CSE 5544: Project Proposal

Ron Davies (davies.404@osu.edu) & Logan Frank (frank.580@osu.edu)

No group members are graduating this semester.

12/09/2020

Title: Visualizing the U.S. Presidential Election

Introduction: The United States (U.S.) presidential election (including the debates and months leading up to the election) is a major event that occurs every 4 years. Arguably, this year was one of the most important and influential elections in recent history. In each election, there is ample amounts of data produced from the debates, marketing campaigns, Twitter, Google searches, and more. With so much data being produced from these sources, it is hard to ingest all of this data and no adequate visualization dashboard exists for ingesting it. It is important that we be able to visualize all of this data and it can often reveal insights about the candidates that the media does not choose to report.

Thus, we aim to create a web-based dashboard that presents factual information from a variety of data sources through easy-to-understand visualizations of election-related data. The time range of data we plan to visualize will start with the day the primary candidates for the Democratic and Republican parties were decided until 11/17/2020 (2 weeks after Election Day). Although the 2020 election has already been decided, the dashboard we plan to create could be easily modified to work using streaming data for the next presidential election, providing real-time information as data is updated and provided and giving voters information beyond media outlets to help them make an educated decision on who they will vote for.

Datasets: Our data came from the three sources listed below. The Kaggle dataset came with all the necessary files to complete our visualizations while the Google Trends and Twitter data required manual querying to retrieve the data. Our final data is available at this Google Drive link and through the link in the README file of our project (which goes to the Google Drive).

- Kaggle US Election 2020 - Presidential Debates
- Google Trends (different keywords were searched and their respective csv was downloaded)
- Twitter data using the Twitter API

Data Preprocessing: We employed various methods for preprocessing each of our datasets. For each dataset, the exact preprocessing used to parse the original data into a format useful for visualization will be described.

Word Clouds. As this data was very noisy and contained many inconsistency and faults, much of the preprocessing done was to address these issues. First, I made sure that the names of the speakers (disregarding the moderators) was consistent amongst the debate csv files. Then, the beginning time for when a speaker spoke reset in the middle of the csv (e.g., one entry started at '33:00' and the next started at '00:00' even though it occurred immediately after the other). This involved adding the previous time to the incorrect one, and adjusting for seconds over 60, etc.

Spectrograms. In order to generate these, we used pyplot to compute the Fourier transforms and generate the figure. Higher frequencies were excluded during the figure generation process to reduce the noise in the output. Due to computational constraints, each of these had to be generated and saved for access later.

Google Trends. Related words csv files were retrieved from the Google Trends website. I manually parsed the csv files to select only a small amount of keywords that were very relevant to the overall topic. From this parsed csv, I used pytrends to automatically query Google Trends for the interest over 2020 for each of the related words. This gave a DataFrame that I was able to parse through and combine all the DataFrames for a given general topic. Each DataFrame would be processed further during event handlers by filtering by time and related word, and then the columns would be summed for each row and normalized to get an interest metric that represents all selected related words.

Detailed Analytical Questions:

1. What vocabulary is most prevalent for each candidate during a given time segment?
2. How clearly is each candidate speaking during their allotted time?
3. How easy is it to distinguish between multiple voices when the candidates talk over each other or the moderator?
4. How does the general United States population's interest change over time in regards to various political topics?

Visualization Requirements:

1. We need a visualization that can show off which words are used with a greater frequency than others and not take up a large amount of space like a Zipf's law graph would has a bar chart.
2. To see how clearly candidates are speaking, we should provide a different way to visualize the audio to reduce the biases that individuals bring with them.
3. When candidates are speaking over each other, we need to see whether they have shifted their pitch or volume from their usual pattern with the intention of being heard over the other.
4. Answering this question requires several options for filtering data to touch on select topics that were poignant throughout this election year. In particular, there need to be clear trends displayed in order to provide an interpretation of the how the population's interests have changed over this sensitive time period.

