Plan Overview

A Data Management Plan created using DMPTool-Stage

Title: #Covid-19 hashtags Tweets digital curation

Creator: yi wang

Affiliation: Non Partner Institution

Principal Investigator: yi wang

Data Manager: yi wang

Funder: Digital Curation Centre (dcc.ac.uk)

Funding opportunity number: 52202

Template: Digital Curation Centre

Project abstract:

This repository focuses on the topic of Coronavirus (Covid-19) on microblogs Twitter, Sina Weibo (or more if time permitted), in the time period of the beginning of the outbreak (December 2019) in China and till December 2020. The repository must contain Corona relating posts of twitter, Weibo Sina, which are selected by a combination of a group of # (hashtags) such as #covid-19, #corona …#□□□/#□□□ to make sure the content really and trustworthy about corona relevant, and usable for future research.

Last modified: 07-11-2020

Copyright information:

#Covid-19 hashtags Tweets digital curation

This repository focuses on the topic of Coronavirus (Covid-19) on microblogs Twitter, in the time period of the beginning of the outbreak (December 2019) in China and till December 2020. The repository must contain Corona relating posts of twitter, which are selected by a combination of a group of #(hashtags) such as #covid-19, #corona … to make sure the content really and trustworthy about corona relevant, and usable for future research.

Data will be collected and edited by the Tweepy, Twarc, Numpy, Pandas libraries… in the end are saved in Jason format text files, since they could be easily transformed and used in a lot of programming languages. The collection is streamed daily around 23.00 to ensure the fully download, with approximately more than 5 G volume of data daily. All those data will be saved on standard hard drive and project webpages. For historical tweets, we generated form GNIP portal Power Track (HPT) to download the tweets since December of 2019.

Question not answered.

Profile privacy related information or any sensitive data would not be collected, such as:
"profile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"profile_link_color":"1DA1F2","pro

Those data are not meaningful so far to collect, but if there are locations which users are posted with, languages they are written in or more information on bio, number of likes, retweets and so on would be collected, they could have potential usages. And all twitter account would be identified meanwhile anonymized.

Any users need to give a brief description of their usage of this dataset, after confirmation their usage only for academic and research purpose, they can download. Content consumer could use them and request the information as long as their project finished, and give us a brief feedback, summary, comments and so on, also if they published, they need to cite our websites.

Data would be backed up directly after download, and afterwards once a year with another hard drive backup all those files

And the data would be preserved by our own experts, and administrators.

There's a lot of potential usages of epidemic data for social media sentimental analysis, comparisons among the public from different regions, not just for medical scholars, virus scientists and institutions but also individual who want to get engaged in social media or digitalization world. To ensure legal and ethical purposes, nothing personal information would be curated for a longer term, only the basic account information, geological locations, tweets counts, and text would be curated. To be more specific the following attributes would be curated from tweets, Created_at, id, text,source, in_reply_to_status_id, in_reply_to_user_id, user,name, location, description, follwers_count,friends_count, listed_count, favourites_count, statuses_count, created_at,time_zone,lang, coordinates, place, quoted_status_id,quote_count,reply_count,retweet_count,fvorite_count,entities,withhelf_copyright,withheld_in_countries, withheld_scope, geo.

All the datasets would be held in hard drives and get frequent copy version every half year to ensure the preservation, meanwhile update the information content, to adapt to latest technologies. Public tweets data are open to selected, streamed lively and downloaded but regarding to whose historical tweets form GNIP would be caused some money, but I have not done enough research like how much is needed.

To spread our newest collection, we would also cooperate with museum, digital libraries, and some other databases, and in return to also get some financial support. Users could get access only through our webpages, and make a request, with a brief summary of their research goals, methods, research questions or get verified by those users who already the access to our datasets had, could download Covid-19 Archives. But after those users used our datasets, they need to give us a short feedback and results, will help us to enhance the management and usability, and users also need to cite our dataset on the end of their publication.

However, there's still a slightly potential that users would demand us to delete their information, then would be a huge cost for us to find and get rid of all the relating information.

Project team members are a small number of experts, who has a user role as admin in the system to teach and give advises back on our web forum and ensure the dataflow, volunteers and students could help to verify the motivation of using our datasets.

Tweetvan: https://github.com/iipc/twittervane

TweetView

TweetStreamAgent

TweetAnalyser