Project Report
CS294-1: Behavioral Data Mining
Chronological and Geographical Visualization for Tweets

Hanzhong (Ayden) Ye, Hong Wu, Rohan Nagesh

Abstract
In this report, we present Tweet Visualizer, an end-to-end platform designed for advanced visualization and data mining on Twitter. We begin with a motivation for data mining on the Twitter platform and proceed to a discussion of our system design, including our front-end client, database logic, and data mining techniques of Naive Bayes and SVM. We conclude with several interesting chronological and geographical visualizations produced with our platform.

Motivation
Twitter has grown dramatically over the past few years. Now boasting upwards of 300 million users and a burgeoning network of brand advertisers as well, the company has inched closer towards becoming the real-time information network for the digital age.

However, for both non-paying users and advertisers, we believe there is a tremendous value proposition in providing enhanced data mining and visualization capabilities on Twitter. With the shear volume of tweets generated each day, it is quite difficult for both users and advertisers to obtain an accurate sense of public sentiment given a query term and related trends/queries given a query term. Additionally, filtering attributes such as geography, timeline trends, and dynamic visualization are currently not provided by Twitter and may prove useful for more rigorous analysis. In this project, we utilize data mining and visualization techniques such as Bayes classification, linear regression, sentiment analysis, and natural language processing to provide these enhanced features for both users and advertisers.

System Design and Implementation
We implemented an automated pipeline as seen in Figure 1 that begins by prompting users to specify a search keyword and time duration through our web interface. The web host will then send a request query to MarkLogic and its associated tweet database on iCluster, and the query will be processed on iCluster and return the desired tweets in XML format. Next, we utilize Scala and Matlab to evaluate sentiment value (positive/negative) using a Naive Bayes classifier and linear regression model. Lastly, we visualize both frequency and sentiment by time and geography through the Google Charts API. We will describe each of these components in more detail below.

Figure 1: Overview of system design

Web User Interface

We designed a simple web interface to allow users to visualize a set of keywords over a user-specified date range. The front-end was implemented utilizing a combination of HTML, CSS, and Javascript. The parameters are then packaged and parsed by server-side code implemented in PHP. We also included a convenient calendar utility to enable quick entry of dates, and this was implemented using the jQuery library. A screenshot of our web interface can be seen in Figure 2.

Figure 2: Web-based Interactive User Interface

MarkLogic Database Layer

We utilized MarkLogic to perform queries on a dataset of approximately 650 million tweets housed on iCluster. In order to accomodate timeline and geography visualization efficiently, we had to optimize our MarkLogic queries to return results as quickly as possible. Although this was quite challenging, we observed that frequency-only queries returned very quickly, and for queries that necessitated the raw text of the tweets, we limited our result set to 100 tweets at a time. The following is our MarkLogic query to generate XML results for a timeline visualization for the keyword “christmas”:

```
<timeline xmlns="http://www.bid.berkeley.edu/statuses"> {
let $startDate := xs:date("2011-11-21")
let $endDate := xs:date("2012-01-30")

let $numDays := fn:days-from-duration($endDate - $startDate)

for $i at $j in (0 to $numDays)

return
<record> {
 <date>{
   <year> {fn:year-from-date($startDate + xs:dayTimeDuration(fn:concat("P", xs:string($i), "D")))} </year>,
```

```
  <month> {fn:month-from-date($startDate + xs:dayTimeDuration(fn:concat("P", xs:string($i), "D")))}
</month>,
  <day> {fn:day-from-date($startDate + xs:dayTimeDuration(fn:concat("P", xs:string($i), "D")))} </day>
 }</date>,

 <number> {count(
        cts:search(//status, cts:and-query((
        cts:element-range-query(xs:QName("created_at"),
">=" ,xs:dateTime(fn:concat(xs:string($startDate + xs:dayTimeDuration(fn:concat("P", xs:string($i), "D"))),
"T00:00:01-08:00"))),
        cts:element-range-query(xs:QName("created_at"),
"<=" ,xs:dateTime(fn:concat(xs:string($startDate + xs:dayTimeDuration(fn:concat("P", xs:string($i), "D"))),
"T23:59:59-08:00"))),

        cts:element-word-query(xs:QName("text"),"christmas"))),"unfiltered")
        )}
 </number>



}</record>
}</timeline>
```

It is worth noting that we performed range queries on the created_at timestamp to leverage the range index built on the variable. Since we do provide date filtering functionality, optimizing such a range query went a long ways towards optimizing the entire trip time with MarkLogic.

