**COVID-19-TweetIDs**

The repository contains an ongoing collection of tweets IDs associated with the novel coronavirus COVID-19 (SARS-CoV-2), which commenced on January 28, 2020. We used the Twitter’s search API to gather historical Tweets from the preceding 7 days, leading to the first Tweets in our dataset dating back to January 21, 2020. We leveraged Twitter’s streaming API to follow specified accounts and also collect in real-time tweets that mention specific keywords. To comply with Twitter’s [Terms of Service](https://developer.twitter.com/en/developer-terms/agreement-and-policy), we are only publicly releasing the Tweet IDs of the collected Tweets. The data is released for non-commercial research use.

The associated paper to this repository can be found here: [#COVID-19: The First Public Coronavirus Twitter Dataset](https://arxiv.org/abs/2003.07372)

**Data Organization**

The Tweet-IDs are organized as follows:

* Tweet-ID files are stored in folders that indicate the year and month of the collection (YEAR-MONTH).
* Individual Tweet-ID files contain a collection of Tweet IDs, and the file names all follow the same structure, with a prefix “coronavirus-tweet-id-” followed by the YEAR-MONTH-DATE-HOUR.
* Note that Twitter returns Tweets in UTC, and thus all Tweet ID folders and file names are all in UTC as well.

**Notes About the Data**

A few notes about this data:

* We will be continuously maintaining this database for the foreseeable future, and will be uploading new data on a weekly basis.
* There may be a few hours of missing data due to technical difficulties. We have done our best to recover as many Tweets from those time frames by using Twitter’s search API.
* We will keep a running summary of basic statistics as we upload data in each new release.
* The file keywords.txt and accounts.txt contains the updated keywords and accounts respectively that we tracked in our data collection. Each keyword and account will be followed by the date we began tracking them, and date we removed them (if the keyword or account has been removed) from our tracking list.
* Consider using tools such as the [Hydrator](https://github.com/DocNow/hydrator) and [Twarc](https://github.com/DocNow/twarc) to rehydrate the Tweet IDs. Instructions for both are in the next section.
* Hydrating may take a while, and Tweets may have been deleted since our initial collection. If that is the case, unfortunately you will not be able to get the deleted Tweets from querying Twitter's API. Ed Summers ([edsu](https://github.com/edsu)) hydrated the Tweets in release v1.0, taking approximately 25 hours to complete, and found that there was an approximate 6% of the Tweets that were deleted at the time of hydration, with final gzipped data size of 6.9 GB.

**How to Hydrate**

**Hydrating using** [**Hydrator**](https://github.com/DocNow/hydrator) **(GUI)**

Navigate to the [Hydrator github repository](https://github.com/DocNow/hydrator) and follow the instructions for installation in their README. As there are a lot of separate Tweet ID files in this repository, it might be advisable to first merge files from timeframes of interest into a larger file before hydrating the Tweets through the GUI.

**Hydrating using** [**Twarc**](https://github.com/DocNow/twarc) **(CLI)**

Many thanks to Ed Summers ([edsu](https://github.com/edsu)) for writing this script that uses [Twarc](https://github.com/DocNow/twarc) to hydrate all Tweet-IDs stored in their corresponding folders.

First install Twarc and tqdm

pip3 install twarc

pip3 install tqdm

Configure Twarc with your Twitter API tokens (note you must [apply](https://developer.twitter.com/en/apply-for-access) for a Twitter developer account first in order to obtain the needed tokens). You can also configure the API tokens in the script, if unable to configure through CLI.

twarc configure

Run the script. The hydrated Tweets will be stored in the same folder as the Tweet-ID file, and is saved as a compressed jsonl file

python3 hydrate.py

**Data Usage Agreement**

This dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ([CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)). By using this dataset, you agree to abide by the stipulations in the license, remain in compliance with Twitter’s [Terms of Service](https://developer.twitter.com/en/developer-terms/agreement-and-policy), and cite the following manuscript:

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. #COVID-19: The First Public Coronavirus Twitter Dataset. arXiv:cs.SI/2003.07372, 2020

**Statistics Summary (v1.6)**

Number of Tweets : **109,013,655**

Language breakdown of top 10 most prevalent languages :

| **Language** | **ISO** | **No. tweets** | **% total Tweets** |
| --- | --- | --- | --- |
| English | en | 71,984,701 | 66.03% |
| Spanish | es | 12,149,916 | 11.15% |
| Indonesian | in | 3,826,448 | 3.51% |
| French | fr | 3,340,808 | 3.06% |
| Portuguese | pt | 2,928,843 | 2.69% |
| Thai | th | 2,630,420 | 2.41% |
| (undefined) | und | 2,327,240 | 2.13% |
| Japanese | ja | 2,156,385 | 1.98% |
| Italian | it | 1,484,474 | 1.36% |
| Turkish | tr | 1,165,210 | 1.07% |

**Known Gaps**

| **Date** | **Time** |
| --- | --- |
| 2/1/2020 | 4:00 - 9:00 UTC |
| 2/8/2020 | 6:00 - 7:00 UTC |
| 2/22/2020 | 21:00 - 24:00 UTC |
| 2/23/2020 | 0:00 - 24:00 UTC |
| 2/24/2020 | 0:00 - 4:00 UTC |
| 2/25/2020 | 0:00 - 3:00 UTC |
| 3/2/2020 | Intermittent Internet Connectivity Issues |