

CS 6630

# Visualizing Sentiment & Geoencoding with Tweets

---

Richie Frost  
Neeka Ebrahimi  
Dyllon Gagnier

Process Book



## 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 work together in a lab under Feifei Li and thus each have 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 wholistic 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 criticising 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 web site, 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?**


The main question we would like to answer is 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. 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 as 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, we 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.

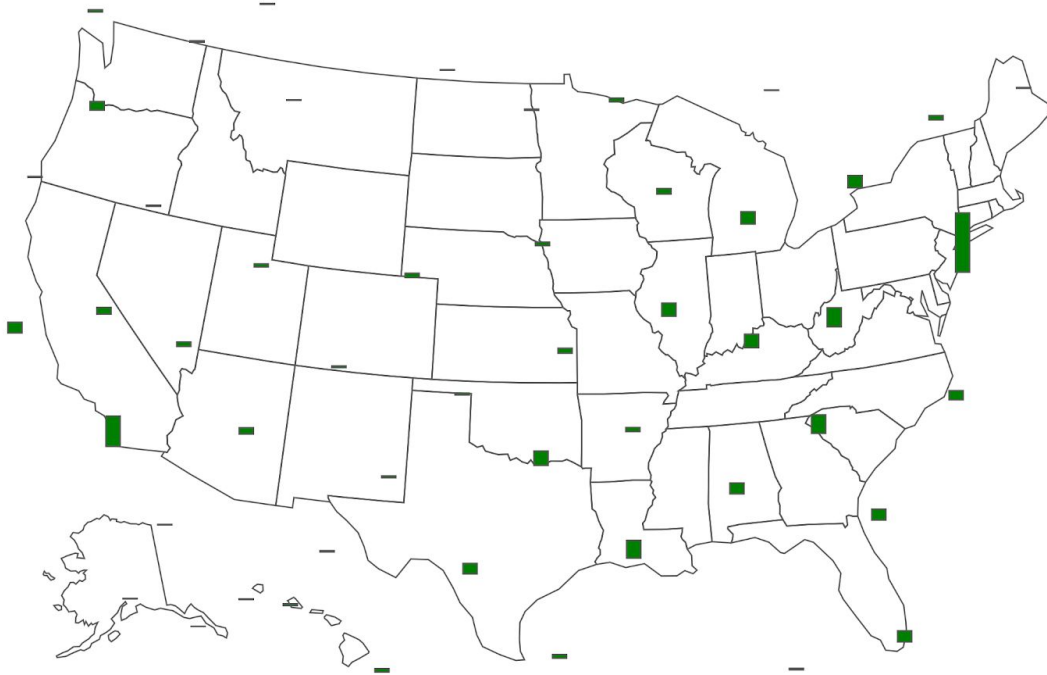
To initially look at the data, we worked on a similar project in 2016 regarding election trends for the United States Presidential Election. We visualized the voter breakdown per county and per state based on the average sentiment score of tweets in that region for democrat and republican candidates, 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.

## 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.

## Tweet Breakdown

We added a simple breakdown of the twitter data, so that users can get a quick overview about 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


Total tweets: 9199

66% Positive Tweets

6107

34% Negative Tweets

3092



each bar is the number of tweets associated with that sentiment.

## 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.**

## 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?**