CS432/532: Final Project Report

Project Title: Twitter Covid-19 Data Analysis

Team Member(s): Carolina Hernandez Mateo

I. Problem

Covid-19 pandemic has been a topic of relevance in the year 2020 and so has been a buzzword for the last 9 months of the year. Since the explosion of this virus on the US twitter has be en one of the main Social media where users interact about the topic in many ways. The online community of Twitter include s politicians, researches, engineers, students, companies, non-profit societies involved in the conversation. With the analysis of a small dataset of 1.5 million tweets from Sept 12 to Sept 1 9 2020. The amount of tweets collected up to date is of more than 700 million tweets related to Covid-19. Given that I am not using a high performing server architecture to process the 700 million tweets the data had been churned by 1,578,205 fr om the dataset provided by IEEE [1].

1.1 Research questions and the study aim to find

As the final outcome I want to respond the following:

RQ1. What are the most used hashtags and more correlated to coronavirus conversation?

After the measurement the popularity of the hashtags on Twitter. I aimed to analyze the frequency of usage of the word or hashtag with this step the data is revealing what are the most used.

RQ2. Who are the top users in datasets tweeting related to Covid-19?

What user has more tweets & retweets associated the Covid-19 trend.

RQ3. What are the top keywords used in the tweets related to Covid-19 topic?

What words are associated with tweets from different dimensions.

Keywords

Covid-19 (Coronavirus), Twitter, Social Media, Keywords, Hashtags.

1.2 Motivation

Given the global pandemic issue affecting the normal life of every human of planet earth and as a way to contribute with community of researchers to include a process for quick understanding of the Covid-19 pandemic I brainstormed to include a Covis-19 data analysis project for CS532 Database systems in order as well to practice the usage of a NoSQL application.

II. Software Design and Implementation

A. Software Design and NoSQL-Database and Tools Used

A.1 **Design of the process**

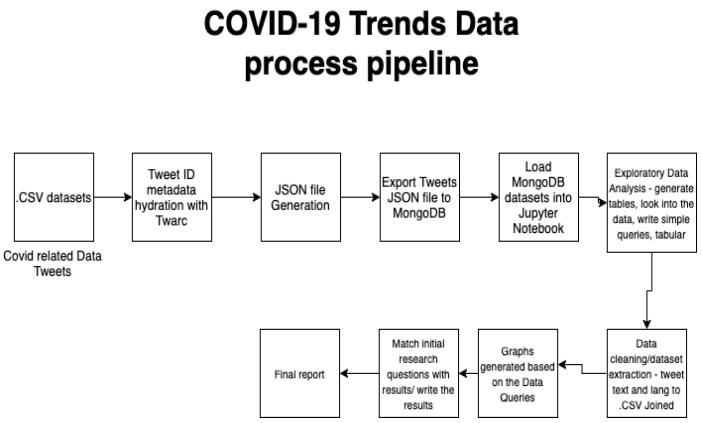


Image.

Data Collection and extraction. As mentioned in the problem description the dataset I used was a public dataset [1] that has been published daily from the IEEE. The process above describes the steps I went through clean the datasets and converted the provided *Tweet IDs* in the dataset with my Twitter Developer Account Tokens and *Twarc Python Library* to retrieve the associated tweets metadata which I imported into my MongoDB for Data Exploration in tabular mode.

1. The Tweet IDs where downloaded from IEEE[1] repository.
2. The Tweet IDs where hydrated with Twarc and converted from .CSV to JSON for each dataset with the following command:

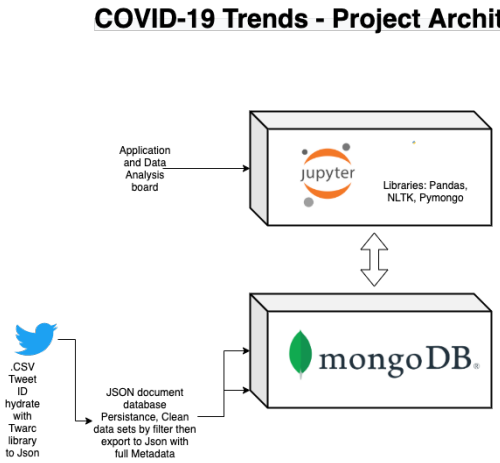
```
twarc hydrate ready_tweets_179.txt > ready_metada/tweets179.json
```

