Twitter Scraper Tool Specifications Document

# Purpose

Twitter posts are a treasure trove of valuable and insightful information about a large number of subjects related to social good, discussion of social problems, demographics, etc. Any research that we do at Social Impact Analytics Institute that does not take in this social media platform’s content increases the possibility of us falling short in our analyses. And in our tool development R&D phase, it would behoove us to spend time developing this data acquisition tool, rather than wait until we need the tool and we have to scramble to put something together.

We want to maximize our R&D time right now because we are not under pressure at the moment. We have the luxury right now to build quality tools for future use. As Abraham Lincoln put it, we have time to “sharpen our axes”.

# What currently exists

There are several Twitter scraper libraries already, but none of them have the full functionality that we want in a Twitter scraper.

* Twitter-scraper 0.4.4 – [https://pypi.org/project/twitter-scraper/]
* Tweepy – [https://www.tweepy.org/]
* Twint – [https://medium.com/towards-artificial-intelligence/how-to-scrape-tweets-without-twitters-api-using-twint-797b196b951c]

# Input parameters

* Twitter account name or hashtag subject
* Inclusive scraping dates
* Number of tweets to download up to “all”

# Output fields and formats

## Output Fields

| Key | Type | Description |
| --- | --- | --- |
| tweetId | string | Tweet's identifier, visit twitter.com/USERNAME/ID to view tweet. |
| userId | string | Tweet's userId |
| username | string | Tweet's username |
| tweetUrl | string | Tweet's URL |
| isRetweet | boolean | True if it is a retweet, False otherwise |
| isPinned | boolean | True if it is a pinned tweet, False otherwise |
| time | datetime | Published date of tweet |
| text | string | Content of tweet |
| replies | integer | Replies count of tweet |
| retweets | integer | Retweet count of tweet |
| likes | integer | Like count of tweet |
| entries | dictionary | Has hashtags, videos, photos, urls keys. Each one's value is list |

## Output Format

Output format will be any of the following, with a choice of all of the formats being the optimal functionality.

Output file name will reflect the search term (username or hashtag term) and date of scrape. For example, an output for a username search would be ‘benpmeredith2020-01-23.csv’ or for a hashtag search term would be ‘shooting2020-01-23.json’ (file extension appropriate).

* Csv
* Json