RUN FOR PRESIDENCY

PROCESS BOOK

DATA VISUALIZATION (COM-480 @ EPFL)

Introduction

In this project, we visualized how the events leading up to and including the Capitol Riots on January 6th 2020 are reflected by the posts, comments and users on Parler. The social network became an echo chamber for right-wing extremists repeating their leaders' statements and conspiracy theories.

Our main objectives were that readers could gain insight into the content of Parler during 2020, who were its users, and how this led to the Capitol Riots.

Path to our final visualization

The path we took to our final visualization started with data analysis. As our datasets were very large, we started processing the different chunks of data in Python, and throughout the process went back and forth from Python to JavaScript and D3.js in order to arrive at our final result.

Textual Data

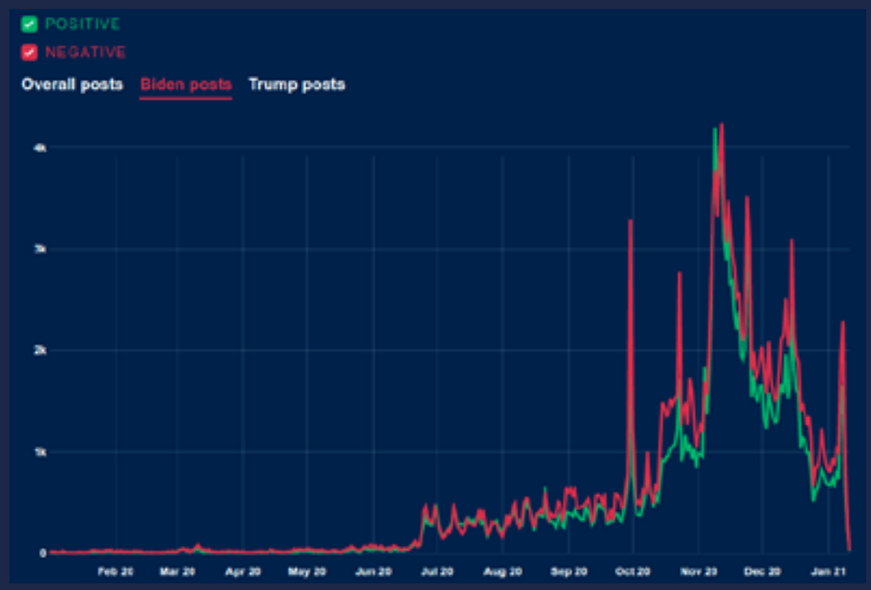
One of the most important stages of the data analysis was performing natural language processing on the Parler posts and comments. Our aim here was to extract insights to pinpoint what interesting visualizations we could create that would convey meaning related to real world events. For example, what types of users were on Parler, their political views and how influenced by political events and statements they were. This analysis was done in Python using libraries for data manipulation, visualization, NLP and statistical modelling such as: pandas, matplotlib, sklearn, nltk and gensim.

For this reason, we processed the posts to identify the most common terms, hashtags and topics. This analysis revealed that the most discussed topics on Parler were related to politics, in particular the support of president Trump. Famous phrases such as 'Make America Great Again' or 'Stop The Steal', among many others, were encoded in some of the most frequently used hashtags i.e. #maga and #stopthesteal. The top terms also included tokens such as 'trump' and 'election'.

In addition to that, we created several time-series of the numbers of posts, comments and hashtags posted per day. These plots revealed the fact that the users were most active around key events, as expected, such as Trump's campaign rallies, the 2020 elections and the Capitol assault.

Finally, we performed global and local sentiment analysis on the collection of posts and comments. The global sentiment analysis was performed on the raw posts/comment and revealed the fact that the general tone was predominantly positive, in particular around the election day when people were expressing their support for Trump. However, the only date when negative sentiment posts dominate in terms of number occurs on the 8th of January, the date corresponding to Trump's ban from Twitter and people discussing the failed Capitol assault and its repercussions.

The local sentiment analysis was performed by selecting popular terms such as 'Trump' and 'Biden' and evaluating the sentiments only on the sentences of posts/comments that contained the terms. The visualization showed that the sentiments related to 'Trump' have highly similar trends to the global ones, once again confirming that the majority of posts/comments are made by Trump supporters. On the other hand, although much fewer in number, the posts related to 'Biden' are predominantly negative, even the positive ones pointing to jokes and sarcasm as discovered following further exploration.



