Real-time Twitter NYC

FINAL PROJECT DESCRIPTION

Yang Liu	Jinglong Li	Yuan Zhou
yl3710@nyu.edu	jl7152@nyu.edu	yz2996@nyu.edu
yl3710	jl7152	yz2996

Project page (on Github):https://github.com/NYU-CS6313-SPRING2016/Group-6-Twitter-NYC
Video: https://vimeo.com/167764465
Working demo: https://twitter-nyc.com/ (IP: 107.182.225.198)

What is the problem you want to solve and who has this problem?

For users curious about what is happening on Twitter in New York City, our project is exactly what they want.
The problem is what is happening/what are people talking about in real time and where of these topics distribute among New York City.

What are the driving analytical questions you want to be able to answer with your visualization?

1. What topics are people in NYC mostly talking about right now? What are details of these topics?

When users open the website, the first information they want to know is “what”. So we need to extract the hot topics from tweets posted in NYC and then visualize them. After that, we need to show the corresponding tweets’ contents of these topics to let them understand better.

2. Where are those hot topics distribute in NYC?

After knowing what are hot topics, users may also want to know the distribution of these topics. So we need to show the tweets on map.

3. What are people talking about the topics that users interested in? What are details of tweets?

Users may want to know the details of a specific topic being talked about. So there should be a search function to let them interact with this page. Users need to input a topic they interested in and see the distribution among city and the contents people talking about in Twitter.

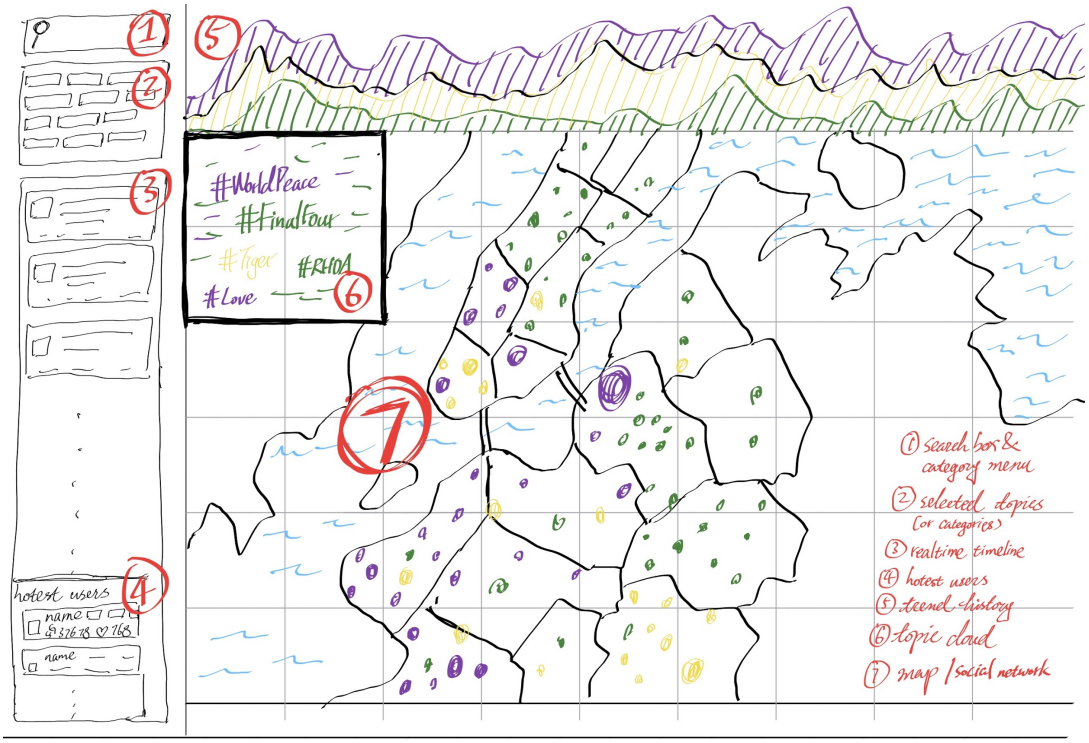
What does your data look like? Where does it come from? What real-world phenomena does it capture?

Attribute names	Attribute type	description	Value range	Derived
geolocation	Quantitative	The location(longitude and latitude)of tweets	[-74,40] - [-73,41]	N
user	Categorical	The username of twitter account		N
text	Categorical	The content of a tweet		N
hot_topic	Categorical	Hot topic; collect from hashtags and text of tweets, count the words by frequency and then extract the top 10.		Y
create_time	Ordinal	When did the tweets be posted	2016 Mar 24 - present	N
topic_correlation	Quantitative	The correlation between two topics	0.0 - 1.0	Y

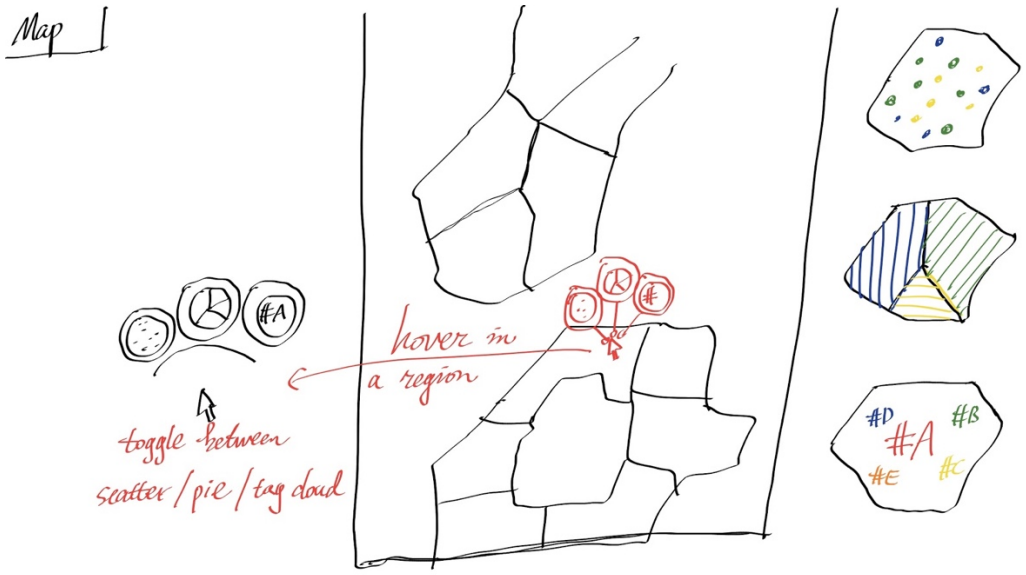
What have others done to solve this or related problems?