Data Mining Techniques

Dataset Used

We use a Twitter dataset consisting of about 3 months of the 1% sample stream. This data set is approximately 50GB compressed and a little over 0.5 TB indexed. To train our sentiment classifiers, we utilize a Twitter emoticon dataset from a research project at UC Berkeley in which users were prompted to input their moods at the time of their tweets. We randomly selected 10,000 tweets from this emoticon dataset for training and tested on a separate randomly-selected set of 1,000 tweets. Some samples of tweets (positive in blue, negative in red) from this twitter emoticon dataset are shown below:

@uniqueLEEtiff happy birthday sunshine! have the best day ever :-)
RT @makisigmorales2: good afternoon...blessed Sunday guyz :-)
RT @RyeRye: We can find heaven if we go look together :-)

bittersweet. nice that im going back to australia but there are so many things to leave here in the philippines. :-(
What the hell am I doing up at this time on a Sunday? :-(
Dad talkin about medical check ... scares shit outta me!!! :-(

Before beginning data mining analysis, we tokenize the input tweets and eliminate stop words such as “a”, “an,” “the”, “is”, “at”, etc. We also utilize the classic Porter Stemming algorithm to stem words down to their root. “Running” is trimmed to “run” for instance. For our data mining

component, three techniques, Naive Bayes, Linear Regression and SVM, are used to train and test the sentiment value of each tweet.

Naive Bayes

Using Bernoulli classification, we built a dictionary mapping each token to its associated count. Laplace smoothing is used to model term conditional probabilities. we utilize the probability model gleaned from our training data to classify new and unseen data. Our decision rule is as follows

$$c_{\text{map}} = \arg\max_{c \in C} [\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c)].$$

As described in the formula, we compute the log-likelihoods of each class being attributed to the data. Then, if the log-likelihood for the positive class for a given document is greater than the log-likelihood for the negative class for that same document, we'd classify that document (tweet, movie review, etc.) as positive and vice-versa.

Linear Regression with SVD

Linear regression is utilized to predict a response variable (y) from a vector of inputs (X). The formulation of this problem is as follows:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The Beta matrix is the set of weights we need to calculate in order to classify unseen data. X is our feature set, and in this case, consists of word counts, the number of times each word in the lexicon appears in a given review. Y represents the ground truth values, which are smiley or frown faces obtained from our Twitter emoticon dataset.

Support Vector Machine (SVM)

We implement SVM by using different kernels. The following is a summary of the SVM kernels we have used.

Linear Kernel:	$k(x,y) = x^T y + c$
Sigmoid Kernel:	$k(x,y) = \tanh(\alpha x^T y + c)$
Polynomial Kernel:	$k(x,y) = (\alpha x^T y + c)^d$
Gaussian RBF:	$k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$
Intersection Kernel:	$k(x,y) = \sum_{i=1}^{n} \min(x_i, y_i)$

Data Mining Results

In Table 1, we list the most significant 10 words for the classification of positive and negative tweets and compare these words with the most significant 10 words for an Amazon movie review dataset in Table 2. We can see that tweets contains more words involving emotion such as "love", "thank", "lol" (laugh out loud).

Twitter Dataset

Positive	Weight	Negative	Weight

love	0.0085	miss	0.0035
thank	0.0066	i'm	0.0028
lol	0.0061	que	0.0021
follow	0.0059	just	0.0019
good	0.0054	like	0.0018
just	0.0048	la	0.0018
gui	0.0045	imu	0.0016
i'm	0.0045	feel	0.0015
like	0.0038	im	0.0015
gt	0.0037	want	0.0012

Table 1: The most significant words in positive and negative tweets

Amazon Movie Review Dataset

Positive Review	Weight	Negative Review	Weight
film	0.0157	film	0.0141
movie	0.0082	movie	0.0109
like	0.0054	like	0.0061
charact	0.0053	it'	0.0053
It’	0.0051	charact	0.0052
make	0.0045	just	0.0046
time	0.0043	make	0.0046
scene	0.0037	time	0.0043
stori	0.0037	scene	0.0039
just	0.0036	good	0.0036

Table 2: The most significant words in positive and negative Amazon movie reviews

In Table 3 and Figure 3, we compare the accuracy across these various techniques and find that Linear SVM achieves the best performance with 83.10% accuracy.

Method	**Accuracy**
Naïve Bayes	79.66%
Linear Regression	81.32%
Linear SVM	**83.10%**
Sigmoid SVM	31.40%
Polynomial SVM	79.00%
RBF SVM	79.00%
Intersection SVM	82.70%

