Data and Descriptions

Our project, “Fantastic News and How to Fake Them”, has used multiple datasets, and within those datasets, have manipulated them in many ways to make them usable for each visualization. This document describes the datasets used and details the columns for each of the files that were used in the visualizations.

Disinformation 6 Dataset

The “Disinformation 6” Dataset was extracted from Twitter using the list of IDs of tweets made by the Disinformation Dozen, available on GitHub. Since Twitter’s Terms of Service do not allow sharing of complete JSON datasets of tweets, only the Tweet IDs were available on the Github repository. Then, we needed to use a tool called Twitter Hydrator to obtain the data currently available on Twitter for our project. Half of the members were suspended from the platform, which is why we could only obtain the interactions of the remaining ones. This information was then used to create three files, each being used in a different visualization.

misinfo6.json

After cleaning, this dataset is used as a JSON file in the force-directed graph titled “Who are the Disinformation 6 Contacting?”, and its columns are:

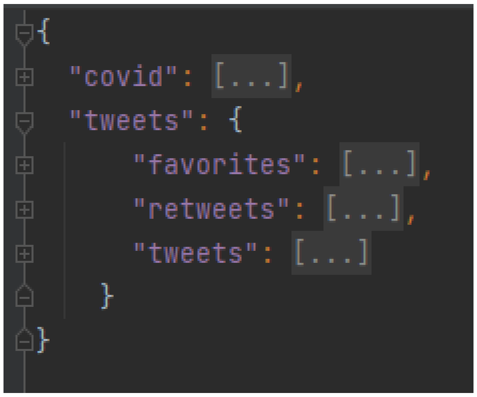
"username": the user’s username
"name": the user’s real name
"verified": boolean variable if the user is verified on Twitter
"description": the user’s self-description in their “Bio” section
"protected": false,
"location": the user’s self-described location.
"created_at": date/time of the creation of account.,
"followers_count": how many users follow this account.,
"following_count": number of users this account follows,
"tweet_count": number of tweets,
"listed_count": number of Twitter lists,
"group": number used to group the user according to the interaction of the Disinformation 6 members.

covid_vs_tweets.json

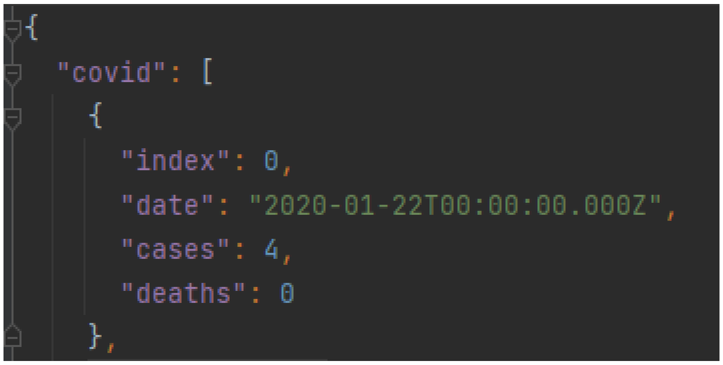
The timeline of COVID-19 deaths vs Disinformation 6 tweets was used by combining the Covid-19 deaths dataset provided to us for Homework 10 and the tweets by the Disinformation 6. The file is organized as follows:

“covid”: contains a list with an index, a date, the number of cases, and the number of deaths for that day.
“tweets”: contains lists for the number of favorited tweets, retweets, and tweets for each of the Disinformation 6 members for each day of the pandemic in the dataset.

