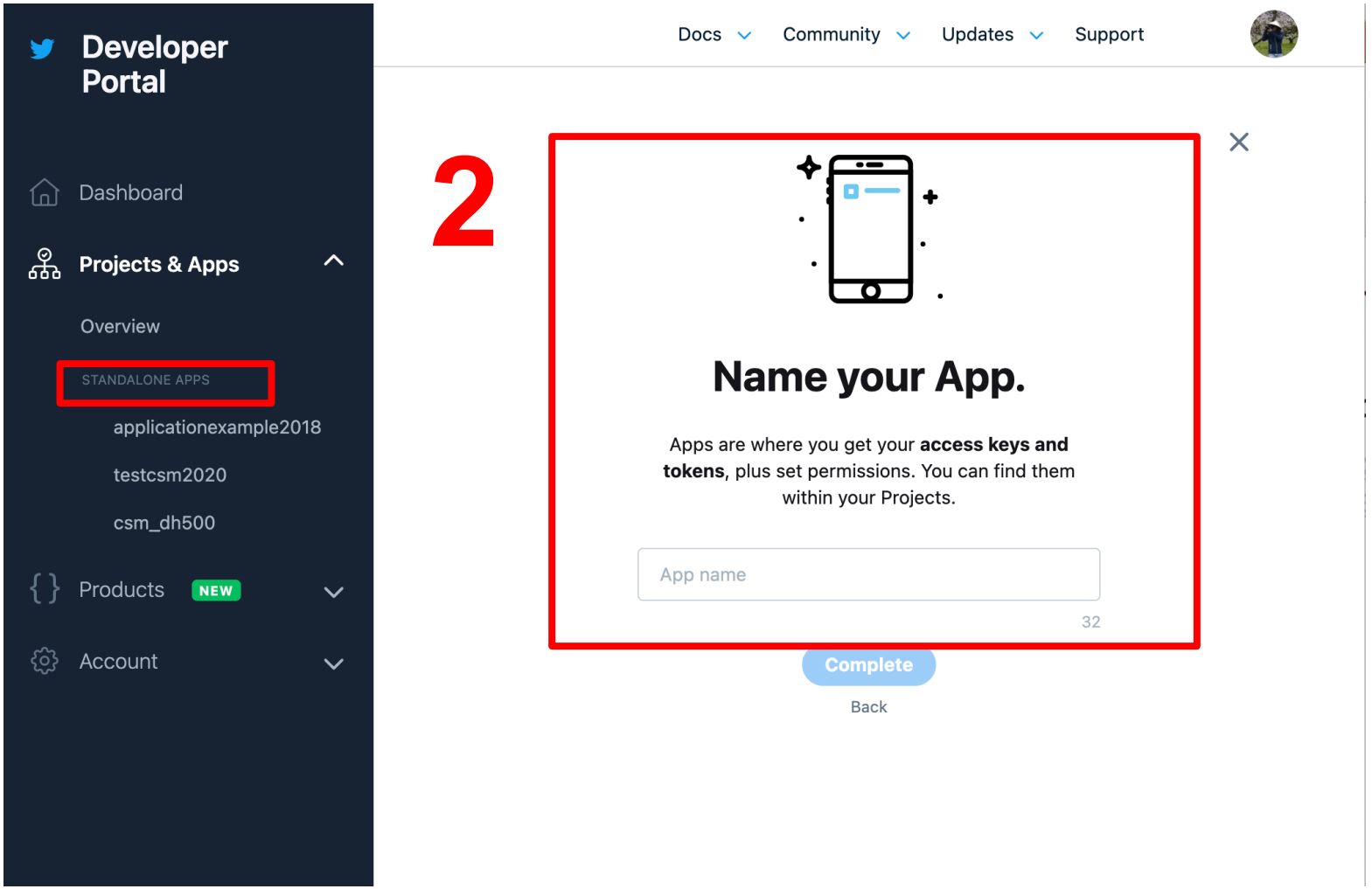
- App registration at
  - Oauth Authentication Protocol (Only for Standard)
  - https://developer.twitter.com/en/portal/apps/new
- Application level authentication
  - Consumer API Key
  - Consumer API Secret Key
- User level authentication
  - Access Token
  - Access Token Secret