Table 3: Accuracy of our data mining techniques

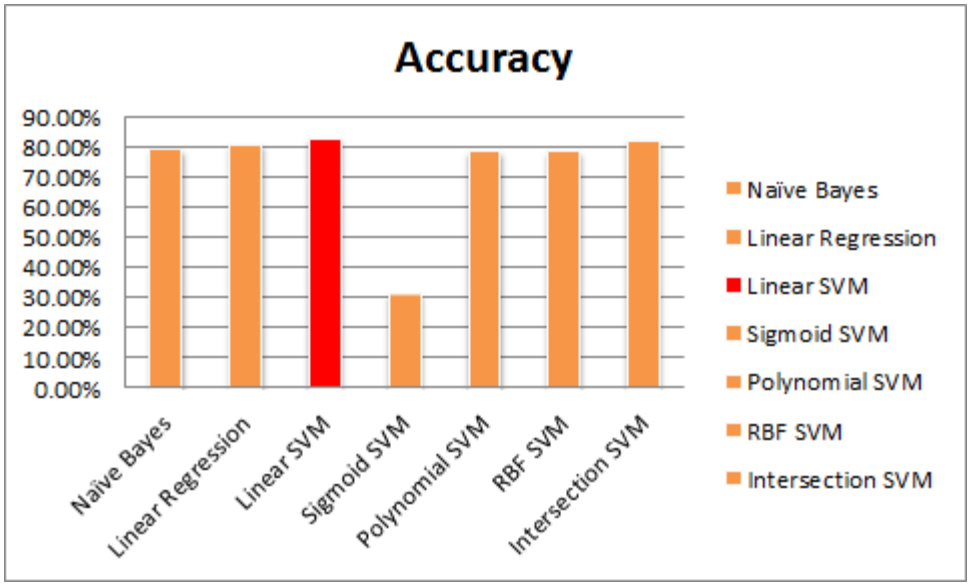

Figure 3: Accuracy by Data Mining Technique

Visualization Layer

After obtaining sentiment values through our data mining analysis, we turn to creating our final output--timeline or geographical visualization. We utilize a combination of Javascript and the Google Charts API to produce our timeline and geography visuals. Example documentation of the timeline chart can be found on Google's code playground here:

https://code.google.com/apis/ajax/playground/?type=visualization#annotated_time_line

Results and Discussion

In this section, we offer several sample visualizations and their corresponding analyses produced with our platform to provide readers a taste of the possibilities of Tweet Visualizer.

Chronological Visualizations

Our first plot (Figure 4) is a timeline-frequency visualization of the keyword “Christmas”. Frequency peaks on Christmas Eve, dips a bit into Christmas Day, and almost immediately falls off December 26th and onwards. Although we expected peaks near Christmas Day, we were a bit surprised by how sudden the drop-off is Dec. 26th and onwards. We also noticed that people talk about Christmas much more before Christmas than afterwards.

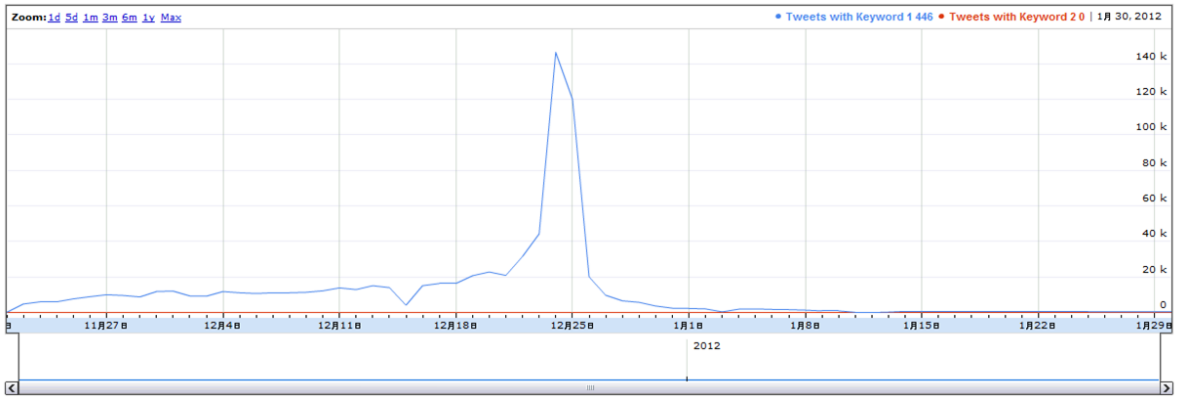

Figure 4: Timeline visualization for word “Christmas”from November 2011 to January 2012

Our next chronological visualization is an event-driven frequency analysis of “Romney” and “Santorum”, two candidates from this year’s Republican primary elections. The time span we visualize is from late February to early March in 2012, during which several key elections occurred. As shown in figure 5, the frequency over this time span aligns nicely with the primary election dates (as listed in figure 6), with the frequency peaking at two pivotal days (February 28th and March 6th, known as “Super Tuesday”). Comparing the tweet frequency of the two candidates, we find that although at beginning Satorum is leading, on two important days Romney is ahead in the frequency, indicating that he is more “popular” during and between the two election days.

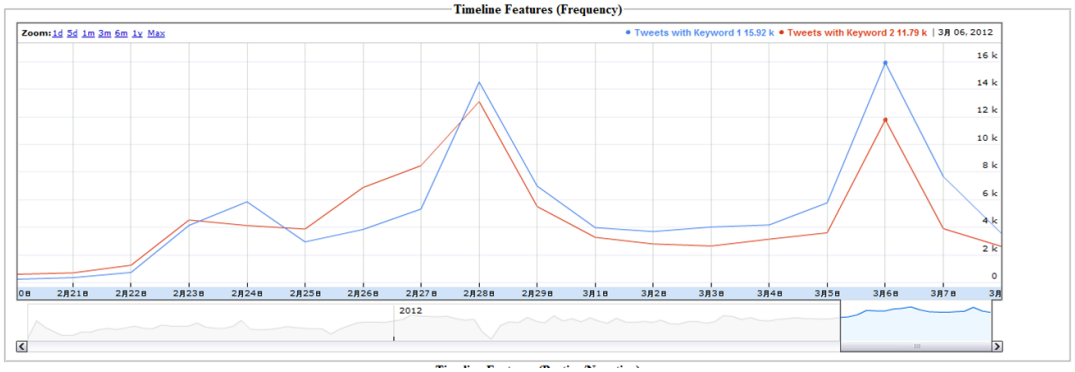

Figure 5: Timeline visualization for Romney (blue) and Santorum (red) in late February and early March, 2012

