Keywords Over Time

Weikai Li wl990@nyu.edu wl990
Chang Liu cl3695@nyu.edu cl3695
Kaiwen Peng kp1804@nyu.edu kp1804

Project page (on Github): https://github.com/NYU-CS6313-Fall16/Keywords-Over-Time-15
Video: https://www.youtube.com/watch?v=-0vNBap9--k&feature=youtu.be
Working demo: www.nemoleoliu.com

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

Information is updating over time. There are lots of new books, articles, publications, etc. exposed on the Internet every day. In this technique project, we are focused keywords which describe the articles. And we want to explore the design space of solutions and provide a critical review/assessment of alternate designs, to visualize how the keywords extracted from a document collection change over time, how their ranking changes over time and how the words appear/disappear at any given time. We provide a solution for the users who are interested in the changing pattern of keywords, for the journalists who want to know the changing pattern of current hot topics, and for the researchers who want to discover further research topics in this area.

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

1. How does appearance number of keywords change over time?
   A keyword may appear in different articles. By extracting the keywords from the articles database, we will get the current keywords list with the frequencies. Users may be interested in current popular keyword list, and how do the appearance number of keywords changing over time.
2. How does the ranking of keywords change over time?
   By extracting the keywords from the articles database, we will get the current keywords list with the frequencies. The list can be described in a ranking mode. Because sometime users may only be interested in how ranking of keywords change over time.
3. What are current top 5 popular keywords?
   The top K popular is always an interest topic to users. For example, users may want to know the top 5 popular keywords in this week. These keywords may also be considered as popular trends. And they can further discover how these trends change over time.
4. I have a poll of keywords, and I want to know how these keywords change(frequency and ranking) over time?

User may only be interested in a few keywords which may not be the popular one. For example, they may only be interested in keywords “breakfast”, “lunch” and “dinner”. We need to provide a tool for user to check the keywords whichever they want.

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

We used two different datasets. One contains information on IEEE Visualization (IEEE VIS) publications from 1990 to 2015. The other is the twitter dataset which we crawl based on a top 100 popular keywords from the twitter.

In IEEE Visualization Publications Dataset, we focus on the attributes as below:

Attribute Name	Attribute Type	Meaning	Values	Derived? (if yes explain how)
Published Year	Quantitative	Published year of the article.	1990-2015	No
Author Keywords	Categorical	The keywords of the article.	Visualization, Tools ...	No
Conference Name	Categorical	Conference Name	Vis,InfoVis...	No
Article Name	Categorical	Article Name	Information ...	No
Keywords List	Set of Attributes	A set of attributes which describe the properties of keyword including frequency, and timestamp	N/A	Yes

For the keywords list, we calculate the articles which contains the keywords in different year, and create a set of attribute as: Keyword, Year, Frequency.

In Twitter dataset, we create a crawler to get the tweets from twitter. While developing the crawler, we create a top 1500 popular keywords list, and only get the tweets with the listed keyword. By doing this, we will get all the tweets contain the keywords. In the dataset, we only focus on the attribute as below:

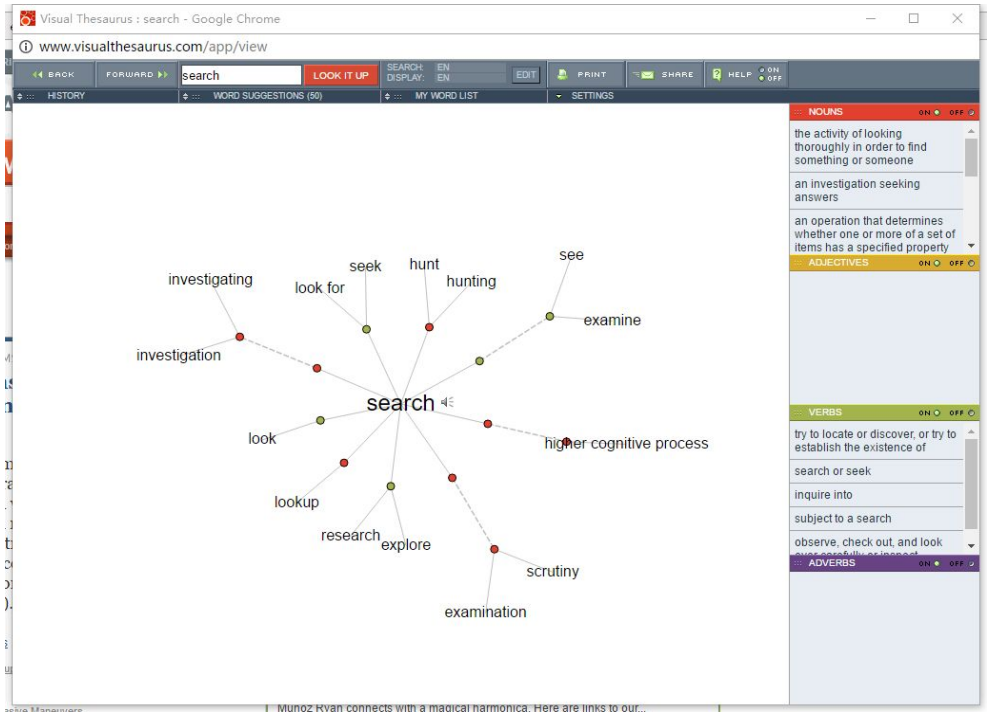
Attribute Name	Attribute Type	Meaning	Values	Derived

				? (if yes explain how)
Post time	Quantitative	Post time of the tweets	Streaming Unixtime	No
Keywords	Categorical	The keywords of tweets	Breakfast, Lunch, ...	Yes
Tweets Content	Categorical	The Content of tweets	Today we have...	No
Keywords List	Set of Attributes	A set of attributes which describe the properties of keywords including frequency, and timestamp	N/A	Yes

In our project, we provide a general technique which shows how keywords change overtime. Thus we have defined a general description of a dataset, which is presented as Keywords List before. If a dataset can be described in a set of attributes which contains the timestamp and keyword frequency, the dataset can be applied in our system. By doing this, our system can be applied to different data with preprocessing.

What have others done to solve this or related problems?

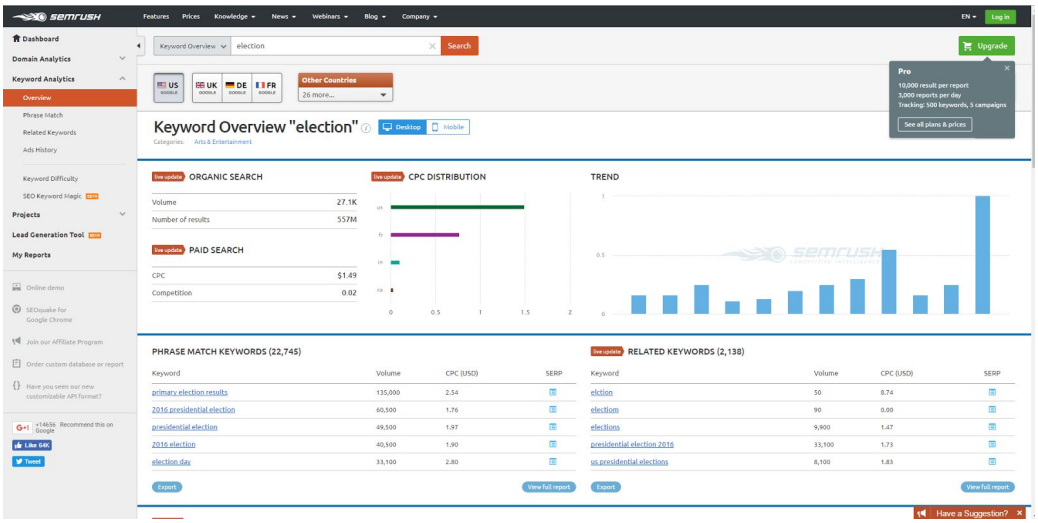
1. Visual Thesaurus



http://www.visualthesaurus.com/

Visual Thesaurus visualizes lexical relationships between various word meanings. It provides an interesting way to visualize relationship between words. In our project, we also provide a solution to analyze the relationship between words, but more focusing on the change overtime.

