CS 6630

How happy is America?

Richie Frost
Neeka Ebrahimi
Dyllon Gagnier

Process Book
https://richiefrost.github.io/dataviscourse-tweetsentiment/

Overview & Motivation

Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

Each of us works together in a lab under Feifei Li and thus each has a lot of experience in working with Twitter data. However, while we have worked towards visualizing this data, these efforts have been more focused on the systems/machine learning techniques used to analyze this data. Therefore, this project presents a unique opportunity for us to spend more time coming up with more useful ways to visualize tweet sentiment, location, and time rather than on the other research aspects. It may be a bit of a challenge to come up with useful visualizations for this data, but we each have extensive experience in this area (turning unstructured/semi-structured data into structured data).

Sentiment analysis is a specific NLP task which involves characterizing text as being either mostly positive, mostly negative, or neutral. Additionally, while basic methods provide these sentiment scores at a holistic level, more advanced models also provide sentiment scores concerning certain entities such as companies, political candidates, etc. since text will often have different sentiment scores concerning different entities. A tweet by Hillary Clinton might on one hand compliment former President Barack Obama while also criticizing the current Trump Administration. While This is still an open area of research in NLP, this project will not be focused on how best to generate these scores, but will instead mostly use sentiment analysis as a black box which we will then visualize the output of. For instance, it might be interesting to view how overall sentiment varies over the course of a political debate in order to get a better idea of how each candidate performed over the debate.

Related Work

Anything that inspired you, such as a paper, a website, visualizations we discussed in class, etc.

We really liked the interactivity of the GapMinder-related homework, as well as GapMinder itself. Their interactions are so intuitive and make you feel like you have so much control over the data that you can find all sorts of patterns you weren't expecting to see. We also love the visualizations of FiveThirtyEight. They're often political, but they do also visualize different topics like sports, economics, and culture.

Questions

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

When we started this project, the main question we wanted to answer was if there is an intuitive and informative way to visualize tweet sentiment, location, and time. Accomplishing this would allow us to determine how people in various locations feel about a given topic over a period of time. This could help answer many questions like, “How did people’s sentiment vary during the Kavanaugh hearing?” or “What is the sentiment of the population in regards to climate change, and how has that changed over the years?”, and much more. However, as our project progressed, we realized exactly what question we were trying to discover: “How happy is America?”. This is the question we ended up trying to answer with our visualization. The benefit is that we would be able to find meaning and view trends in twitter data that has been difficult to see up to this point.

Data

Source, scraping method, cleanup, etc.

The data is a large collection of tweets collected using Twitter’s public API. Tweets are notoriously short, which makes them easy to read like a human, but also quite difficult for most natural language processing methods, particularly using the built-in models available out of the box using common libraries. There are also other noise factors to consider in the data, such as the presence of special characters like emojis and different languages; for instance, while English is the most common language in the United States, Spanish is also spoken and thus might appear as the language of a tweet in the US. Furthermore, it is even possible for there to be tweets which mix multiple languages, particularly slang terms. Our machine learning model handles these unique cases with character level embeddings using Facebook Research’s FastText. For the purposes of this project, are only going to be using tweets written in English, but this could be extended to other languages in the future with our customizable data processing pipeline.

We use Twitter's streaming API to gather data with our lab's data processing pipeline, Compass. During the process of gathering tweets, we noticed that there were a significant number of tweets that didn't have any geolocation on them. For this reason, instead of using only current tweets that were streamed within the last few minutes, we needed to get tweets from our archive to supplement the total number of tweets. After that, we could filter down more easily by topic.

However, in the future, we still plan to use the live streaming method to get tweets. The only caveat with the live streaming is that geolocation may not always be available, especially when filtering on certain keywords, handles, and hashtags that don't show up very frequently on Twitter.

Cleaning up the data is mostly a matter of projecting the right attributes from the original JSON retrieved from Twitter. Each tweet is about 1 KB in size, but we only need a small subset of the dozens of attributes - namely, the text, the date the tweet was posted, the geolocation, sentiment score (which we compute), and possibly a few others, depending on the types of visualizations we'd like to do down the road.

Exploratory Data Analysis

What visualization did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

The data seems to be pretty large at about 1 KB per tweet. It may be necessary for more complex visualizations that look at aggregates to be performed on a server (either beforehand or through a live server) in order to keep data sent to the client low. Another option we have is to drop the attributes that aren't needed at the time. For instance, there is a lot of metadata about the user that we don't need that takes up a significant chunk of memory per data record. We can still keep the raw data in a database for future reference, but we only need to retain a small subset of the attributes for this visualization.

To initially look at the data, we worked on a similar project in 2016 regarding election trends for the United States Presidential Election. We visualized the voter breakdown per county and per state based on the average sentiment score of tweets in that region for Democrat and Republican candidates, and how it changed over time. There was significant data processing for that project that we won't go into here, but it showed us the power of visualizing sentiment towards different topics in the context of geographic location. Because of this, we decided to try to visualize sentiment towards different topics, not just specific political candidates, in the context of geographic location.

What we realized, after going through our data, was that getting tweets with different topics, even when filtering on specific hashtags and handles, meant getting very few tweets overall for a given topic. So few, in fact, that it didn't make sense to put them on the map anymore.

Instead, we decided to change the way we looked at sentiment by trying to figure out which states were the happiest, and which were the saddest. Most of the averages were the same overall, so we normalized by changing the domain to the minimum/maximum of the average

sentiment per state, not 0/1, as it was previously. This did show some more drastic difference in overall sentiment, as can be seen in the version 1 sentiment map in the design evolution section.

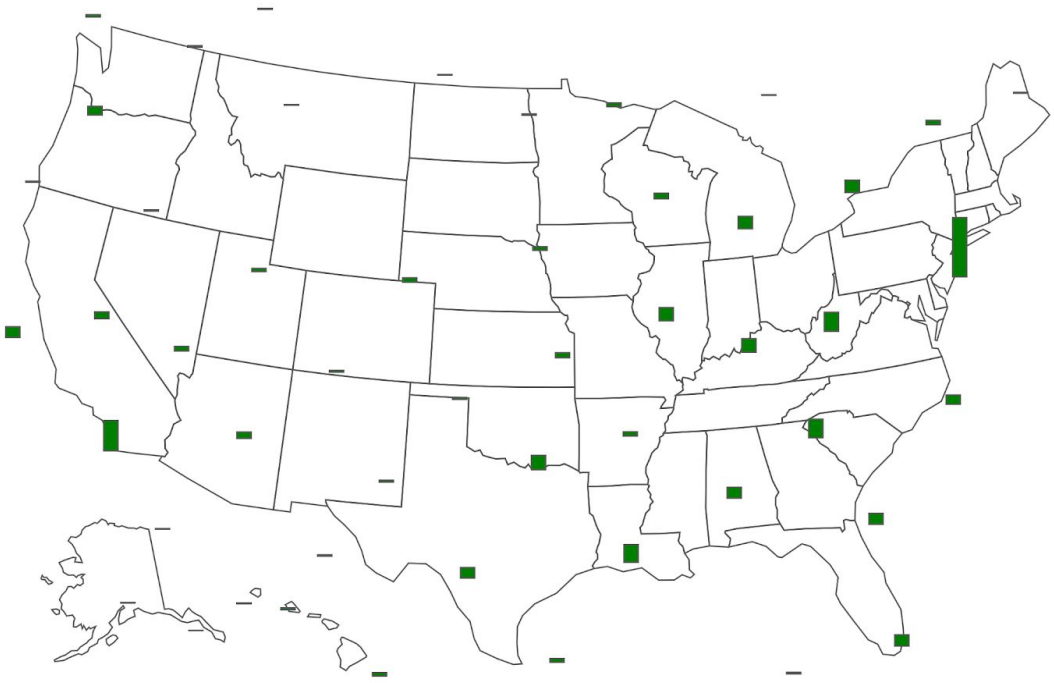
Another issue we had was that some states, such as California and Texas, had a pretty big chunk of the overall number of tweets. So we had to figure out how to normalize the data values to show a more reasonable comparison. Some of the ideas we tried were converting the actual value to the rank of that state compared to other states, a log scale, and dividing each state by the maximum value. After trying the multiple aforementioned methods, we felt that the most informative way of displaying the happiest/angriest states was to order them by rank.

Design Evolution

What are the different visualizations you considered? Justify the design decisions you made during the perceptual and design principle you learned in the course. Did you deviate from your proposal?