February 11, 2012	Maine	Results	24	Caucus	Caucus Information from Maine GOP
February 28, 2012	Arizona	Results	29	Primary	Primary Information from Arizona Department of State
	Michigan	Results	30	Primary	Primary Information from Michigan Department of State
March 1, 2012	Wyoming	Results	29	Caucus	Caucus Information from Wyoming GOP
March 3, 2012	Washington	Results	43	Caucus	Caucus Information from Washington GOP
March 6, 2012 (Super Tuesday)	Alaska	Results	27	Caucus	Caucus Information from Alaska GOP
	Georgia	Results	76	Primary	Primary Information from Georgia Department of State
	Idaho	Results	32	Caucus	Caucus Information from Idaho GOP
	Massachusetts	Results	41	Primary	Primary Information from MA Sec. of Commonwealth
	North Dakota	Results	28	Caucus	Caucus Information from North Dakota GOP
	Ohio	Results	66	Primary	Primary Information from Ohio Department of State
	Oklahoma	Results	43	Primary	Primary Information from Oklahoma State Election Board
	Tennessee	Results	58	Primary	Primary Information from Tennessee Department of State
	Vermont	Results	17	Primary	Primary Information from Vermont Department of State
	Virginia	Results	49	Primary	Primary Information from Virginia Board of Elections
					-Only Mitt Romney and Ron Paul will appear on the VA ballot, see this report
March 10, 2012	Kansas	Results	40	Caucus	Caucus Information from Kansas GOP
	U.S. Virgin Islands	Results	9	Caucus	Caucus Information from VI GOP
	Guam	Results	9	Caucus	Caucus Information from KUAM News
	Northern Mariana Islands	Results	9	Caucus	Caucus Information from Saipan Tribune

Figure 6: List of Republican primary elections and their dates [1]

The following visualization shows the average sentiment value of the keyword “Microsoft” from March 1st to March 8th, (Figure 7). The reference graph we use is the stock price of Microsoft (NASQ: MSFT) from March 1st to March 8th, as shown in Figure 8. Although there is no direct relationship between people’s sentiment attitude towards the name of the company and its stock price, it is reasonable to assume that people’s attitude towards a company will influence (or will be influenced by) the company’s stock price. As shown in these two figures, the drop in sentiment value is in accordance with a stock price drop on the same day, and the same holds for the recovery afterwards. However, we believe this relevance is not very strong in this case.

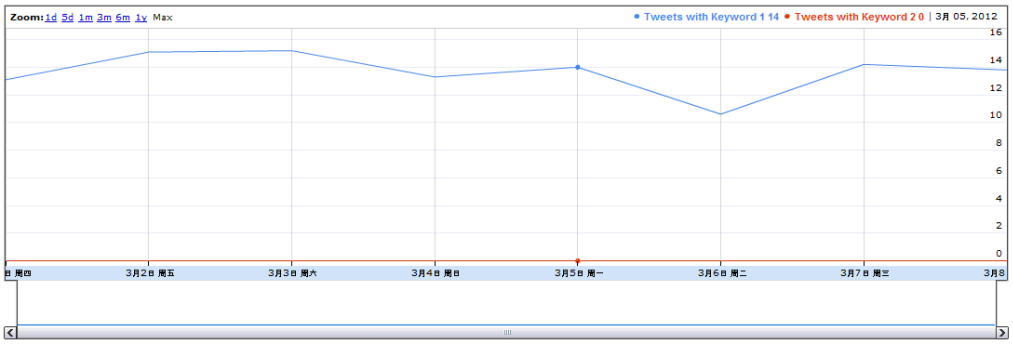

Figure 7: Sentiment value of Microsoft from March 1st to March 8th

Figure 8: Stock Price of Microsoft (NASDAQ: MSFT) from March 1st to March 8th [2]

We use the previous example again to produce another visualization of sentiment value over time. The following visualization (Figure 9) shows the sentiment value of “Romney” and “Santorum” in late February and early March, 2012. Compared with the frequency distributio plots of Figures 7 and 8, this visualization reveals more about people’s sentiment attitude, which could have a stronger influence on their voting behavior. On Feburary 28th, Santorum has higher overall sentiment value and on March 6th, two candidates are even, but afterwards, Romney won out. This is all in accord with the ground truth of the voting results, as shown in figure 10.

