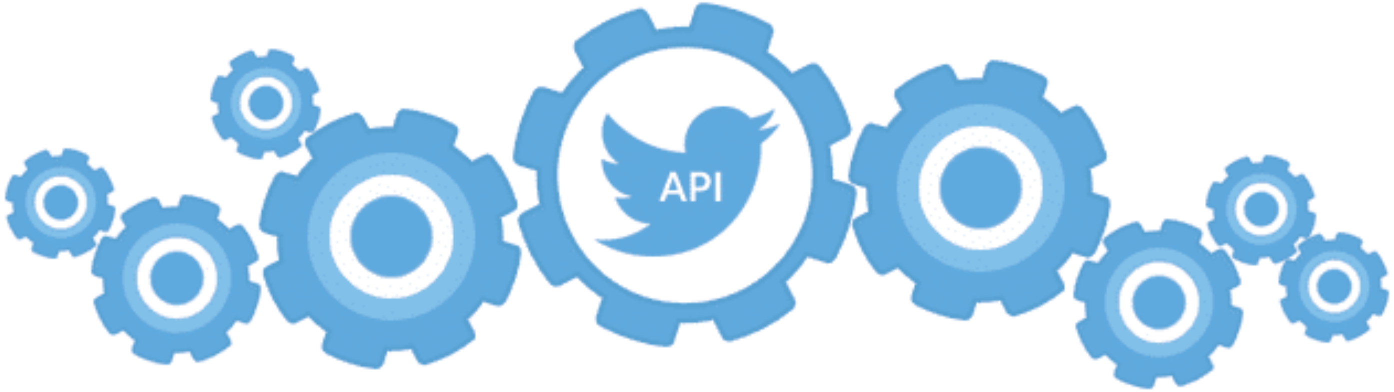
Extracting Twitter data by API



MIE1624 Group 6

Chloe Qi

Max Ng

Jiaying Zhu

Yingzheng Ma

Yirui Chen

Zheng Gong

Objectives

01

Using API to extract timeline and stream twitter data

02

Searching tweets by hash tags

03

Extracting metadata such as a number of retweets

04

Discussing the limitations of Twitter APIs and its alternatives

Background

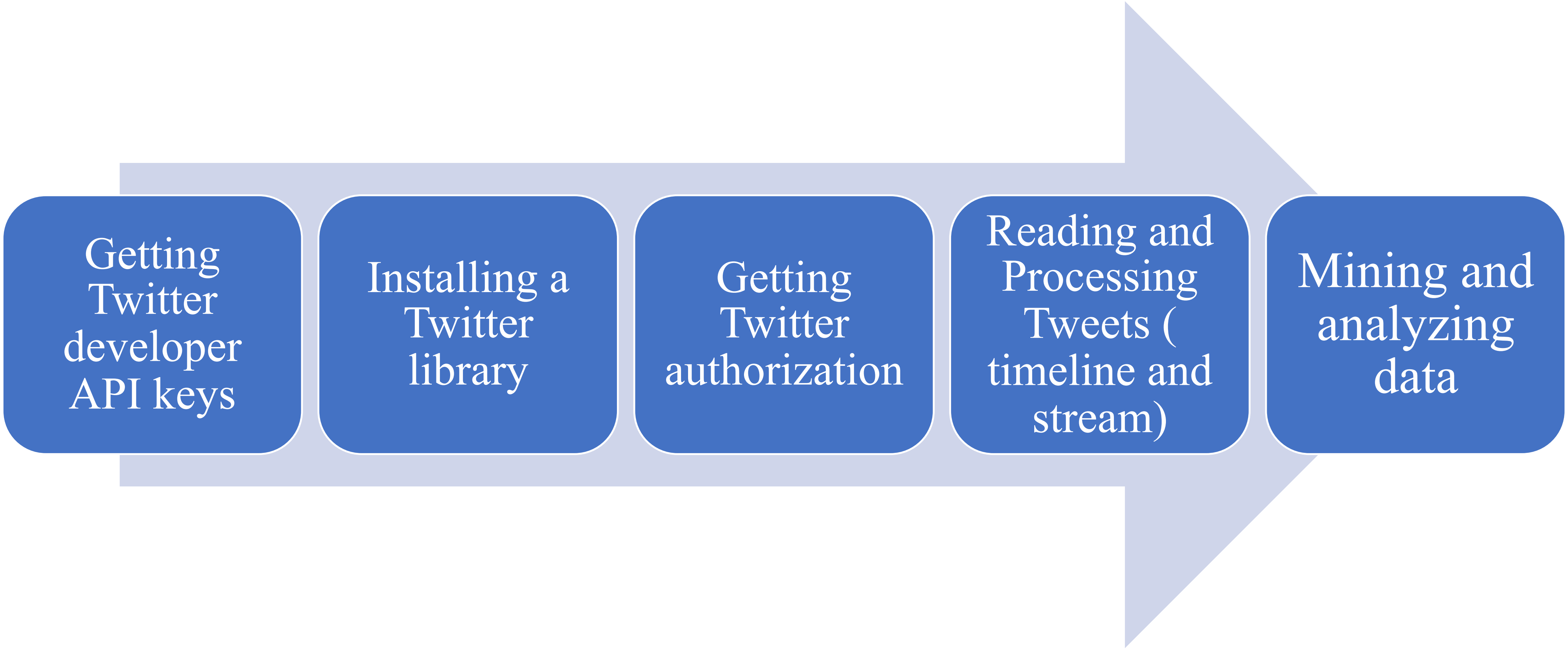
- API(application programming interfaces): using endpoints to link specific types of information
- Twitter API platform provides broad access to public Twitter data and users' own non-public information under authorization
- The Twitter APIs can be looked up by

https://developer.twitter.com/en/docs/api-reference-index

There are many possibilities with the Twitter APIs

- **Tweets**: searching, posting, filtering, engagement, streaming etc.
- **Ads:** campaign and audience management, analytics.
- **Direct messages**: sending and receiving, direct replies, welcome messages etc.
- **Accounts and users** :account management, user interactions.
- **Media**: uploading and accessing photos, videos and animated GIFs.
- **Trends**: trending topics in a given location.
- **Geo**: information about known places or places near a location.

