CS 6630

Project Proposal:
Visualizing Sentiment and Geoencoding with Tweets

Fall 2018
University of Utah

Richie Frost
Neeka Ebrahimi
Dyllon Gagnier

Basic Info:

Project Title: Visualizing Sentiment and Geoencoding with Tweets

Names:

Richie Frost:

UID: 00997512

Email: richie.frost@utah.edu

Neeka Ebrahimi:

UID: 00671080

Email: neeka.ebrahimi@gmail.com

Dyllon Gagnier:

UID: 0708264

Email: dyllongagnier@gmail.com

Repository link: https://github.com/richiefrost/dataviscourse-tweetsentiment

---

Background and Motivation

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 this data source 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 on each question during the debate.

Project Objectives

The main question we'd 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:

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 language, 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.

Data Processing:

The tweets will be processed with a novel auto-generated data processing pipeline developed by our lab called Compass. We use the Twitter streaming API to gather the data and filter it based on predetermined filters (hashtags, handles, keywords, etc), which is fed directly into an LSTM recurrent neural network to get a sentiment score on the text. The scores in this system range from 0.0 to 1.0 inclusive

where 0.0 is as negative as possible and 1.0 is as positive as possible. Twitter also geotags tweets whenever possible, but when they do not, we use an API to get the approximate latitude and longitude coordinates for tweets based on the city name provided in the tweet's location. In order to prevent all these tweets from being in the exact city center, we introduce a small amount of noise to the coordinates. All of this is done in a streaming fashion, but for the purposes of our initial idea, we will save the output of our processing in a flat JSON file to be processed with D3. We hope to be able to get to the point where we can stream the data directly from Compass to our visualization web page in order to see changes in sentiment in real time for specific filters, but for now, we will just use a static subset of the overall stream.

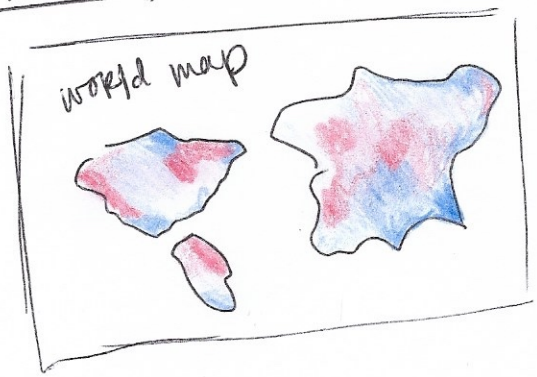
Visualization Design:

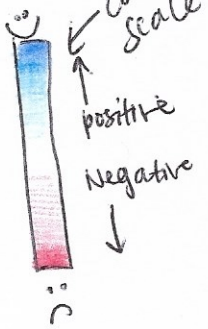
The most important part of our data is the geolocation and sentiment of each tweet, so we designed our visualization with that in mind. For the geolocation, the most obvious visualization is to show a world map displaying the locations of the tweets. The map will be colored according to a heat colormap that describes the sentiment value (positive or negative) and the number of tweets from each location. A dual-color heatmap is ideal in this scenario since there is a meaningful zero point (i.e. neutral in disposition). Ideally, we would like to be able to implement some sort of selection/zooming mechanism on the map with brushing thereby allowing users to focus on the areas they are most interested in. With that, we will be able to show the zoomed-in portion of the map with some details about the region selected. Another

interesting aspect of our data is the time of the tweets. We would like to be able to visualize the sentiment of a topic for a location over a period of time. Therefore, it would also be beneficial to have a trendline with the time period over the X-axis and the sentiment score (0.0 to 1.0) as the Y-axis. The trendline should update accordingly as a region is selected in the map. If we are able to stream the tweet data from Compass in real-time, this trendline will stream as new data is analyzed. During our brainstorm, we thought about a vertical bar chart to show the sentiment score for each region (broken down by either continent or country) but ultimately decided that it was redundant and didn't add any extra value to the visualization. The following five pages are the sketches from performing the Five Design Sheet Methodology.

BRAINSTORM

world map colored by sentiment score





← break down by country or state?