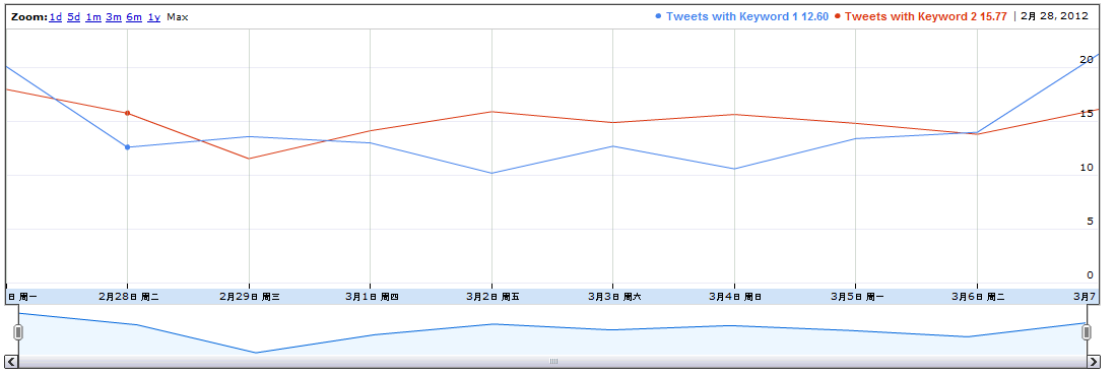

Figure 9: Chronological sentiment value of Romney (blue) and Santorum (red) in late February and early March, 2012

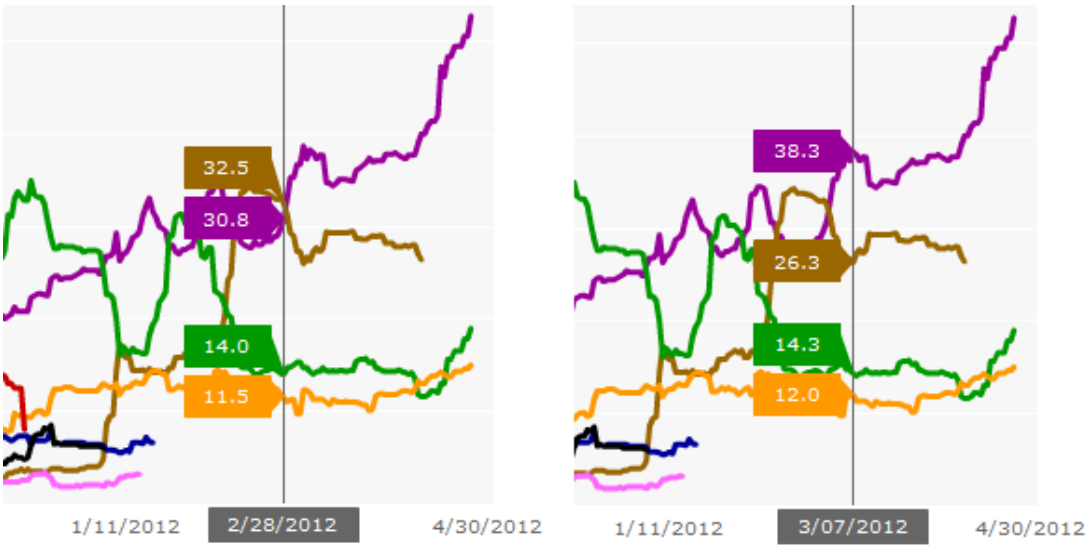

Figure 10: Voting Rate of Romney (brown) and Santorum (purple) in late February and early March, 2012

Geographical Visualization

Our first geographical visualization visualizes the frequency distribution of the keyword “startup” from November 21th 2011 to March 8th 2012, as shown in Figure 11. Darker (redder) regions indicate higher affinity. Not surprisingly, California has the highest frequency which illustrates Silicon Valley’s leading position as the national high-tech center. Other states such as TX and NY also show moderate levels of affinity towards “startup”.

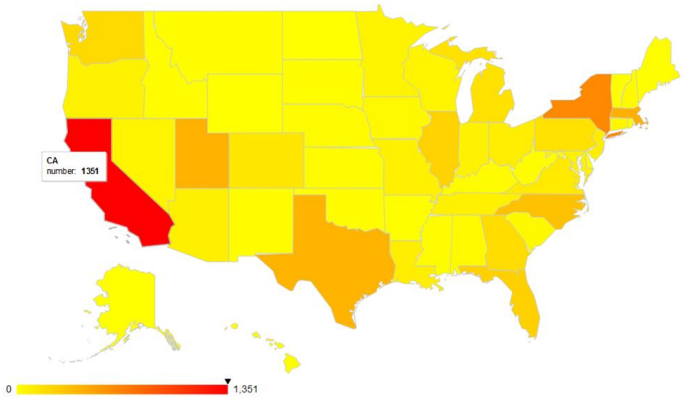

Figure 11: Geographical visualization for word “startup” (Nov 21st, 2012 to Mar 8th, 2012)

Perhaps a more interesting example is our visualization of natural disaster keywords, for example, “tornado” (as shown in Figure 12) and “earthquake” (as shown in Figure 13), during November 21st, 2011 to March 8th, 2012. Compared with the geographical statistics of average yearly tornadoes and USGS’ earthquake hazard map, the frequency visualization is in accordance with the numbers of occurrences.