3. Exported each converted dataset into my MongoDB database *Covidfall20* and into the collection *tweets.*

```
mongoimport --db covidfall20 --collection tweets --file ~/Documents/TwitterCovid-19Analysis/tweets178.json --jsonArray
```

4. Loaded the database into Jupyter notebook. Did Exploratory and queries to extract the data I am looking to respond to my research questions and saved the tabular data.
5. Exported the MongoDB dataset into .CSV but with full joined from exporting face with Twarc for a deeper exploration and analysis graph generation from data.
6. Tokenized with a Python Script the .CSV a converted into Mini Json for Graph purposes.

A.2 **Supported tools and architecture**



The toolchain I used consistent of python libraries like Pandas, Maplotlib, Seaborn, PyMongo and other extensions with Jupyter notebook as the interface layer and MongoDB as a persistence layer to store the initial Tweet dataset.

B. Supported Queries

The reason to persist the Twitter data with MongoDB as No-SQL system is because by default the Twitter data is a JSON file and the database that fits the best this model and I have some experience with is MongoDB CovidFall20 which contains just one collection that is Tweets with all project datasets.

Note: For size purposes I exported the MongoDB from my local database as follows. If want to import must use tweets.json.

```
mongoexport --host="localhost" --port=27017 --collection=tweets --db=covidfall20 --out=tweets.json
```

B.1 Dataset description

1. The main Dataset is in the folder /ready_metada which I imported into Mongodb.
2. I extracted a Mini Tweet JSON and . CSV from the MongoDB after Querying and exploring the main Dataset in the Database. With the Parameters of "Full_text" or full tweet and "Language which is "En".
3. For Graphs purposes used a clean Mini Tweet JSON to just process the Keywords in the full tweet without hashtags still all data was exported from the main datasets in MongoDB.

B.2 Query description

Mostly support queries to Join, Group and Sum up all the keywords in hashtags and Sort by most relevant or most found in the document based on the JSON data structure. Query processing was possible by the *Aggregation Framework* provided by MongoDB. It was a better choice the Map/Reduce in Mongo given that the processing for the Query was much faster for a large dataset as the one I was working. The most relevant Queries implemented with the aggregation framework.

In my project I converted this was one of the Query I tested in MongoDB console with python code in order to be able to work with PyMongo in Jupyter Notebook. During Datasets exploration phase.

```
db.tweets.aggregate(

{ $project: {
_id: 0,
"entities.user_mentions" : 1
}},
{ $unwind: "$entities.user_mentions" },

{ $group : {
_id: "$entities.user_mentions.screen_name",
count: { $sum : 1 }
}},

{ $sort : {
"count" : -1
}}
)
```

III. PROJECT OUTCOME

RQ1. Top 20 hashtags in the dataset

	_id	count
0	COVID19	51731
1	coronavirus	11958
2	VMAs	8619
3	BTS	8434
4	HowWeGotHere	6006
5	China	5645
6	Covid_19	5637
7	COVID	5564
8	Covid19	5372
9	COVID—19	4180
10	covid19	3919
11	Coronavirus	3466
12	TrumpTownHall	3375
13	TrumpLied200KDied	3162
14	PostponeACF	2194
15	BREAKING	2156
16	pandemic	2080
17	Corona	2011
18	Rahul_Ji_Help_Raj_Students	1982
19	Covid	1926

Data Source: Jupyter notebook Twitter Exploratory Notebook

In this tabular list I could subtract with an aggregation Query similar to Map Reduce what are the top hashtags in the Dataset. Notice it's not the same as Keywords.