In a visualization project named “Twitter Monitor” that was implemented in Fall 2015, the author use the map of Manhattan as the main view, divided in rectangular subareas, which can be zoomed by the User. The rectangles have entities inside it which are sized based on the relevance of entity in the tweets from that area and colored based on age of entity.
This can provide the information about different hot tweets trends in different areas. They also applied an entity list to show the list of entity based on area where the user has clicked. A Tweets list are used to show the tweets based on the entities the user has selected, and they are sorted by the followers of the user posting the tweet, which displays the top topics in real-time. An user list is applied to show the users who posted tweet related on the selected entity, which is also sorted based on the followers.
This project’s topic is similar to us, and also implement “what are hot topics and where are topics distribute among Manhattan”.

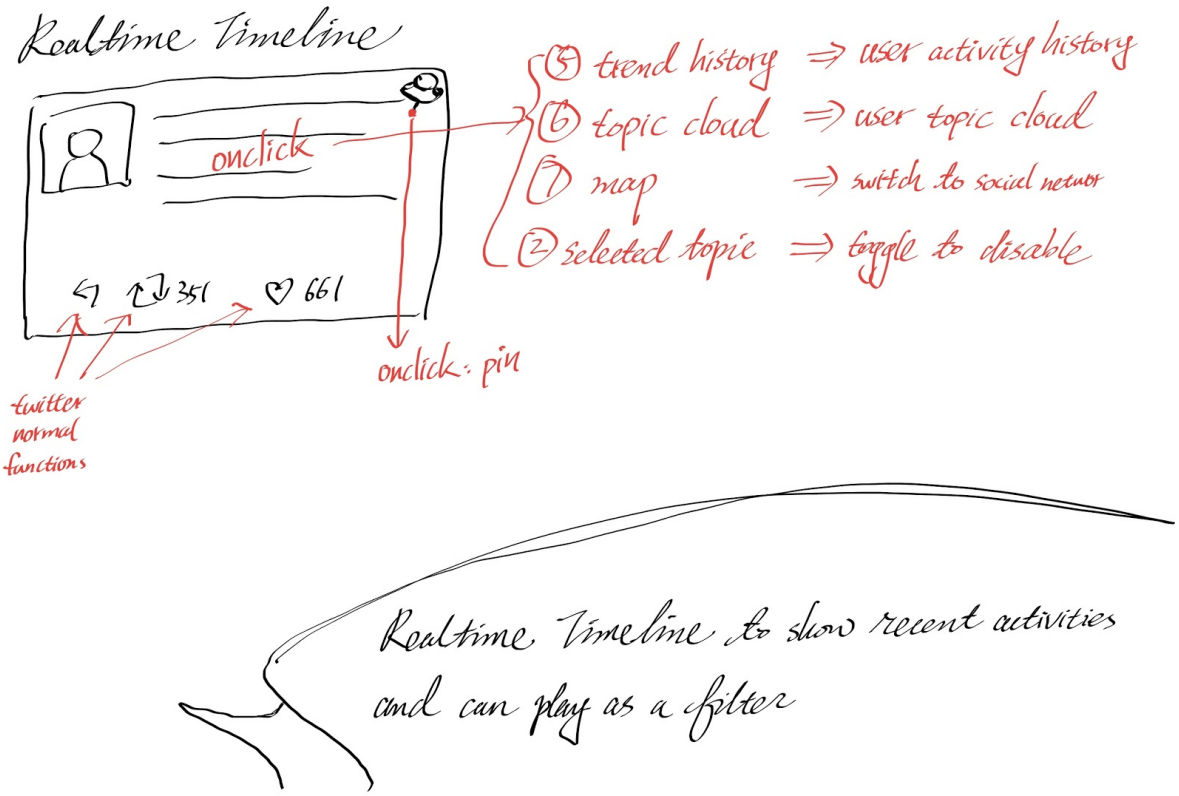
Initial Mockup



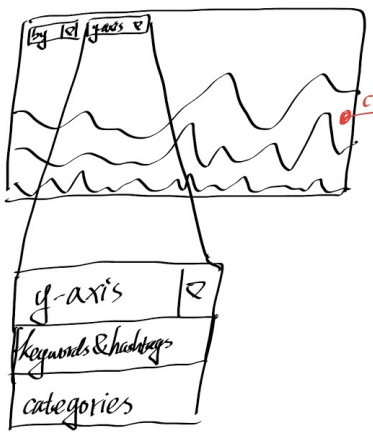
1. How to discover recent hot topics in NYC? By observing the topic cloud, users can find out the most discussed topics by the size of
the topic tag. By observing the trend history, users can also find out what was on the trend in a recent
past.
2. How to find the difference of hot topics between different area in NYC? Our map provides 3 way to represent tweets. In the scatter mode, we can distinguish the
concentrating of certain topics. In the pie chart mode, we can see the proportion of categories discussed in that area. At last, we can find out the specific trending topics from investigating in the embedded tag cloud mode in each area.