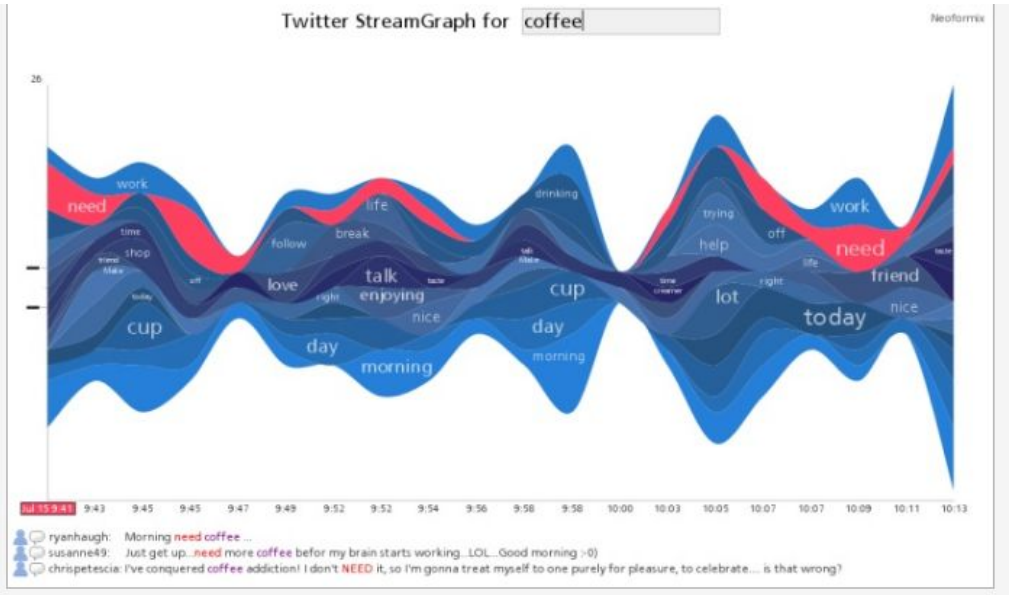
2. SEMRush Trends



https://www.semrush.com/info/election?db=us

SEMRush Trends is a tool that is specific to finding profitable keywords. It is a program many new and older bloggers are using to optimize their sites, create the content their audience want and create a better experience for their visitors. It provides an overview of keyword with many statistics.