Visualization Design:

1. Here we used a word cloud where the largest words represent the most frequently used word by each candidate for a given time period in the debate. The marks are the words themselves allowing for our users to gain an immediate understanding of the vocabulary preferences of each candidate. Each plot is laid out side by side to simplify comparisons between each word cloud. Additionally, we provide interaction in the form of a slider which enables our users to see the specific changes in vocabulary usage each candidate applied during their answer to one or multiple questions.
2. We generated the spectrograms for every 30 seconds of the debate. The color scheme was set to shades of gray with lighter grays and white meaning greater intensity. We provide a slider for times in the

debate so they may select the time segment they would like to analyze more closely. Having the ability to scan through the debate to different segments provides the user with the ability to find times where only one of the candidates is speaking. This is a vital component to include particularly since the candidates often speak over one another.

3. This question is answered with the same spectrograms as discussed above. The multipurpose applicability of the spectrograms is one of its key values that it brings to our dashboard. Unlike the previous question, to answer this one, the user will seek out moments in the debate where heated exchanges occur or where the candidates are deliberately talking over one another.

4. We provide several levels of interaction and a line graph to show the change in term usage over time. The main drop down list allows users the choose between highly relevant terms from this election year. Additionally, we provide some related search phrases to enable a more specific view on what has been searched for under each of the major topics. Our line graph is simple and interactive. It allows our users to zoom in on any time period they choose and even save copies of the figures if they would like to. Making this figure highly interactive is the most important part for answering our analytical questions because search terms and trends can have highly complex patterns over time.

Software: We used exclusively Python with the Dash and Plotly packages for creating our dashboard and its associated visualizations. To assist with creating the visualizations, we used the following Python packages: wordcloud, tweepy, pydub, matplotlib, ffmpeg, scipy, numpy, and pandas.
Here is how to run it:

1. Download all of the files and be sure to visit Logan Frank's GitHub to get the data folders.

2. All python files should be on the same level in a folder and the data folder should be listed there as well.

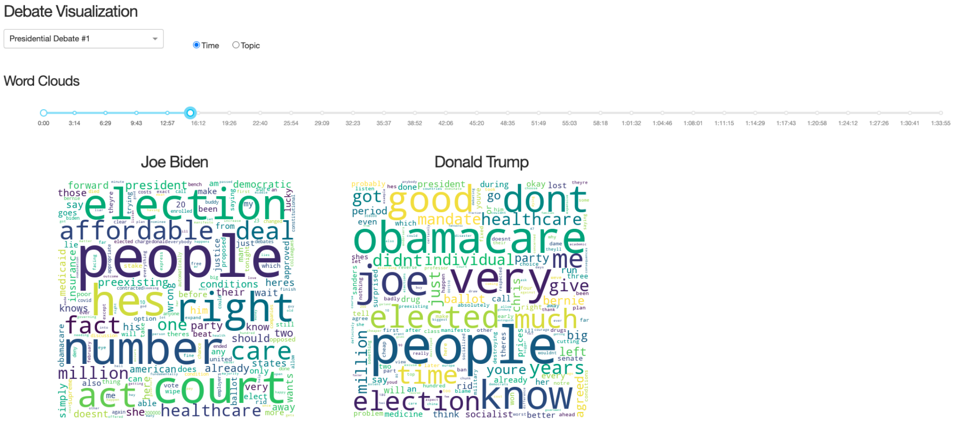
3. Open and run the dashboard.py file

4. If your dashboard has not appeared in an open browser, visit this link to connect to one of your local ports where it should be connected.