3. How to find the biggest contributors of hot topics? By counting the times of retweet, quote or favorites of tweets, which related to a specific
hot topic, we can find the author of these hot tweets, and then find those big contributors. So users can see the top 5 contributors/ recommend users in left bottom corner. Default
are those users with biggest retweet/quote/favorites count. And users can also select one or more categories to see the biggest contributors of this topic.
4. How to show the trends change over time? From the trend history stream graph, a topic is/was trending when its extent is relatively large within a certain time range.

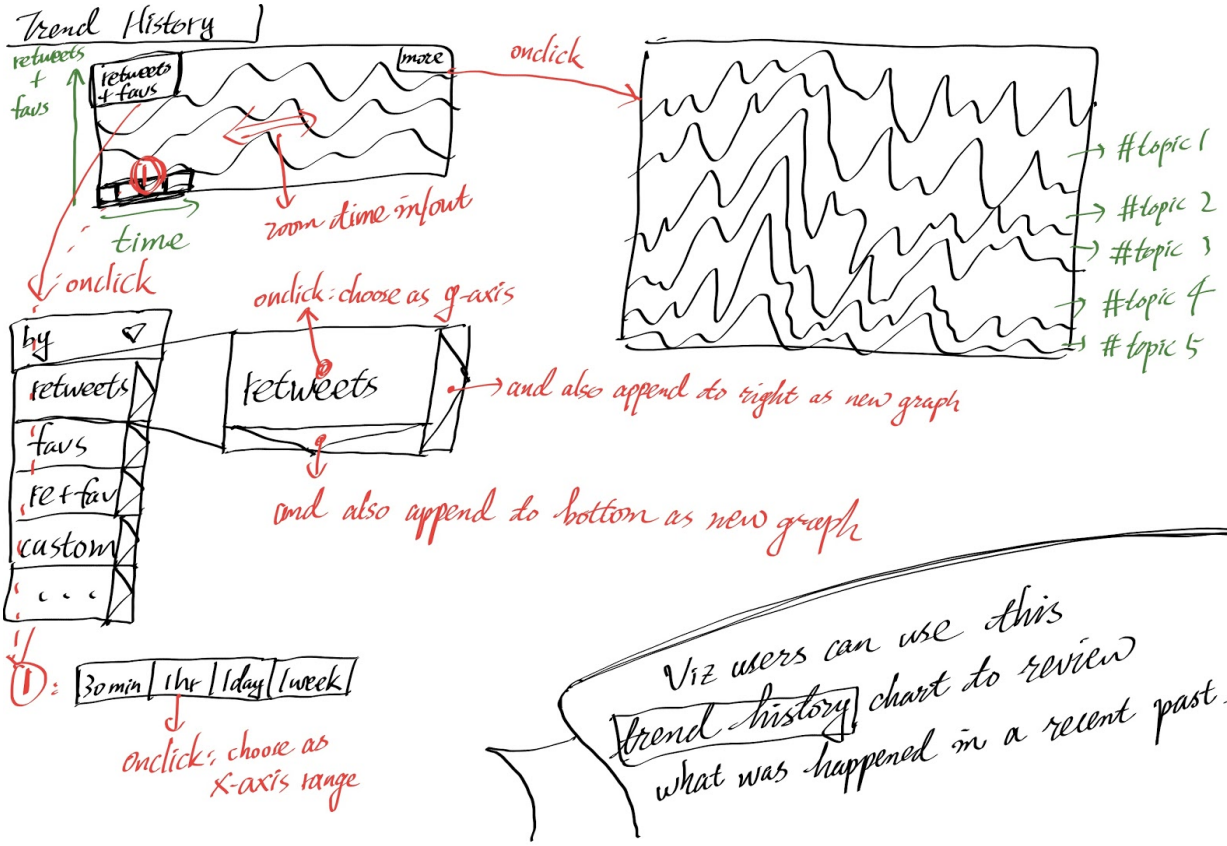

Realtime Timeline
onclick
351
661
twitter
normal
functions
onclick: pin
⑤ trend history ⇒ user activity history
⑥ topic cloud ⇒ user topic cloud
① map ⇒ switch to social netwo
② selected topic ⇒ toggle to disable
Realtime Timeline to show recent activities
and can play as a filter

Trend History [2]



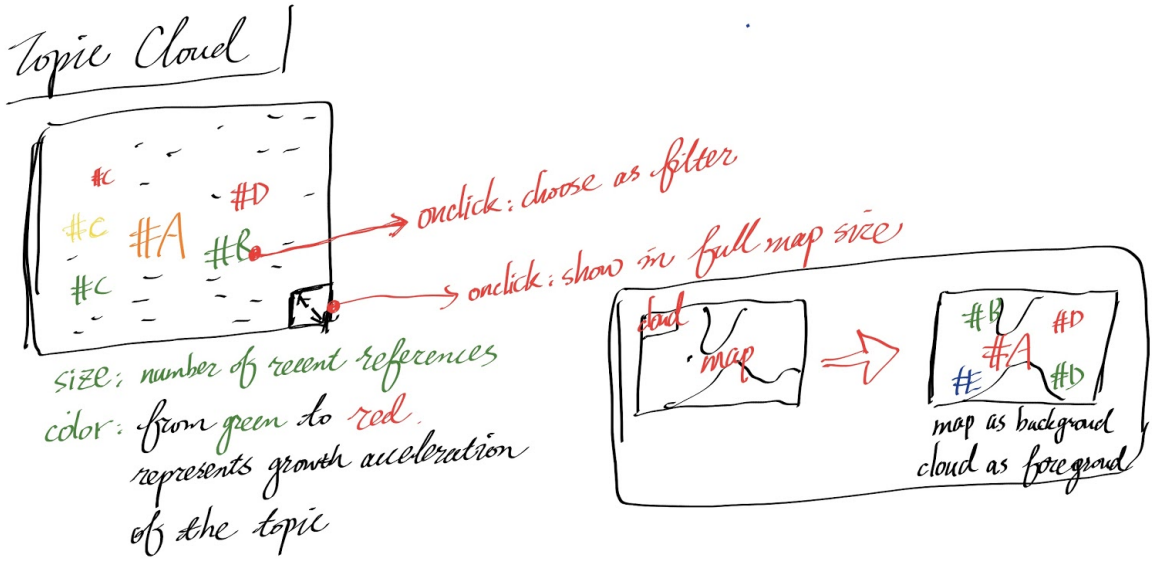
click on single stripe →
- Trend history under this category (if y-axis = categories)
- Relative topics' history (if y-axis = keywords & hashtags)

