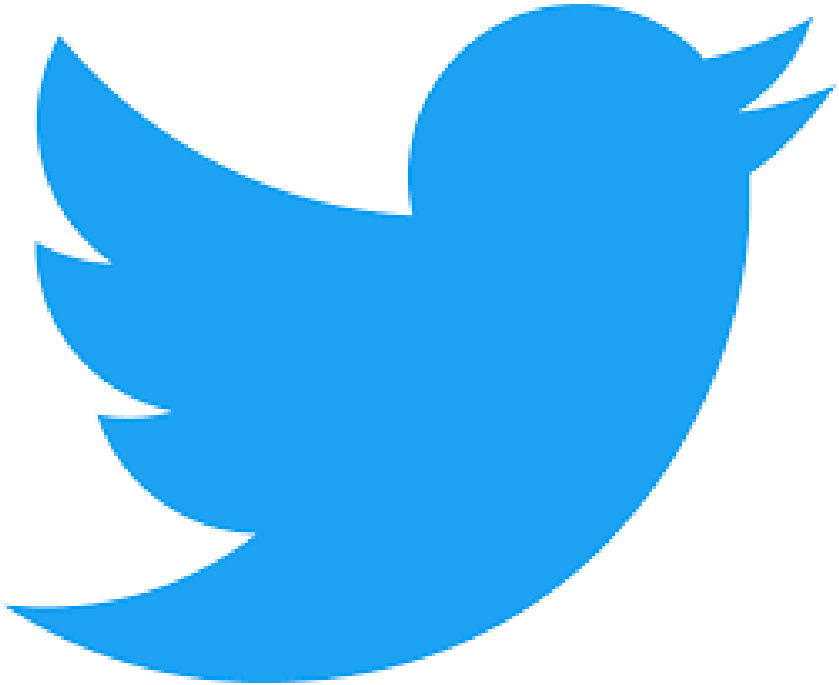
Twitter



Covid Trending Data Analysis

BDP Final Project

December 10, 2021

Wanxuan Zhang

wanxuanzhang@uchicago.edu

Agenda

- Executive Summary
- Data & Methodolody
- Clean-up & EDA
- Author Identification
- Location Analysis
- Timeline Analysis
- Message Uniqueness
- Conclusion

Executive Summary

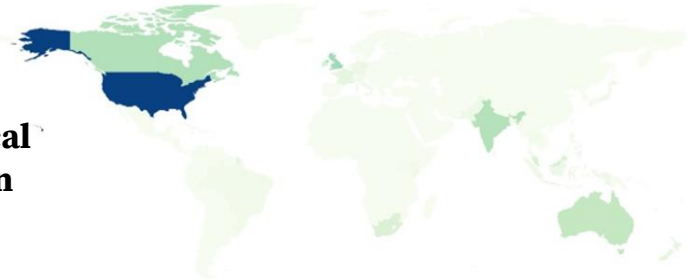
Twitter as A Source of Pandemic Progression Reflection

2 million tweets are generated in nearly a month (10/2021 -11/2021)

Among influencers, health organization generated the most tweets, **71%** of which are original

Most influential covid twitterer has been retweeted **93k** times

Tweets Geographical Distribution



Data & Methodology

Methodology

- Platform: Google Cloud Platform
- Dataproc Hub Instance: msca-bdp-dphub-students
- Language: PySpark
- Format to store the intermediate results: parquet (maximize speed and efficiency)

Source Data Overview

- Format: Nested Json
- Rows: **25,191,000**
- Content: tweets information

```
[11]: tweets.printSchema()

root
 |-- id_str: string (nullable = true)
 |-- in_reply_to_screen_name: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- retweeted_status: struct (nullable = true)
 |    |-- coordinates: struct (nullable = true)
 |    |    |-- coordinates: array (nullable = true)
 |    |    |    |-- element: double (containsNull = true)
 |    |    |-- type: string (nullable = true)
 |    |-- created_at: string (nullable = true)
 |    |-- display_text_range: array (nullable = true)
 |    |    |-- element: long (containsNull = true)
 |    |-- entities: struct (nullable = true)
 |    |    |-- hashtags: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- text: string (nullable = true)
```