As a result of the sentiment analysis, we decided to include a time-series visualization of the sentiments in our project as we felt it depicted highly interesting links between real-world political events and the Parler users' feelings and opinions. The graph is composed of 3 views as shown in the figures above: one for the global sentiment, and the other 2 for the local sentiments based on the terms 'Biden' and 'Trump'. The visualization was generated using D3.js on the entire collection of Parler posts and comments.

Geolocation

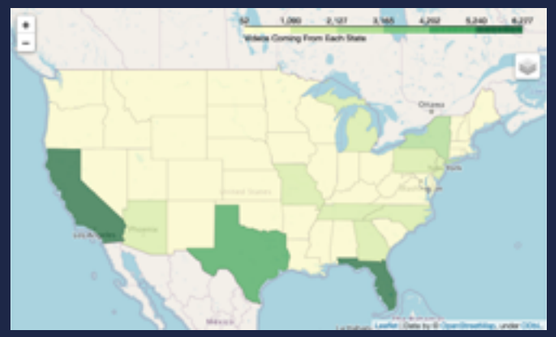
The second part of our dataset consisted of videos uploaded to Parler, as well as metadata associated with them. We started exploring the data in Python, looking at the dates and locations from which videos were uploaded the most. Unsurprisingly, most videos had been updated in 2020 or 2021 (when Parlers popularity increased). We chose to only work with videos published in 2020 or later, as that period of time is what we were interested in exploring.

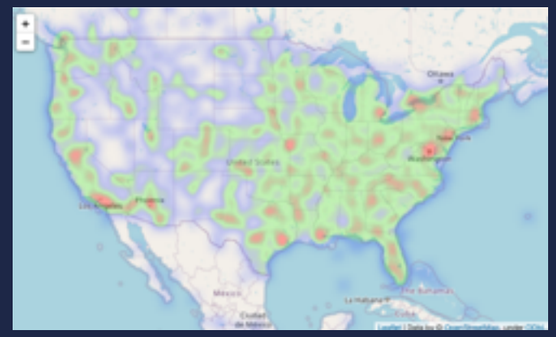
We found a clear outlier: the number of videos uploaded from Washington, D.C. the day of the riots. Any videos uploaded from Parler users who went to the riots and breached the Capitol. It's also what we had seen in other projects working on this same metadata, such as the one made by ProPublica. We noticed a second week during which many videos were uploaded from Washington: the week after the elections (from the 9th to the 15th of November). After investigation, we found that a first Trump rally occurred on the 14th of November.

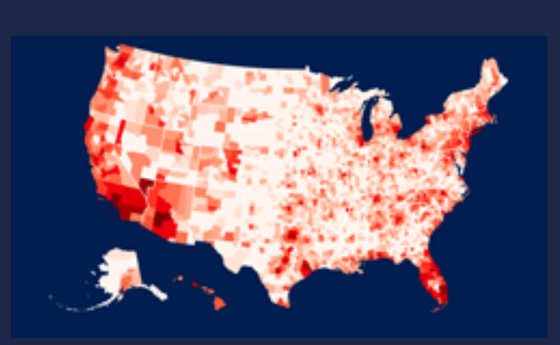
Continuing to work in Python at first, we visualized our data on maps of Washington D.C and of the United States. We started by looking at the States from which videos were uploaded the most, using a choropleth state map. This wasn't really satisfactory, as there was not enough granularity to show where the data really came from. It did give us an intuition of the areas from which users are the most active, but was really too broad. We then decided to create a heatmap, hoping that it would be able to show our outliers better. We found that while it highlighted the cities better, it still wasn't really satisfactory. There is such a large difference between the most densely packed areas and the sparsest ones that it didn't produce the results we were hoping for.

Finally, we decided to use a choropleth map, but having aggregated the data per county instead of per state. As county maps were difficult to do with Folium, we directly made it using D3.js. The biggest issue with this map is that it's difficult to see small counties from which many videos were uploaded (such as the one in which Washington D.C. is located) are difficult to see. We still decided to use this map, as we believe it shows a nice distribution of the uploads, and we dealt with the small area issue in the bubble map described below.

Below, the heatmap and choropleth Python maps are shown next to the final county choropleth map in D3.js.







Then, our attention turned to the evolution of video uploads from the start of 2020 until the riots. We want to see how the Capitol Riots stood out amongst the data. We decided to use a bubble map, so that the size of the data points are truly representative of the number of videos uploaded from that area and clearly visible for all counties.