Bar Graph Map

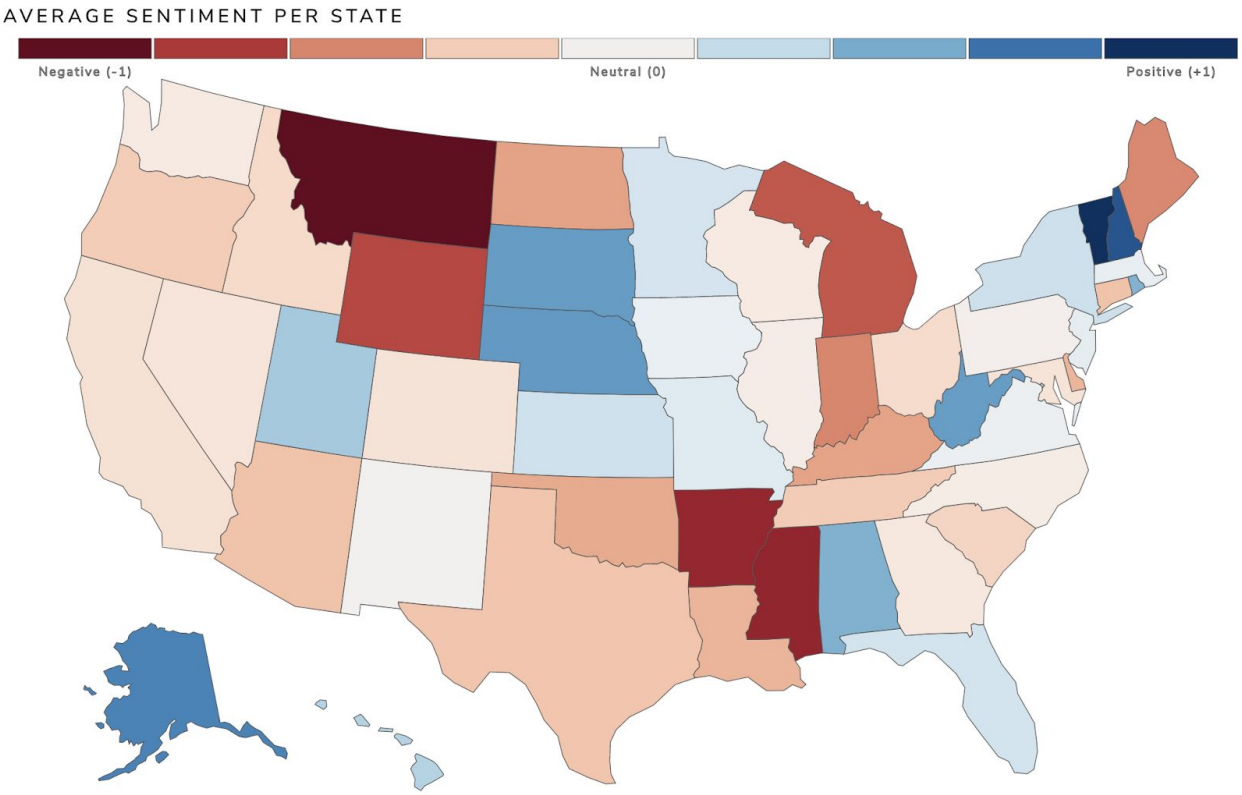
One iteration that we tried was to use length (aka mini bar charts) to visualize quantity over the geographic map. To do this, we computed clusters using the Gonzalez clustering algorithm and then showed the number of tweets from that region using the height of a rectangle.

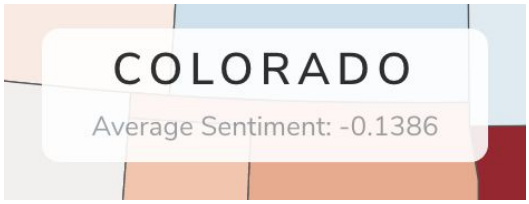


Note that this particular iteration does not have marks for the sentiment, though that could be done through either a quantized scale or through saturation. One flaw with this particular clustering is that it does not provide very good resolution in high-density areas and is difficult to view very small regions.

Sentiment Map

After the bar chart idea, we decided to make some choropleth maps to show sentiment per state in different ways. Our version 1 sentiment map took the average sentiment per state, normalized the values by minimum and maximum average sentiment per state, and colored states with a diverging color scale (red to blue). The results were surprising - Alaska is a really really happy state, as is Vermont (the happiest state), but states like Montana and Mississippi were relatively miserable. The map can be seen below:

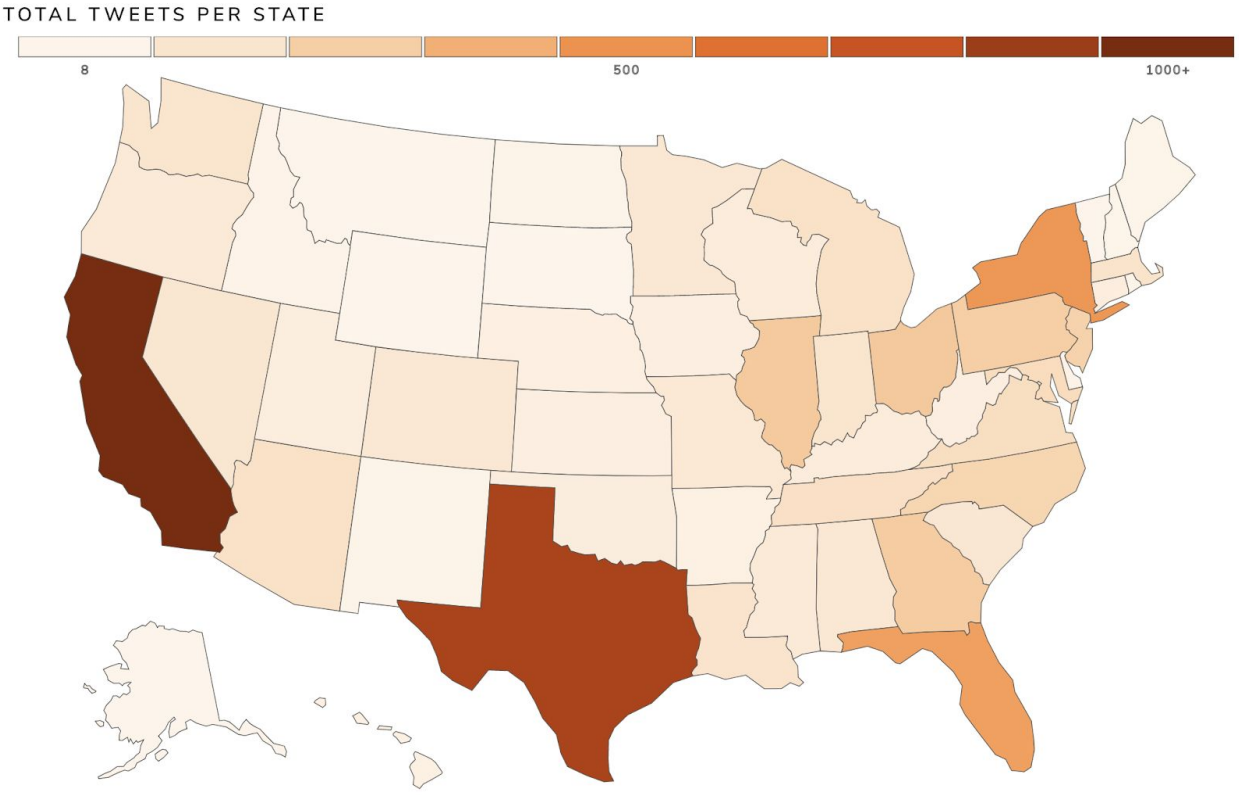




Once we had this map complete, we added a tooltip on hover to show the name of the state and the actual average sentiment value.

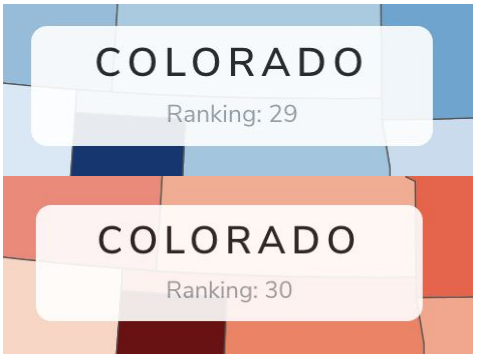
Total Tweets Map

This map ended up being an evolution from our original bar graph map showing the number of tweets per state. We essentially combined the bar graph map data with the same style as the sentiment map. However, for this map, we used a sequential color scheme with just oranges rather than a diverging color scheme. This map also has the same tooltip as the sentiment map showing the total number of tweets per state as you hover over each state.