5. The dashboard should be fully interactive in most browsers; however, we only tested it in google chrome.

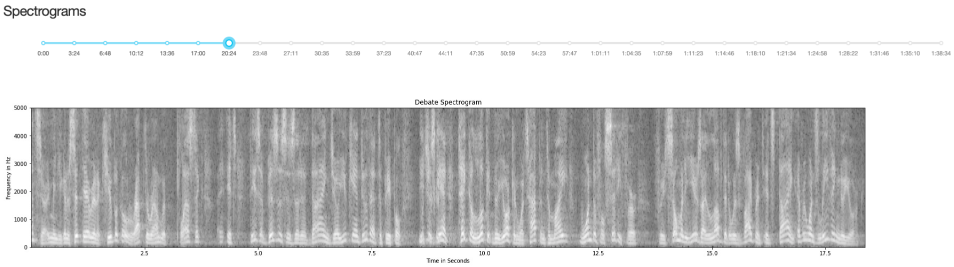
Results and Evaluation:

Question 1 Our figure provides a direct answer to this question in a single glance after a set time period has been selected. For example, during the first 15 minutes of the debate, President Trump referred to Joe a significant number of times whereas Presidential Nominee Joe Biden did not refer to his debate opponent beyond the use of "he's" in a substantial way. Our users can infer from this information what they would like based on their political affiliations. Not only is this visualization valuable for answering our first question, but also it is highly intuitive to interpret for even the most casual user.

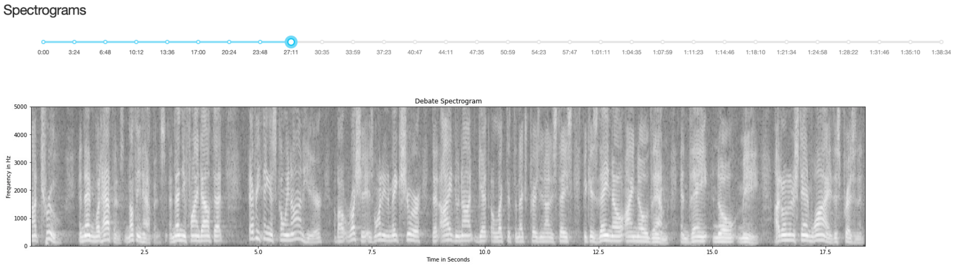


Question 2 In the opening remarks to the first question asked for the 1st Presidential debate, there is a marked absence of visually apparent vocal inflection delivered by President Trump whereas his counterpart used it only once to deliver a specific remark. This is something that can easily slip the attention of a voter that is trying to listen along specifically for the sake of content or entertainment even though it may be a compelling feature of speech. One may make the argument that the lack of specific, directed emphasis on particular points enables one of the debaters to appeal to a more general population. A figure like this allows our users to investigate these differences for themselves without being exposed to layers of rhetoric. That is one of the key reasons that this figure is essential for voters in the next major election. This top image below shows Trump's vocal response to answering a question about Dr. Fauci, often maintaining a weak vocal intensity and producing light harmonics. The bottom image below shows Biden's vocal response when refuting a claim made about him retrieving money from foreign countries. For Biden, you can see the stacked harmonics providing a stark contrast to Trump's response to a difficult question.