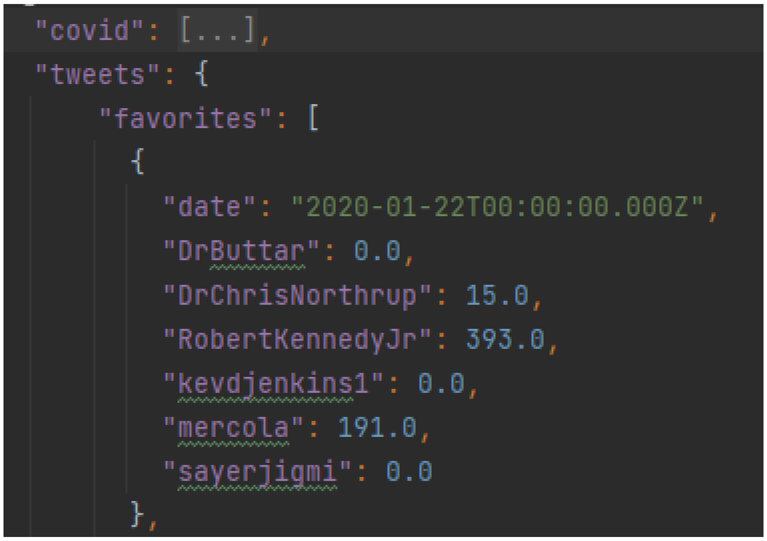
Screenshots of the nested data structure of this JSON file are added for further clarification:



```
{
  "covid": [...],
  "tweets": {
      "favorites": [...],
      "retweets": [...],
      "tweets": [...]
    }
}
```



```
{
  "covid": [
    {
      "index": 0,
      "date": "2020-01-22T00:00:00.000Z",
      "cases": 4,
      "deaths": 0
    },
```



```
  "covid": [...],
  "tweets": {
      "favorites": [
        {
          "date": "2020-01-22T00:00:00.000Z",
          "DrButtar": 0.0,
          "DrChrisNorthrup": 15.0,
          "RobertKennedyJr": 393.0,
          "kevdjenkins1": 0.0,
          "mercola": 191.0,
          "sayerjiqmi": 0.0
        },
```

wordcloud.csv

The top 150 words on the tweets made by the Disinformation 6 were made into a word cloud, in which the bigger words were the ones used most frequently. This file is used for that visualization, and only has two columns:

“key”: the word tweeted.
“value”: the number of times it was tweeted.

ESOC COVID-19 Misinformation Dataset

This dataset, available here, contains data collected by the researchers with misinformation efforts on social media relative to COVID-19. It has categorical data relative to the motives for this misinformation, the narrative being pushed, and the region that misinformation is related to. This data was then cleaned and separated into 3 separate .json files for the “Misinformation Over Time” bubble visualization. Information from the ESOC dataset was also used in the “Disinformation: A Story Across the Globe” visualization, in a CSV file.

motive.json, narrative.json, region.json

For the “bubbles” visualization, the ESOC dataset was broken down into three JSON files, named motive.json, narrative.json, region.json, each of which shares both a date and a count of tweets for that date, as follows:

"date": "2020-01-27T00:00:00.000Z",
"count": 1

Depending on the file chosen, the JSON object will contain one of the columns below, which are used in the category selector of the visualization and serve to aggregate the data:

"motive": "Downplay Severity"
"narrative": "COVID-19 status of individuals"
"region": "Africa"

fake_news_story.csv

This file is used in the visualization to count the number of documented fake news cases across the world. It has the following columns:

“country”: the country from which the fake news originated
“country_code”: a three-letter abbreviation of the country name. Example: Brazil = BRA.

“lat”: the country’s latitude
“long”: the country’s longitude
“num_fake_news”: the number of fake news events recorded for that country

Historical Fake News Dataset

This dataset was manually created by us by collecting images and information from several online sources. It contains a list of historical fake news ordered by year, with images, a describing text, and sources for the images and text. It is available as a tab-separated document.

historyfake.tsv

This dataset, in tab-separated format, contains the following columns:
“id”: used for uniquely identifying each fake news event
“year”: the year in which the event took place
“name”: the name of the country in which the event occurred. Called “name” to match how the country names are referenced in the GeoJSON data used in the globe for this visualization.
“image_code”: the name of the .jpg file to be loaded in the visualization.
“Img_source”: the source of the image.
“source”: the source of the news content.
“alt”: the description of the image used in the “alt” property of the HTML element.
“event”: the name of the event.
“news”: a text describing the event.