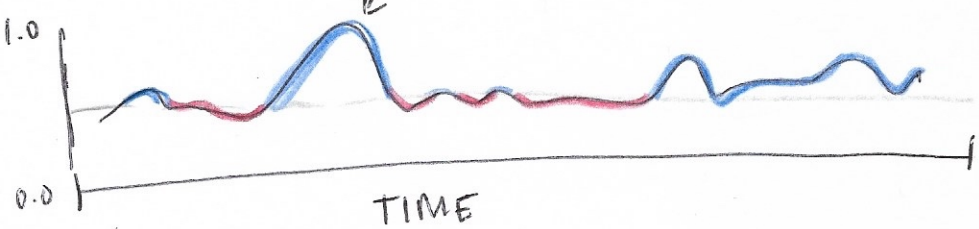
Or single overall heatmap?

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

combination of both?

Trend line

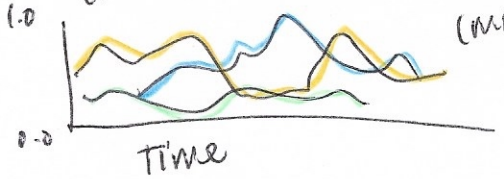
trend for overall sentiment across the world or for a selected region



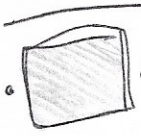
* makes use of time, but not location

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

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

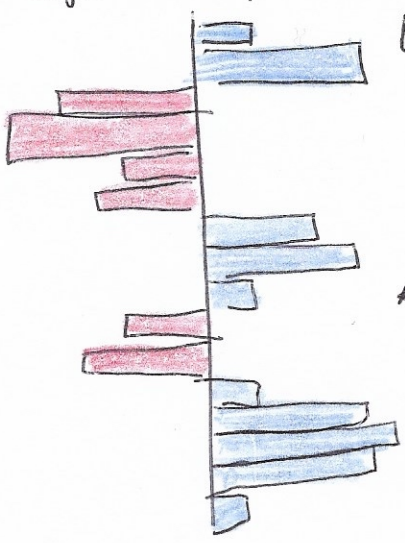


other additions

-  ← 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?

