Data Collection for COVID-19

Steffy Dias
Binghamton University
Binghamton, NY, USA
sdias1@binghamton.edu

Sagar Sahani
Binghamton University
Binghamton, NY, USA
ssahani1@binghamton.edu

Arun Reddy Gummala
Binghamton University
Binghamton, NY, USA
agummal1@binghamton.edu

Figure 1: Covid-19 Pandemic

ABSTRACT

The hour of the pandemic in 2020 has brought intensified efforts and experimental analyses by researchers and scientists alike to analyze the dynamics of COVID-19. This has led to great importance to the related data that is readily available in huge amounts these days on social networking sites. The pouring amount of data available every week, every day and nearly every hour can be of great importance once brought to its correct use.Collecting this data to make meaningful insights can help a great deal.

KEYWORDS

Twitter, Reddit, MongoDb, datasets, MongoDb, Twitter api

ACM Reference Format:
Steffy Dias, Sagar Sahani, and Arun Reddy Gummala. 2020. Data Collection for COVID-19. In *Binghamton '20: ACM Symposium on Data Collection for COVID-19, November 03–05, 2020, Binghamton, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Data Collection is the process of collecting relevant data for experiments, analysis and for scientific research needs. This data can then be used to answer relevant questions and eventually evaluate outcomes. There is a rapid growth in online communities in these recent years. Enormous amount of data is generated by many social networking sites everyday and every single hour. This same data can go a long way and help us as a source for a variety of analysis purposes. Twitter and Reddit are one of the many popular networking sites with millions of users. This includes users with a variety of views, reviews and opinions that can be gathered with the help of tweets and reddit posts.In this project we studied the importance of data collection and all the difficulties that can come across while collecting data. Data collection process as proposed was done from two social media sites Twitter and Reddit.

The COVID-19 pandemic has led to many efforts being taken to analyze the data collected in many countries and cities to study and predict its growth so that it can help plan for healthcare resources and socio-economic decision making. The pandemic has caused huge loss in public health, affecting millions of people world wide throughout countries.

2 MOTIVATION

The data collected through this project regarding the COVID19 situation can help predict the conditions that can come up and how the precautionary measures that are being carried out can help us in the long run and what more can be done at various levels likewise country, state , county or even at a university level.

3 PROBLEM STATEMENT

To do extensive data collection about COVID-19 from popular social media websites that can be used for insightful impactful analysis.

4 LITERATURE SURVEY

Campan,Atnafu,Truta and Nolan [3] proved the sampling filtering process to collect tweets on Twitter with the key-matching tweets. And also showed how tweets collected for popular filtering terms led to biased results due to the non-random sampling process. They also analysed how these tweets could provide reliable results for research purposes. They collected data through the free version of Twitter Streaming API collected between June 24th to July 15, 2018 .

Tao, Hauff, Houben, Aben and Wachsmuth [5] conducted analytics

Binghamton '20, November 03–05, 2020, Binghamton, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

on the data generated by social networking sites like Twitter and how challenging it can be due to the enormous volume, variety and velocity of the data generated through it. They further propose their Twitter Analytical Platform for conducting analytic tasks with Twitter data.

V. Z. Marmarelis [6] in his study mentions how the Suitable Model form and the Robust Estimation and Adaptive Modeling when applied to Covid-19 daily time-series data of US confirmed cases represented the four waves that passed the time-point of peak infection rate. This data collected since March 2020 through June 2020. Therefore, it can thus be seen how the data collected that is the number of cases in those months helped analyze how the covid-19 epidemic can be controlled if no new wave emerges.

He [4]through his paper explains the risk prevention and control decision-making model proposed due to the incompatible characteristics of the pandemics risk prediction. Further mentioning how the collection of data on the personnel flow helped analyze the attribute of epidemic risk.

5 IMPLEMENTATION FOR TWITTER DATA COLLECTION

Initially we use the requests module to send the http requests. Creating a header with the bearer token and type of the content we need. Passing the Url and the header to the endpoint function. Using the requests library in Python we call the get method on the passed URL with the header. Then opening a file with extension .ndjson which will interpret the json data properly and writing the received response data into the file. And also an extra file for just storing the timestamp of the tweet occured and number of tweets for a rough estimation about tweet count per minute as well as plotting the same.

5.1 Amount of data Collected

We initially collected 1 lakh unfiltered tweets, which is around 100 MB of data and based on that we drew the plot. We plan to collect 1 million tweets for the future analysis.

5.2 Changes since Proposal

The proposal specified that the data collection from Twitter would be done with the help of the Twitter api [1] directly, and so does this implementation of the code that collects data via this Twitter api.

5.3 Challenges Faced

Deciding the expansions in the tweets which we need in the future for collecting the data according to the topic we need has taken a considerable amount of time. Formatting the json such that the entire tweet must be saved in a single line so that while importing into the mongo db the fields will be captured properly. Plotting the data according to the timestamp using pandas in Python.