Important Steps of Getting Data



Getting Twitter developer API keys

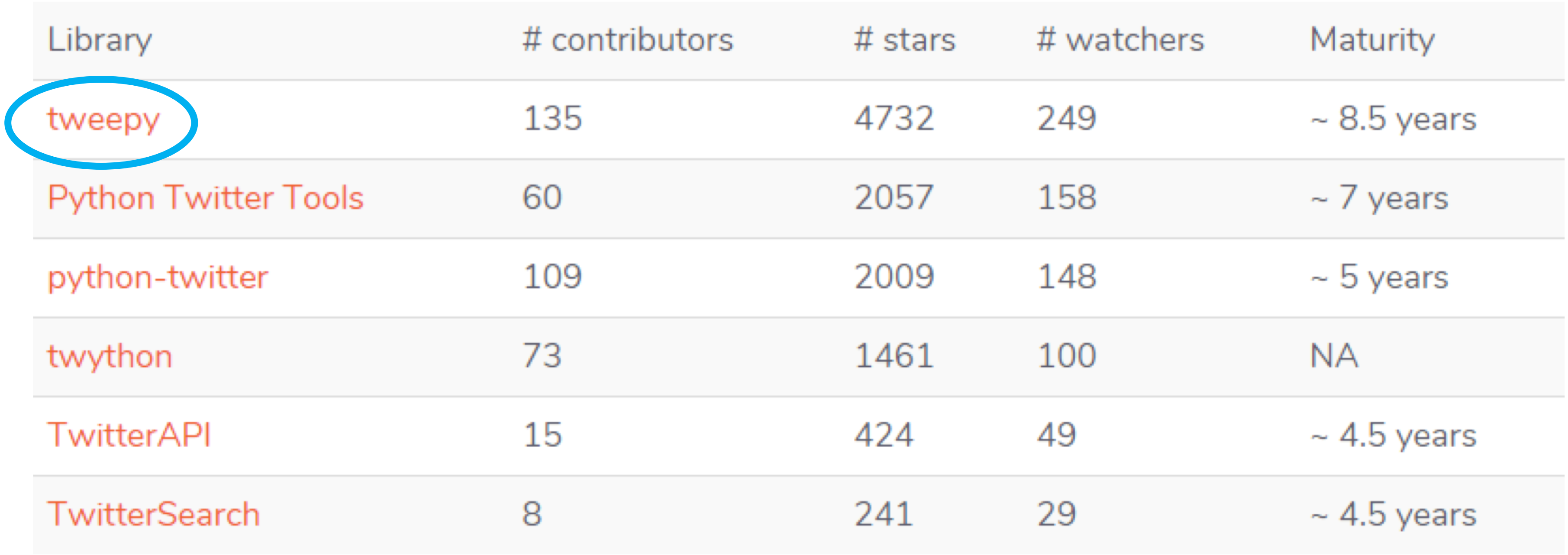
Installing a Twitter library

Getting Twitter authorization

Reading and Processing Tweets (timeline and stream)

Mining and analyzing data

Accessing the Twitter API with Python



Library	# contributors	# stars	# watchers	Maturity
tweepy	135	4732	249	~ 8.5 years
Python Twitter Tools	60	2057	158	~ 7 years
python-twitter	109	2009	148	~ 5 years
twython	73	1461	100	NA
TwitterAPI	15	424	49	~ 4.5 years
TwitterSearch	8	241	29	~ 4.5 years

Tweepy is the most popular **open-source** library for python.

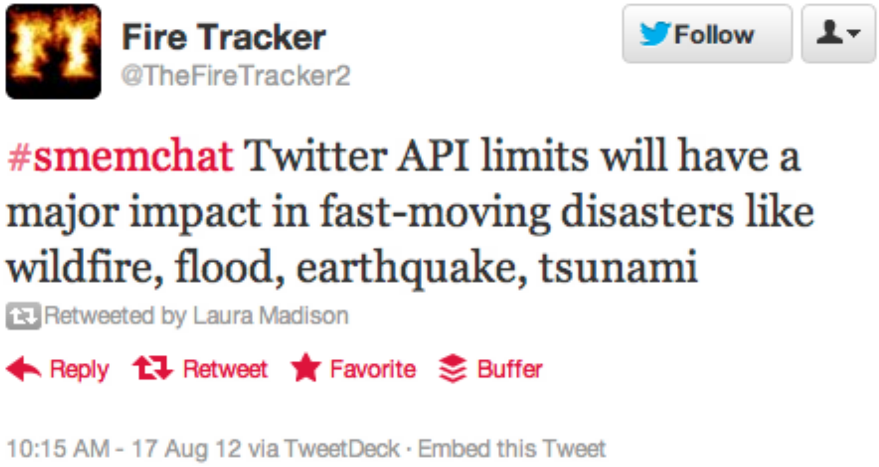
Limitation of Twitter API

- Rate limit
  Standard API: 30 requests/min
  Premium API: 60 requests/min

- Costly

- Too many connection tries are not allowed as well and will result in a user being blocked.

- Biased data
  1% ~ 40% of the entire volume of all tweets sent within a particular interval will be provided.
  Not choose randomly! Not a uniform sample!