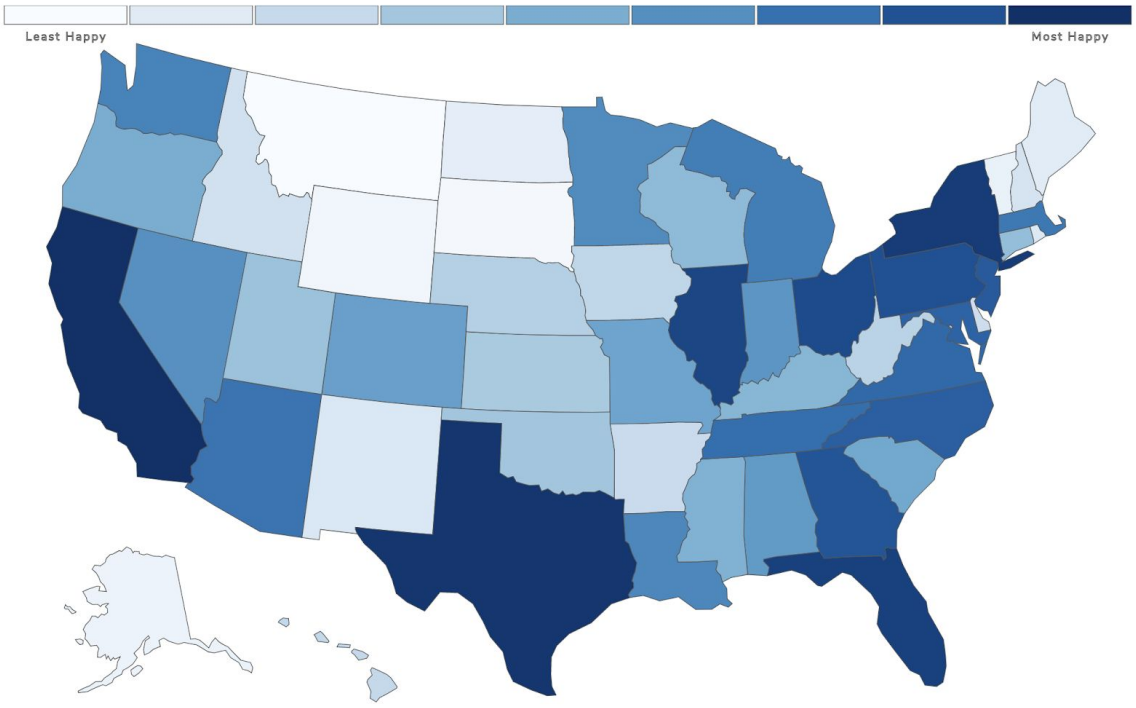


Happiest & Angriest States

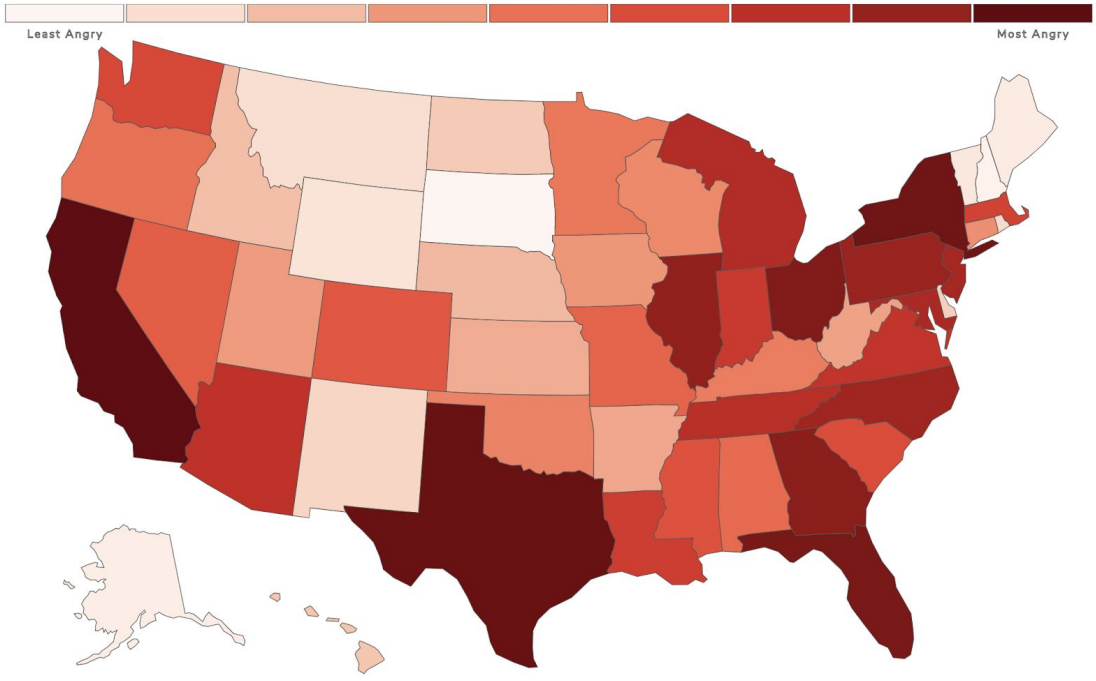
For our final two maps, we decided to display the ranking of the happiest and angriest states. They have the same style as the previous two maps with the same tooltip hovering mechanism. The tooltip displays the ranking of the state in terms of happiest/angriest. The higher the ranking, the happier/angrier the state. The ranking is based on the total number of happy/angry tweets per state (not the average).



HAPPIEST RANKED STATES BASED ON NUMBER OF HAPPY TWEETS



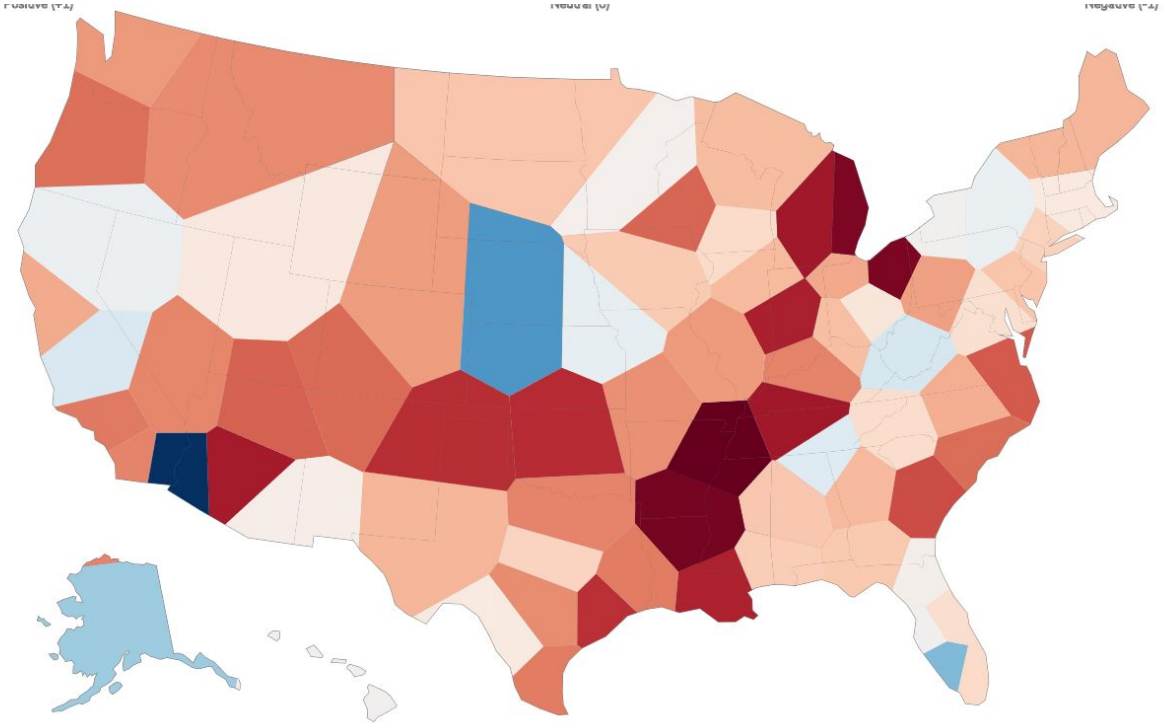
ANGRIEST RANKED STATES



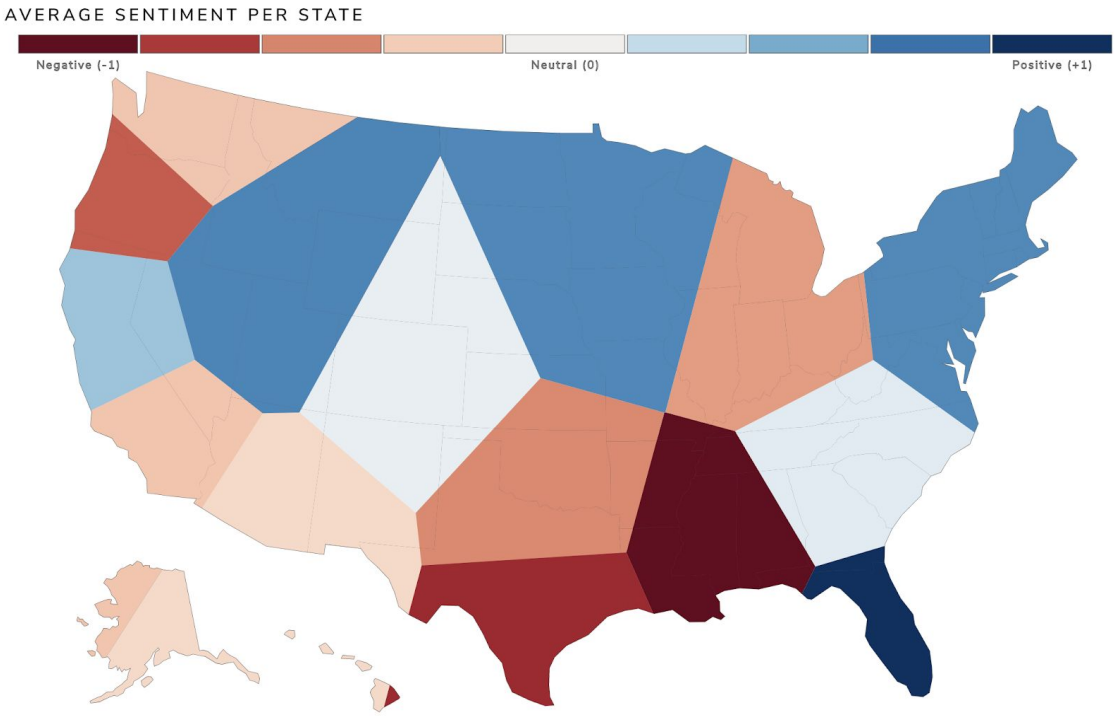
Clustering Tweets Into Regions

We thought it might be interesting to see what overall geographic sentiment looks like when using a clustering algorithm to find regions instead of aggregating by state alone.