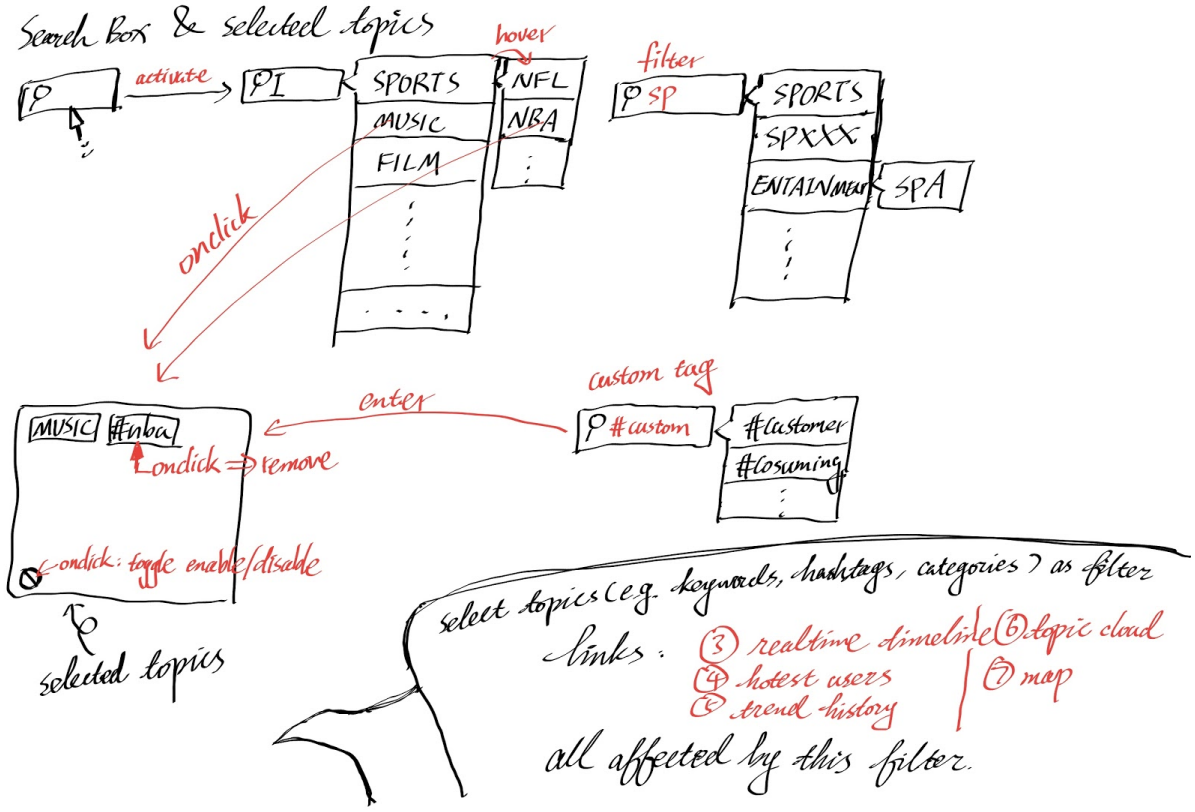
Users can also further filter by category; or show comparisons between relative topics



Viz users can use this trend history chart to review what was happened in a recent past.

5. How to guess the future hot topics in recent time? By observing the tag color in the topic cloud graph, we can tell its growth tendency
where the most conspicuous color (e.g. red) implies the fastest growth acceleration.
6. How to find the place the users with specific interest may want to go? From the map filtered by user's topics in interest, he/she may find some clusters on it
where he/she might haven't been discovered.



