**COVID-19-TweetIDs**

**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:

* There may be a few hours of missing data due to technical difficulties. I have done our best to recover as many Tweets from those time frames by using Twitter’s search API.
* I 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.
* 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.

**Used Keywords:**

Coronavirus

Koronavirus

Corona

CDC

Wuhan

N95

Kungflu

Epidemic

outbreak

Sinophobia

China

covid-19

corona virus

covid

covid19

sars-cov-2

COVIDー19

COVD

pandemic

lockdown

lock down

stay at home

stay home

stayhome

**Sample of gathered Twitter IDs’:**

1245229407721213952

1245229407683264512

1245229407725219840

1245229407226040320

1245229407662305283

1245229407675068418

1245229407465172993

1245229407557443585

1245229407767150593

1245229407696019462

1245229407683280896

1245229411944693760

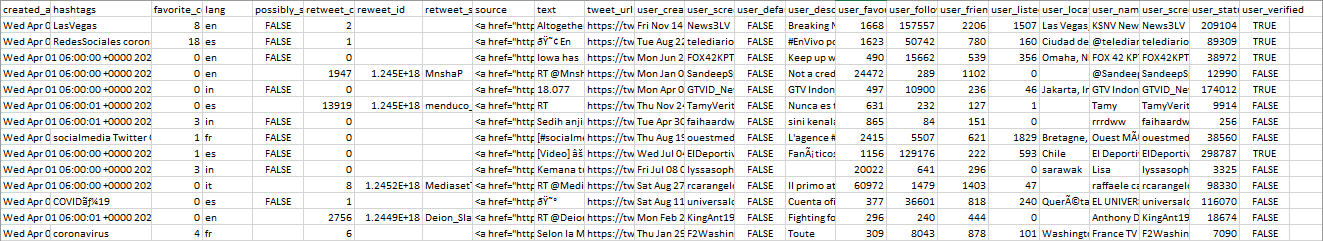
1245229412045524993

1245229411932266496

**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.

**Sample of gathered hydrated info for the above Twitter IDs’:**

**Statistics Summary**

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

| **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 |