The clusters were generated by running k-means++ to find good geographic centers. This clustering algorithm is more appropriate than something like k-max (Gonzalez) because cluster location is less prone to fitting outliers instead of the actual density of tweets.



We initially thought that a large number of clusters would be more helpful. However, this seems to have been a bit messy and hard to interpret. This also took a couple of seconds to compute due to the time complexity of k-means++. We reduced the number of clusters to 15 which ended up being much more interesting since it gives one the ability to reason about larger geographic regions more easily than is possible either with many clusters or with state grouping



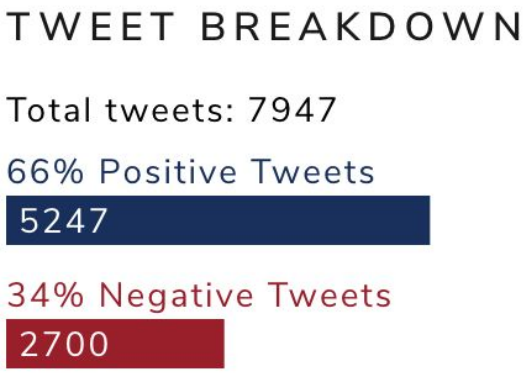
.

This is a randomized algorithm, so clusterings are slightly different each page load, but this graphic allows users to learn some interesting things not as apparent under the other approaches. For instance, the south seems to very positive while the midwest is very negative.

As a final note, it would be interesting research question to consider specialized clustering algorithms for visualizations of this type. k-means++ does a relatively good job at equalizing density in the clusters and Gonzalez works well for ensuring clusters are geometrically small. However, an ideal clustering algorithm for clustering data like ours might be to minimize the variance of sentiment in the clusters. This would be equivalent to generating something like a variance optimal histogram, but with additional constraints imposed by the geography of points.

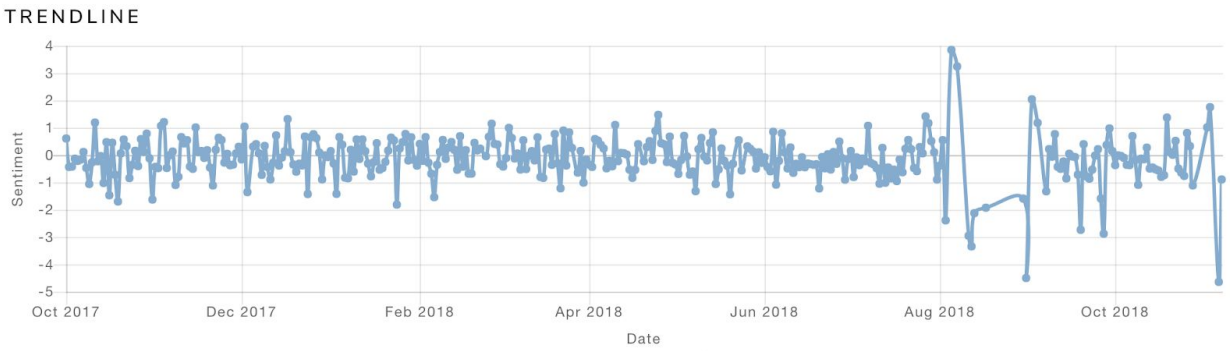
Tweet Breakdown

We added a simple breakdown of the Twitter data so that users can get a quick overview of the data. We used bar charts similar to the popular vote bar charts in homework 6. This section simply shows the total number of tweets and then has a bar graph with the percentage of positive tweets and the percentage of negative tweets. Inside of each bar is the number of tweets associated with that sentiment.

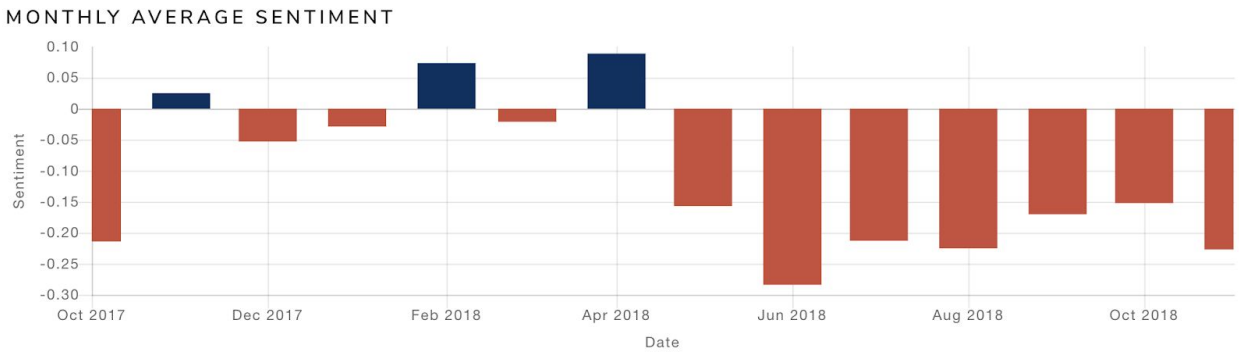


Tweet sentiment timeline

Since we have the time data for each tweet, we wanted to display a timeline of the average sentiment of tweets for each day. Initially, we did a line chart with the date as the x-axis and average sentiment of tweets for each day on the y-axis. The graph ended up like this:



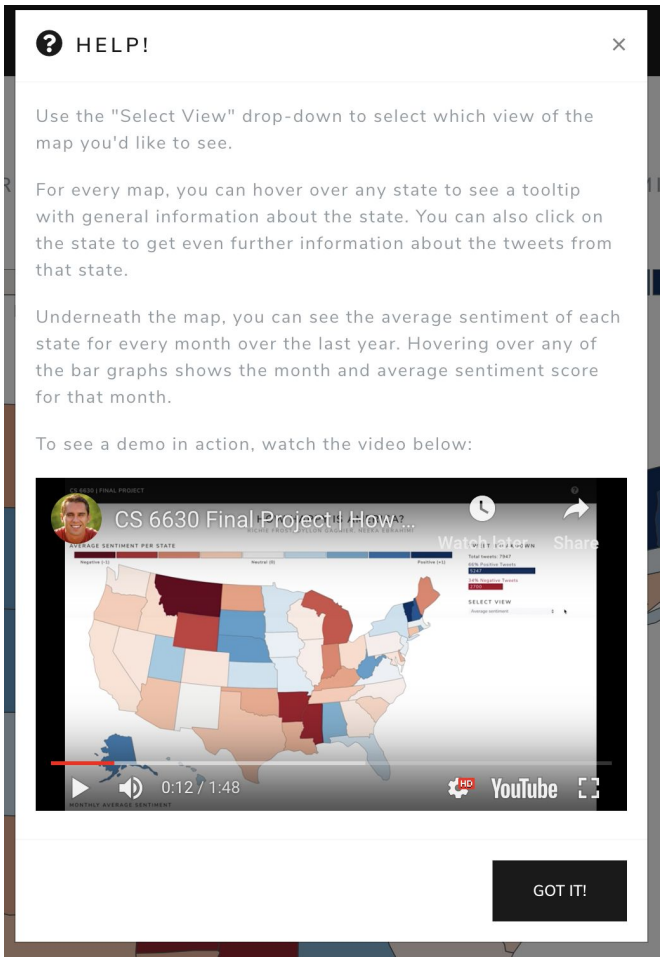
This graph is too cluttered to get good insight from it. Therefore, we decided to get the average sentiment over each month rather than each day and display it as a bar graph. This made the data much more readable and informative. We ended with the following graph:



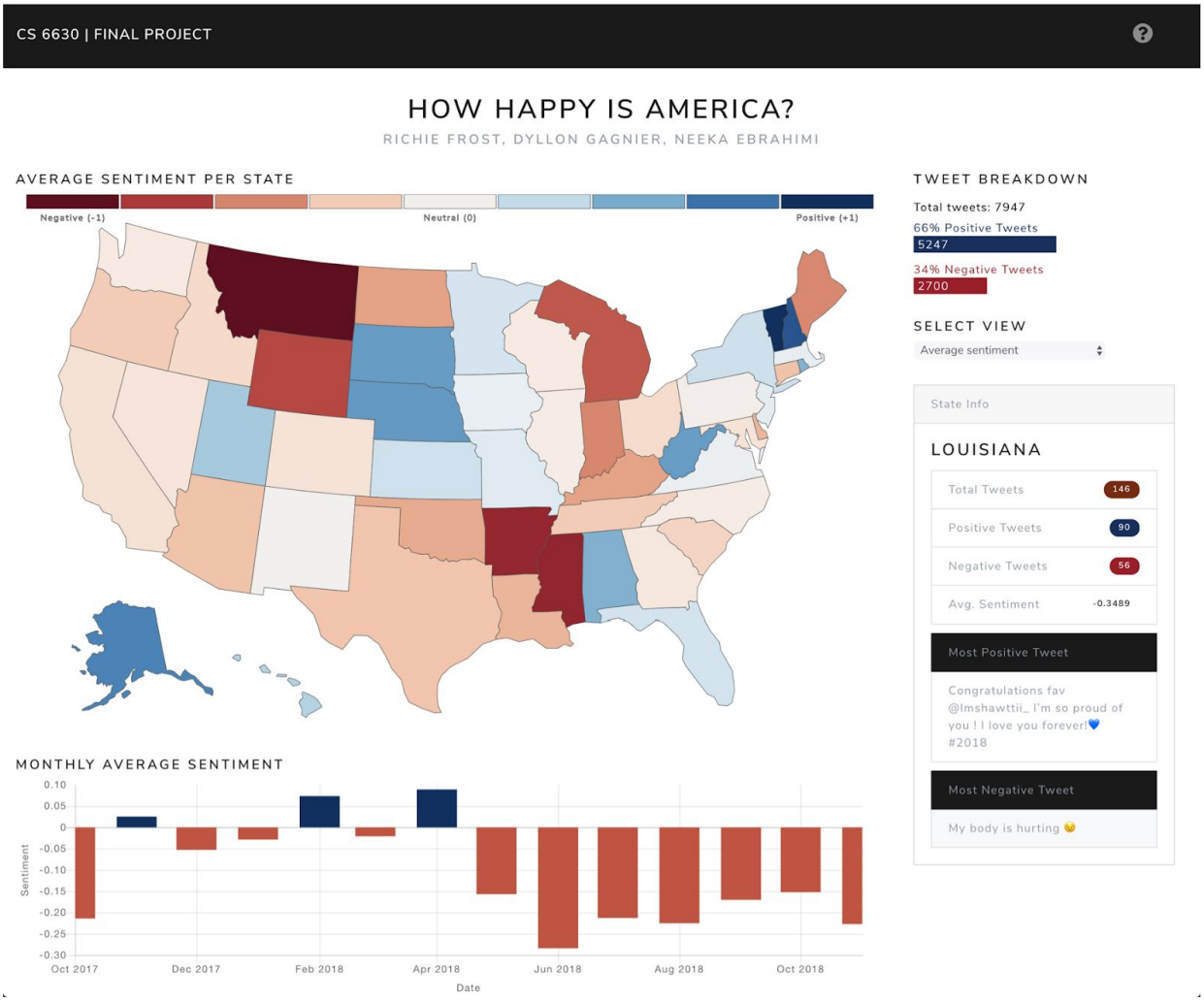
Help Popup

We added a very simple and short help popup. Clicking the question mark in the upper right corner of the page gives general instructions on how to use the site.

This is also where we've embedded the video for our project, as can be seen in the image below.



Final Display:

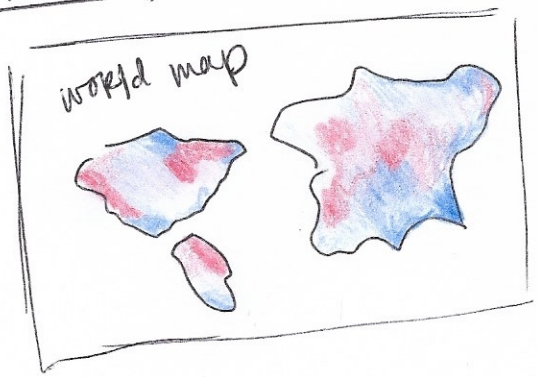


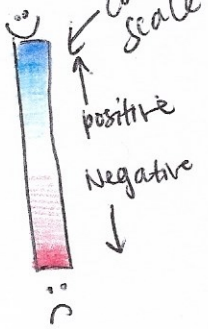
Changes From Proposal

The following five pages are the sketches from our Five Sheet Design Methodology in our proposal. Our design has differed a bit from the original final design we proposed. One change was that we used a map of the United States rather than a world map. During the data collection phase, we realized that it would be more interesting and easier to find trends in a narrower region rather than the whole world. We also removed the zoom section of the page since that was no longer necessary without such a large map. In place of that, we used hovering and clicking functionalities to give more detailed information about each state in question. Lastly, we used a bar chart instead of a line graph for the display of sentiment over time. As previously stated, we tried the line graph and felt that it was too cluttered to gain any real insight from it. Using the bar chart showed us stronger, more interesting trends.

BRAINSTORM

world map colored by sentiment score





← break down by country or state?

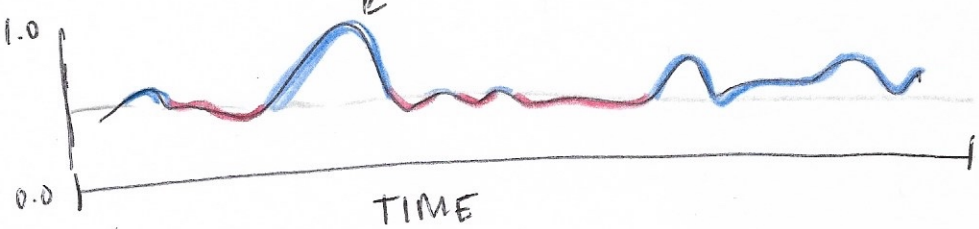
Or single overall heatmap?

* makes use of location, but can't see values shift over time

combination of both?

Trend line

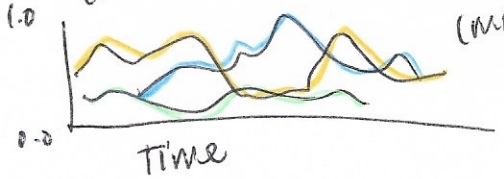
trend for overall sentiment across the world or for a selected region



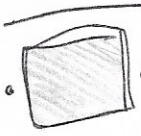
* makes use of time, but not location

↳ increment by? (hours, seconds, day, year?)

could break up over continents or countries (might get cluttered)

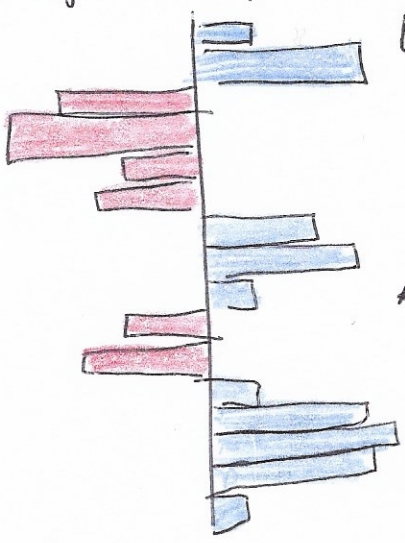


other additions

-  ← add brushing over world map that corresponds with the trend line.
- add tooltips on trendline to get more detailed info
- add brushing on ~~trend~~ trendline

stacked bars

negative positive



← stacked bar chart. Each bar is a single country's sentiment score

* not sure how to show time with this?