Fire Tracker
@TheFireTracker2

Follow

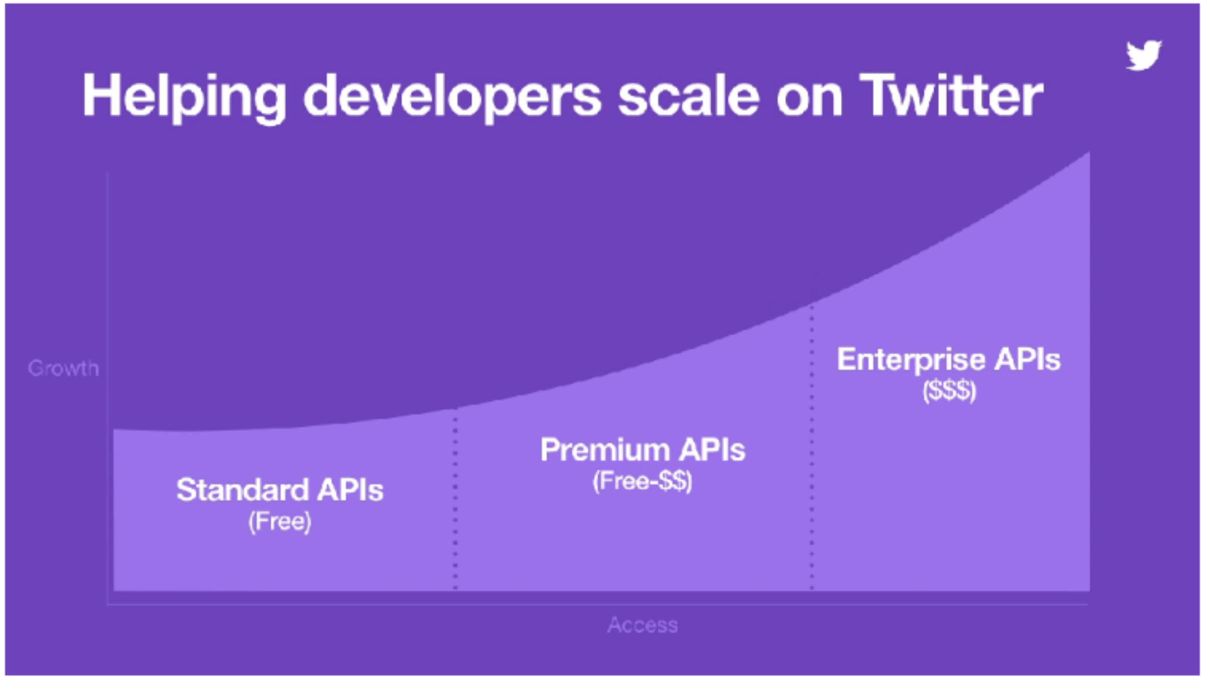
#smemchat Twitter API limits will have a major impact in fast-moving disasters like wildfire, flood, earthquake, tsunami

