INF552 Project KALUGIN
US 2020 Elections Tweets

Dmytro KALUGIN

December 2020

1 Data set

The data and the code for the project are available by the link
https://github.com/KaluginD/INF552-Data-Visualisation.

As a date set for visualisation I have chosen US 2020 election tweets from a kaggle competition. The data set is available by the link below:
https://www.kaggle.com/manchunhui/us-election-2020-tweets

The time frame covered by the data set is from 15.10.2020 to 04.11.2020. It consists of the following information for both candidates:

- Date and time of tweet creation
- Unique ID of the tweet
- Full tweet text
- Number of likes
- Number of retweets
- Utility used to post tweet
- User ID of tweet creator
- Username of tweet creator
- Screen name of tweet creator
- Description of self by tweet creator
- Join date of tweet creator
- Followers count on tweet creator
- Location given on tweet creator's profile
- Latitude parsed from user location

- Longitude parsed from user location
- City parsed from user location
- Country parsed from user location
- State parsed from user location
- State code parsed from user location
- Date and time tweet data was mined from twitter

I decided to track the number of tweets and views over all of the states during this period as well as their ratio. Comparing those results to the elections results we could track a dependence. So initially we are interested in the data about the states, the time of the tweet creation and the number of the user's followers.

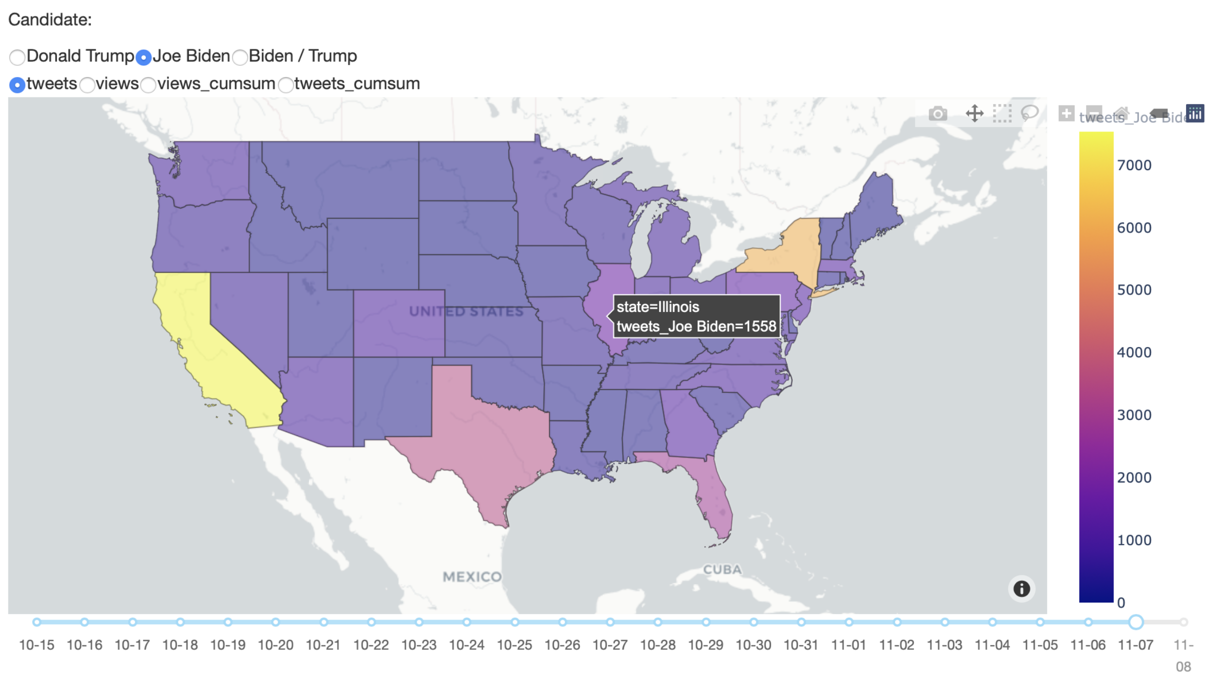
2 Data Processing

A considerable part of work was to prepare the data to the form that would allow it. The pipeline of the data processing was the following:

1. Select the tweets from the US
2. Separate them by the state
3. Group the tweets by the appearance date
4. For each group count the number of the tweets and cumulative number of followers as views
5. Calculate cumulative number of tweets and views over time
6. Join the results for both candidates for each date and state
7. Calculate the ratio of each parameter

3 Visualisation

The visualisation has the following form:



Candidate:

