ETC1010 PROJECT | TEAM KNITR

DATA DESCRIPTION & PROJECT PLANS

OVERVIEW

- Tweets were downloaded from a Twitter API in '01-downloading-tweets.Rmd'.
- Additional data was sourced about the state of COVID-19 in different countries.
- Extra information was extracted from the tweet data frame and mutated on as additional columns.
- The tweets were assigned different levels of sentiment based on the tidy text package and the results were saved as 'tweets.rds'.
- Location information was added and saved to 'geo_characteristics_of_tweets.rds'.
- The Rmd files are also annotated with additional information.

DATA DESCRIPTION

The bulk of the data we have cleaned is stored in 'tweets.rds'. It consists of 104,674 English language tweets containing the term 'covid' that were posted around the 15th of May. Each tweet has been assigned an index for ease of use and interpretability. There is an arbitrary numerical classifier for each user which is also useful for identifying user behavior patterns. The time the tweet was created is also present. We have also added in columns representing the number of hashtags, user mentions, and URL links in each tweet.

There is a column which contains the text of the tweet, as well as multiple columns containing the number of words in each tweet that expressed different types of sentiment, based on the 'nrc' lexicon. These sentiment columns can potentially be used to discover patterns in the ways people have been discussing COVID-19 online. Additionally, there is a column containing the mean positive sentiment of words in each tweet, calculated using the 'afinn' lexicon. There are also columns representing the number of words and number of characters in each tweet.

For every unique user who made a tweet in our dataset, we have information about the number of followers they have, the number of friends they have, and whether their account is verified or not. We also have the number of likes and retweets each tweet has. It will be interesting to see how the characteristics of a Twitter user affects the level of engagement of the tweet and the sentiment expressed in it.

Additionally, the 'geo_characteristics_of_tweets.rds' file contains information about the state of COVID-19 in the country where approximately 2700 of the tweets in our dataset were posted. The country was extracted using longitude and latitude information stored in a 'bbox', but unfortunately only a limited amount of the tweets contained this valuable information.

The estimated latitude and longitude of the tweet are stored in the 'x' and 'y' columns of the dataframe. The 'closest_country' contains the closest country centre to the tweet from which the rest of the information is derived from. Additionally, the user's self-described location is present, largely to subjectively confirm that the closest country determined by our function is relatively accurate. Each tweet also has the corresponding new cases, total cases and deaths, both nominally and per million. There are also columns relating to the median age and total population of the countries assigned to the tweets.

DOWNLOADING DATA

The bulk of the data we have used consists of tweets. These were sourced through the Twitter API and are mostly from the 15th of May. The data was downloaded in 7 batches, each with slightly under 15,000 tweets. Each batch of downloads was saved as a file called 'tweets_n.rds' which can be found in the 'data' folder. The format of the downloads is captured in '00-downloading-tweets.Rmd'.

Note that using the Twitter API requires a free private developer license, so the code is not usable without a private key. It is also only possible to download tweets up to a week in the past, so the exact dataset that we have used in our project cannot be reproduced.

Additionally, the geo location data of countries was downloaded from:
https://raw.githubusercontent.com/albertyw/avenews/master/old/data/average-latitude-longitude-countries.csv

And the data about the state of COVID-19 in different countries was downloaded from:
https://ourworldindata.org/coronavirus-source-data

CLEANING TWEETS

The cleaning process happens in '02-cleaning-tweets.Rmd'.

The first step was loading all the previously downloaded sets of tweets and then binding them into one dataframe of more than 100,000 tweets. Then an index column was added for ease of use and interpretability.

The tweet data contains several nested columns with lists related to hashtags, URLs, and mentions. Tweets without the specific element are denoted by a list containing a single NA value. It would be useful to have a variable representing the number of hashtags, mentions and URLs in a tweet.

To solve this problem, we unnested the relevant columns, filtered out the NA values so that only actual mentions were remaining and then summarized by 'n()' to get the count for each applicable tweet. We right joined the counts back into the tweet data and then imputed the missing values with 0, giving the count of mentions.

We also added in a column with whether the tweet had any likes. About half the tweets had no likes, likely because a lot of them were from bots which would make their posts less engaging.

Then, we selected only the columns we felt would be relevant to our analysis.

ADDING SENTIMENT INFORMATION

The next step was to assign sentiment information to the tweets so we could analyze their content.

First, we created a separate data frame where we tokenized the words. Then, we summarized this data frame so that we could determine the amount of words in each tweet.

To add sentiment information to each tweet, we used the tidy text library and the 'get_sentiments('nrc')' function to produce a list of English words and the sentiments expressed with those words. We then counted the number of words with each type of sentiment in each tweet and joined that information with the data. We did a similar thing with the 'get_sentiments('afinn')' function to determine the average level of 'positivity' expressed in each tweet. This data was saved as 'tweets.rds'.

ADDING LOCATION DATA

Unfortunately, the ‘location’ variable that was included with the downloaded tweets featured strings that appeared to be manually entered by each user. Sometimes the users entered in a country such as “India” or “Australia”, but many times users entered locations which don’t exist such as “citizen of the world” or “Here. Sometimes I here.”. Because of the lack of consistency in this variable, it was very difficult to use this variable to extract the country where each tweet was posted.

Instead, to get information about the locations of tweets, we used the tweets ‘bbox_coords’ which provided 4 sets of latitude and longitude coordinates for each tweet. Unfortunately, only around 2700 tweets featured these coordinates. To process these coordinates, we first unnested the coordinate boxes to get 4 sets of latitude and longitude coordinates. Then, we calculated the average of these coordinates and assigned them to each tweet. Finally, to identify the country from these coordinates, we created a function that takes in the averaged latitude and longitude coordinates and outputted the country with its centre closest to thes coordinates.

We then downloaded another dataset with the state of COVID-19 for a range of countries. This dataset included variables such as total cases, total deaths, population and population density. This new dataset was joined with the previously identified countries in the tweet data so that any tweet with information about its country also had accompanying information about the state of COVID-19 in that country as well.

This information was saved as a separate file titled ‘geo_characteristics_of_tweets.rds’

PROJECT PLANS

Based on our cleaned dataset, we aim to focus our analysis on the following areas:

- Relationship between the number of shares and the number of likes of a tweet.
- The other factors that may impact the number of likes of a tweet (followers, friends, URLs, hashtags).
- How the number of likes on a tweet relates to the source of the tweet (mobile, desktop, etc.).
- Relationship between number of followers and number of friends for verified and unverified accounts.
- The most common words used in tweets which reference ‘covid’.
- The most common positive words and negative words in tweets which reference ‘covid’.
- Relationship between number of favourites and number of retweets for a tweet referencing ‘covid’ that expresses a strong level of positive or negative sentiment.
- How positivity, trust, disgust and anger are related to the state of COVID-19 in the country where a tweet is posted.