Project Proposal

Data analysis from 2020 US Election Tweets

Name : Yue Zhang

Student ID : 30976316

Tutor : Benjamin Lee, Xiaojiao Du

Questions

1. What is the difference of the number of related discussions, like or retweet of tweets between Trump and Biden in general or in terms of time, location, and source?
2. What are the frequent words in tweets content? Can we analyze and compare the sentiment inside the content?
3. Compare the vote by state with the tweets by state and see the difference.

Data sources:

a. US Election 2020 Tweets. Oct 15th 2020 - Nov 8th 2020, 1.72M Tweets (two dataset inside)
b. US Election 2020 Race to Presidential Election 2020 by County

The data source a will allow me to answer question 1 and question 2.

The data source b will allow me to answer question 3.

I may need additional information to help me do the visualization, such as some important date in 2020 US Election, the final vote for the two people.

Description of data sources:

- Tabular data: Two separate tabular data. Those two datasets are tweets about two people during the US 2020 Election time

(https://www.kaggle.com/manchunhui/us-election-2020-tweets)

1. donaldtrump: related discussion tweets about Donald Trump

csv format
21 columns, 971121 rows

2. joebiden : related discussion tweets about Joe Biden

csv format
21 columns, 777042 rows

- Tabular data: president_county_candidate.csv. This dataset contains the state votes for all candidates.

(https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv)
csv format
6 columns, 3007 unique values