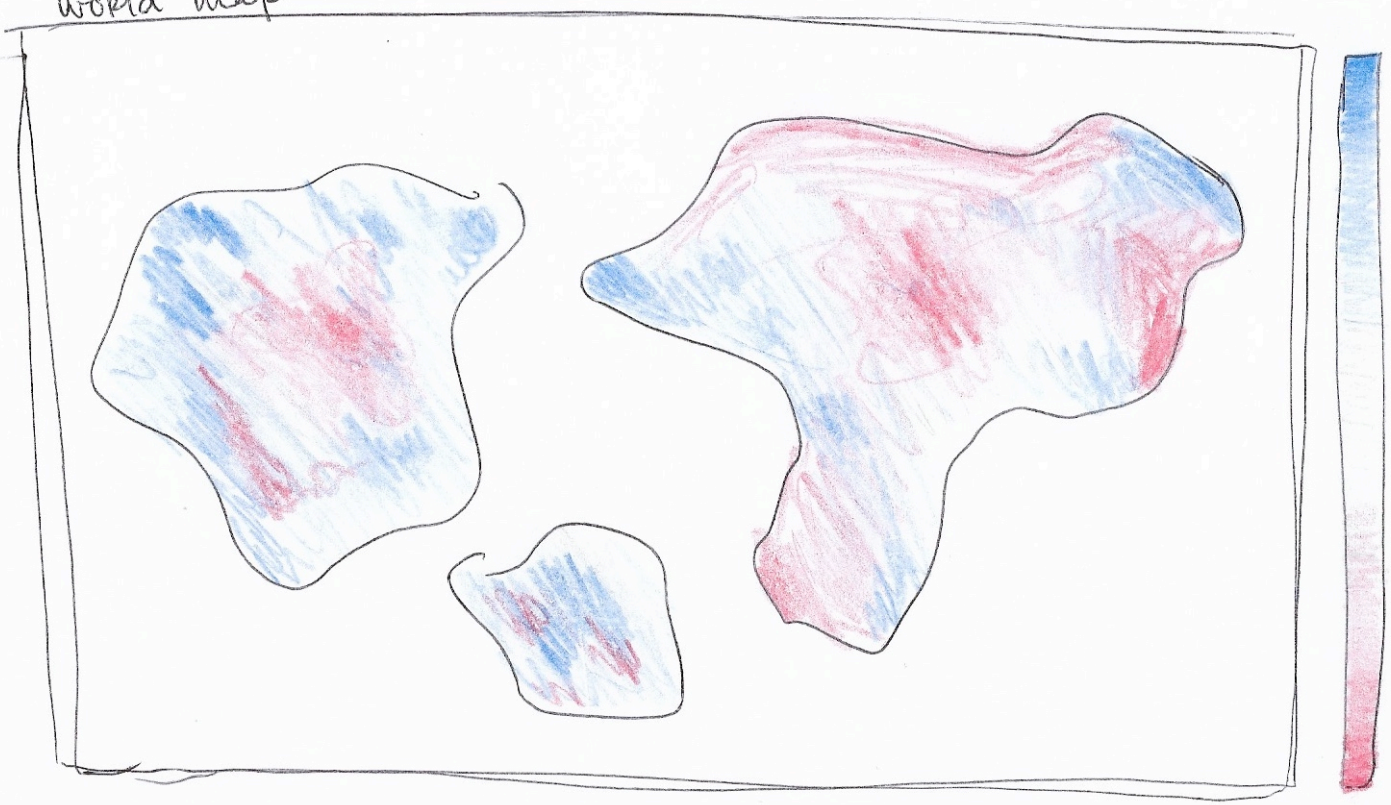
- doesn't make use of geotags

DESIGN 1

TITLE

authors / description

world map

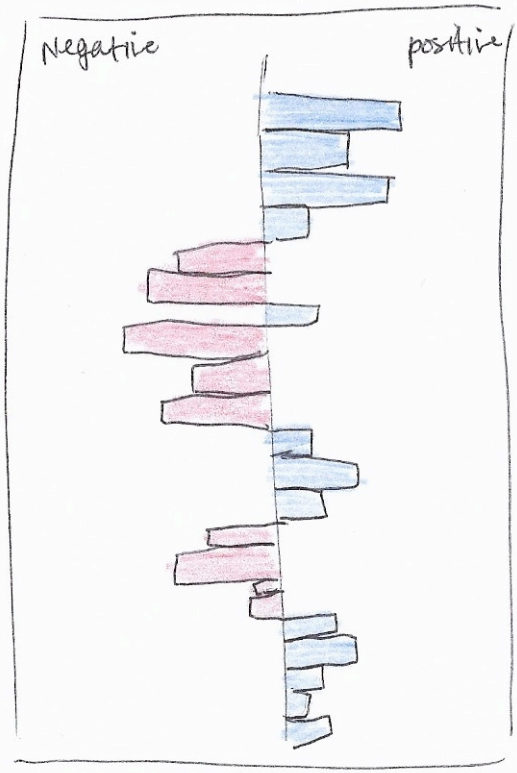


Overall Sentiment

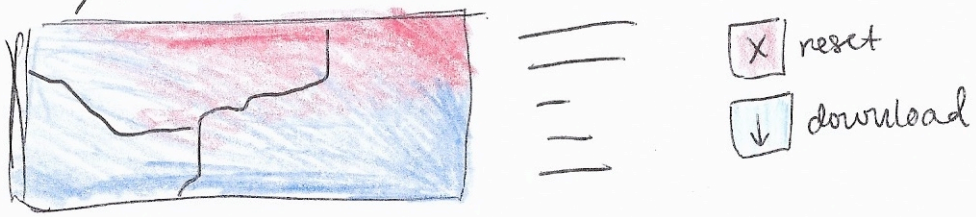
20% negative

80% positive

Stacked Bars



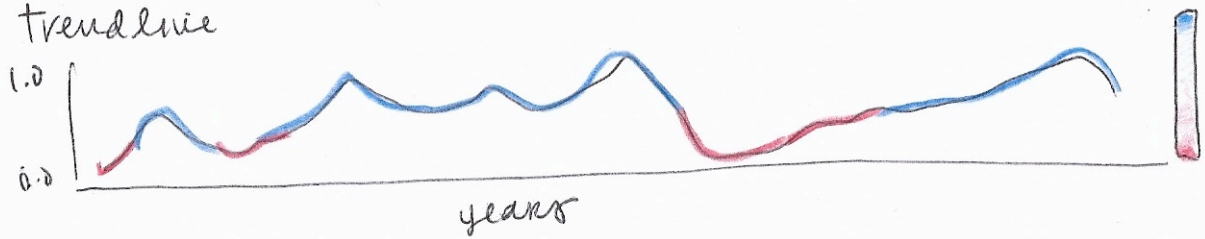
zoom / selection



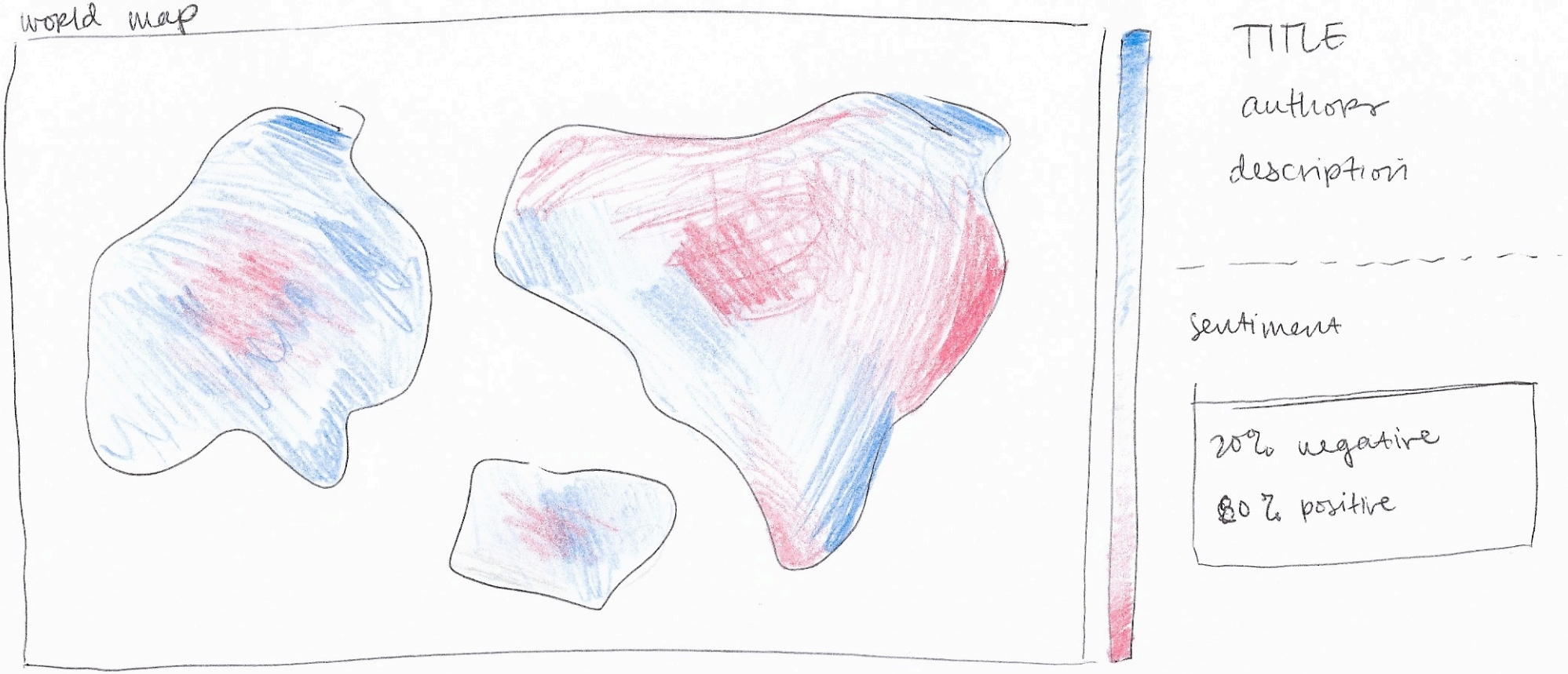
reset

download

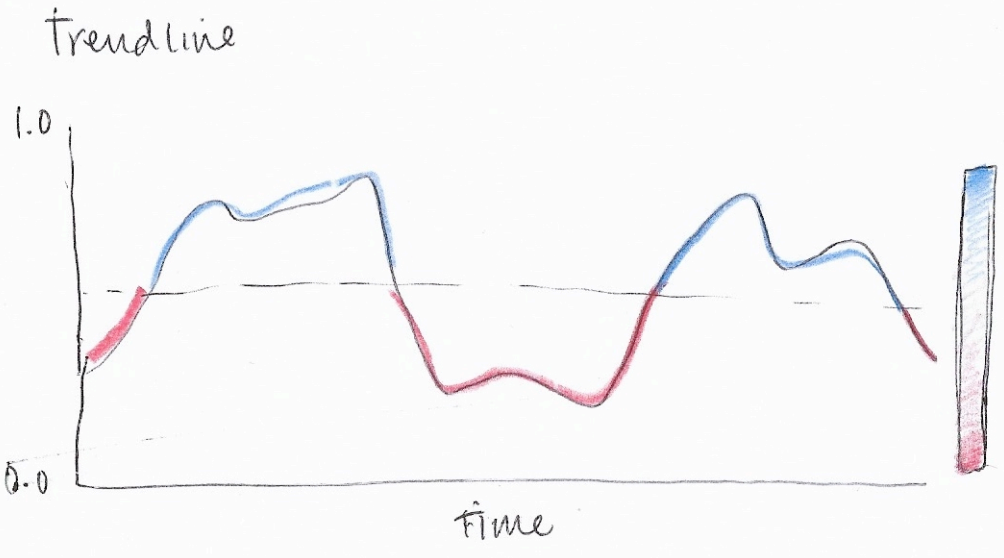
trendline



DESIGN 2


world map
TITLE
authors
description
sentiment
20% negative
80% positive

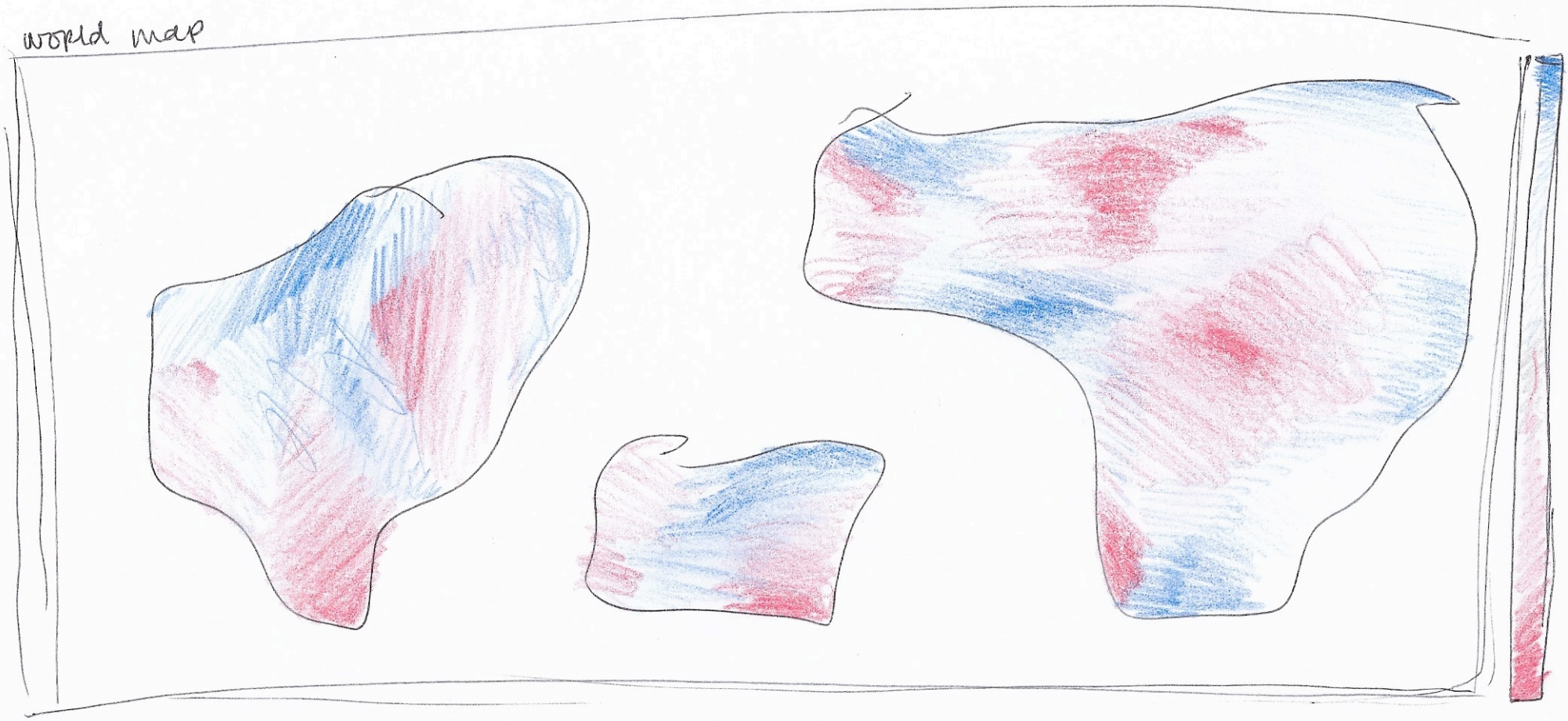

zoom/selection
reset
download


trendline
1.0
0.0
time

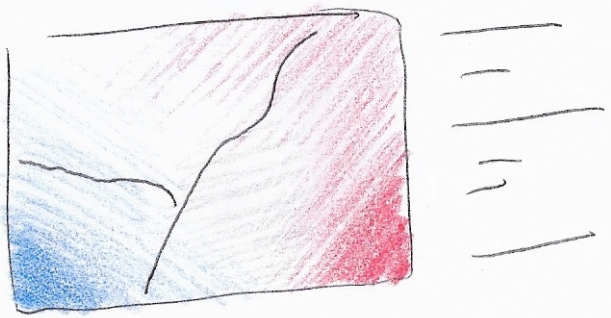
DESIGN 3

TITLE

author / description

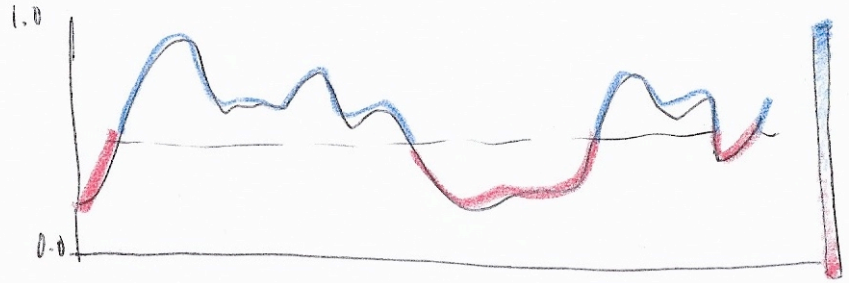


Zoom / selection



reset download

Trendline



Sentiment

20% positive

80% negative

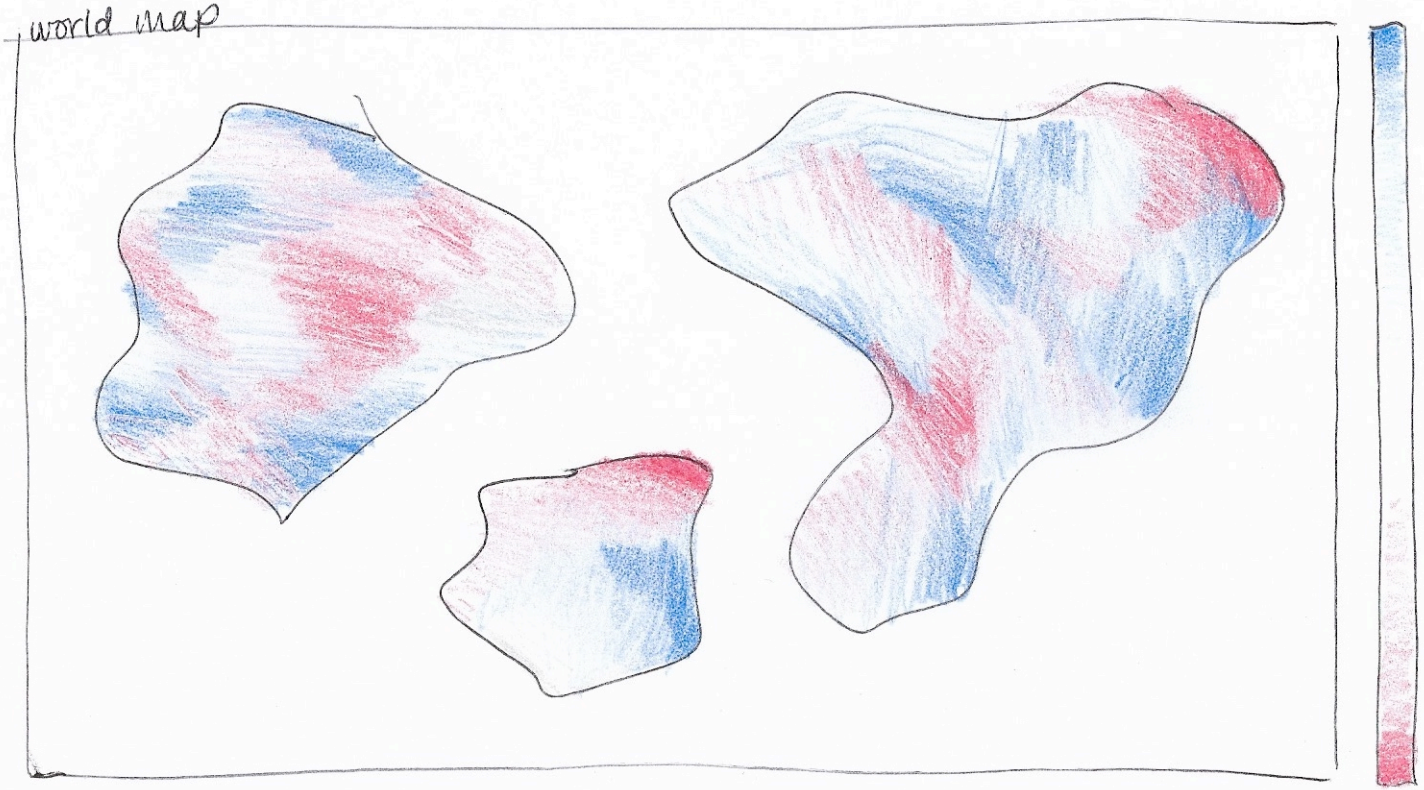
FINAL DESIGN

TITLE

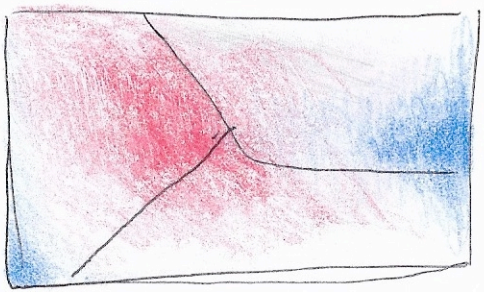
Authors
short description

help

world map



Zoom / selection



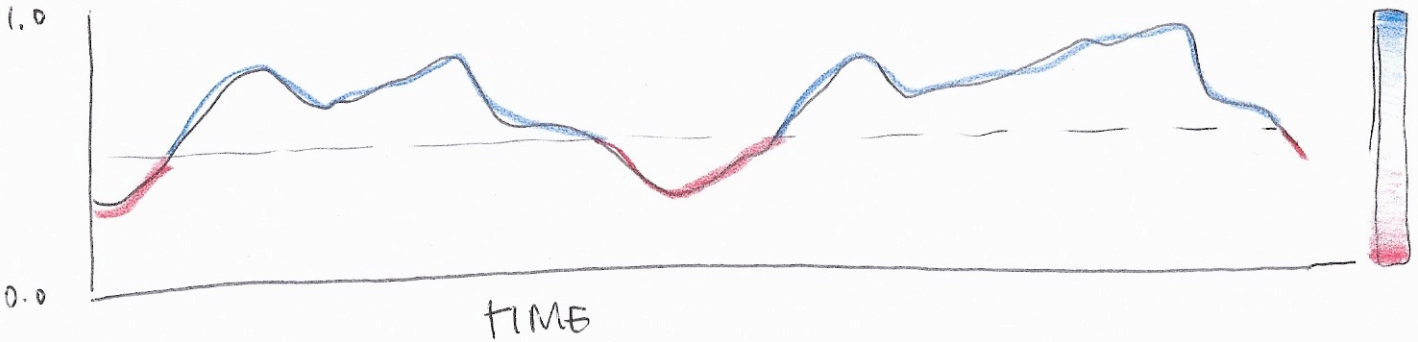
selection details

Sentiment

20 % negative

80 % positive

Trendline

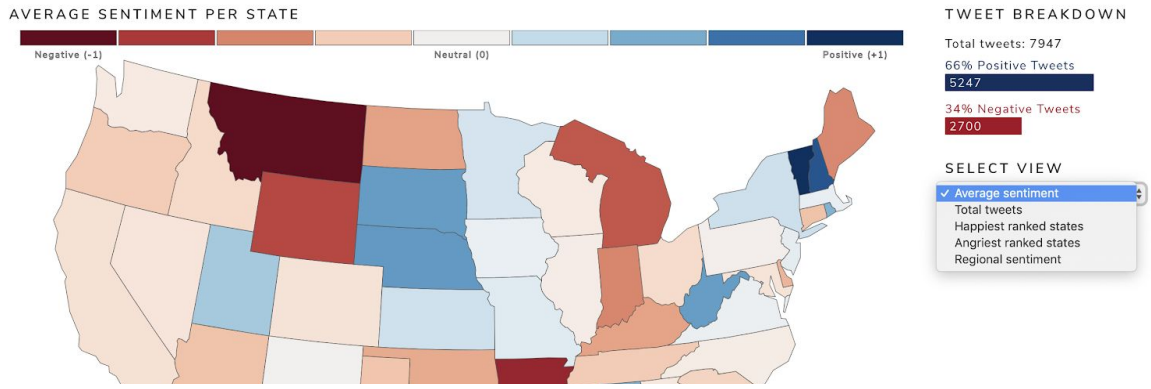


Implementation

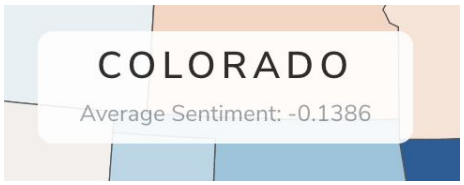
Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

We noticed that after working on the visualizations a bit, we had to change some things and go a different direction than we had originally realized. We still ended up doing things in the JavaScript/HTML/D3 ecosystem, but we didn't get around to the streaming part. Instead, we ended up having to adapt our visualizations based on some of the different insights we