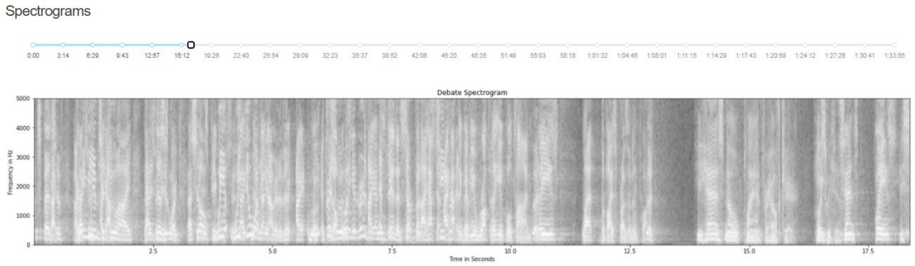
Trump



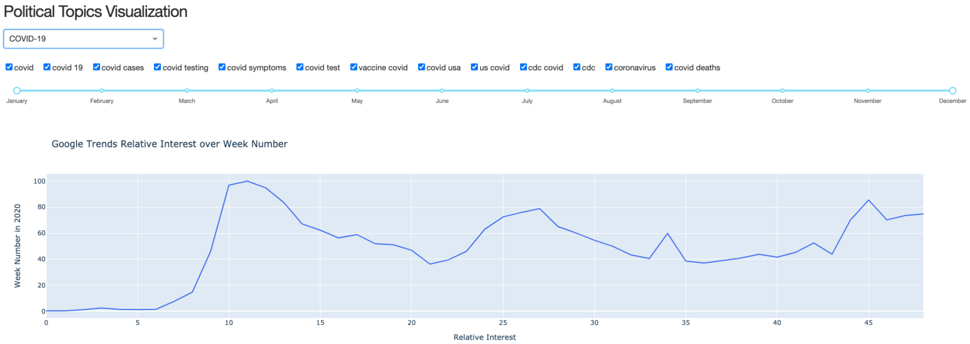
Biden



Question 3 The spectrogram does an excellent job of showing how noisy the debate got when all 3 speakers were talking at the same time around the 18th minute in the debate. The figure is almost entirely marked up with noise and there are no immediately obvious pauses for breath when each individual is trying to speak over the others. We can also see that there are more heavily stacked harmonics and more vocal inflections occurring in that brief amount of time relative to the parts of the debate where only one person was speaking. This figure is essential even during these moments of intense energy at the debates because of the unique, grounded perspective it provides during chaotic moments. It is clear to see that we can understand these individuals even as they talk over each other, but it is definitely much harder to do so. The increased number of harmonics suggests that the speakers are likely aware that this is happening. To counteract the chance that they may not be fully heard, they add more energy and enunciate more clearly which is what creates these structures in the spectrogram figures.


Spectrograms

Question 4 Users can select one of the different political topics using the dropdown menu, then within each political topic, users can filter out different subtopics for determining what contributes to the interest metric. Using the visualization, we can see that COVID related Google searches tended to increase when there were spikes in cases in the United States. For economy, many people were worried and searching the internet frantically towards the beginning of COVID around the time of lockdowns. Racial equality related searches spiked largely around the death of George Floyd and the subsequent riots. Many of the spikes in searches for different political topics can largely be attributed to a particular event, rather than the election itself, which seems like a reasonable expectation.

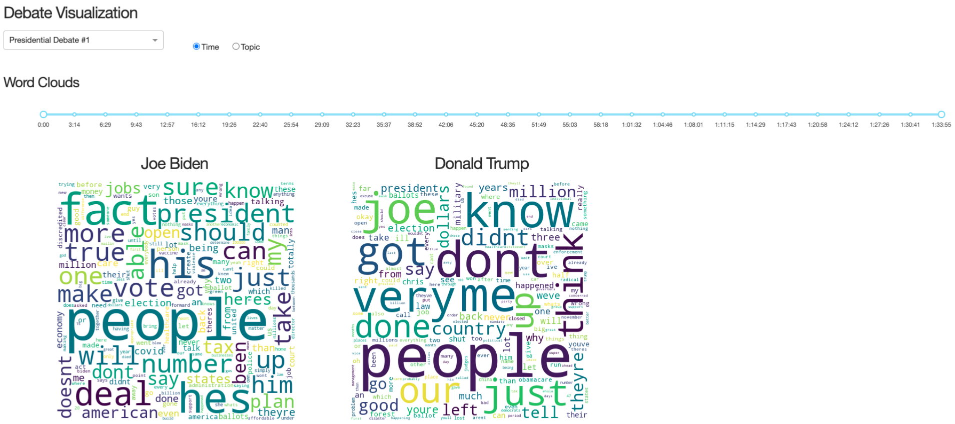

Political Topics Visualization

COVID-19

☑ covid ☑ covid 19 ☑ covid cases ☑ covid testing ☑ covid symptoms ☑ covid test ☑ vaccine covid ☑ covid usa ☑ us covid ☑ cdc covid ☑ cdc ☑ coronavirus ☑ covid deaths

Visuals: In this section, we provide visuals along with a description of the various interactions a user can have with our dashboard.

