Process Book

US Election 2016 Twitter Visualization

Haiyan Liang, Jiaqi Liu, Zhiyuan Liu

{hliang, jliu6, zliu7}@wpi.edu

Final Project, CS573-16F

1. Introduction

1.1 Updated Project Proposal

After many discussions about the project, we decided not to visualize the real time Twitter data, instead we will work on the data that we already have or easy to get, and still work on the Twitter. Luckily our classmates Andy Nie has scraped Tweets about the hot 2016 US elections on the Election Night using stream Twitter API. We all think it is a very interesting topic so Andy provided the data to us and we finally decided to visualize the election night tweets. The data we will talk later in the data description part.

1.2 Overview and Motivation

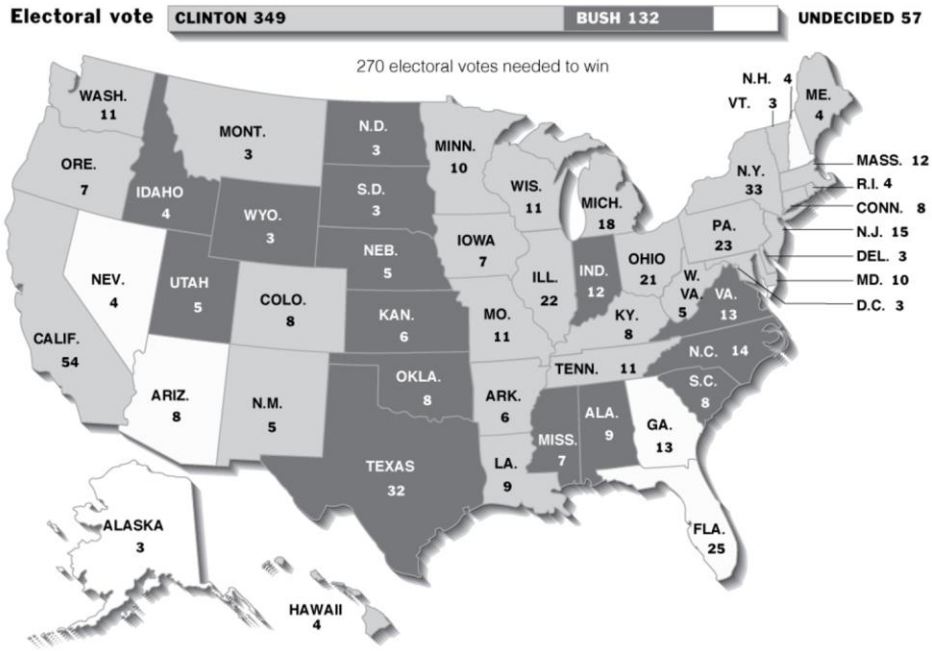
The most important even of USA in 2016 is the president campaign relative to Hillary Clinton and Donald Trump. Before the due time for the election, no one knows the exact result, since the process was so dramatic. so we want to use the data of night on November 18th from Twitter to visualization the aptitude of the social media about the two president candidates, in order to give people a better sense of the current situation.

Second, we want to know people's idea about this election through the tweets. One way is we can visualize the tweets content. We have presented before in the class about the SentenTree by Mengdie Hu al et. In this part we will use word cloud to visualize tweet contents. And we also show the most frequency words in different time of election night.

Another way is using sentiment analysis. We first score every tweets on their sentiment. And we show the sentiments on Tweets contains "Trump" or "Hillary" related words, and some topic words like "economy". We also add a slide to control the time. By doing so we show the trend of Twitter idea about Trump and Hillary and some key topics as well.

1.3 Related Work

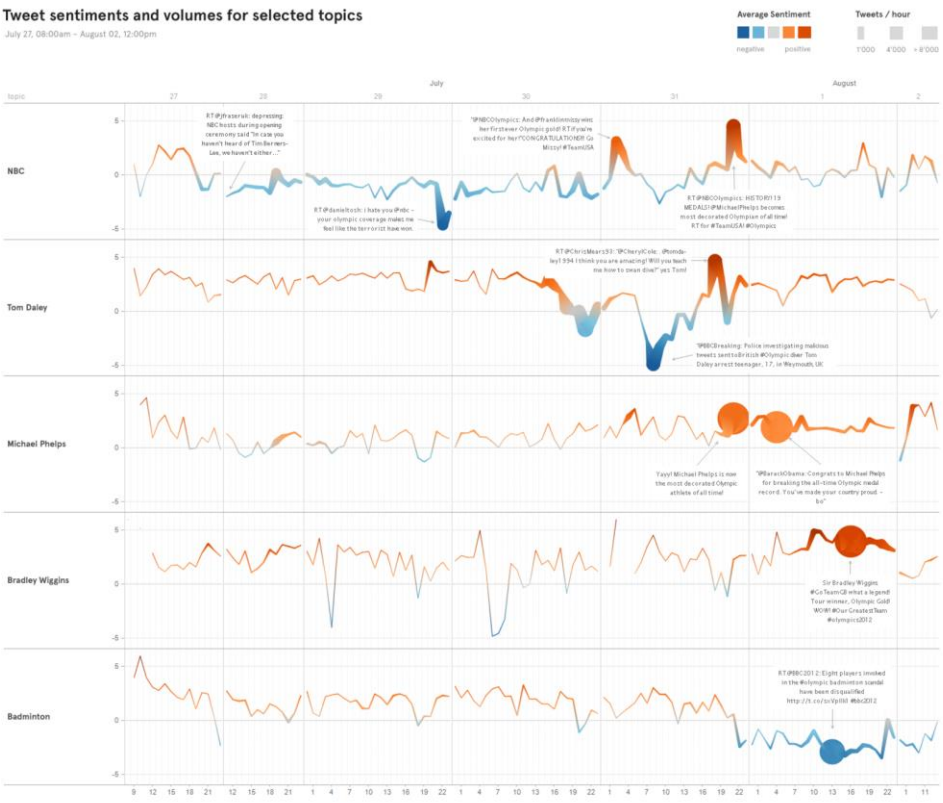
New York Time has made many well designed data visualization, for example the following graph. This data visualization clearly illustrates the situation for president campaign between Clinton and Bush. People can easily get the poll number from the numbers under the state names. At meantime, the different color for the two candidate can be well distinguish between each other. This inspire us that map is the best way for the event.