Clean-up & EDA

The selected useful variables:

Column	Type	Usage
id_str	String	Primary key for tweets
in_reply_to_screen_name	String	Indicator of reply tweets
retweeted_status	Tweet object	Obtain the original content of retweets
text	String	Filter out the tweet that related to covid
timestamp_ms	String	Obtain the time of the tweet
user	User object	Obtain information of twitterers such as name and location

Clean-up & EDA

Filter covid-relevant tweets
(*text* or *retweeted_status.text* contains 'covid')

25,191k rows → **2,730k rows**

Exploratory Data Analysis:

summary	id_str	in_reply_to_screen_name	text	timestamp_ms
count	2730097	718651	2730097	2730097
mean	1.453720815279429...	3.1451517720571427E9	null	1.635429025846028E12
stddev	2.945222894559343E15	1.452600254647355...	null	7.021958576582956E8
min	1448850337998192642	000000001404072	! An extremely s...	1634267813573
max	1459021442469666820	zzzzzz_cz17	□ Due to rising ...	1636692793849

Author Identification

Top 5 Twitterers By Original Content

(Original means not retweet or reply)

screen_name	num_original
ClareKuehn	445
jcurlycurls	396
CatbearMoggy	387
42Sz40	315
CovidHelpBot	261

Top 5 Twitters By Messege Retweet

(Found from the user of the original tweets)

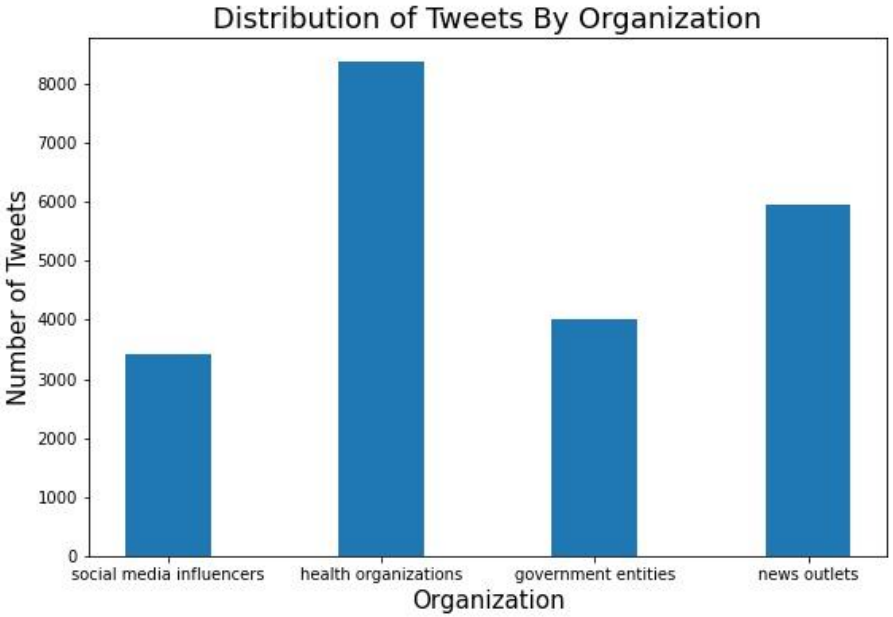
screen_name	num_retweet
PeterSweden7	93776
abcdrih	45440
ClayTravis	31321
8EightPillars	28519
Sasyity	21781

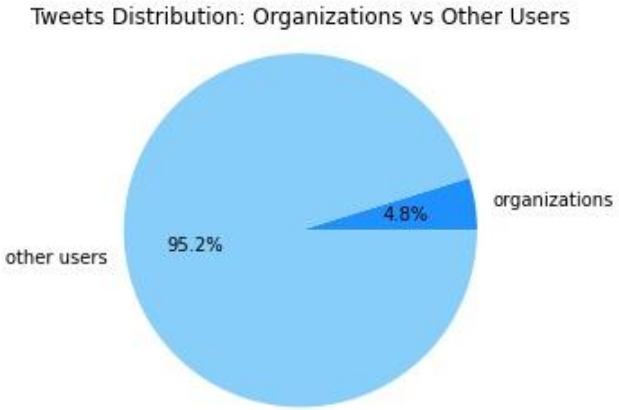
Top 5 Most Influential Twitterers

(Since the scale of retweets are much larger than originals, assign weights to define influential score: **Influential = 0.99 Original + 0.01 Retweet**)