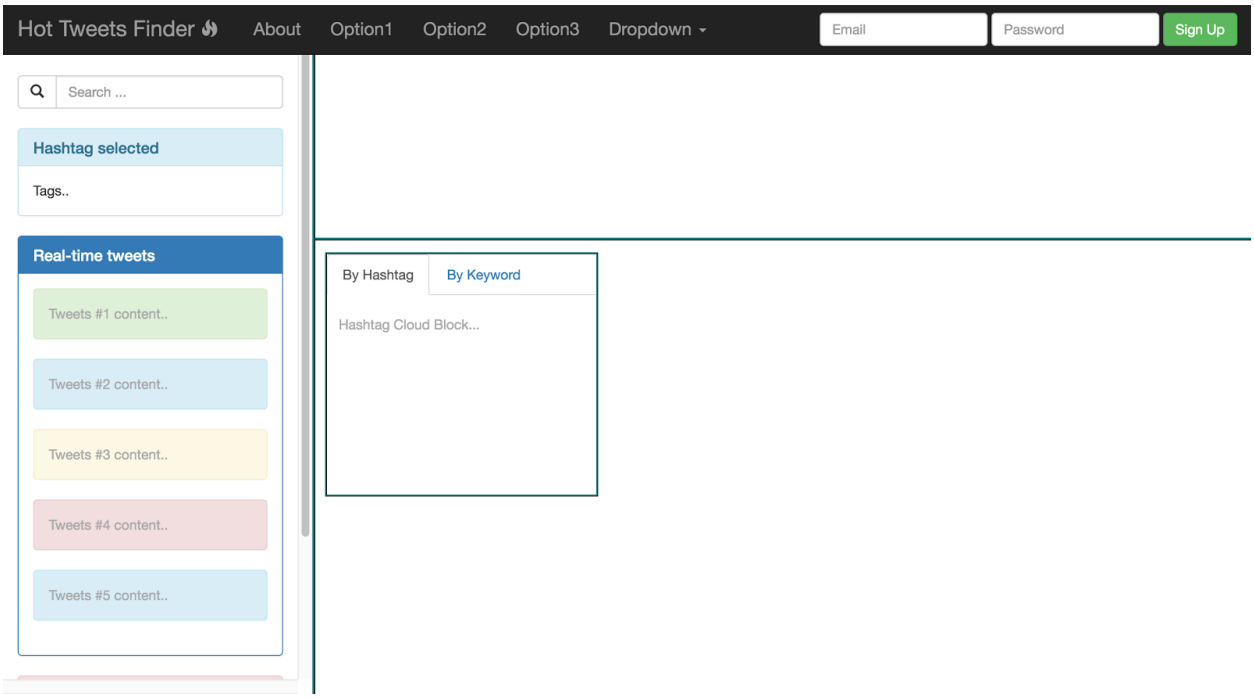


What did not implement/how we changed

Implement categorization of hot topics. There are too many topics among twitter, listing all of them may be a mass. So we think we can use categorization and then display by different categories.

Project Update

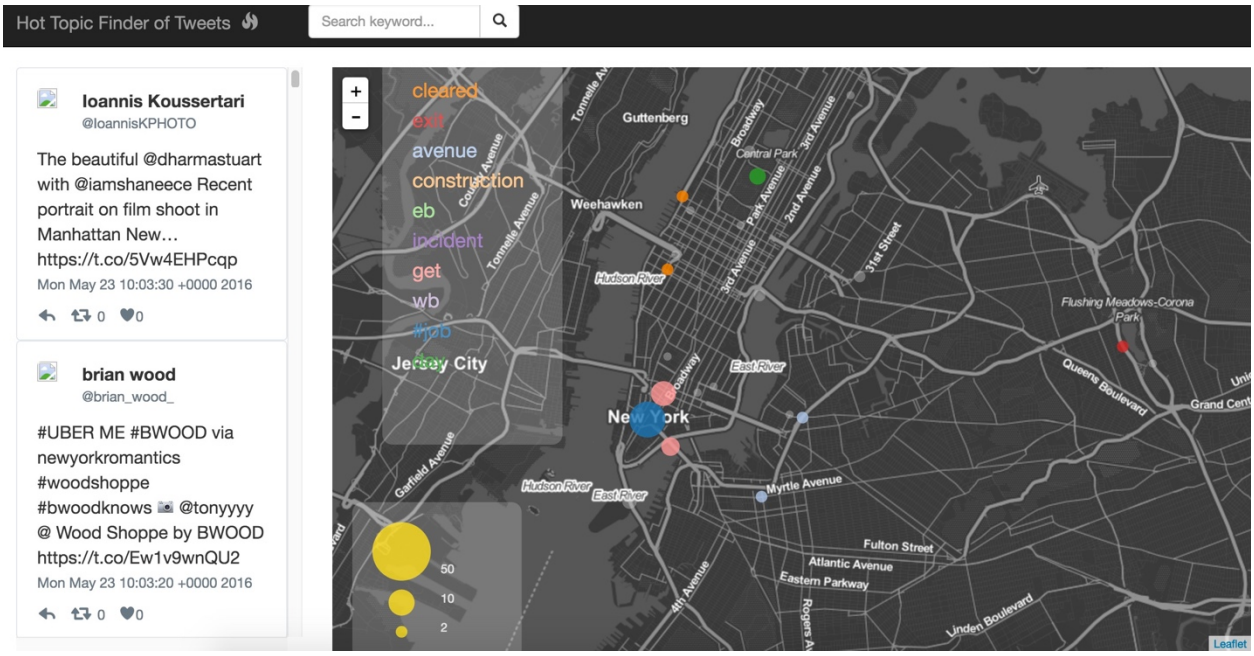
Insert images of your project update, describe how to read it, and describe what did not work and how you changed it to improve it (use a bullet list).



Explain how to read them:
Navigation bar: user can click “About” to read the introduction for this project and questions we focus on. Option 1~3 and dropdown menus are remained for changing the vis chart or other functions in the future.
Left block: search bar is for filtering the tweets users want. Real-time tweets block displays the real-time tweets content. Top 3 hottest users would be calculated on-the-fly and shown to users. (not implemented yet)
Top main block: this block is used for several vis charts to explore the data. (not implemented yet)
Bottom main block: this block contains NYC map and a relative small sub-block holding hashtag cloud/keywords cloud (not implemented yet). In the map we can find where are the real-time tweets generated and explore the distribution and more details about them.

Describe what else needs to be done until the end of the project (e.g., what is missing to complete the project):
At first visualize the hotkeys cloud in the front left corner of map module. This cloud can help users to know what is the most popular hot trends in the first time.
In the front left corner of the whole page, realize the search function. Users can search one or several keywords, by hashtag, hot keys or mixed, to know the distribution of these topics among NYC.
In the bottom left corner of whole page, list the biggest contributors of hot topics.
List the tweet content in left. And it can also be filtered by specific keywords or area.
Combine these parts together.

Final Visualization



How to read:

Left block:

Real-time tweets list block displays the all the history tweets content and does real-time update. It has a scroll bar and let users to see the details of those tweets in history.

Main block:

There are many circle shown on the map. In the left front corner, it shows the top ten popular topics by different color, the list content can be added from search bar, and the circle represents these topics by same color. The circle represents the cluster of tweets, the size of circle is corresponding to the number of tweets in neighbor, the exact size is referenced by a yellow legend bar showed in left bottom corner. If users use mouse to zoom in or out, the cluster would change its size correspondingly.