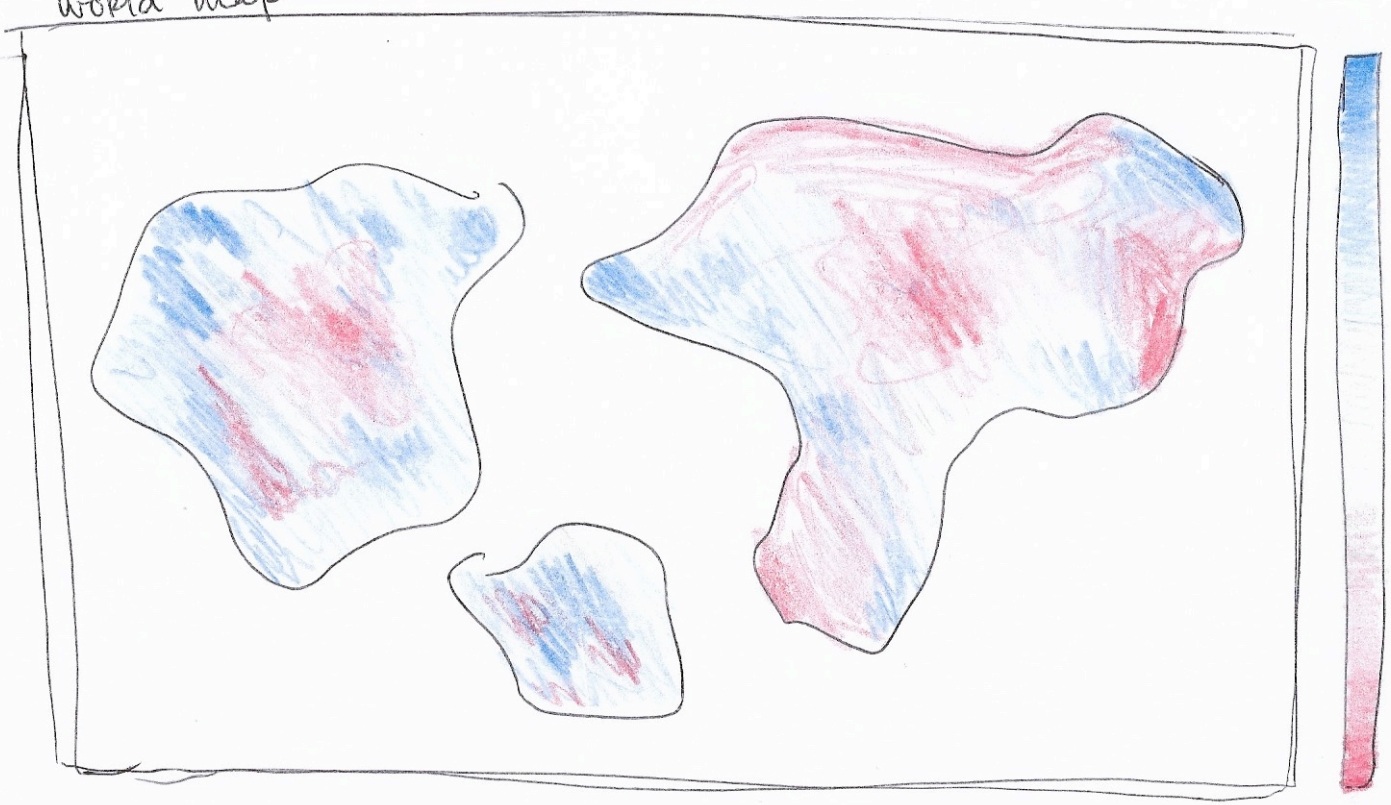
- doesn't make use of geotags

DESIGN 1

TITLE

authors / description

world map

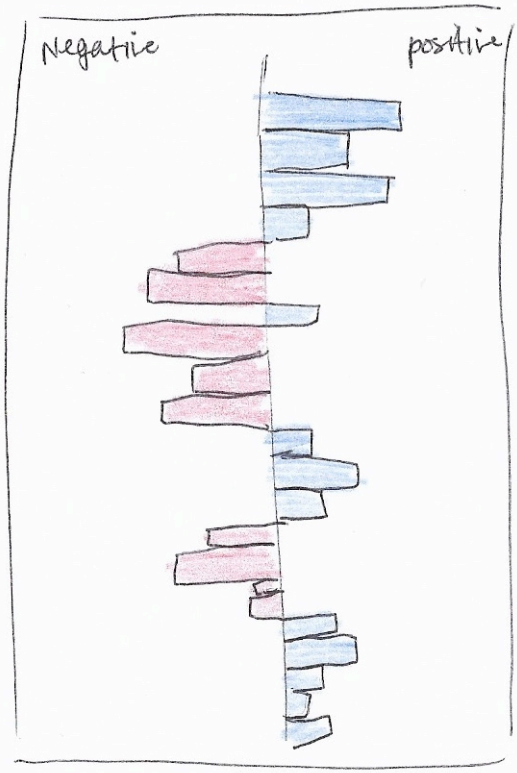


Overall Sentiment

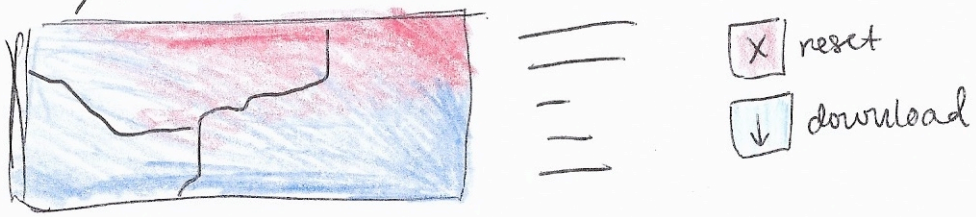
20% negative

80% positive

Stacked Bars



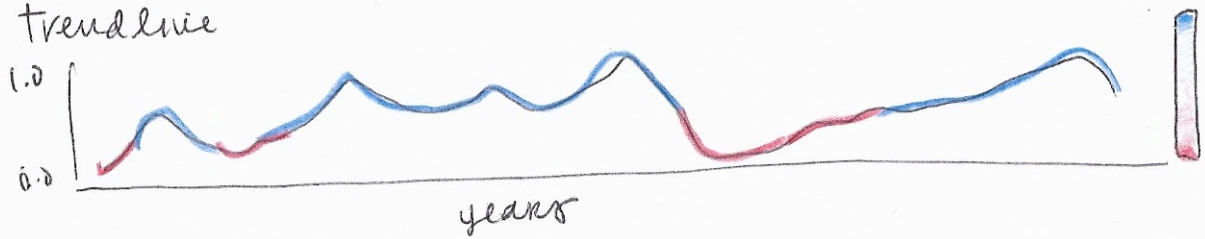
zoom / selection



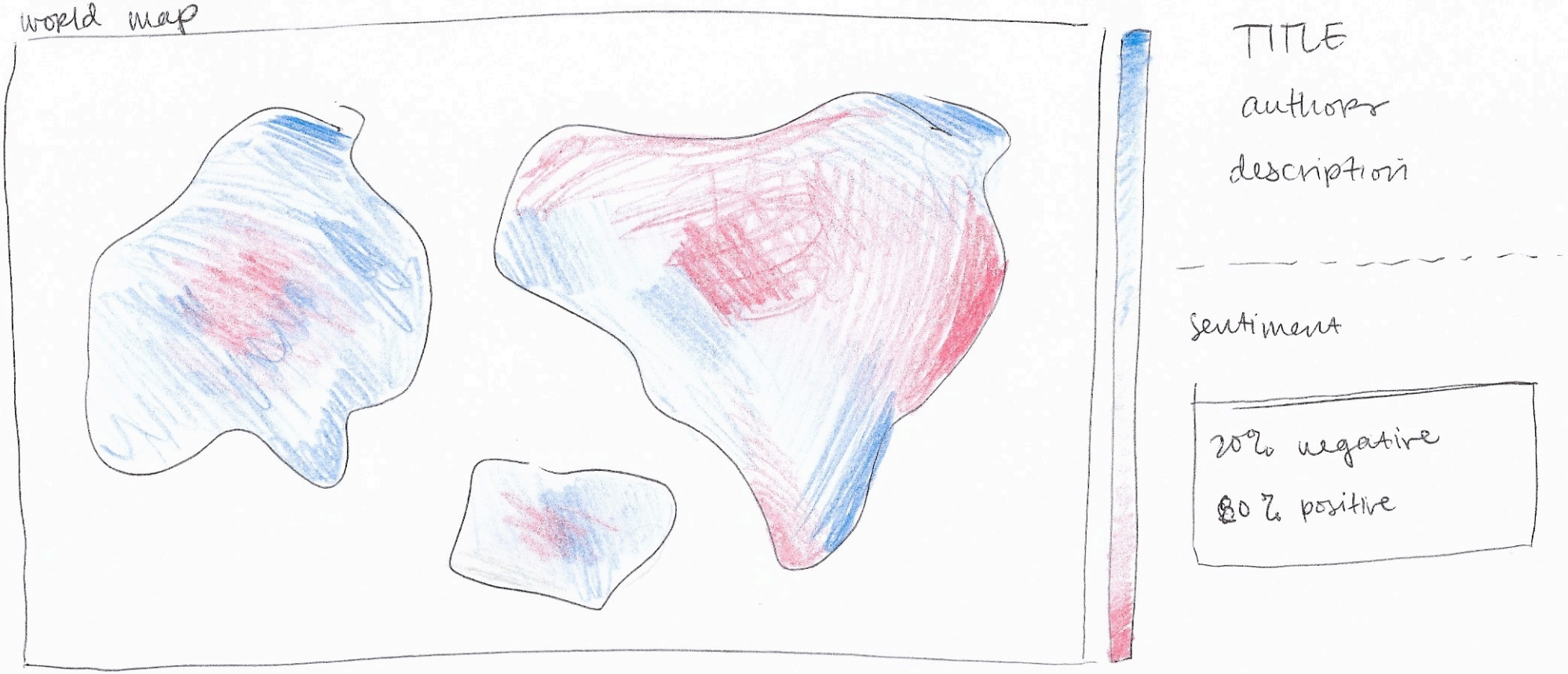
reset

download