screen_name	num_original	num_retweet	influential
PeterSweden7	29	93776	966
abcdrih	1	45440	455
ClareKuehn	445	418	444
jcurlycurls	396	6	392
ClayTravis	37	31321	349

Author Identification



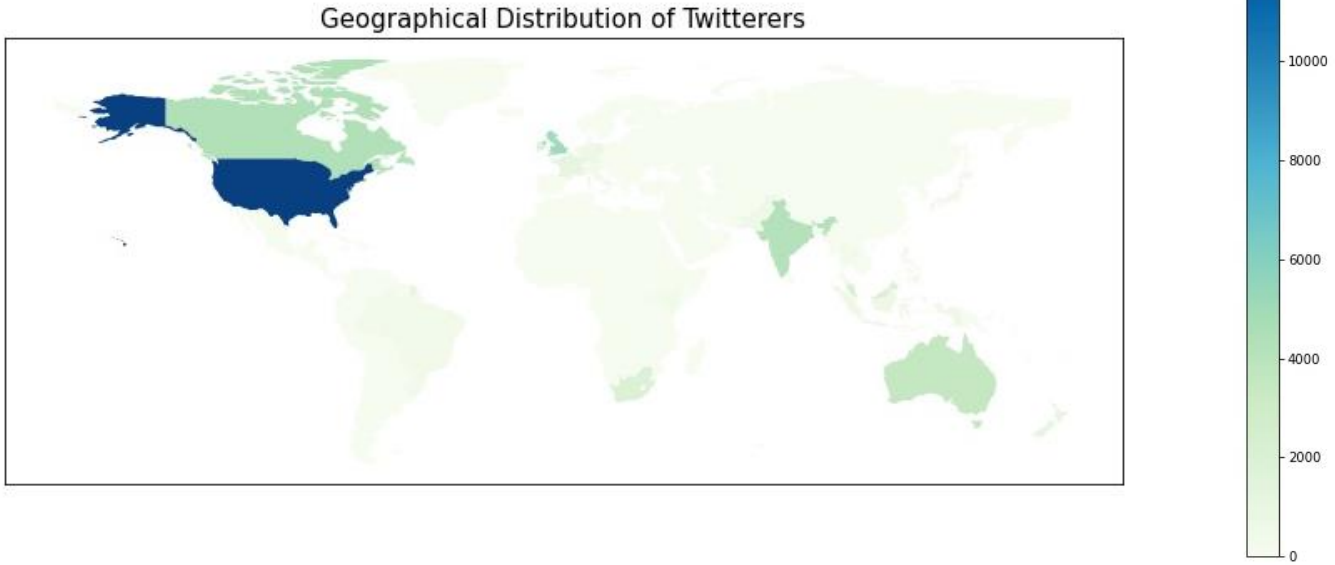


Method:
Filter from *user.description* by keywords

Findings:

- Health organizations generate most volumes of tweets (**8,358**) among the 4
- The volume of government entities is only half of that of health organizations (**4,000**)
- Organization tweets only consist of **4.8%** of total tweets, which means most of the tweets are sent by random users
- Can conclude that **influence of organizations are limited**

Location Analysis



Method:

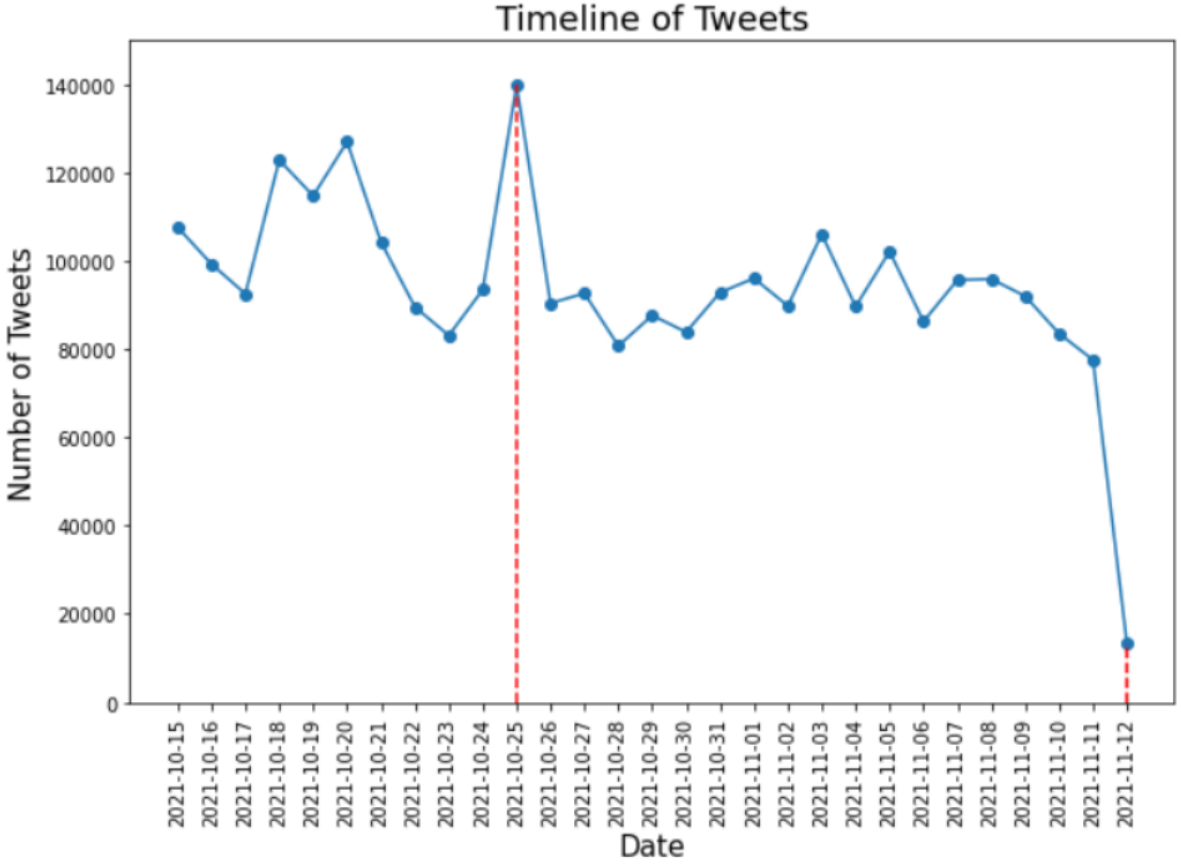
Data cleaning for *user.location* and visualize by *geopandas* libary

Findings:

- Most twitterers are from **United States** (**12,750**), followed by UK (**5,052**) and Canada (**4,360**)
- Positive correlated with pandemic severity (US has the most cases)
- Can be seen as a reliable source of **geographic pandemic progression**