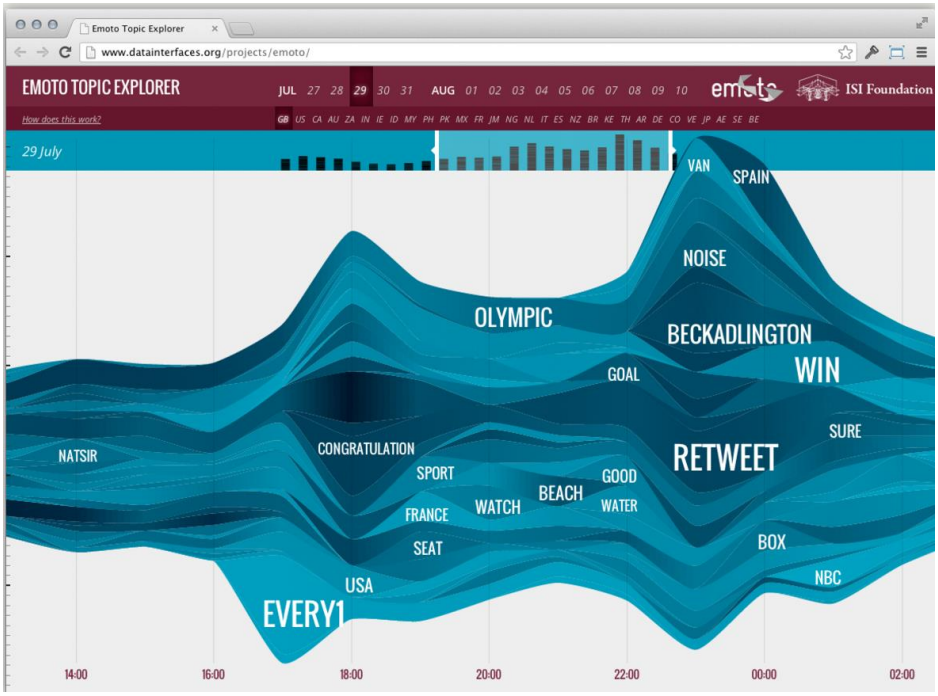
As for visualize the tweet contents, there are many methods and researches on that. First is the SentenTree. It is kind of advanced visualizations designed by Twitter group. It can show the words in a tree like shape which also form the sentences. So readers will know the most frequent words as well as the sentences formed by these words. And also there are some other ways to visualize words like word tree and word cloud.

In this SentenTree Example, readers can easily see the tweets about World Cup.

For analyzing the tweets like sentiment, there are many websites and researches have done a lot. For example the website www.emoto2012.org visualizes tweets about 2012 Olympics Games.



This example visualizes tweet sentiments and volumes for topics

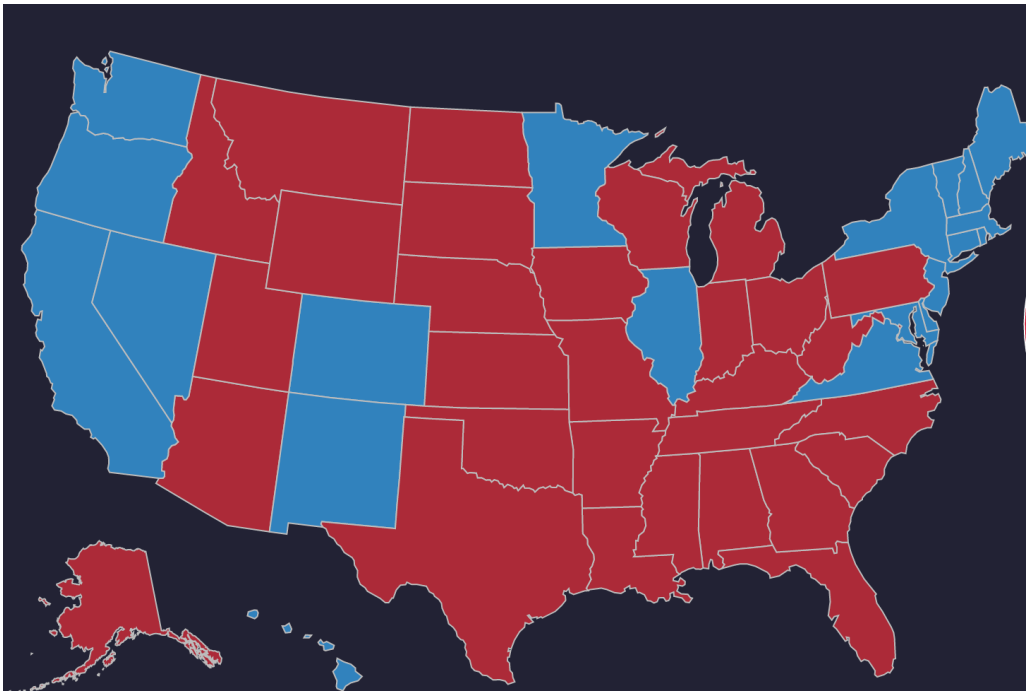


This example shows topics were discussed on Twitter during the Olympics

1.4 Questions

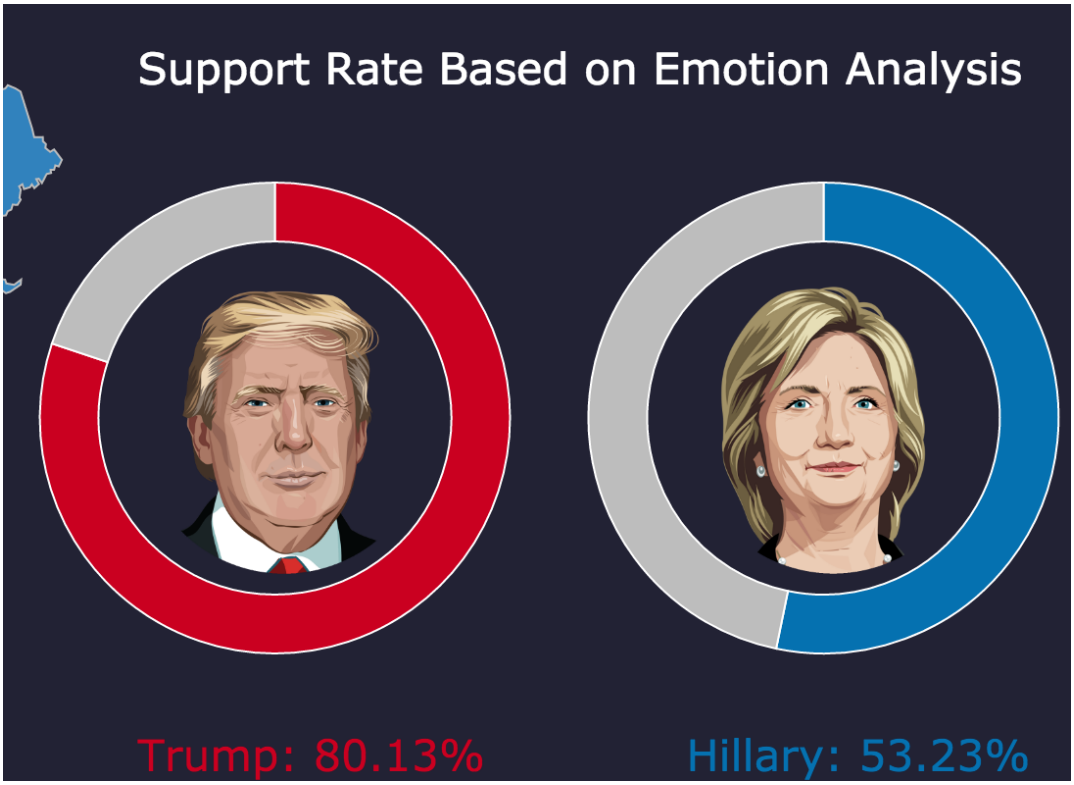
what information can we get from the exist dataset?