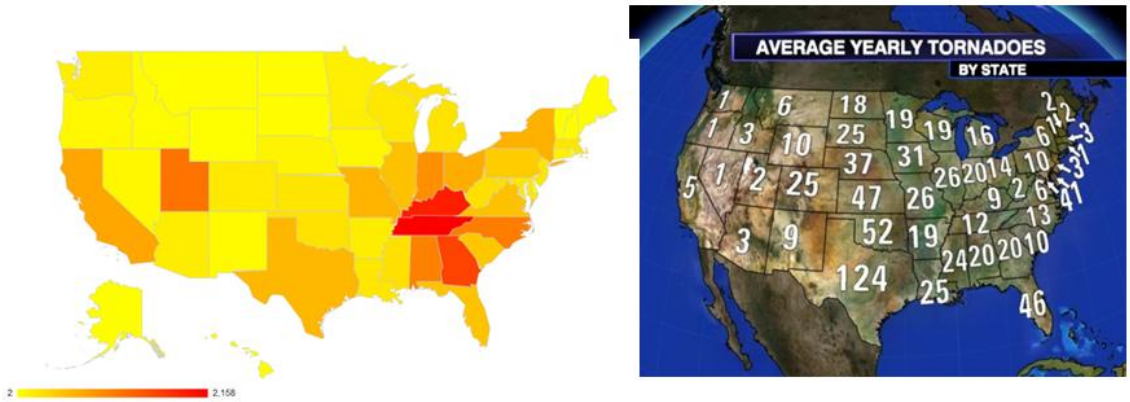

Figure 12: Geographical visualization for word “tornado” and ground truth (right image) [4] Nov. 21st, 2012 to Mar 8th, 2012)

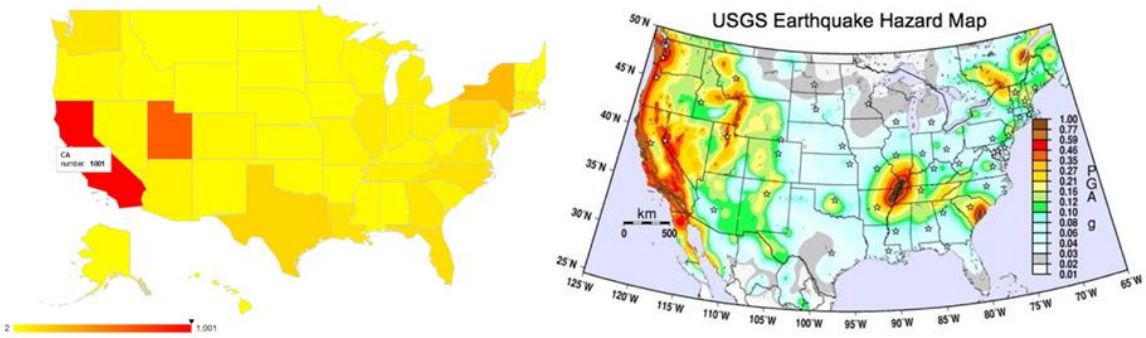

Figure 13: Geographical visualization for word “earthquake” and ground truth (right image) [5] Nov. 21st, 2012 to Mar 8th, 2012)

The following image displays a geographical visualization of the keyword ‘sailing’ (as shown in figure 13). Darker colors indicate high interest in the keyword while lighter regions represents less interest. Compared to the topographic maps of the U.S (which represent the ground truth). we can see that the states near bodies of water, such as California, Utah, Florida, New York, Maryland and Texas, achieve higher affinity scores. We also take into consideration the population as a confounding variable, since states with more population tend to have higher tweet frequency generally. To eliminate this effect, we normalize over all states, importing the population data (using number of election seats as a proxy for state population). We find that after normalization, Utah and Washington show significantly higher affinity scores, meaning more tweets per person on “sailing” than other states.