Source: https://developer.twitter.com/en/docs/authentication/overview

Authentication

Login to your Twitter account and go to the link

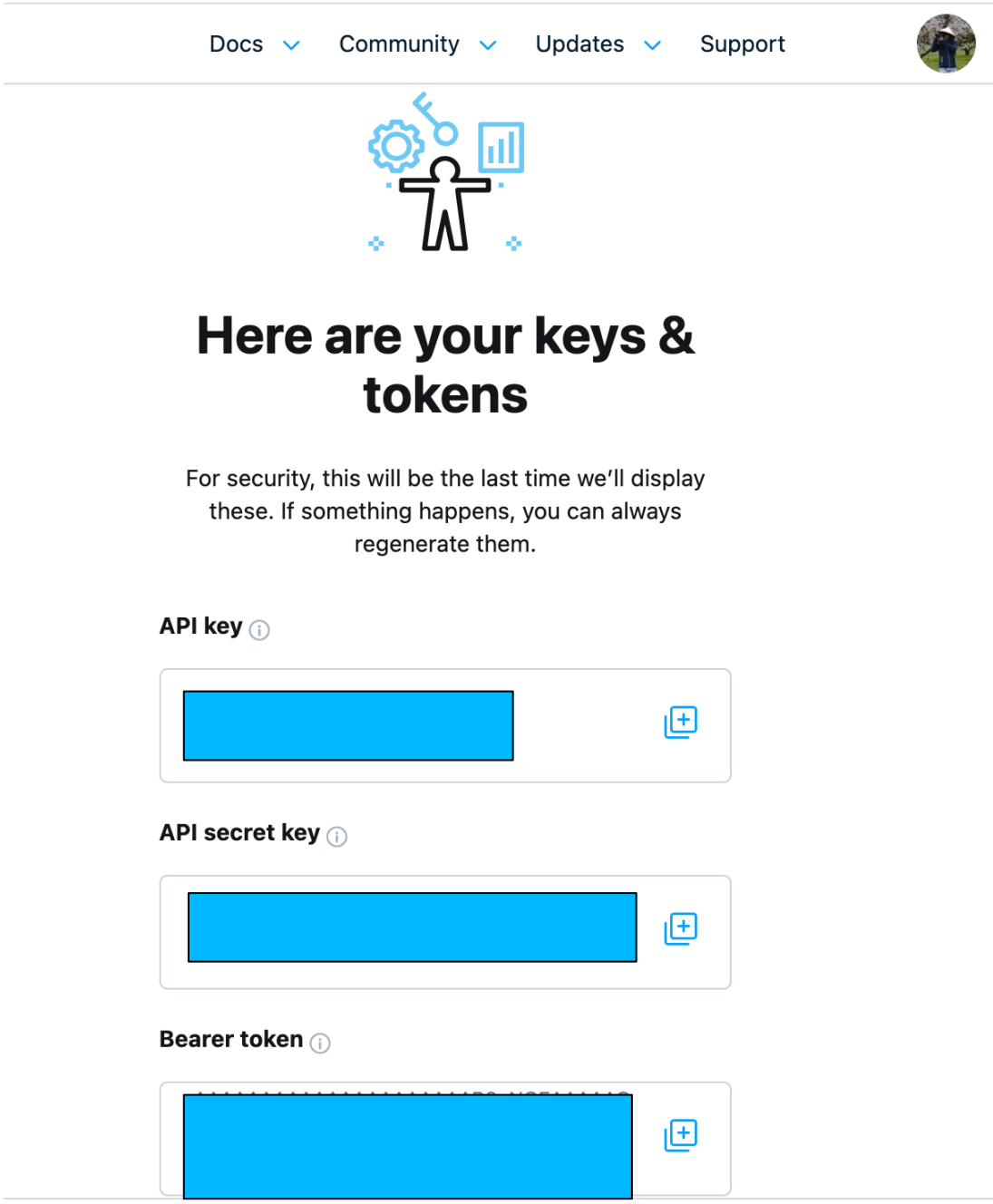
https://developer.twitter.com/en/docs/basics/authentication/oauth-1-0a/obtaining-user-access-tokens