president election result can be reflected by the outcomes of each state of USA. So The most important feature is the state-level result for the two candidates. Based on our search for data visualization elections, we found that map on state-level is mostly widely used for illustrating the process of the campaign and giving people a better sense to know the current situation. We think this is the best way to visualize the event. For this part we decide also to mimic the approach using D3.



How to show social media users' emotion for the two candidate?

Usually, for each president candidate, people have no more than 3 kinds of emotions, positive, neutral and negative. If we illustrate the data in this three dimension, that would be too messy. The most important information about the election is whether the people support the candidate or not. So we combine the neutral and negative attitude into one group which is for non-support group. For Trump, the red arc comparing with the gray part is the proportion of number of support V.S. the number of non-support. The pie chart for Hillary is using the same idea, but the support part is colored blue



We also want to know people's idea about this election through the tweets. Like to find out who they support during the election processing, Trump or Hillary? What's their most concerned topics about the election? Even can we Twitter know the result before the real world?

During the project we learn to know we can use bubbles to show the top words in different time, and we even can interactively let users select different time to show them accordingly the information, like sentiments, using D3.

1.5 Data

Using Twitter Stream API, we get around 190,000 tweets about 2016 US president election from November 8th 21:00 to November 9th 5:00.

2. Process

2.1 Exploratory Data Analysis

We get 190,701 tweets about election on that night. Each tweet contains a lot of information in from the API. After discussions, we extract some key information we may need: tweets creation time, user names, user images, tweets locations (in coordinates) and tweet texts.

	created_at	text
0	Tue Nov 08 21:15:39 +0000 2016	RT @NPR: We hear you, America. No more 'I Vote...
1	Tue Nov 08 21:15:39 +0000 2016	RT @OccupyWallStNYC: Shame on the #Texas GOP f...
2	Tue Nov 08 21:15:39 +0000 2016	RT @campneil: Trump!! I voted!! https://t.co/G...
3	Tue Nov 08 21:15:39 +0000 2016	RT @tkinder: Chuck Norris: Clinton presidency ...
4	Tue Nov 08 21:15:39 +0000 2016	RT @AnitaDWhite: ⚡Polls are OPEN,Have U✅yet? ...

Examples we extract key information

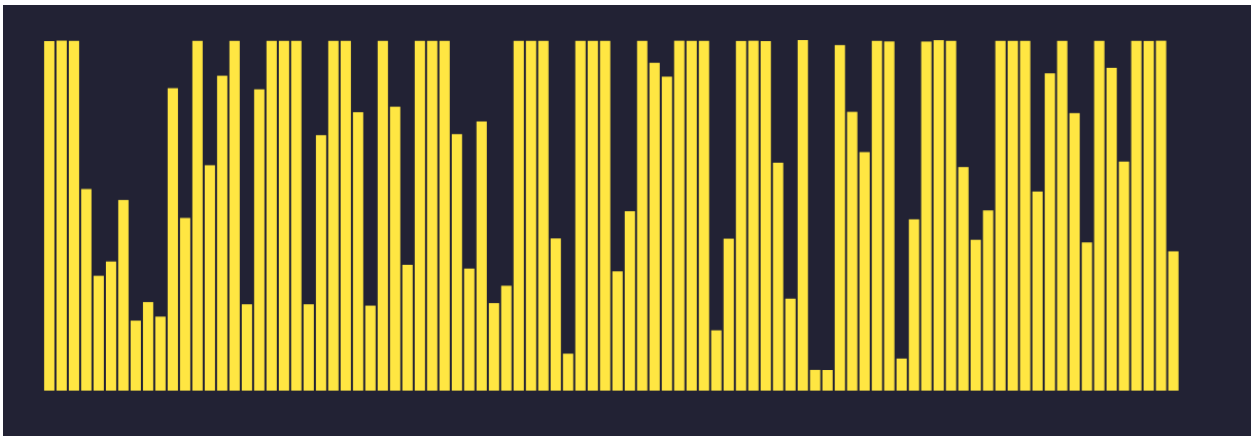
We have found several facts about our data. First is that there are so few tweets have locations; Next, the data are sampling from the Stream API, which means there are time no tweets recorded. In other words, our data are sparse in time, and this problem will be shown in our visualization. In addition, our data is kind of large, so we use 1/5 data as our samples to visualize. And finally, the tweets text need to be preprocessing by some NLP methods.

2.2 Design Evolution

Now our project is divided into two parts: visualizing tweets and sentiment analysis.

How to show the volume of each Tweet in certain window?

At first, we thought the volume of Tweet that generated can be different as the final time of election was going due. So we tried to use bar chart with certain time interval to illustrate the trend of volume. However, after completing the chart, we found that the trend did not change as we thought, and there is no regular pattern in it. Hence, in the end, we abandon the visualization for volume, since this is not the best feature for our project.