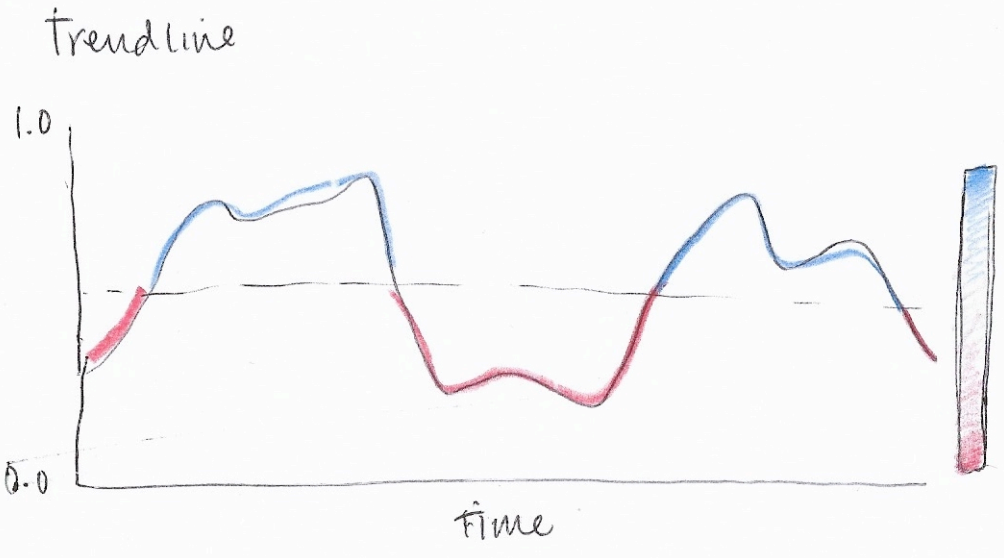
trendline



DESIGN 2


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

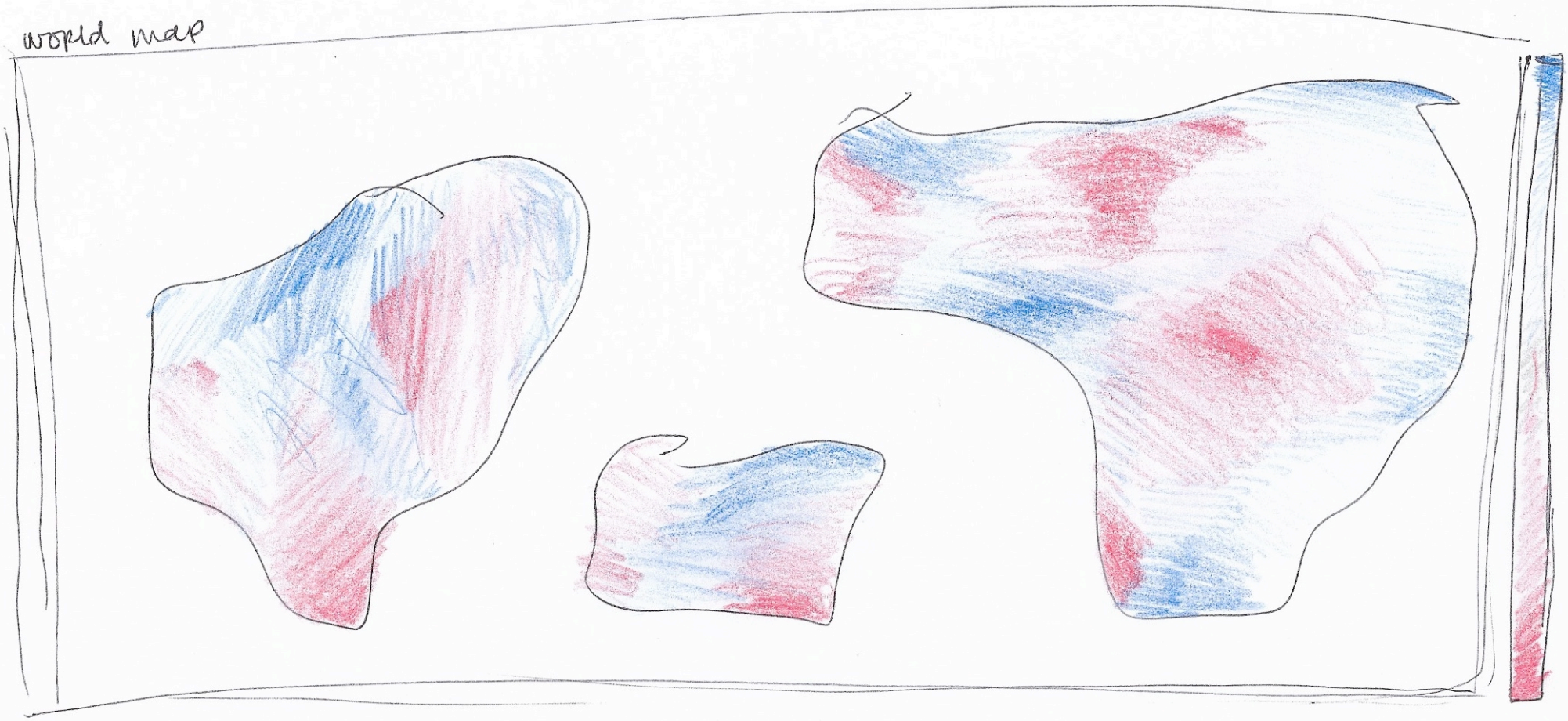

zoom/selection
reset
download


trendline
1.0
0.0
time

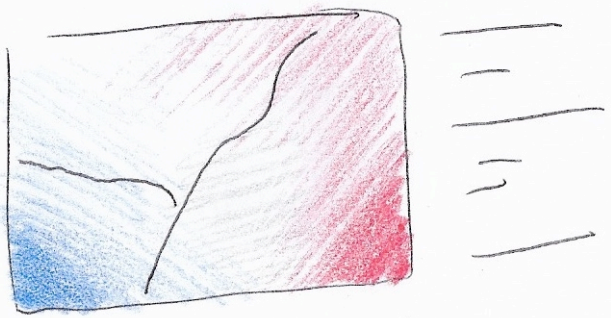
DESIGN 3

TITLE

author / description

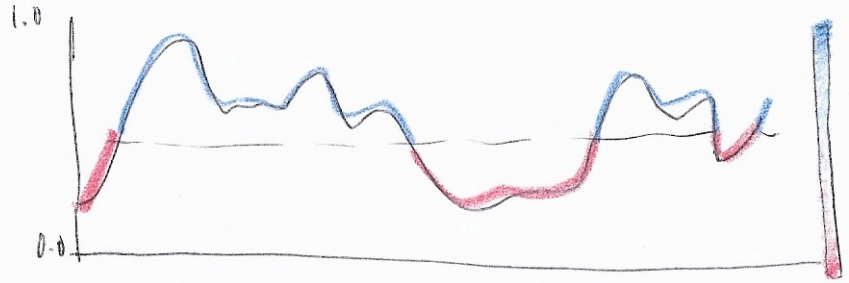


Zoom / selection



reset download

Trendline