Navigation Bar:

Our website's name is showed in the left. After that there is a search bar. Besides the top ten keys, we can add other keys and it will be arranged with a new color and showed on map. We can add five keys in maximum.

What what did not work and how you changed it to improve it:

1. Hot key clouds cannot provide accurate information about "how popular" this topic: This is because every hot key with different length of letters, it is not good to just list them in cloud by different total area according to word frequency. So instead, we just list the most popular hot keys by order of decreasing frequency, and it is more intuitively.

2. Instead of just list top five hot keys, we list top ten: Since we change the data history showing on map (from all the data to just one hour's data), it is not so many tweets here. If we keep on using top five, it may just show too little color on map. After experiment we decide to keep top ten. And it does not bring any confuse to users' view and also can show more data.

3. Separation of area --use cluster instead. Classifying different area of city by latitude or longitude may have trouble for tweets in the border line between two. Instead, we use cluster to

merge the tweets in neighbor, which is also more natural. And the cluster would change their size based on users' current view.

4. Add a legend bar in bottom left. We use cluster to represent a gather of tweets. And this cluster would be changed from users' zoom in or out. In order to let users know the exact number of tweets in a cluster, we add a legend bar, as a reference, to show how many tweets here.

5. Biggest contributors. Same as time flow, it is not so interesting for users, so we spend more time for those 'interesting' functions.

Data Analysis

[NOTE: This section is very important for your final submission! If you have doubts on how to do it right contact me as soon as possible!]

Describe the results of your data analysis performed with your application. What interesting information did you extract? What are your major discoveries? What did you learn about the data? What interesting data stories can you show using your developed tool?

Please add as many screenshots as needed and describe how to read them and what interesting information can be extracted from them.

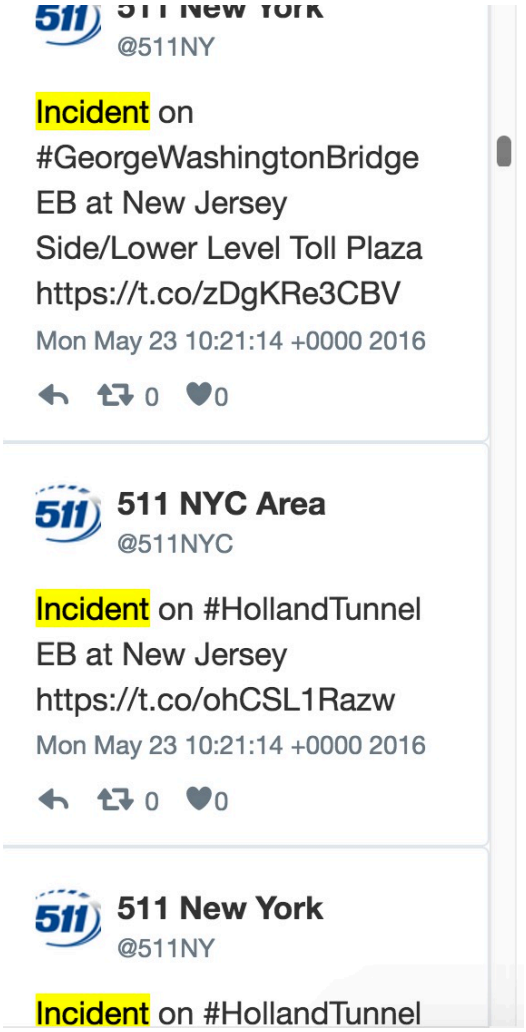
I suggest you to organize this as progression of questions and screenshots. Start with a question and show it can be answered with your tool. Start with one that has an overview, describe what can be observed there, then how this leads to a new question, then how this leads to a new question, then a new screenshot showing how this question can be answered and so on …

1.What are people talking about now?

From this screen cut we know now people are talking about incidents most. And they are also interested about exit, construction, road, …. Furthermore, there are a lot of tweets cannot be tagged whit high frequency key words. Now let's focus on the first hot key – incident.

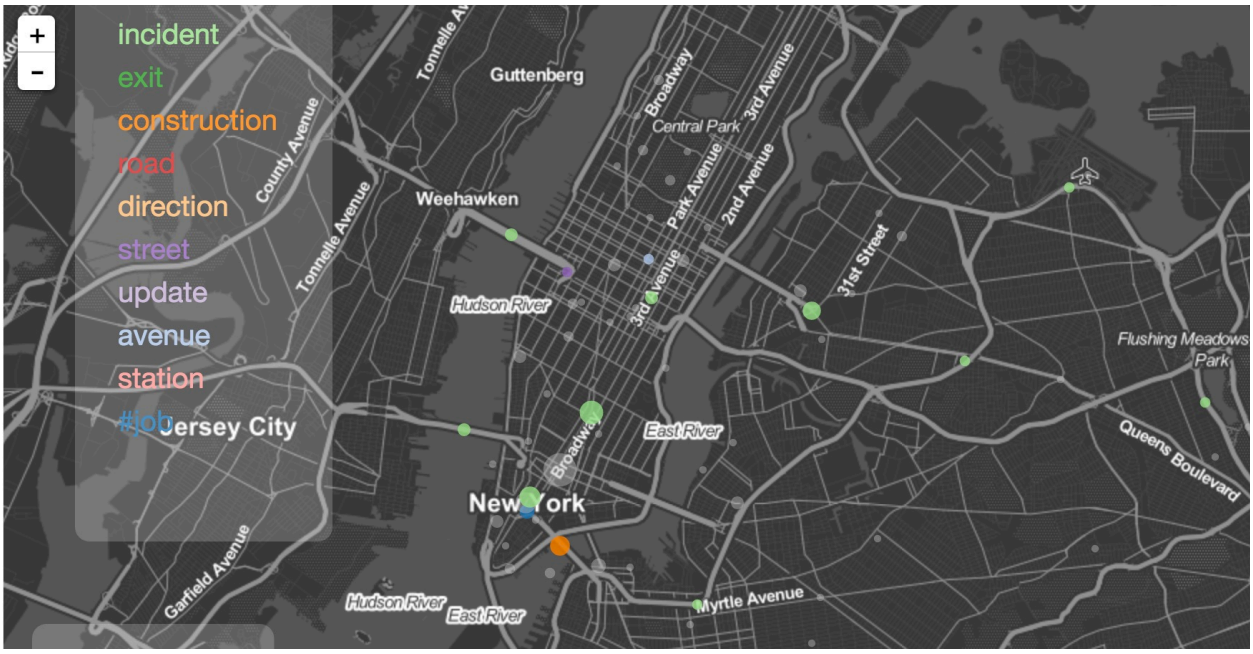
2.So what exactly people talk about?