Map

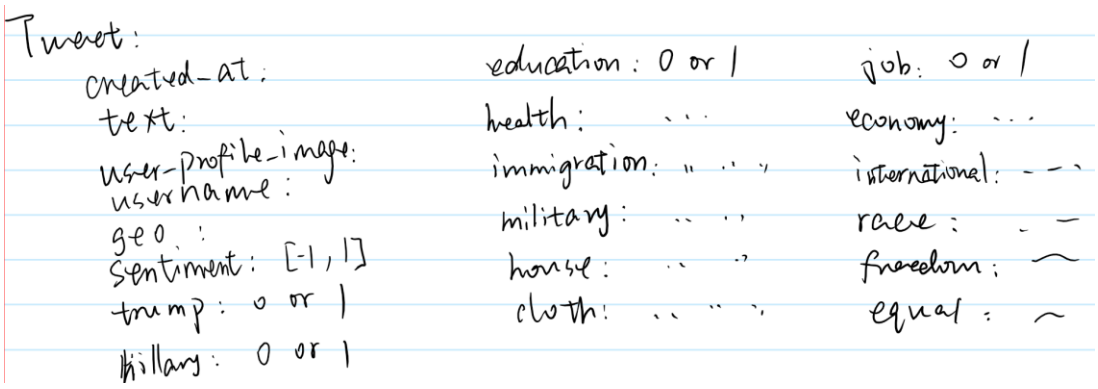
At first, we tried to use Google Map and its API instead of D3 Map. However, we found it is hard to change the background of the Google map and to draw polygons on the map to point out each American state. So we change to D3 map with the help of map json file. After solving this problem, we learnt that D3 has more functionality to control every element on the web page; however, if we use Google map, the map is well packaged and it is too hard to change the element inside of the map. A Google example is following.

Data processing

We first need to preprocess the data. For future use we add some features to our original data. Now every tweet has features like below. For example, if “trump” is 0, this tweet doesn’t contain “trump” and some similar words in its text. Otherwise if “trump” equals 1, it contains the “trump” so this tweet has highly chances commenting Trump.

At first we only add “trump” and “hillary”. But later we consider some topic words like “job”, “race”. So if a tweet contains these words, it is about these topics. So we will know Twitter users’ idea about Trump and Hillary and the topics.

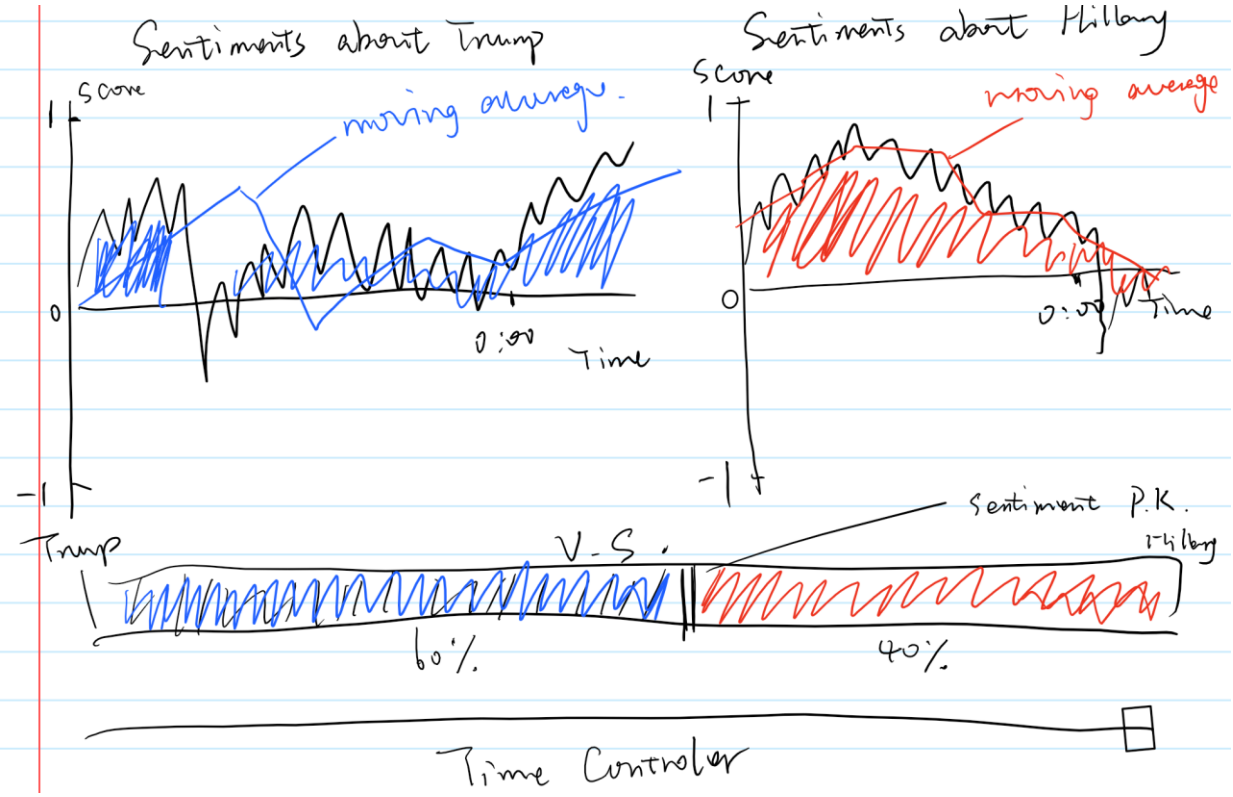
We also add a feature named sentiment. Which is the sentiment score we calculated from the tweet text. It is a decimal from -1 to 1. The larger the more positive it is. Negative scores mean negative feelings. About the sentiment method, we use the TextBlob package in Python. See here: https://textblob.readthedocs.io/en/dev/