Our first visualization allows users to see what were the most frequent words a candidate used when they spoke in a debate. Users can select between the first presidential debate, the second presidential debate, and the vice presidential debate. In each debate, users can also choose to narrow the debate by time or by topic, leading to a more focused visualization. The image below shows the default view for this visualization: the first presidential debate for the full time amount.



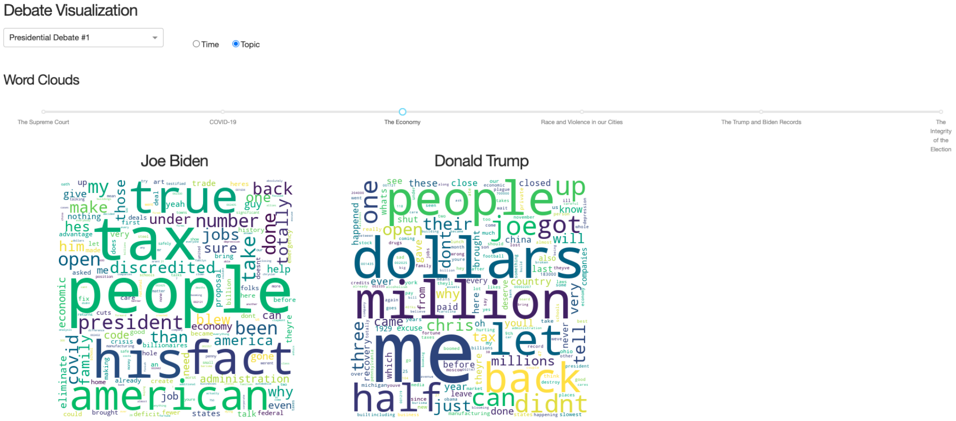
Debate Visualization

Presidential Debate #1

● Time ○ Topic

Word Clouds

As mentioned, users can switch to narrowing the debate by topic. In the image below, the debate has been narrowed to only show the words that each candidate said during the economy topic of the first presidential debate.



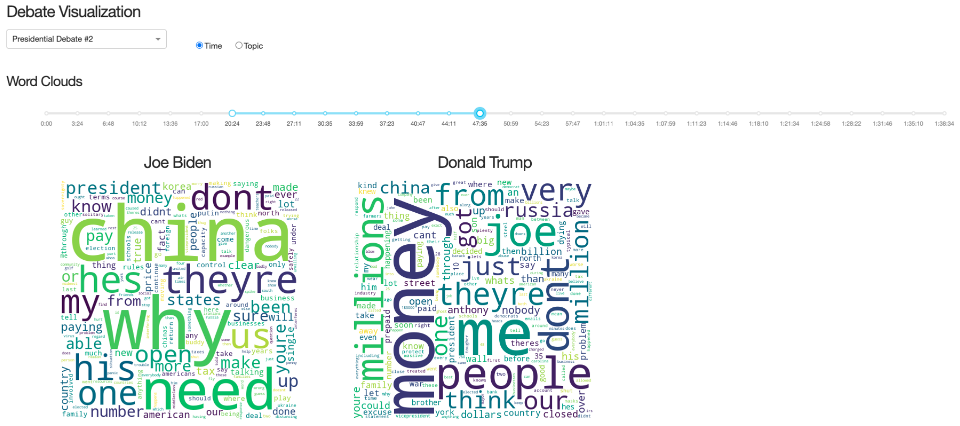
Debate Visualization

Presidential Debate #1

○ Time ● Topic

Word Clouds

As also mentioned, users can narrow the debate by time. The image below shows the words each candidate used during (roughly) the second quarter of the second presidential debate.