Since we have all the history tweet list, I just use browser find function to see the tweets with word "incident". The results are as follow:



Now we know that “511 New York” updates all the incidents happening in New York. And from this we can know what is happening in details. It seems like there’s a lot of accidents happening now, what a pity.

3.So how about theses incidents distributes among the city? Is there something interesting about it?



(Now the light green represents “incident”)

Pretty obvious and interesting, the incidents happens in some particular area – the traffic route. The most obvious two are in the bridges between Manhattan and New Jersey. Now users can know that the incidents happen here is more about traffic, and also in a large amount.

Besides the traffic route, we all know that incidents happen much more in those area with dense people. From map we can see in neighbors of Broadway, it has two big green bubble – users can judge from it that there are a lot of people here. And also we know it is the truth, Broadway is pretty crowded.

For verify, I input “traffic” in search bar and insert this word into hot keys list.

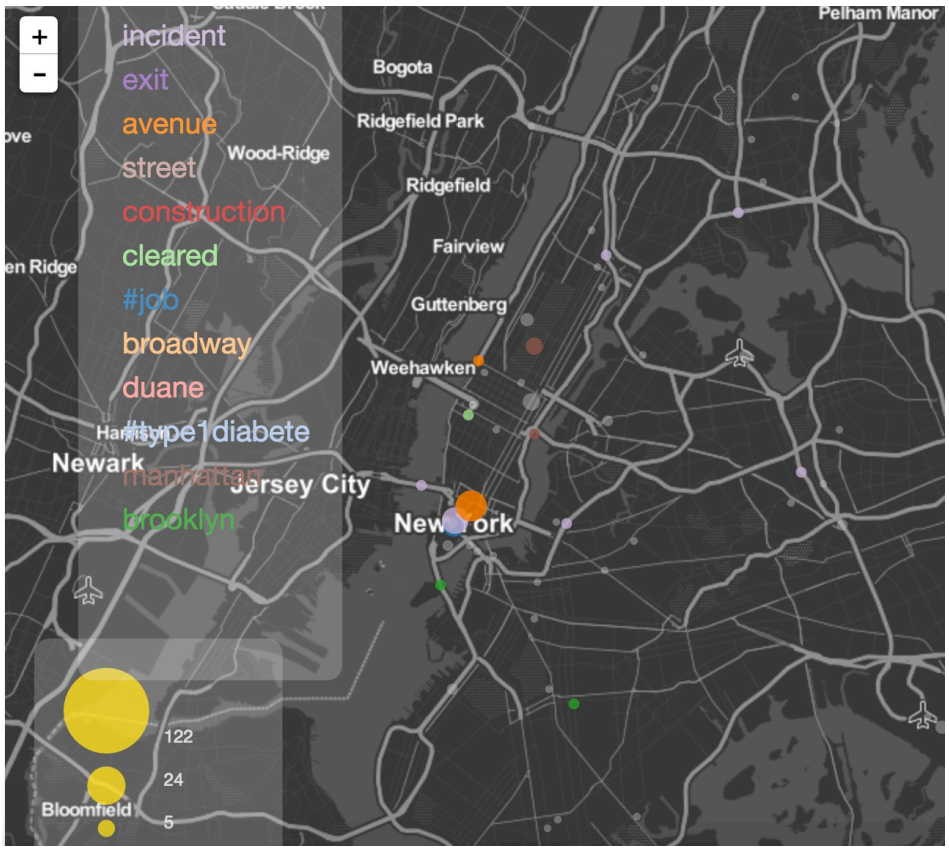
As expected, surrounding of Broadway shows a big light purple bubble. That prove what I guess – the incidents happen about traffic issues, and it also happens in those area with dense people.

4.Other discoveries?

1). For the link between density of people and tweets size, it can be verified very simply – there are tons of tweets bubble shown in Manhattan, but much less shown in Queens and Brooklyn.