Sentiment lines

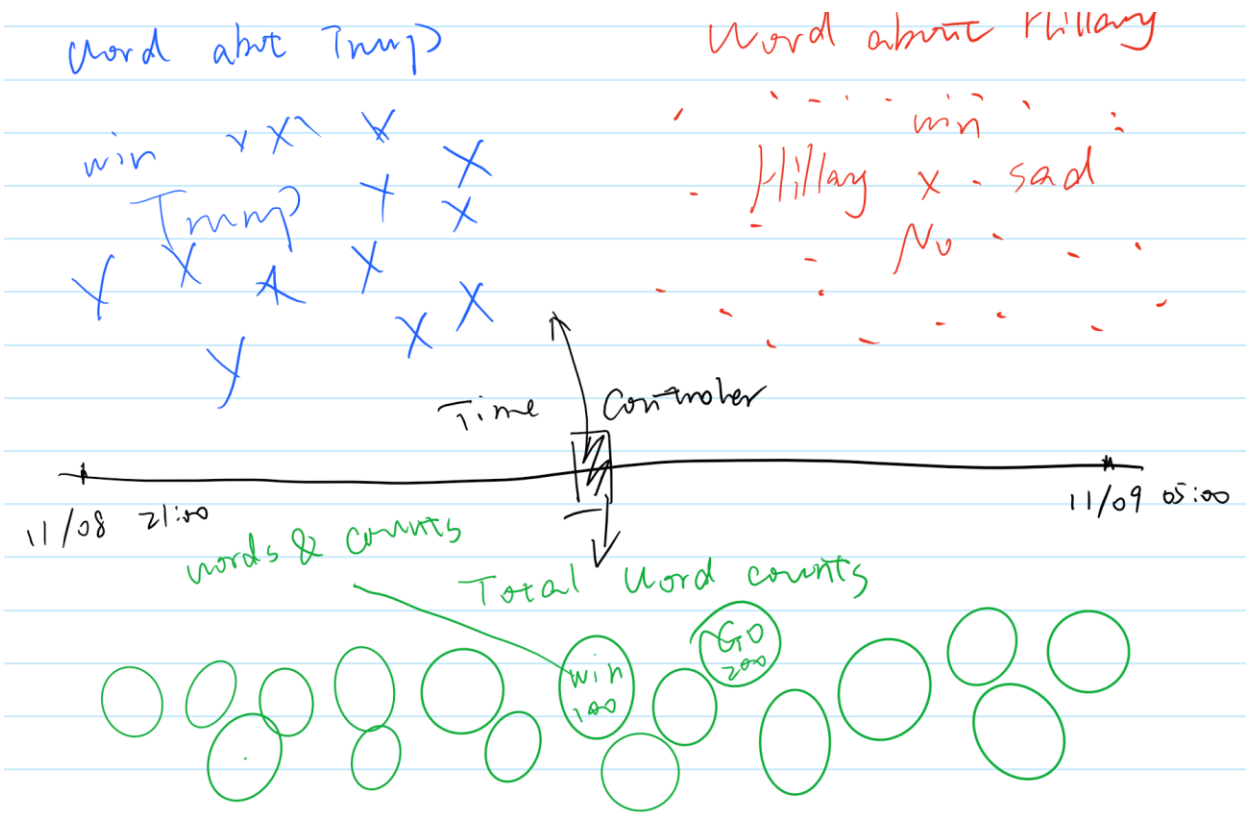
Here we first want to visualize sentiment scores. We like to draw two plots, one is the sentiment on tweets about topic “trump”, another is about “hillary”. We want to use the area line plot. So we can see the sentiment changes in different time. And we also want users can control the time. So we can compare the different emotion about Trump and Hillary in different time.

Because every seconds there are a lot of tweets. So every seconds the average sentiments are fluctuated a lot. So we also want to show the moving averages of the sentiments.



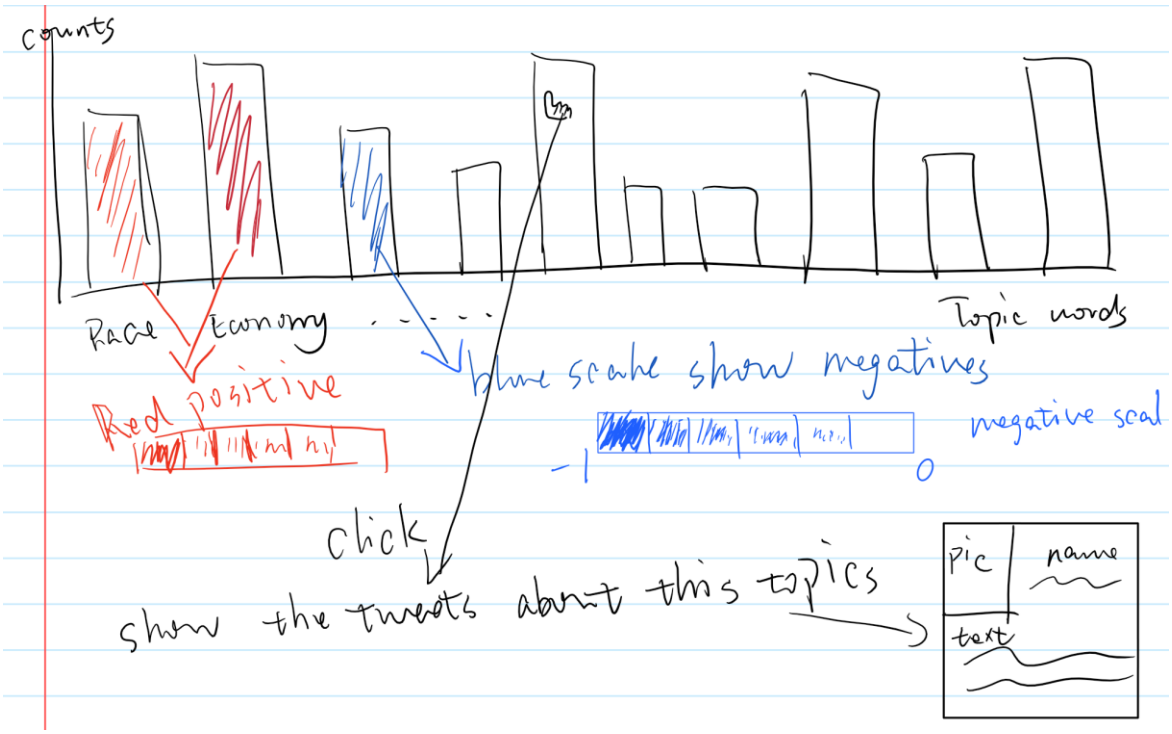
Word visualize

We also want to visualize top words in tweets. First we come up with is the word cloud. And we hope it also can be controlled by time. So we will know in a specific time what is the top words the tweets use. Later we also design a word bubble chart for better visualization.



Topic words sentiment

In this part we visualize tweet sentiments on topics. We also want to show the counts of topics. So we can know which topic the user mentions most in the election and how their feelings about that during election. Like if Trump will win, they perhaps show their angry on race things.