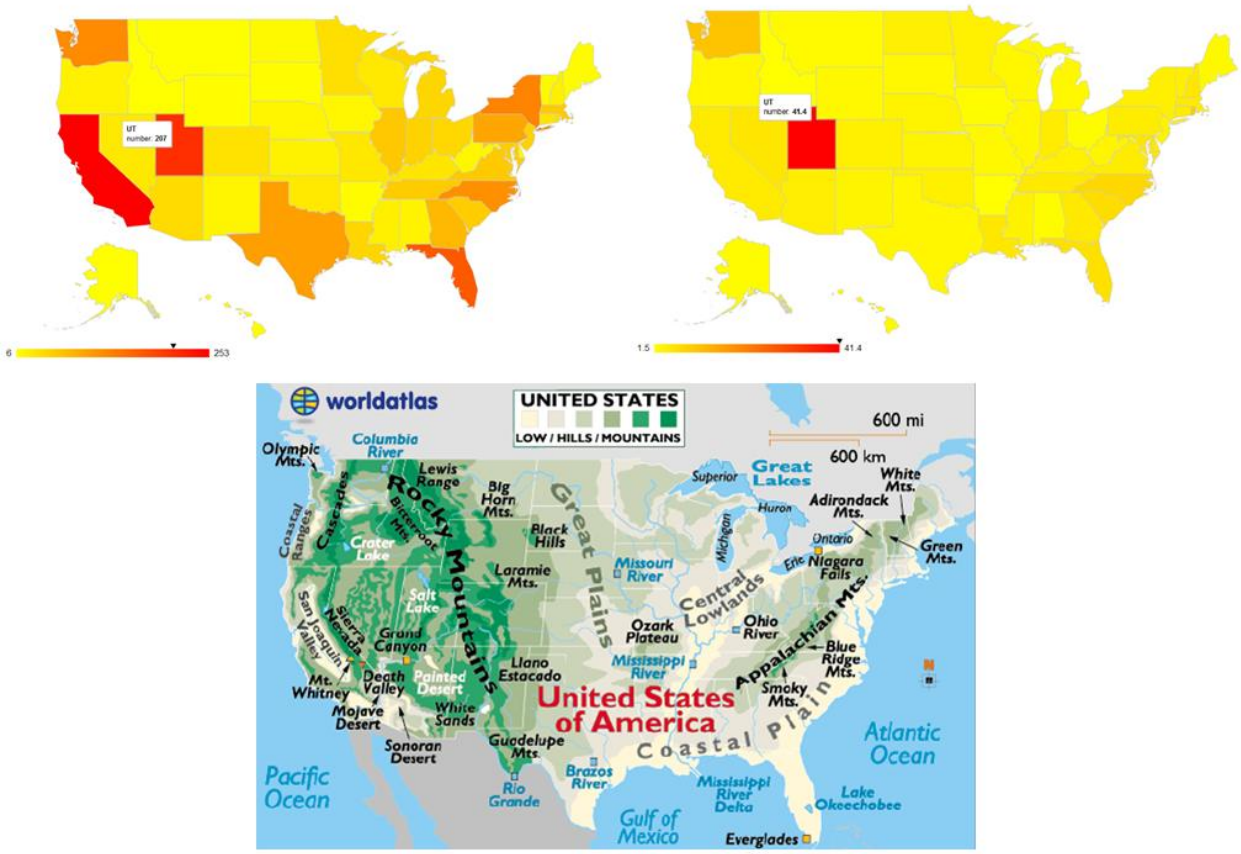

Figure 14: Geographical visualization for word “sailing”, from Nov 21st to March 8th 2012. Top-left: Raw frequency visualization , Top-right: normalized frequency visualization , Bottom: Ground truth topology map [6]

We also visualized the sentiment value of the weather, as seen in Figure 15. We made this visualization because we believe people’s attitude towards weather varies across different geographies. Our dataset used here is one week over the winter, and our results show that people show a more favorable attitude to the weather in states like California, Florida, Hawaii, while exhibiting a more negative attitude in states such as Arizona, Massachusetts, New York, and Washington. This aligns with people’s common sense of weather in different states.

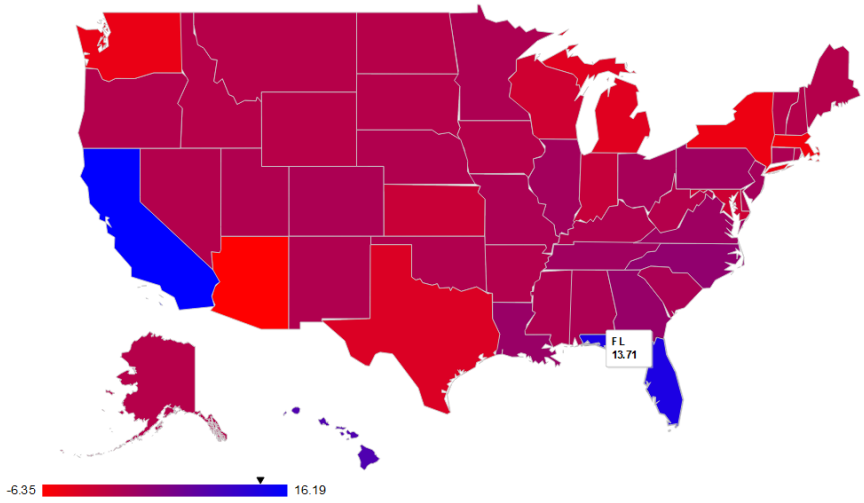

Figure 15: Geographical visualization for the sentiment value of word “weather”

Our final visualization displays affinity for the keywords “Democrat”, as shown in figure 16. Blue states show strong indications of supporting the Democratic Party while we believe red states align more closely with the Republican Party. The states with purple color are inconclusive states both parties are fighting for. States like CA, NY, PA, MN, IL show higher positive attitude towards the Democratic party and states like MT, ND, WY, etc. show more negative value, indicating support for the rival party republican party. The right image is a prediction of the election in 2012. States in green are the competing areas, and we believe we may be able to offer election predictions using our visualization.