Retweeted by Laura Madison

Reply Retweet Favorite Buffer

10:15 AM - 17 Aug 12 via TweetDeck · Embed this Tweet

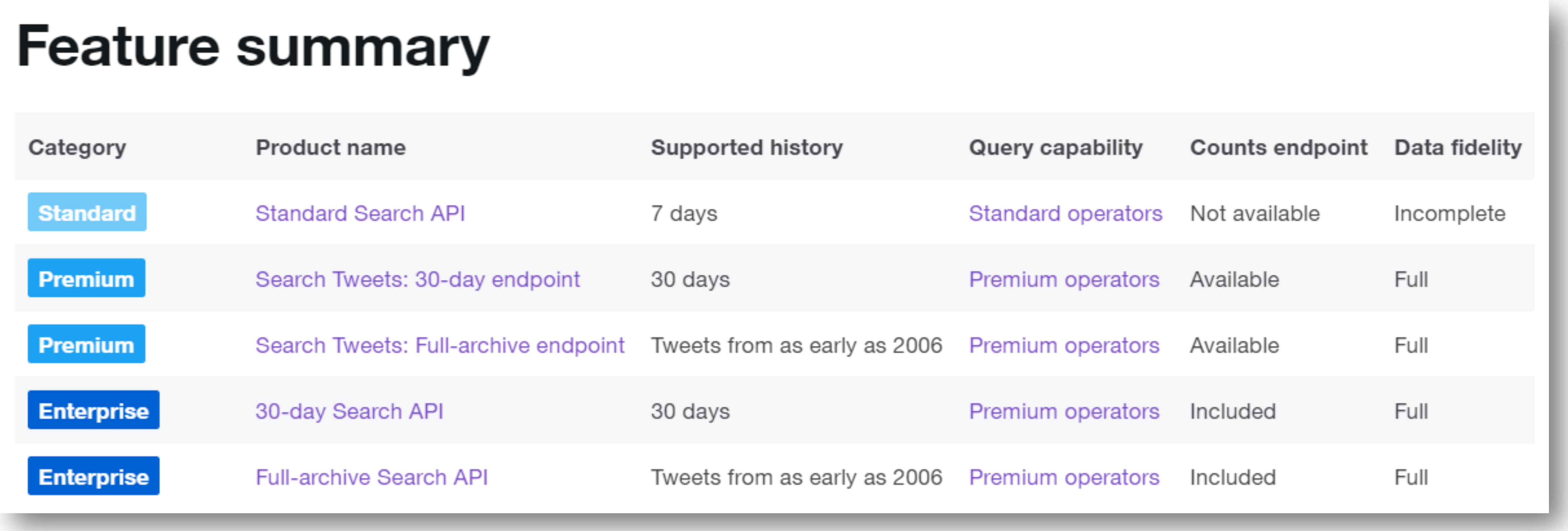
Ref [1]



Ref [2]

Limitation of Twitter API

- Take Search API as an example



Feature summary

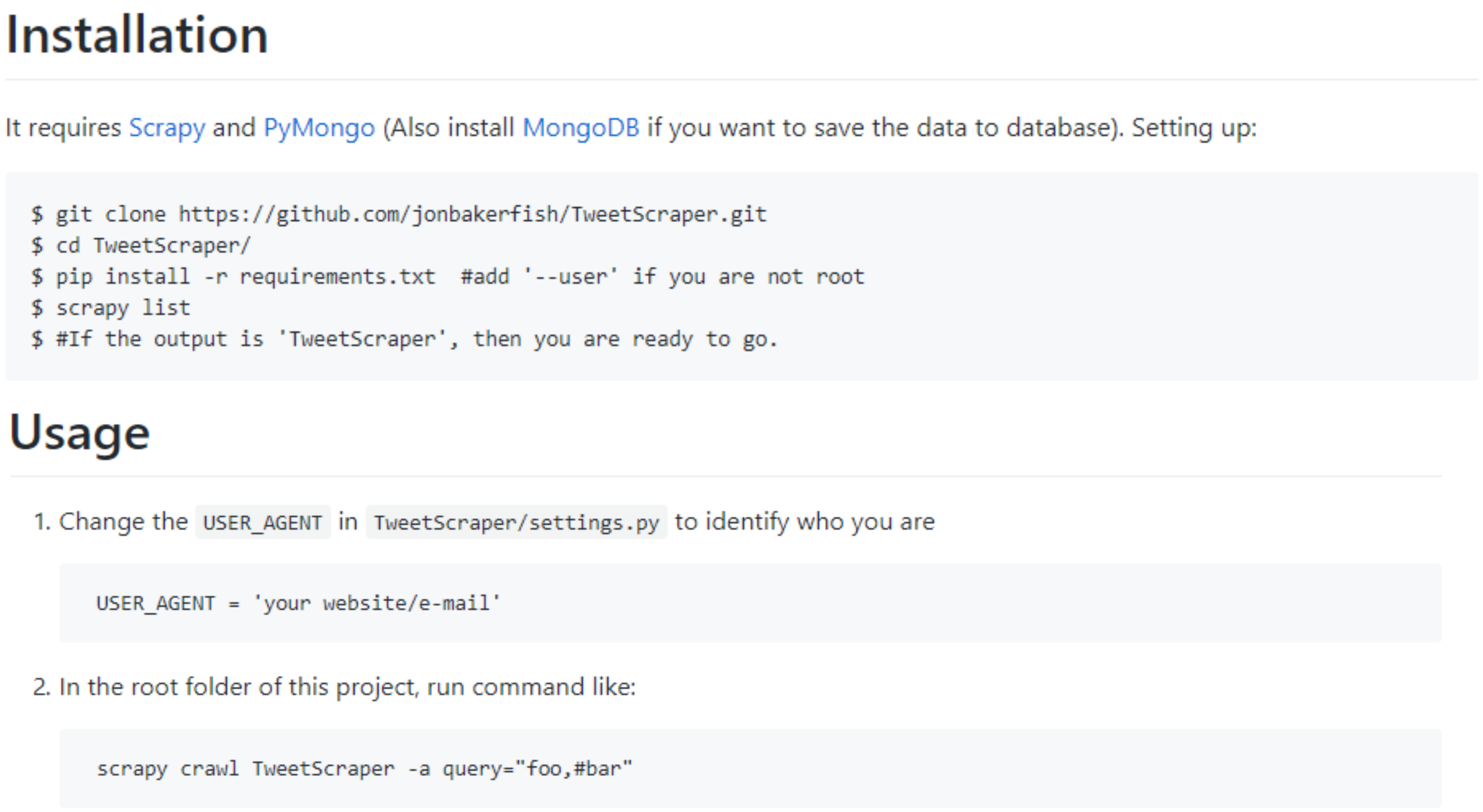
Category	Product name	Supported history	Query capability	Counts endpoint	Data fidelity
Standard	Standard Search API	7 days	Standard operators	Not available	Incomplete
Premium	Search Tweets: 30-day endpoint	30 days	Premium operators	Available	Full
Premium	Search Tweets: Full-archive endpoint	Tweets from as early as 2006	Premium operators	Available	Full
Enterprise	30-day Search API	30 days	Premium operators	Included	Full
Enterprise	Full-archive Search API	Tweets from as early as 2006	Premium operators	Included	Full