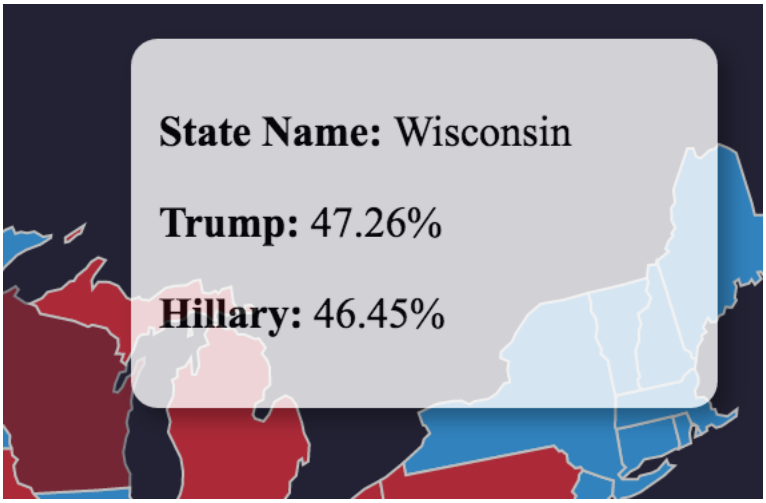


3. Implementation

3.1 Interaction of elements on map

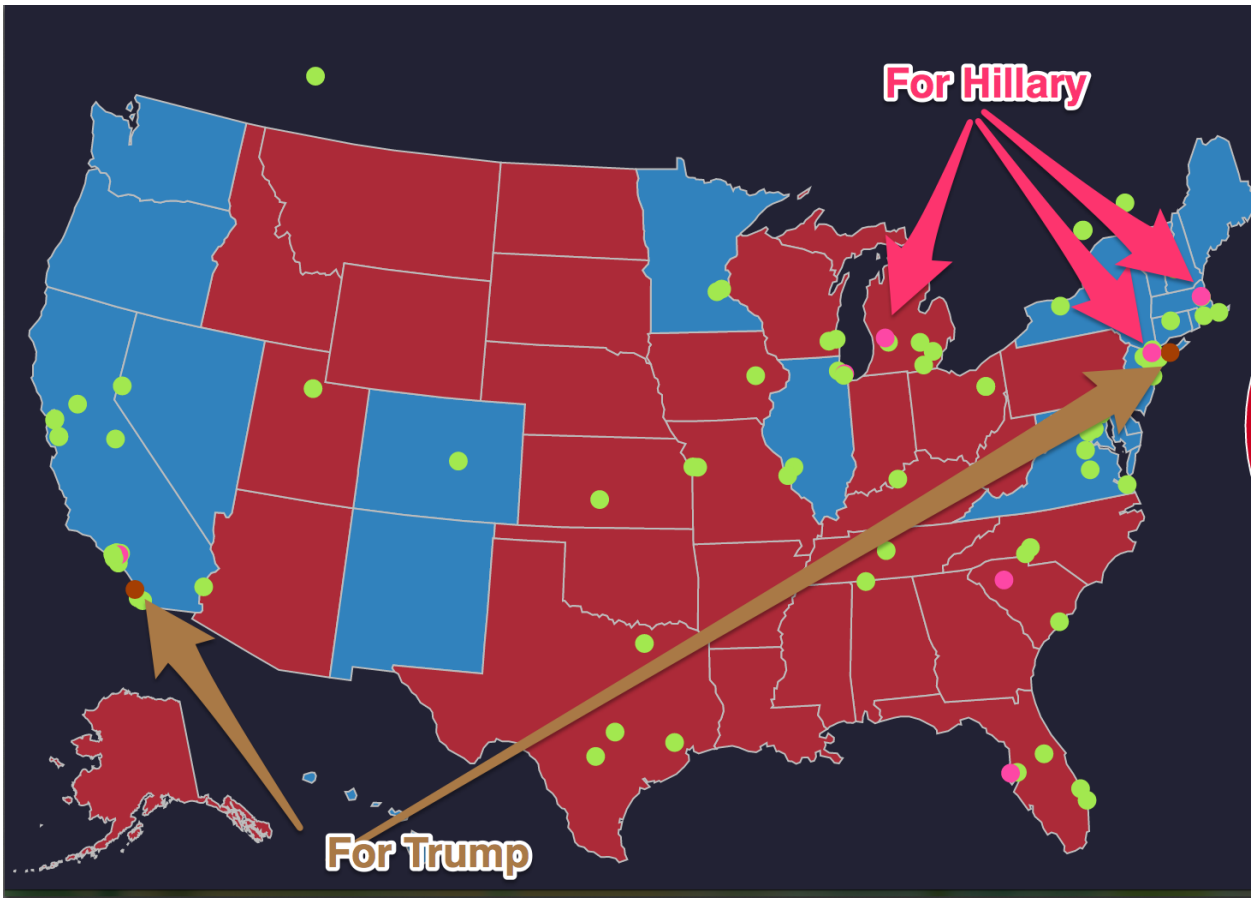
States interaction

If you place your mouse on any state on the map, the detail information of situation of the state will be popped out, like the follow graph. For example, For Wisconsin, about 47.26 percent of the people support Trump, and 46.45 percent of the polls support Hillary.