Reference: https://covid19.who.int/

Timeline Analysis



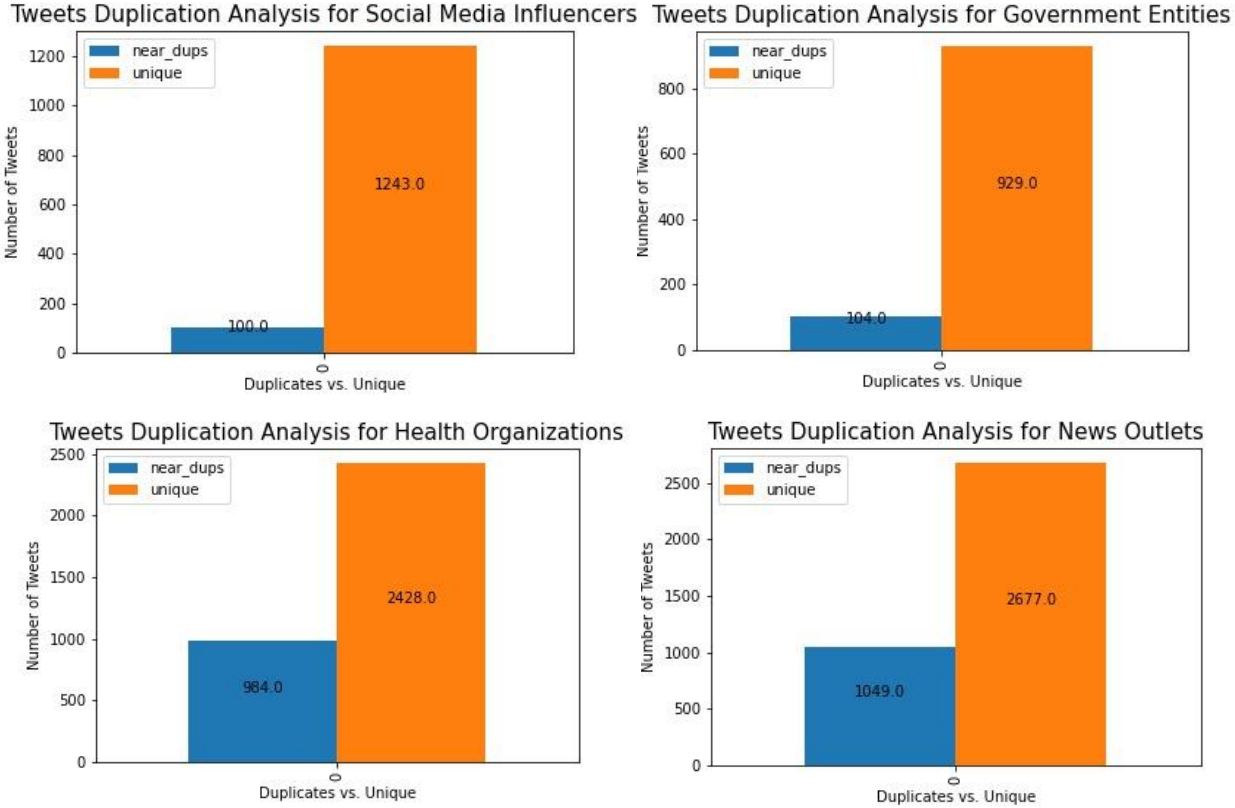
Method:
Covert *timestamp_ms* to date

Findings:

- Timeframe is from 10/15/2021 to 11/12/2021
- Peak at 10/25
- **valley at 11/12 (may due to incomplete data collection on 11/12)**
- Not correlated with worldwide pandemic
- But positive correlated with **US new cases** (decreasing trend in Oct)
- Can be seen as a somewhat reliable source of **temporal covid progression**

Reference: https://covid19.who.int/

Message Uniqueness



Method:

- Use **LSH** to measure original text similarity
- Use Jaccard similarity with threshold **0.7**

Findings:

- Social media influencers and government entities generate more **unique** tweets (near duplicates only consists of **10%**)
- Health organizations and news outlets generate more copy and paste **duplicates** (accounts for **1/3**)
- In general, organizations tend to send more original tweets

Conclusion

Main Findings

- Twitter can be consider as a credible source to reflect the risk of covid **geographically and temporally**
- **2 million** covid-related tweets are generated in nearly a month, most of which are **located at US**, with most influential covid twitterer retweeted **93k** times
- Per twitterer generates **retweets** a lot more than original messages
- Higher Tweets reflects more **US covid cases** as time goes
- Influencers such as government entities and health organizations generate more **original tweets**, but their tweets are smaller in volume w.r.t random users and **influence are limited**

Recommendations

(In order for Twitter to become more reliable in terms of **covid reflection**)

- Encourage more **government entities and health organizations** to create **original content** about covid
- Follow the time trend of covid **in US** and attract more users to **retweet**
- For government entities: share more **government actions**
- For health organizations: share more **availability of vaccines and medications**