○Donald Trump ●Joe Biden ○Biden / Trump

●tweets ○views ○views_cumsum ○tweets_cumsum

state=Illinois
tweets_Joe Biden=1558

Figure 1: Example of visualisation

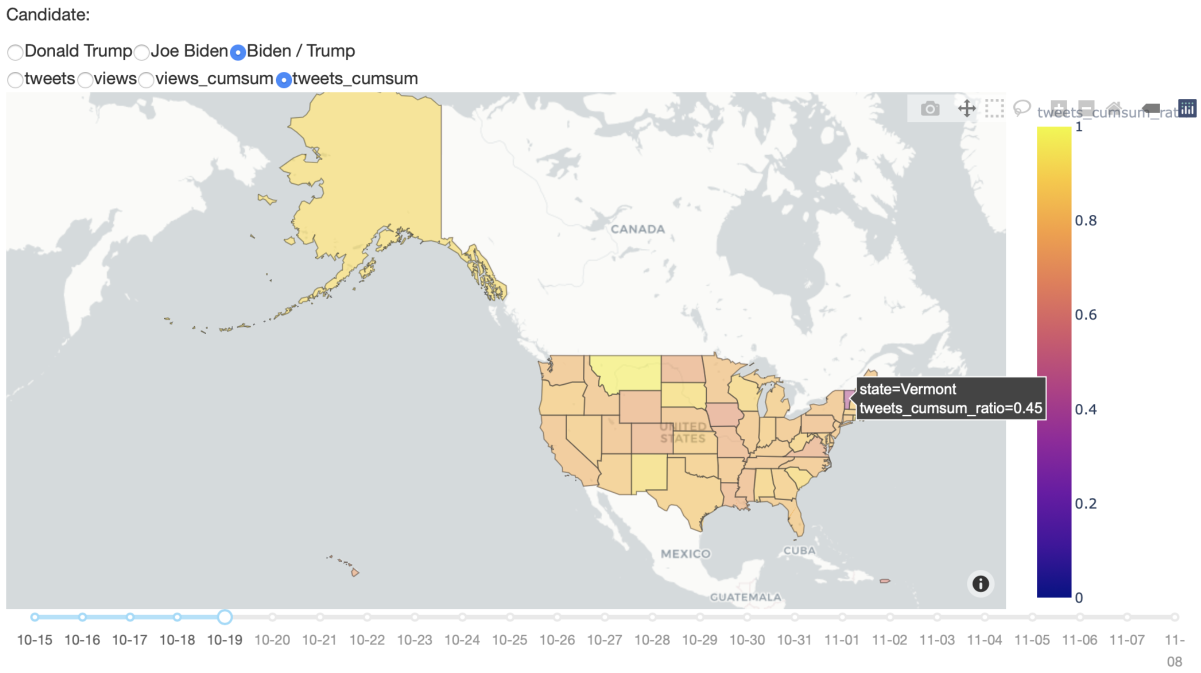
It allows to chose the following parameters:

- One of the candidates or the ratio of their statistics
- The subject of interest: number of tweets in certain day, number of views in certain day, number of all tweets up to this day, number of views up to this day
- The date of the interest

The visualisation was made with an interactive graphing library Plotly and deployed to the local server with the Dash platform. To launch the visualisation those libraries are the principle requirements.

I used Pyplot Mapbox Choropleth Maps to visualisate the map as the one we covered in the course. The color of the state corresponds to its value by the scale on the right.

This is also a multi-variable visualisation problem having different metrics we want to show at the same time.



Candidate:

○Donald Trump ○Joe Biden ●Biden / Trump

○tweets ○views ○views_cumsum ●tweets_cumsum

state=Vermont
tweets_cumsum_ratio=0.45

Figure 2: Another example of visualisation

So at the figure 2 for example we see that as of 2 weeks before the elections people were mainly tweeting about Joe Biden rather then Donald Trump everywhere except for Vermont.

4 Design decisions and insights

Among all the features provided in the data set the location was the most visual one so I decided to base the visualisation on it. I decided to limit the area of interest to the US for several reasons:

- Those are the most important regions because they vote and other countries do not,
- Other countries represent small part of the data set,
- From a practical point of view it was easier to find GeoJSON for the US map.

I was considering including other metrics to the visualisation such as:

- Aggregated number of likes,
- Aggregated number of retweets,

- Top of aggregated hashtags from the tweets,
- Top rated tweets,
- Devices used to tweet.

All of those are possible continuations for the project. Here are some reasons why it was not included:

- Aggregation of hashtags would require more complicated data preprocessing extracting them from the tweets and aggregating them,
- Used devices can be interesting for some narrowly segmented audience research but otherwise it would be superfluous,
- Top rated tweets can give some highlight on the public opinion but because they are not representative over the whole set of the aggregated tweets it can be misleading,
- Aggregated number of likes and retweets could be an easy addition but when I looked into the data I saw that those numbers are very correlated with the number of tweets so it would not be very informative.

Juxtaposing this map with the results of the elections it may help us partly interpret the results. Although having the results of the election very close to 50-50 in the majority of the states we can not expect that 20 days of twitter tweets statistics would be able to reflect such sensitive difference.

As for more broad result we can see that the elections were much more discussed in New York, California, Texas, Florida. And the least discussed in Montana, South Dakota, Wyoming and Vermont. The biggest number of views was on November 7th, 3 days after the elections which aligns with vote counting delays. We can also see that Joe Biden was dominating Tweeter by both number of tweets and number of views all over the US for the last 2 weeks before the elections. On the views especially from which we can conclude that influencers were mainly posting about Joe Biden.