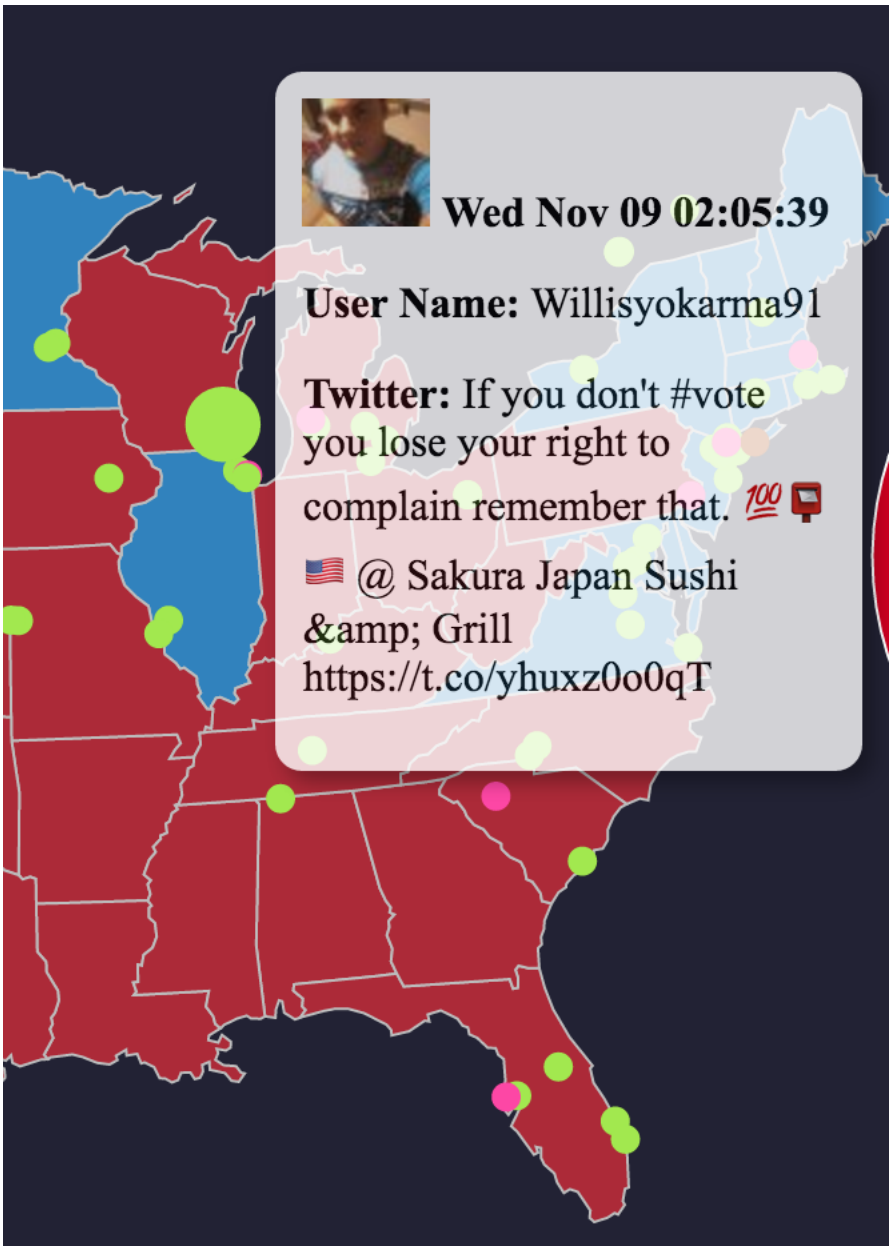


Points on map

Our project is using emotion of Twitters to find the correlation of the result of the election. So it is very important to pinpoint each Twitter on the map based on the location information when the Tweets generated. Unfortunately, in our dataset, there are only 0.2% of the data own the geographical data, which are longitude and latitude. However, we still need to draw the data which have the information on map. We give points with different color based on our sentiment analysis. Green points are the Twitters that do not mention or do not explicitly support one of the two candidates. Pink points are relative Hillary, and brown points are the Twitters that mentioned Trump.

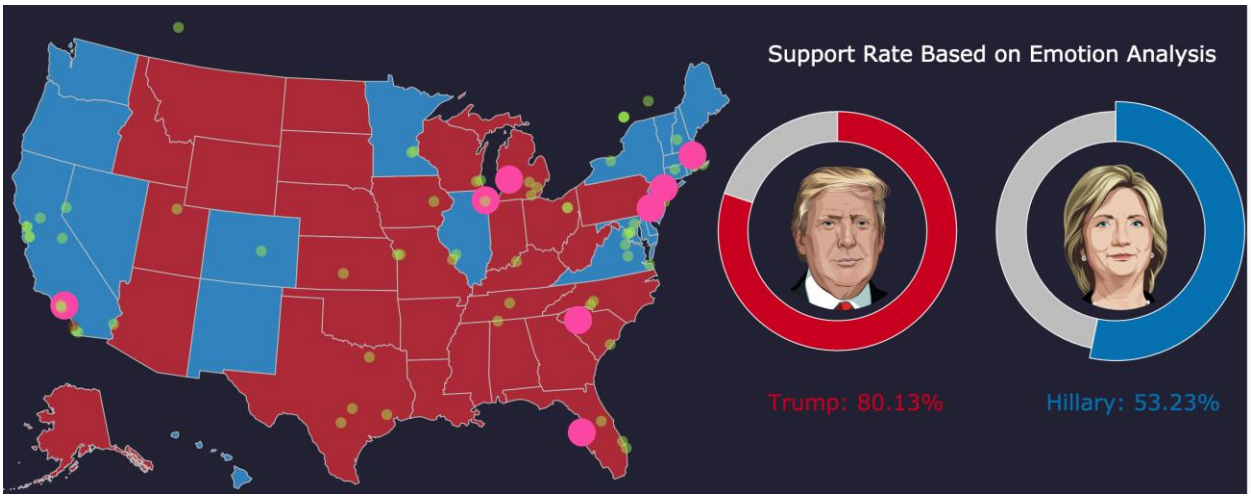


If you put your mouse on any of the points, the point will become bigger to distinguish current point with other points. And a toolkit will be shown up to tell the user the detail of the tweets, such as the profile image of the Twitter, the date and time the Tweet generated, the user name of the Twitter, and the content of the Tweet.