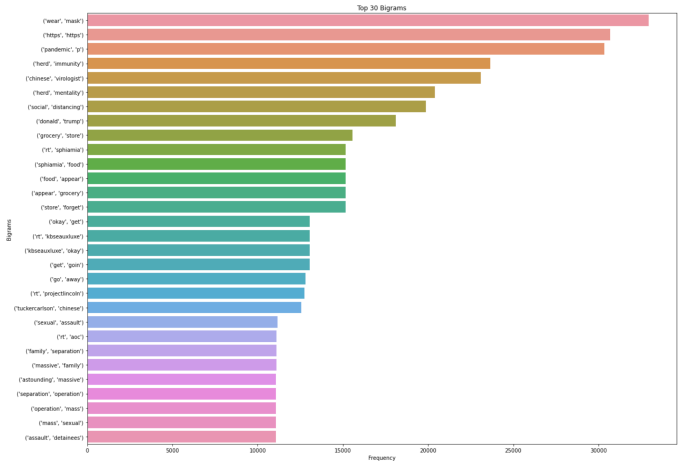
RQ2. Top 25 User in the Datasets related to public tweets and retweets and mentions.

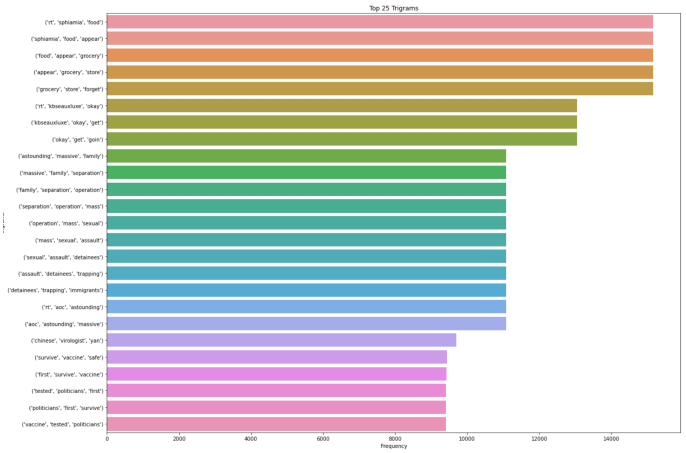
In the Next tabular list we can understand who are the top users related in mentions, tweets, retweets involved in the Covid-19 conversation. They were from the covidfall20 DB

	_id	count
0	realDonaldTrump	31552
1	TuckerCarlson	16741
2	Sphiamia	15165
3	RahulGandhi	14355
4	ProjectLincoln	14173
5	KBSeauxluxe	13059
6	AOC	11280
7	donwinslow	9802
8	RealJamesWoods	9771
9	DanRather	9339
10	AuthorMonika	9332
11	kylegriffin1	9012
12	flors_les	8923
13	voguemagazine	8659
14	bts_bighit	8645
15	BigHitEnt	8286
16	NYGovCuomo	7396
17	atrupar	7338
18	Iam_Afrodisiac	7334
19	JoeBiden	6895
20	DrEricDing	6341
21	JasmineLWatkins	6299
22	MohanadElshieky	6243
23	CNN	6005
24	RBReich	5779

Data Source: Jupyter notebook Twitter Exploratory Notebook

RQ3. Top Keywords in form of Bigrams and WordCloud





Bigrams, Trigrams and WordCloud Data analysis found in Jupyter notebooks.

GitHub repository
https://github.com/codecella/cs532

REFERENCES

[1] **IEEE 2020 Covid-19 Datasets.**
https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

[2]Seaborn documentation for plotting code samples.
https://seaborn.pydata.org/generated/seaborn.barplot.html

[3]Matplotlib code samples.
https://matplotlib.org/gallery/lines_bars_and_markers/filled_step.html#sphx-glr-gallery-lines-bars-and-markers-filled-step-py

[4]NLTK corpus API.
https://www.nltk.org/api/nltk.corpus.html

[5]NLTK tokenize examples.
https://www.nltk.org/_modules/nltk/tokenize.html

[6] Pulling Data from MongoDB to Pandas Dataframework.
https://github.com/alysivji/blog-notebooks/blob/master/mongodb-bson-numpy-twitter-analysis/001-pulling-data-all-together.ipynb

[7] Pandas JSON conversion
https://pandas.pydata.org/docs/user_guide/io.html#json

[8]**Note.**
All diagrams from design where designed by me in Draw.io the file is included in Github link.

Word Cloud of Top Data Source: Jupyter notebook.