Ref [3]

Alternative Ways of Extracting Data

- **Find an existing Twitter dataset**
- **Purchase from Twitter (Historical Power Track enterprise)**
- **Access or purchase from a Twitter service provider**
- **Scrape tweets without using the API by python**

Scrape tweets without using the API by python



Installation

It requires Scrapy and PyMongo (Also install MongoDB if you want to save the data to database). Setting up:

```
$ git clone https://github.com/jonbakerfish/TweetScraper.git
$ cd TweetScraper/
$ pip install -r requirements.txt  #add '--user' if you are not root
$ scrapy list
$ #If the output is 'TweetScraper', then you are ready to go.
```

Usage

1. Change the `USER_AGENT` in `TweetScraper/settings.py` to identify who you are

   ```
   USER_AGENT = 'your website/e-mail'
   ```

2. In the root folder of this project, run command like:

   ```
   scrapy crawl TweetScraper -a query="foo,#bar"
   ```

References

- https://developer.twitter.com/en/docs/basics/rate-limiting.html
- https://stackabuse.com/accessing-the-twitter-api-with-python/
- https://developer.twitter.com/en/docs/tweets/search/overview
- https://twittercommunity.com/t/what-is-the-rate-limit-for-the-premium-search-api/97314
- https://pdfs.semanticscholar.org/df22/cca197ea12fc2cd5c30b4647802fd6c37536.pdf
- https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/
- http://support.gnip.com/articles/consuming-parsing-and-processing-tweets-with-python.html
- http://docs.tweepy.org/en/v3.5.0/api.html#tweepy-api-twitter-api-wrapper
- https://blog.hartleybrody.com/web-scraping/
- https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data
- http://www.simonlindgren.com/stuff/2017/11/7/scrape-tweets-without-using-the-api