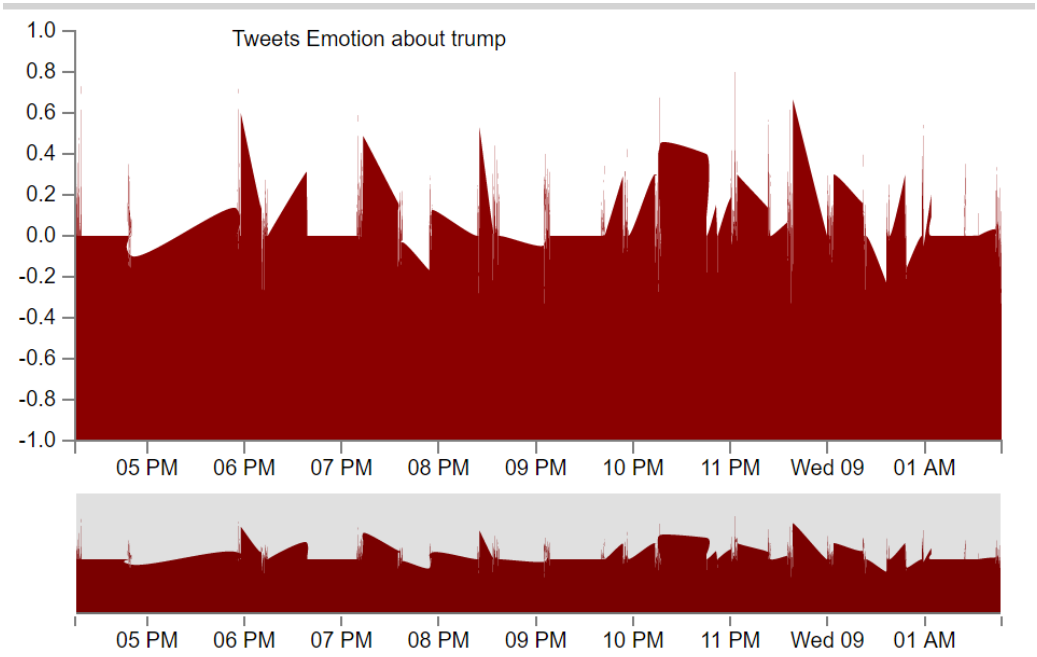


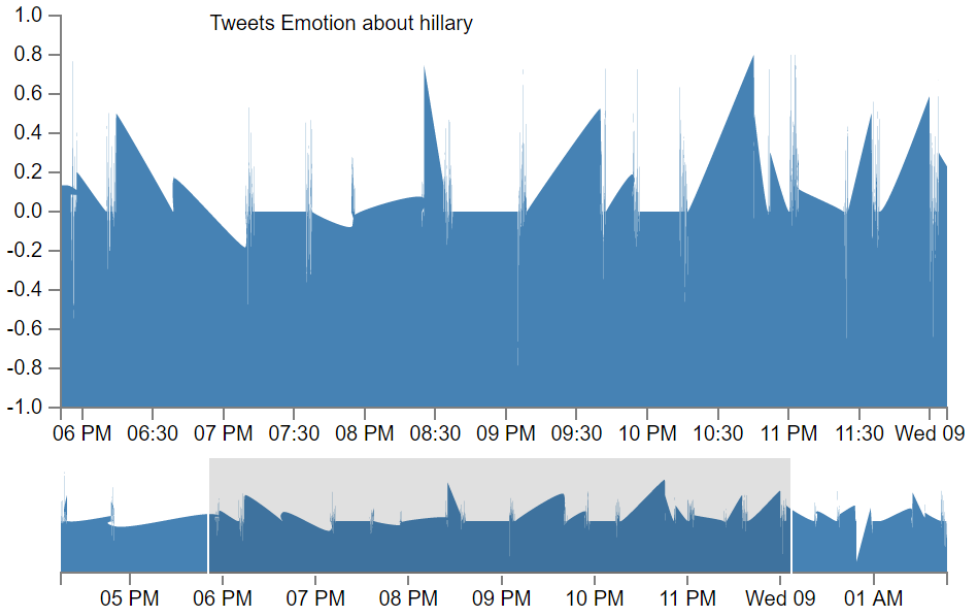
Interaction of Pie chart with the points on map

If users place their mouse on the bar chart of any candidate, the points which are for that person will be pop up, and other irrelevant points would be translucent. This function can help users to quickly find the Tweets which are related with a certain candidate who they are interested in.

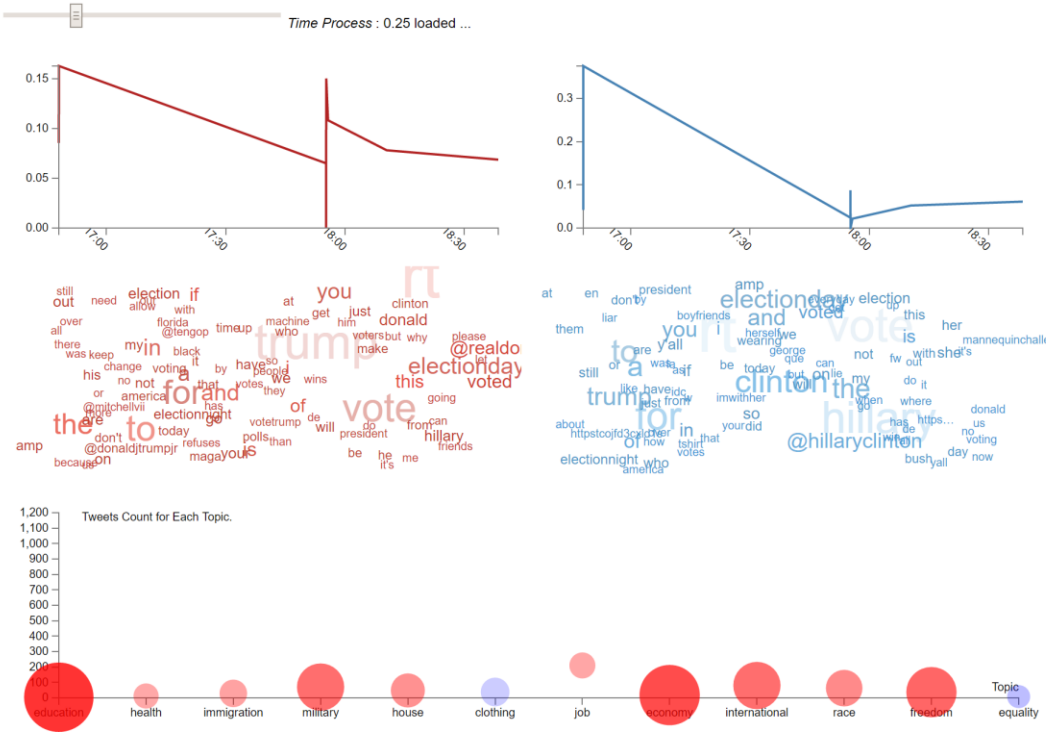


In the sentiment part we almost finish as we designed. First we use the area charts to show every second the average sentiment about Trump and Hillary. We also use a brush below the area charts to control the time. Users can zoom and see details in the time periods they choose.





We can see the area chart is very fluctuated as we mentioned before and the data is sparse because of missing data. And we do not draw the moving average here, we draw it separately instead, together with the word cloud and topic words part.



Here we see three different charts here: the average smoother sentiment lines about Trump and Hillary, the word clouds about Trump and Hillary and the topic word bubbles. They are controlled by a time slide which can show how they change with the time from beginning to the end. In the topic bubble chart, the negative sentiment is in blue and positive is in red. When they are more neutral (near 0), they are lighter and smaller.

popular words part.

4. Evaluation

What did you learn about the data by using your visualizations?

There is really correlation between the emotion on social media and the final result of the campaign. Now, we know Trump get the president position for next four years, and the data visualization shows that for Trump, there are more people have positive attitude to him. Based on our data visualization, we realize that few people shared their location then post tweets.

And from the sentiment analysis, we learn to know that in the midnight of the day, Twitter users may know the results that Trump will win, so there is a going up for Trump and going down for Hillary on their sentiment scores. And in the same time some key words like “Florida” mention a lot and popped out in our visualization if you look in detail.

How to improve it?

For future other big event, we will start to collect data from social media earlier so that we can get more information about the event, and some more interesting pattern probably come out. And also our sentiment model maybe not that accuracy. As for the visualization design there are much space to improve, like we can visual the tweets in a fancier way like tweets walls or SentenTree. And our update every time especially in the sentiment analysis part is kind of slow, which is really bad for visualization.