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—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

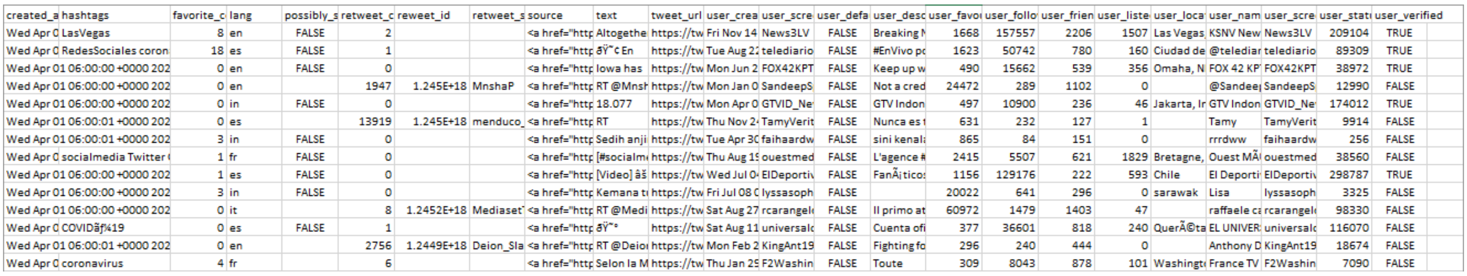
1245229411932266496

How to Hydrate

Hydrating using Hydrator (GUI)

Navigate to the Hydrator GitHub repository 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':



created_a	hashtags	favorite_c	lang	possibly_s	retweet_c	reweet_id	retweet_s	source	text	tweet_url	user_crea	user_scre	user_defa	user_desc	user_favo	user_follo	user_frien	user_liste	user_loca	user_nam	user_scre	user_stat	user_verified
Wed Apr 0	LasVegas	8	en	FALSE	2			<a href="http	Altogethe	https://tw	Fri Nov 14	News3LV	FALSE	Breaking N	1668	157557	2206	1507	Las Vegas,	KSNV New	News3LV	209104	TRUE
Wed Apr 0	RedesSociales coron	18	es	FALSE	1			<a href="http	ðŸ˜¢ En	https://tw	Tue Aug 2	telediario	FALSE	#EnVivo pc	1623	50742	780	160	Ciudad de	@telediar	telediario	89309	TRUE
Wed Apr 01 06:00:00 +0000 202		0	en	FALSE	0			<a href="http	Iowa has	https://tw	Mon Jun 2	FOX42KPT	FALSE	Keep up w	490	15662	539	356	Omaha, N	FOX 42 KP	FOX42KPT	38972	TRUE
Wed Apr 01 06:00:00 +0000 202		0	en		1947	1.245E+18	MnshaP	<a href="http	RT @Mnsh	https://tw	Mon Jan 0	SandeepS	FALSE	Not a cred	24472	289	1102	0		@Sandee	SandeepS	12990	FALSE
Wed Apr 01 06:00:00 +0000 202		0	in	FALSE	0			<a href="http	18.077	https://tw	Mon Apr 0	GTVID_Ne	FALSE	GTV Indon	497	10900	236	46	Jakarta, In	GTV Indon	GTVID_Ne	174012	TRUE
Wed Apr 01 06:00:01 +0000 202		0	es		13919	1.245E+18	menduco_	<a href="http	RT	https://tw	Thu Nov 2	TamyVerit	FALSE	Nunca es t	631	232	127	1		Tamy	TamyVerit	9914	FALSE
Wed Apr 01 06:00:01 +0000 202		3	in	FALSE	0			<a href="http	Sedih anji	https://tw	Tue Apr 3C	faihaardw	FALSE	sini kenal	865	84	151	0		rrrdww	faihaardw	256	FALSE
Wed Apr 0	socialmedia Twitter (	1	fr	FALSE	0			<a href="http	[#socialm	https://tw	Thu Aug 1	ouestmed	FALSE	L'agence #	2415	5507	621	1829	Bretagne,	Ouest MÃ	ouestmed	38560	FALSE
Wed Apr 01 06:00:00 +0000 202		1	es	FALSE	0			<a href="http	[Video] âš	https://tw	Wed Jul 0	ElDeportiv	FALSE	FanÃ¡ticos	1156	129176	222	593	Chile	El Deporti	ElDeportiv	298787	TRUE
Wed Apr 01 06:00:00 +0000 202		3	in	FALSE	0			<a href="http	Kemana t	https://tw	Fri Jul 08 C	lyssasoph	FALSE		20022	641	296	0	sarawak	Lisa	lyssasoph	3325	FALSE
Wed Apr 01 06:00:00 +0000 202		0	it		8	1.2452E+18	Mediaset	<a href="http	RT @Medi	https://tw	Sat Aug 27	rcarangel	FALSE	Il primo at	60972	1479	1403	47		raffaele ca	rcarangel	98330	FALSE
Wed Apr 0	COVIDãƒ¼19	0	es	FALSE	1			<a href="http	ðŸ˜°	https://tw	Sat Aug 11	universalc	FALSE	Cuenta ofi	377	36601	818	240	QuerÃ©ta	EL UNIVER	universalc	116070	FALSE
Wed Apr 01 06:00:01 +0000 202		0	en		2756	1.2449E+18	Deion_Sla	<a href="http	RT @Deio	https://tw	Mon Feb 2	KingAnt19	FALSE	Fighting fo	296	240	444	0		Anthony D	KingAnt19	18674	FALSE
Wed Apr 0	coronavirus	4	fr		6			<a href="http	Selon la M	https://tw	Thu Jan 2	F2Washin	FALSE	Toute	309	8043	878	101	Washingt	France TV	F2Washin	7090	FALSE

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%

Language	ISO	No. tweets	% total Tweets
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