Authentication

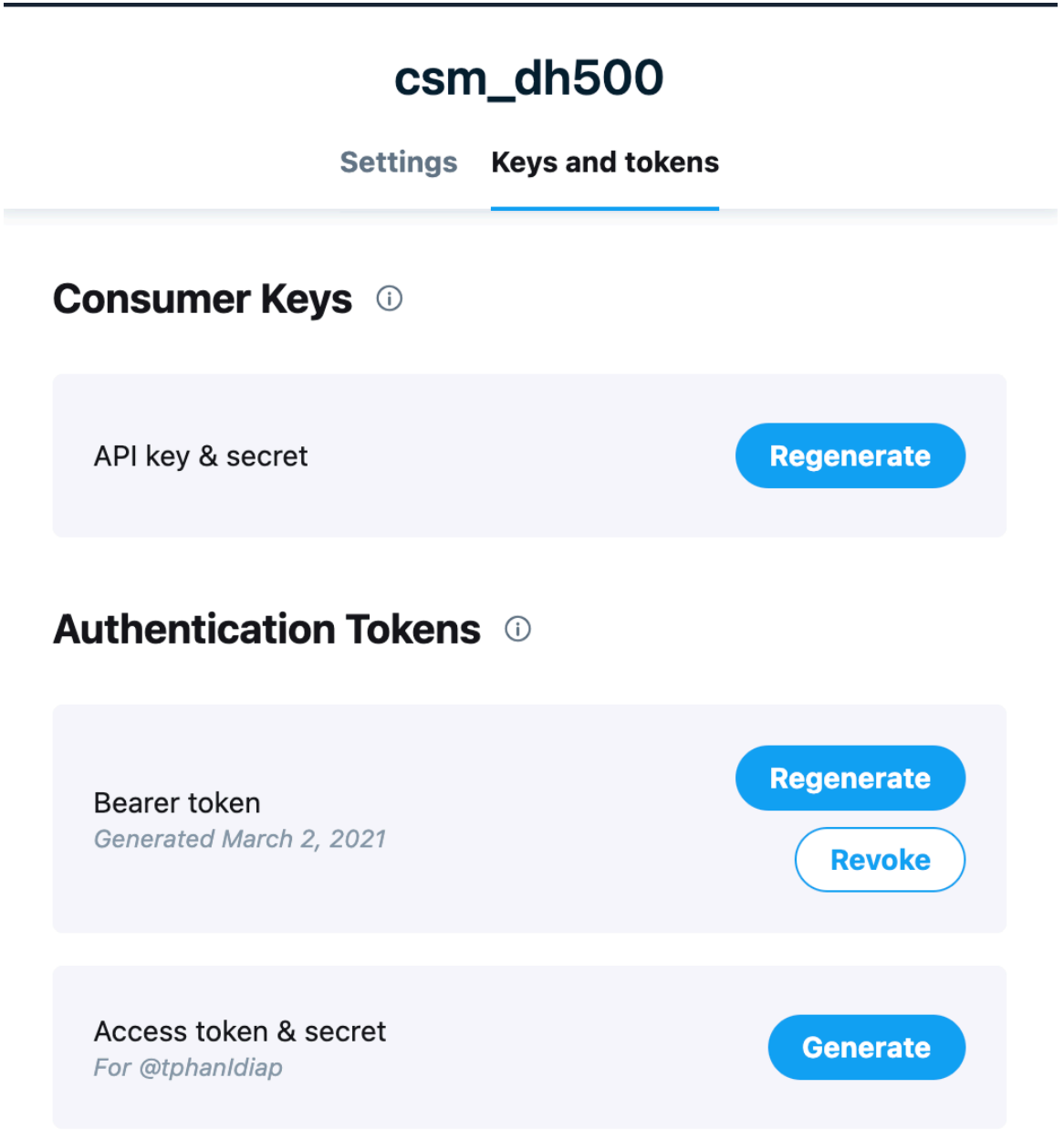
3

We can create application after clicking the button Complete



Authentication

4



Filtered Stream Request Parameters

- locations
  - comma-separated list of {longitude, latitude} pairs specifying a set of bounding boxes to filter tweets
  - southwest corner of the bounding box comes first
  - bounding boxes do not act as filters for other filter parameters