Debate Visualization

Presidential Debate #2

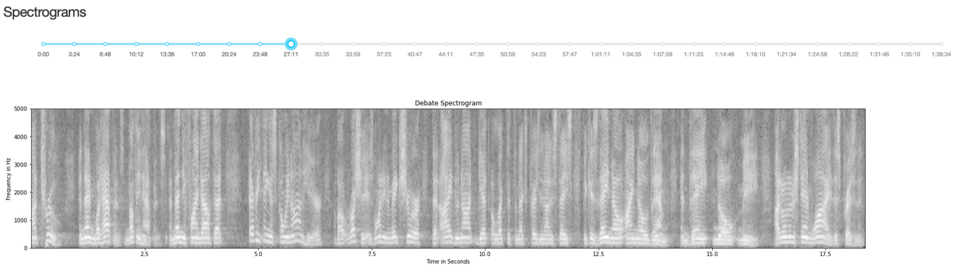
● Time ○ Topic

Word Clouds

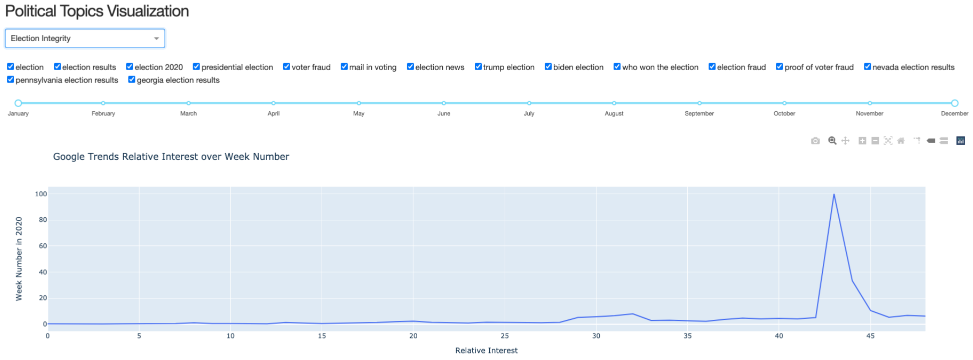
Joe Biden

Donald Trump

Our next visualization was to utilize the audio data provided in the Kaggle U.S. debates dataset to construct spectrograms. This allows a user to better analyze how a candidate speaks and particularly focus on how a candidate answers difficult questions. Users can move a slider to select a 20 second segment for visualization. The image below shows a segment where Biden was refuting claims of receiving money from foreign entities.


Spectrograms

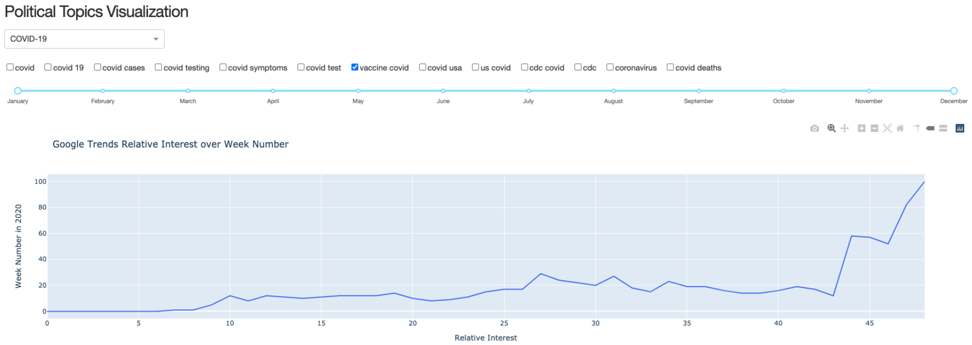
Our final visualization was to understand and view the United States population's interest in various political topics over the course of 2020. We utilized Google Trends data to complete this visualization. For each political topic, there was a list of related subtopics that users could either include or omit in the visualization. Each line is an accumulation of the relative interest metrics for each of the selected relevant subtopics (and normalized). The image below shows the trend line for the election integrity topic with all relevant subtopics included.


Political Topics Visualization

Election Integrity

☑ election ☑ election results ☑ election 2020 ☑ presidential election ☑ voter fraud ☑ mail in voting ☑ election news ☑ trump election ☑ biden election ☑ who won the election ☑ election fraud ☑ proof of voter fraud ☑ nevada election results ☑ pennsylvania election results ☑ georgia election results

Moving to the COVID political topic, from question 4 we could see that spikes in COVID related searches tended to occur with spikes in cases. If we narrow the related subtopics to just ‘vaccine covid’, we can see that people’s interest in this has only recently become more popular. This is shown in the image below.