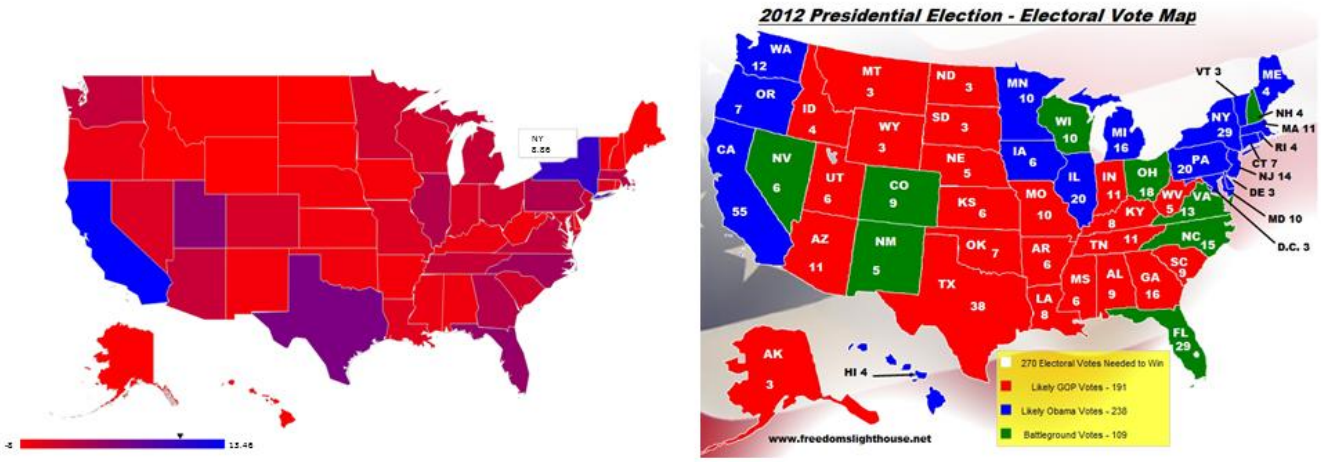

Figure 16: Geographical visualization for the sentiment value of word democrat. Left Image: Tweet Visualizer output for sentiment value across US. Right Image: 2012 Election prediction representing ground truth [7]

Conclusion

In conclusion, we believe very strongly in the value proposition of visualizing the Twitter ecosystem. There is tremendous value to be gained for both users and advertisers alike, and we the results presented in this paper are just a taste of what’s possible with this rich dataset.

Our chronological visualization shows frequency of a given keyword over a specified timeline. We can use this visualization to observe the trending of people's discussion over time of a given keyword, upon which we can detect event-related spikes and lows and draw conclusions regarding tweeting behavior through an event. Our chronological visualization on sentiment value shows people's attitude towards a given keyword in a given duration. We find that this visualization shows some relevance to our ground truth references. We believe can further improve our algorithms and data mining techniques to offer more accurate predictions and real-time analysis.

Our geographical visualization on frequency shows the number-of-tweets containing a given keyword distributed over a given area. We find that this distribution is highly relevant to a ground truth distribution in some cases, such as natural disasters ("earthquake", "tornado") or industry-specific keywords ("startup", "entrepreneurship"). We can also use this frequency map to observe people's behavior (as in our analysis of the keyword "sailing") over an area. By normalizing the raw frequency counts with population data, we can depict a normalized map which shows people's average number of tweets (thus average affinity) to a given keyword within an area. Our geographical visualization on sentiment value shows people's attitude distribution toward a given keyword in different areas. Our visualization of "weather" and "democrat" is in accordance with our common sense and real political environment data.

Future Work

In the future, we will adapt our pipeline into more applications and visualizations and build an open platform to enable users and advertisers to visualize the rich Twitter data stream. Regarding the technical implementation of our algorithms, we hope to obtain improvements in our data mining analysis. As our feature vector has high dimensions, it is more efficient for training by employing dimensionality reduction, which makes data more compact and dense. We plan to use LDA to compress the feature set to increase the speed of training and classification. Furthermore, LDA can improve both recall and precision. We will also try the Learning Vector Quantization (LVQ) classifier to improve the robustness of classification.

References

1. 2012 GOP primary/caucus schedule, http://www.2012presidentialelectionnews.com/2012-republican-primary-schedule/
2. Stock price of Microsoft Corporation from March 1st to March 8th, 2012 (NASDAQ:MSFT) http://www.google.com/finance?cid=358464
3. 2012 Republican Presidential Nomination http://www.realclearpolitics.com/epolls/2012/president/us/republican_presidential_nomination-1452.html
4. Average yearly tornadoes by states http://mikeheard.files.wordpress.com/2010/06/3-3-tornadoes-by-state.jpg
5. USGS Earthquake Hazard Map. http://img.ibtimes.com/www/data/images/full/2011/08/23/149920-usgs-earthquake-hazard-map.jpg
6. United States Landforms map http://www.drkresearch.org/resources/us_landforms_map.png
7. 2012 United States presidential election - electoral vote map http://freedomslighthouse.net/wp-content/uploads/2010/12/2012electoralmap050412.jpg