Finally, one of our goals was to show how the metadata contained in the videos uploaded during the riots were able to help us track the movement of the crowd during the riots. We did this with a heatmap, using Python and Folium.

After talking to Professor Vuillon, he allowed us to integrate this Python map directly in our website. We aggregated the data in windows of 20 minutes, plotting for each window the location of the videos uploaded during it.

We wanted to be able to offer users an insight into the content of these videos. In an ideal world, we would have been able to store all of the videos on a server so users could select videos to watch. As this would have been very expensive, we decided to take frames from a subset of the videos (as this process was incredibly time-consuming) and assembled them into a GIF, which could be placed next to the map of the day of the riots.

Website Design and Organization

We decided to use the ReactJS framework alongside D3.js. While integrating D3.js visualizations would be more difficult (both frameworks like to modify the DOM and don't like it when other elements modify the DOM), it would help us greatly to make the website look more attractive.

Compared to our first milestones, the website changed in a few different ways. First of all, we decided that the initial visualization should be the choropleth map showing how many videos were uploaded from each county. This allows us to introduce the reader to our visualization, explaining what we aim to show in our website. It helps the user to understand where these users are located around the United States.

We then decided to place first our sentiment analysis, and then the race chart. These two visualizations aim to show the evolution of the Parler content, from the beginning of the year when users were certain Trump would win the election, to the end of the year when they believed he had won but that the election was "stolen". This part is really dominated by the conspiracy theories pushed on Parler.

Then, we append our bubble map. This intends to show the change in activity in the last months of the year, most notably the increase after the election, and the two clear outliers that are the Million Maga March and the Capitol Riots.

Finally, the map showing the locations of videos uploaded during the riots and the GIF giving a taste of the content of the videos finish our visualization. We had hoped to be able to give users more freedom in choosing the videos they wanted to look at, but were too constrained by the necessary computational resources to do so.

These changes are fairly significant since milestone 1. We realized we had not contemplated enough the way that the information should be presented, and how readers would be able to follow the website to increasingly build a better understanding of what we are presenting. Our initial idea was to have a timeline guide readers from the beginning of the year 2020, all the way through to the Capitol Riots in January 2021. Our final visualizations aren't specific to a single time of the year. They show the evolution of the data from a different perspective.

Challenges

The main challenge for us was the size of the dataset. Initially, just the Parler posts and comments were more than 30GB, and the videos themselves were too large to download to our computers. This issue kept coming up, as every time we wanted to explore new parts of the data it would take significant time to process. What really helped us was to take a random subset of the posts and comments, and perform our initial analysis on it. Once we had something we were content with, we would re-run the code on the entire dataset.

Integrating ReactJS and D3.js was a real challenge. We faced many issues, from the website being too slow because ReactJS kept reloading D3.js elements, to these elements simply not rendering, to errors we did not know how to solve. But in the end, working with ReactJS and D3.js was beneficial. It forced us to really understand how they both operated, in order for the website to look visually pleasing and be responsive.

Finally, there were certain features that we could not implement due to some of us lacking experience in ReactJS, D3.js and front-end development in general e.g. tooltips on the Sentiment Time-Series Plot.

Peer Assessment

All major decisions were made as a group, meeting regularly on Zoom to discuss the direction in which our website should go. This is true for the data we chose to extract, but also the organization of the timeline, and the way the information is displayed.

Iuliana analyzed the text data from Parler posts and comments through NLP and data manipulation tools. She explored the data through various visualizations in Python and decided to include in our final project a visualization that displays the users' sentiments and how they evolved in time. Thus, she designed and implemented the Sentiment Time-Series Plot which she created by using D3.js on JSON files that she generated in Python.

Jean handled the exploration of the interaction between the Twitter and the Parler datasets. He prepared the data for and implemented the bar chart race showing evolution of the hashtag visualization of the evolution of top hashtags on Parler. He embedded Tweets by creating a ReactJS component for tweets hovering over the visualization. He also took care of setting up the website, the design, and the color palette for the final appearance.

Niels took care of processing and displaying information related to the locations of the uploaded videos. This included cleaning and exploring the geodata available, creating map templates in Python, and implementing the maps in D3.js.

Conclusion

In this project, we visualized how the events leading up to and including the Capitol Riots on January 6th 2020 are reflected by the posts, comments and users on Parler. The social network became an echo chamber for right-wing extremists repeating their leaders' statements and conspiracy theories.

Our main objectives were that readers could gain insight into the content of Parler during 2020, who were its users, and how this led to the Capitol Riots.