Source: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters

Filtered Stream Request Parameters

- track
  - comma-separated list of keywords incl. hashtags
  - a keyword may be one or more terms
  - keywords' order does not matter

- More request parameters: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters

Source: https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters

2. Inside the Data

Response Object (JSON)

- JSON, or JavaScript Object Notation
  - Key-value pairs
  - Human-readable format
  - Platform independent

- Twitter-specific
  - Attributes of a JSON-encoded object are unordered
  - Unexpected or missing fields

Home
Explore
Notifications
Messages
Bookmarks
Lists
Profile
More
Tweet

World Health Organization (WHO)
59.3K Tweets

Follow

World Health Organization (WHO)
@WHO

We are the #UnitedNations' health agency - #HealthForAll.
▶ Always check our latest tweets on #COVID19 for updated advice/information.

Geneva, Switzerland who.int Joined April 2008

1,732 Following **9.1M** Followers

Not followed by anyone you're following

Tweets Tweets & replies Media Likes

Pinned Tweet

World Health Organization (WHO) @WHO · 1h
LIVE Q&A ahead of #WorldHearingDay. Ask your questions about #HearingCare. #AskWHO

Search Twitter

You might like

EDHEC_BSchool @EDHEC_BSchool — Follow
Promoted

CNN @CNN — Follow

NASA @NASA — Follow

Show more

Trends for you

Tweet Attributes

created_at
id
id_str
text
source
truncated
in_reply_to_status_id
in_reply_to_status_id_str
in_reply_to_user_id
in_reply_to_user_id_str
in_reply_to_screen_name
user
geo
coordinates
place
contributors
quoted_status_id
quoted_status_id_str
quoted_status
retweeted_status
quoted_status_permalink
is_quote_status
quote_count
reply_count
retweet_count
favorite_count
entities
favorited
retweeted
filter_level
lang
timestamp_ms

- Entities
  - Urls
  - User Mentions (@)
  - Hashtags (#)
- Place is location with geo coordinates
- Tweets associated with places are not necessarily issued from that location

Source: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

User Attributes

id
id_str
name
screen_name
location
url
description
translator_type
protected
verified
followers_count
friends_count
listed_count
favourites_count
statuses_count
created_at
utc_offset
time_zone
geo_enabled
lang
contributors_enabled
is_translator
profile_background_color
profile_background_image_url
profile_background_image_url_https
profile_background_tile
profile_link_color
profile_sidebar_border_color
profile_sidebar_fill_color
profile_text_color
profile_use_background_image
profile_image_url
profile_image_url_https
profile_banner_url
default_profile
default_profile_image
following
follow_request_sent
notifications

Source: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

1. Retweet

2. Retweet with comments

1. **'retweeted_status'**
2. **'quoted_status'**

Practical Notes

- Pay attention to rate limit
  - https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits
  - Be nice to Twitter !!!
- Twitter libraries for downloading tweets
  - https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries
- MongoDB, SQL, or CSV for storing JSON tweets.
  - https://www.mongodb.com/python

3. Demo

Getting ready to download Twitter data

- Create Twitter account in case you do not have one.
- Follow the steps shown in Slides 9 to 12 to create your own Twitter authentication.
  - Consumer API Key
  - Consumer API Secret Key
  - Access Token
  - Access Token Secret
- This information will be mandatory to collect data for assignment 2. **PLEASE DO IT NOW.**
- In case you have problems, contact Lakmal.

email:
lakmal.meegahapola@epfl.ch