got out of our data. For instance, we thought that the average sentiment per state would have some variance, but most of the states were pretty much the same. So we had to improvise and normalize the data so that differences were highlighted, for example.

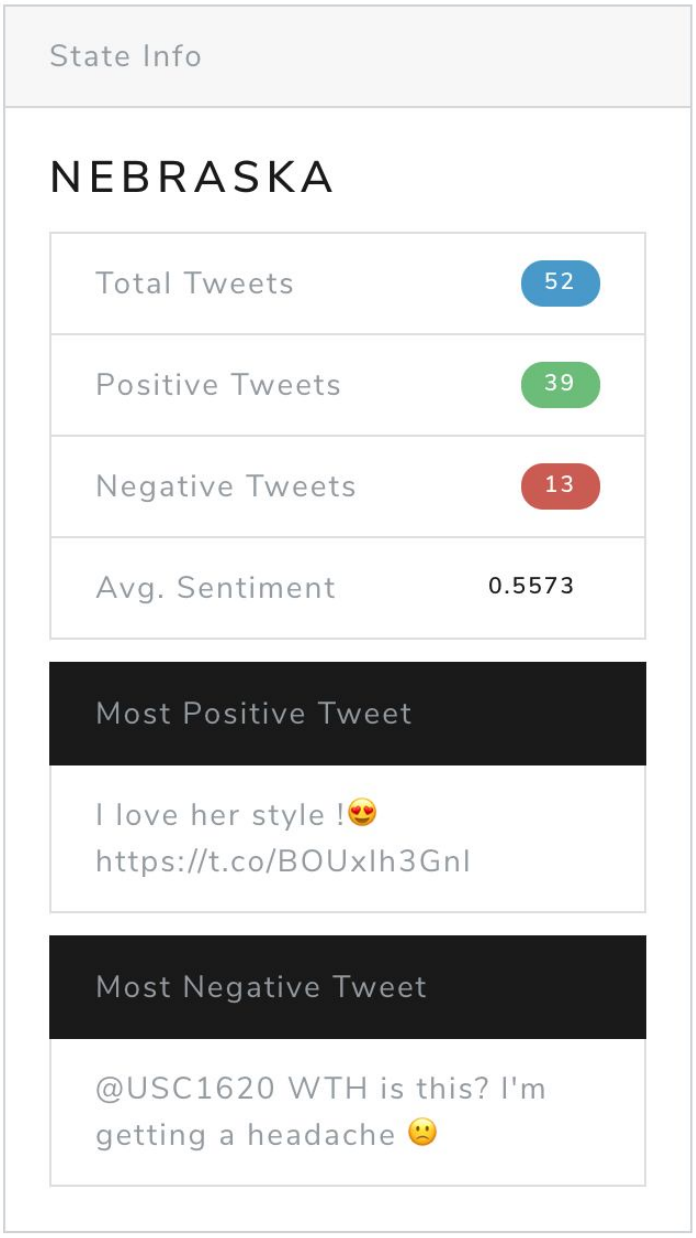
In terms of interaction elements, we added a drop-down menu to be able to switch between any of the maps mentioned above.



As mentioned, we also added a tooltip when hovering over each any state of each map. Depending on the map shown, different information is shown in the tooltip.



We also added the ability to click on any of the states in any of the maps. When clicking on a state, it shows a box with more information about the state underneath the “Select View” drop-down. This box shows the number of total tweets, the number of positive tweets, the number of negative tweets, and the average sentiment of all the tweets from that state. One this was implemented, we decided it would be interesting to show the tweets that had the highest sentiment score and lowest sentiment score for that state. We added to boxes below the tweet breakdown to show the content of the most positive and most negative tweets from that state.



Lastly, there is also a tooltip when hovering over any of the bars in the average monthly sentiment bar graph.

It’s worth noting that unlike the rest of the graphs, this graph was made with Chartjs. Since we had a lot of time to make graphs with d3, we thought it would be interesting and fun to try a different library for this final graph.

For our styling, we used a Bootstrap template from Bootswatch.

Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

There were a few really interesting insights we gained from the visualizations that we didn't expect. The first thing we noticed from the Average Sentiment Per State map was that Montana has the most negative average sentiment from their tweets making them the "angriest" state. Vermont, on the other hand, had the most positive average sentiment from their tweets making them the "happiest" state.

Another interesting insight was that there were significantly more positive tweets (66%) than negative tweets (34%) out of all the tweets. However, the monthly average sentiment graph shows that most of the months had an average negative sentiment for the tweets from that month. This means that the positively rated tweets were not rated *as* positive as the negatively rated tweets, which was also interesting to discover.

The total tweets map was not very surprising because, of course, the states with higher populations (California, Texas, and New York) had more tweets than other states with less population. Aside from that, this map shows that the east coast has a slightly higher occurrence of total tweets per state compared to the west coast (not including California).

Another interesting insight for us was to see the most positive and negative tweets for each state when clicking on the state. Most of the time we found that the tweets our model found were, in fact, actually positive/negative showing us that our sentiment model worked fairly well. Up until we visualized these tweets, we unsure of how well the model was actually categorizing sentiment.