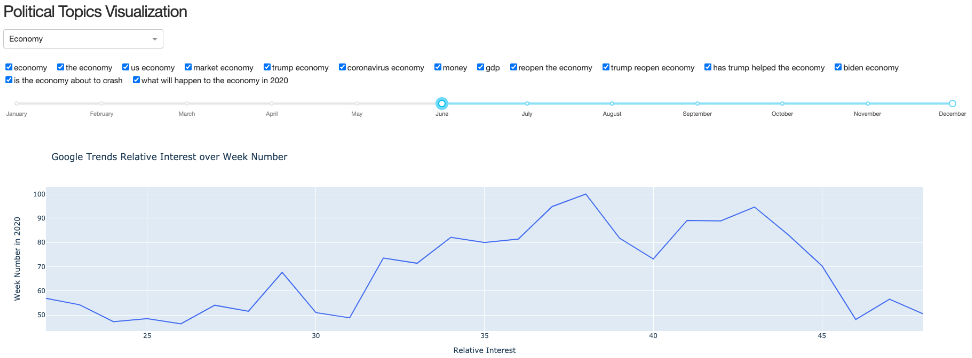


Political Topics Visualization

COVID-19

□ covid □ covid 19 □ covid cases □ covid testing □ covid symptoms □ covid test ☑ vaccine covid □ covid usa □ us covid □ cdc covid □ cdc □ coronavirus □ covid deaths

With this visualization, users can also specify a time range to analyze the data from. For the economy political topic, if we included the full time range, the relative interest reached a maximum around the start of COVID lockdown. However, we can change the time range to start in June (until December), and see the data in a more narrowed way, seeing a spike around election time. This is shown in the image below.



Political Topics Visualization

Economy

☑ economy ☑ the economy ☑ us economy ☑ market economy ☑ trump economy ☑ coronavirus economy ☑ money ☑ gdp ☑ reopen the economy ☑ trump reopen economy ☑ has trump helped the economy ☑ biden economy ☑ is the economy about to crash ☑ what will happen to the economy in 2020

Project Log:

- 11/20/2020: Project topic decided on
- 11/22/2020: Project proposal completed
- 11/29/2020: Data preprocessing code for Google Trends data completed
- 12/01/2020: Data preprocessing code for textual debate data completed
- 12/02/2020: Code for creating the Dash webpage done
- 12/02/2020: Some text and widgets were added
  - Text: Name of project + our names
  - Text: Names for the two different topics of visualizations we are creating - “Debate Visualization” and “Political Topics Visualization”

  - Widgets: In the “Debate Visualization” section, added a dropdown menu to select which debate, two radio buttons to choose whether to select a range (on a slider) based on time or by topic, and sliders to narrow down the time (or topic) to visualize from
  - Widgets: In the “Political Topics Visualization” section, added a slider to narrow down the time to visualize from (time corresponds to each week in 2020)

- 12/03/2020: Some code wrote for retrieving tweets using Twitter API
- 12/03/2020: Phase I presentation slides completed
- 12/05/2020: Spectrogram interactive plot started
- 12/06/2020: Spectrogram interactive plot failed; transitioned to preprocessing figures
- 12/07/2020: WordCloud visualization and interactions completed
- 12/08/2020: Spectrogram preprocessing completed
- 12/09/2020: Spectrogram visualization and interactions completed
- 12/09/2020: Google Trends visualization and interactions completed

Teamwork:

Ron:

- Anything related to spectrograms and audio debate data
- Project proposal
- Phase II presentation slides
- Large portion of final project report

Logan:

- Anything related to wordclouds and textual debate data
- Anything related to Google Trends data
- Project proposal
- Phase I presentation slides
- Large portion of dashboard code
- Small portion of final project report