Sentiment

20% positive

80% negative

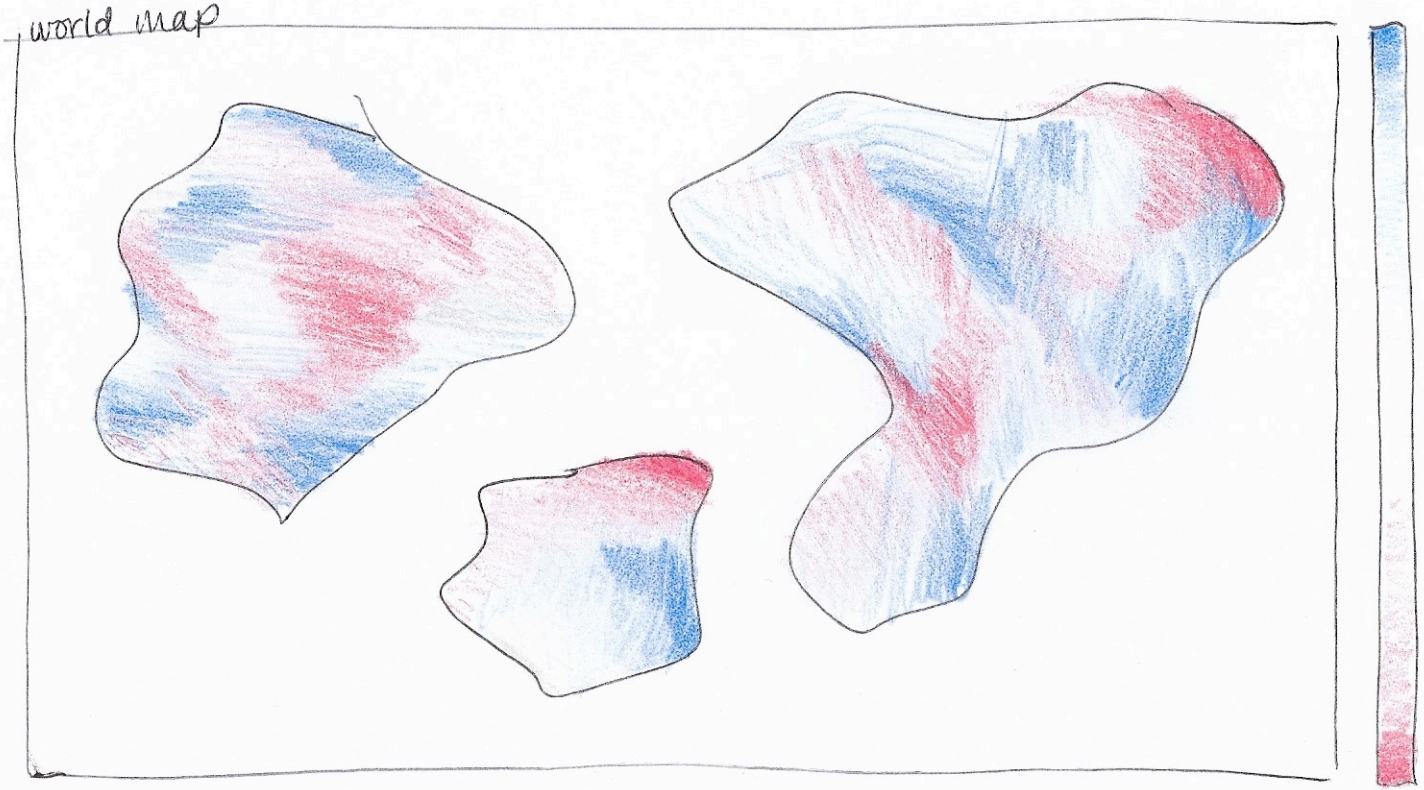
FINAL DESIGN

TITLE

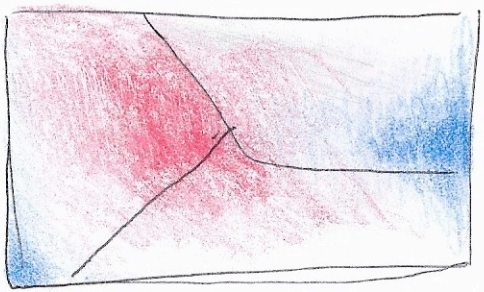
Authors
short description

help

world map



Zoom / selection



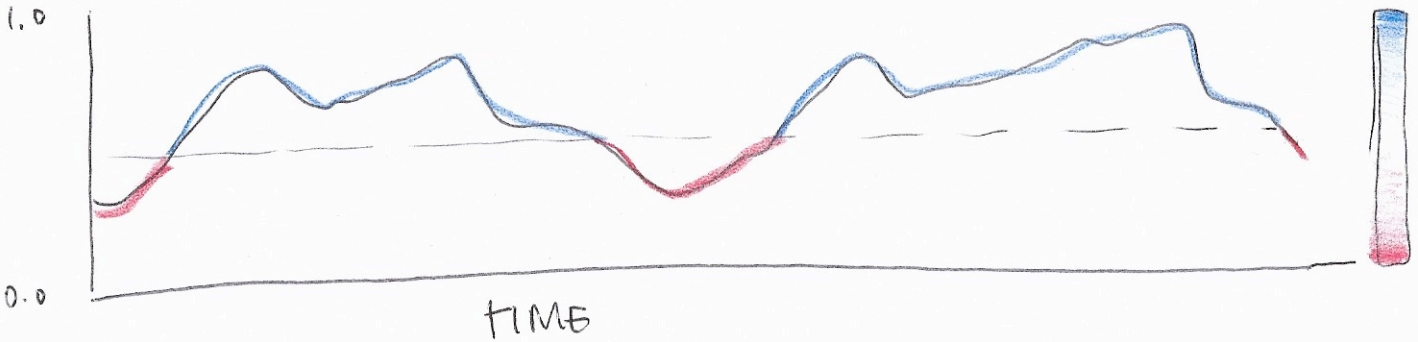
selection details

Sentiment

20 % negative

80 % positive

Trendline



Must-Have Features:

- Map visualization with with heat colormap of sentiment
- Ability to zoom or select regions of the map
- Trendline showing the sentiment of tweets over a time period

---

Optional Features:

- Streaming data with corresponding updates in the visualizations
- Ability to search/filter data on a topic

Rough Schedule:

1. Initial JSON file created with data.
2. Basic JavaScript project is building with all basic libraries such as Vue/React, D3, etc. *Week 12*
3. Website is available on the internet. *Week 13*
4. Have static versions of all non-map visualizations. Project Milestone
5. Create static version of map-basic visualizations. *Week 14*
6. All visualizations fully interactive.
7. Tutorials and visualizations are designed and explained on site so that any user can figure out how to use it without instruction. *Week 15*