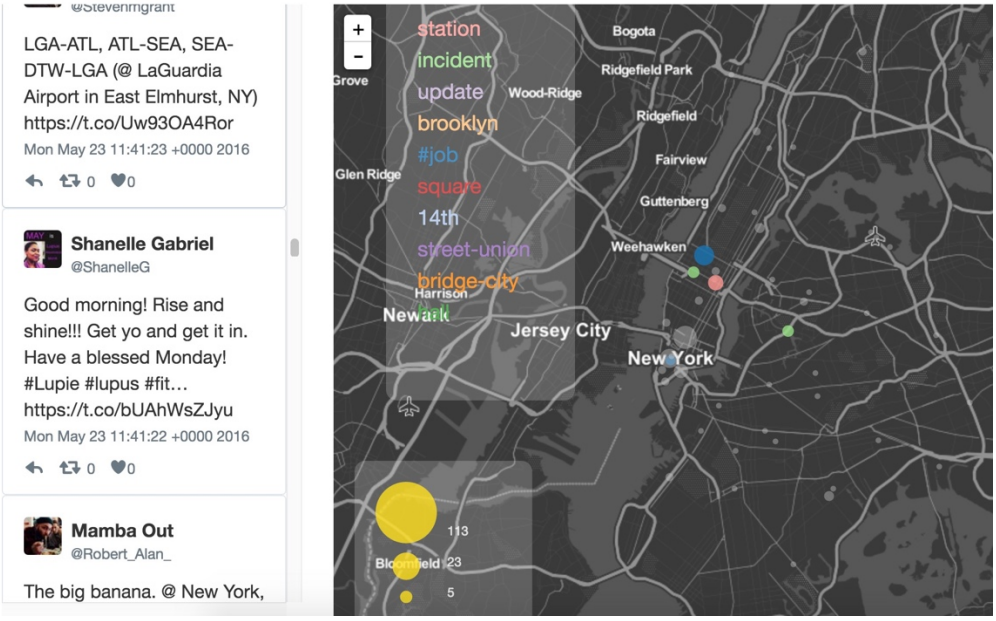
2). For the different hot keys between different area, I select two words to see what’s going on.

I input “brooklyn” and “manhattan” into search bar:

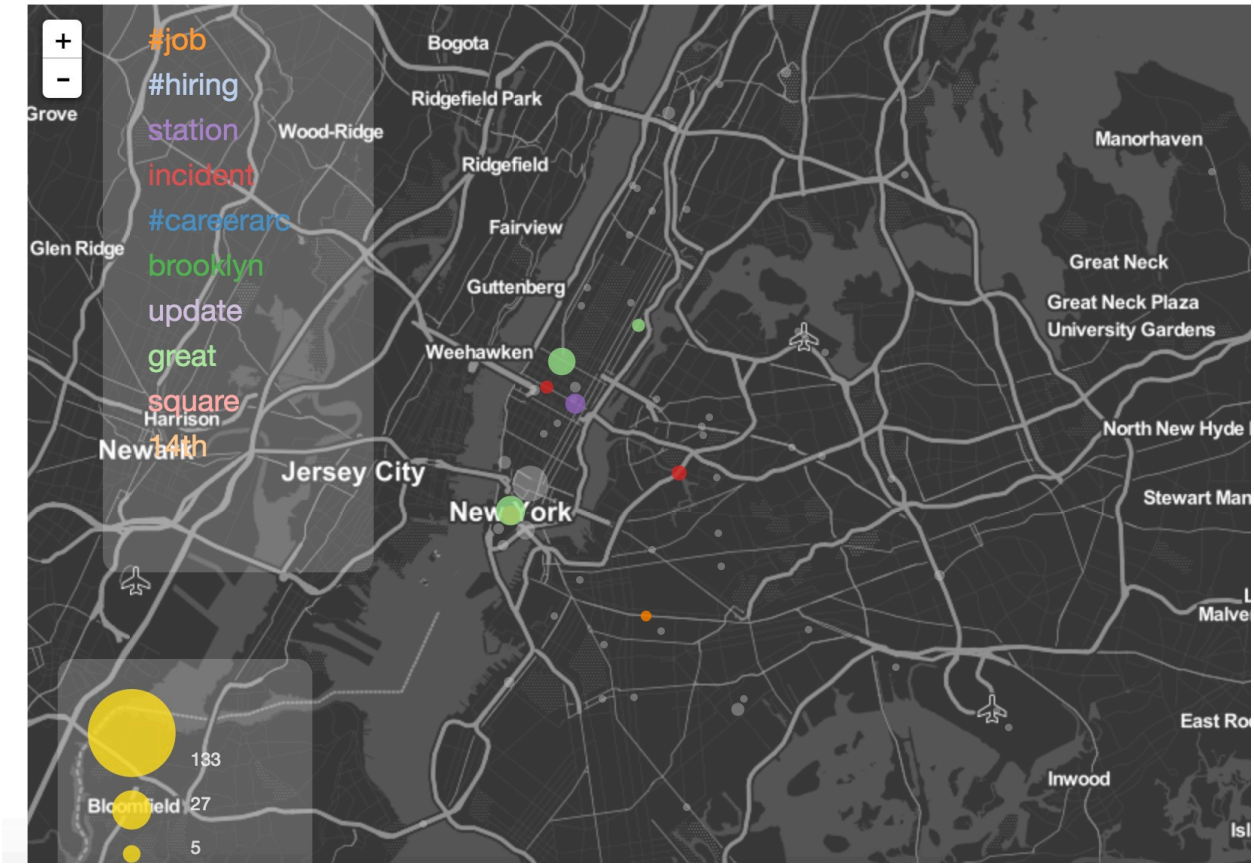


Obviously, people stays in Manhattan talk more about the floor they stand. People in Brooklyn do the same thing, too. From this we can come to a conclusion that what people care is highly related to location they stay.

3). In the processing I am doing this analysis, I found that the hot keys change a lot.



Now very obvious, something happens in 14th street, union square. We can know it even do not need to see the content of tweets. People really like to share what is happening in their daily life in Twitter. But the things changes fast. Just one minute pass, whole world changes.



Now it is 8 am, we know that people are weak up and going to work. They now talking about hiring and job – a eternal topic in every popular city.

Limitations and Future Works

Major Limitation: It's about data source. Since there are few part of tweets has geolocation information, we cannot grasp all information from twitter in NYC. In the end, our visualization cannot provide a persuasive solution.

So if we have more time, we will improve the way we grasp this information, to visualize those more realistic data.

1. links between tweet list, keywords, and map; In the data analysis I just use browser find tool to know the content of tweets, if we have more time, we can add the function that if we click on a word in hot keys list or a cluster on map, it will do some filter to get the corresponding tweets.
2. Better cluster generalization. For instance, use pie chart for big cluster to show the distribution of hot keys within one cluster.