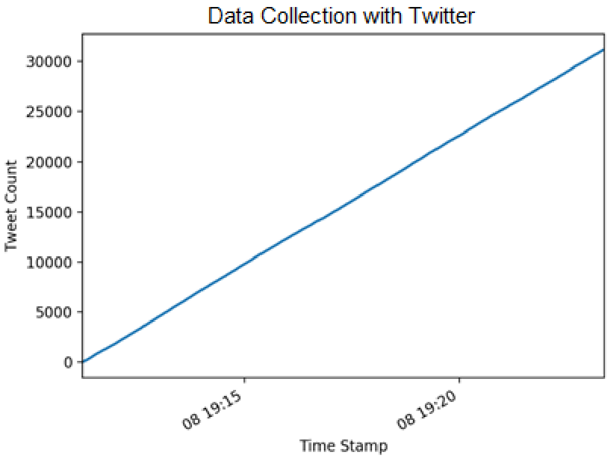

Figure 2: Twitter Data Collection

6 IMPLEMENTATION FOR REDDIT DATA COLLECTION

Firstly, we import a couple of libraries to request, process and store data. As MongoDB is used for data storage, the pymongo package is used in the imports. The data collected are from top 50 subreddits for COVID-19, the script iteratively requests each subreddit with the last post the script gets and if there are any new posts in the subreddit they are fetched.
A separate script was used to get these 50 top subreddits from the search API which queried with search params like covid-19, coronavirus etc. Basic flow of the script follows the examples provided by the official API, i.e. GET authorization with the tokens and the app name, query for a success status and then proceed for collection, here we take authorization and request the data from the same base URL 'https://oauth.reddit.com'. After getting the requested data, the json object that consists of the whole bunch is accessed and is appended with the current timestamp, this timestamp is the system's timestamp, when the data was collected by the system. As there are no API parameters that facilitate this, it should not be inaccurate as the script only collects the newly posted data.
Secondly the collection of new data, this is done with parameters like after before provided by the API, here we use the before param to get the latest data.After attaching the timestamp, a function is created to take the data and insert it into MongoDB. Further, the script then stores the latest id of the post it fetched in the list, to query new posts in the next iteration.

6.1 Amount of data Collected

The amount of data collected from November 1st to November 6th was in total 2.61GB. The number of posts collected on a daily basis along with their timestamp is further plotted in the graph below.

6.2 Changes since Proposal

The proposal specified how the data collection could be done manually by collecting the posts by concatenating the posts url with

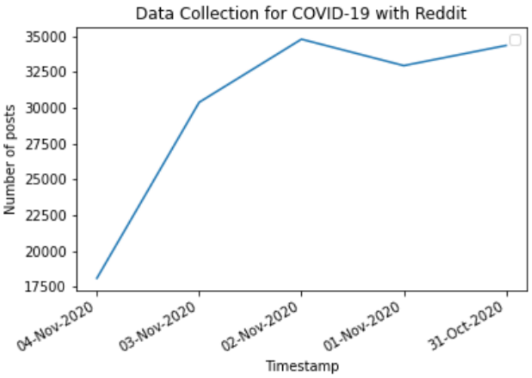

Figure 3: Reddit Data Collection

the .json extension. Wherein here the same implementation is done by creating an App [2] for reddit(Required for Data Collection) and using the oauth for the authentication further collection of data.the subreddits were manually considered to be the choice of collection.

6.3 Challenges Faced

As RedditAPI [2] did not have anything like stream or filtered stream that gets data across multiple subreddits on a single request, this was the initial challenge that was mitigated by its own search API feature that allowed us to get the top 50 subreddits for our topic.

For the main collection script, a starting point had to be decided on for collection, for this in the start of the script functions are used to get the 100 recent posts first, and then with the most recent one's id is used with the 'before' param to keep quiring the subreddits in a loop. There were many authentication failures which were solved by sleep functions in the script.

Also, when there was only one post fetched at a given time, mongo crashed the script, which then was solved by inserting one or many logic. Many subreddits in the list, had new posts after 45mins to 1 hour. The script would easily crash at the extraction part as it came with no data, a quick fix was to have a logic that skips the function.

7 UPDATED PROJECTIONS ON THE AMOUNT OF DATA WE ARE TO COLLECT:

Along with the amount of data collected till date we propose to collect our next set of dataset from November 10th to the 20th of November that would help us further analyse the data better in the upcoming project.

8 CONCLUSION

Data depends upon what assumptions and requirements one has in an attempt to collect it, basically neutral or objective facts. It constitutes a useful, highly accessible and comprehensive resource for study. The theoretical position obtained by users shapes the interpretation that influences the results.

Also, the ease of access with API's that these social media websites have provided made the whole collection process flawless and precise

9 ACKNOWLEDGMENTS

We would like to thank Prof. Jeremy Blackburn for guiding and helping us come up with this project and providing his insights on the various steps that could help us plan this project.

REFERENCES

[1] [n.d.]. https://developer.twitter.com/en/docs/twitter-api

[2] [n.d.]. https://www.reddit.com/dev/api/

[3] T. M. Truta A. Campan, T. Atnafu and J. Nolan. 2018. Is Data Collection through Twitter Streaming API Useful for Academic Research?. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, WA, 3638–3643. https://doi.org/10.1109/BigData.2018.8621898

[4] P. He. 2020. *Study on Epidemic Prevention and Control Strategy of COVID -19 Based on Personnel Flow Prediction*. IEEE, Zhuhai, China. 688–691 pages. https://doi.org/10.1109/ICUEMS50872.2020.00150

[5] G. Houben F. Abel K. Tao, C. Hauff and G. Wachsmuth. 2014. Facilitating Twitter data analytics: Platform, language and functionality. In *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, DC, 421–430. https://doi.org/10.1109/BigData.2014.7004259

[6] V. Z. Marmarelis. 2020. Predictive Modeling of Covid-19 Data in the US: Adaptive Phase-Space Approach. In *IEEE Open Journal of Engineering in Medicine and Biology*, Vol. 1. IEEE, 207–213. https://doi.org/10.1109/OJEMB.2020.3008313