3. Twitter StreamGraph

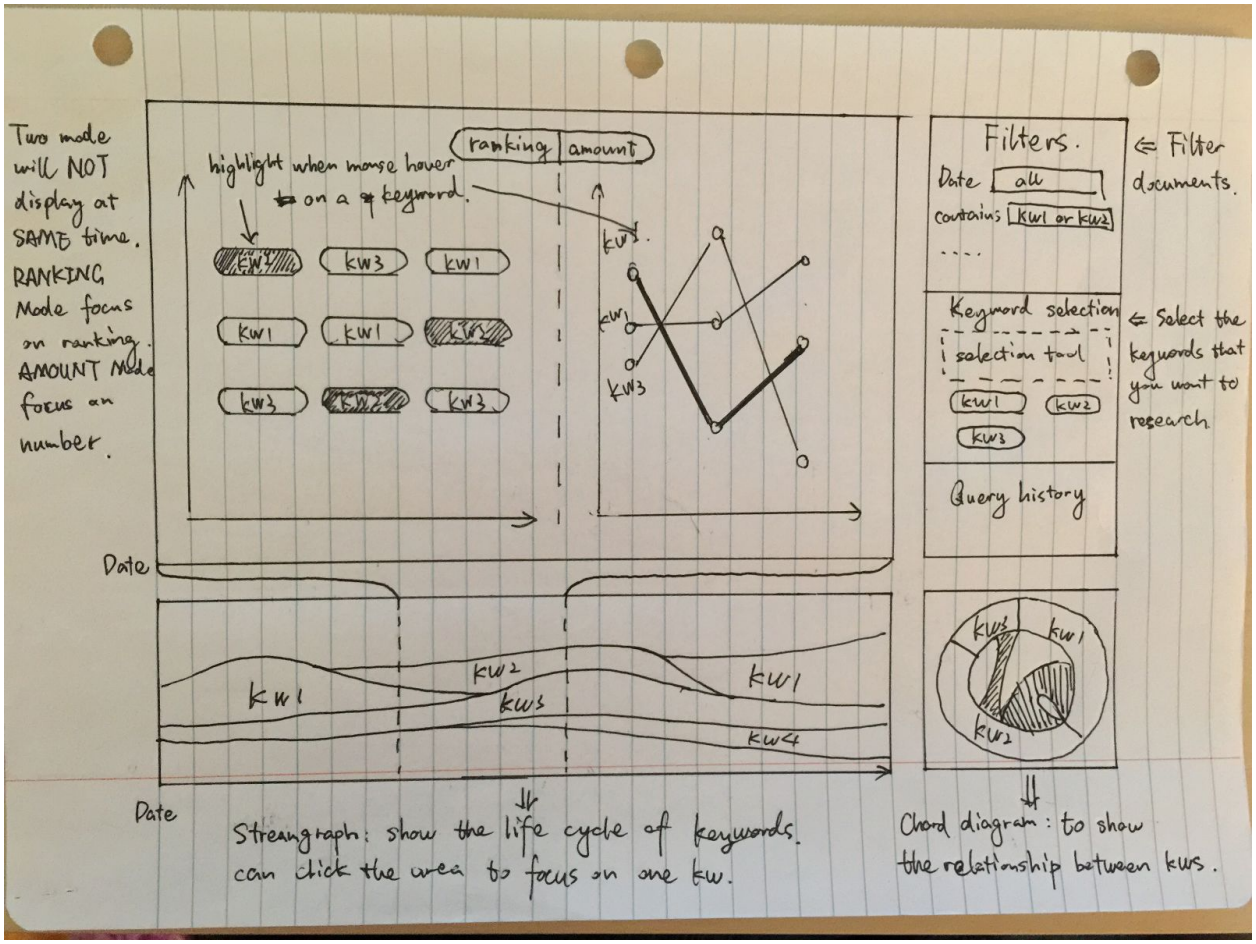


http://www.neoformix.com/2008/TwitterStreamGraphs.html

The Twitter StreamGraph shows the usage over time for the words that most highly associated with the search word. In our mockups, we also tried the stream graph as one of our mockup. We will talk about it later.

Design Iterations

1. **Initial Mockup**



How to read it:
There are 4 different views in the mockup, top left is the main module, bottom left is the navigation module, top right is filter module and bottom right is the relation module.

In the filter view, document filter is used to get a subset of the whole dataset. Because sometime we don't need all the data to answer the question. Keyword filter is used to filter the dataset to get keywords lists which are only interested to the users. In this view, users can type in a keyword and make it as a filter which can help filter out the dataset.

The relation view is described in a chord chart. User can discover the relationship between keywords. If two words are related, there will be a line connecting these two words. Matrix chart is another option in this view.

In the navigation view, the life cycle of keywords will be shown using stream graph. User can easily find when the keywords appear and disappear. Meanwhile user can select the time period in the stream graph. Then the changes of keywords between this time period will be shown in the main chart.

In the main view, user can switch between two modes. Ranking mode is used to show the ranking of keywords. Amount mode is used to show the accurate values.

What worked and did not work:

From the discussion, we decide to keep the following views:

- Two modes which describe the keywords changing pattern: This part is the main part in our project, and it provides a very straightforward visualization of keywords over time.
- Filters view: A filter is necessary if we want to visualize large dataset.

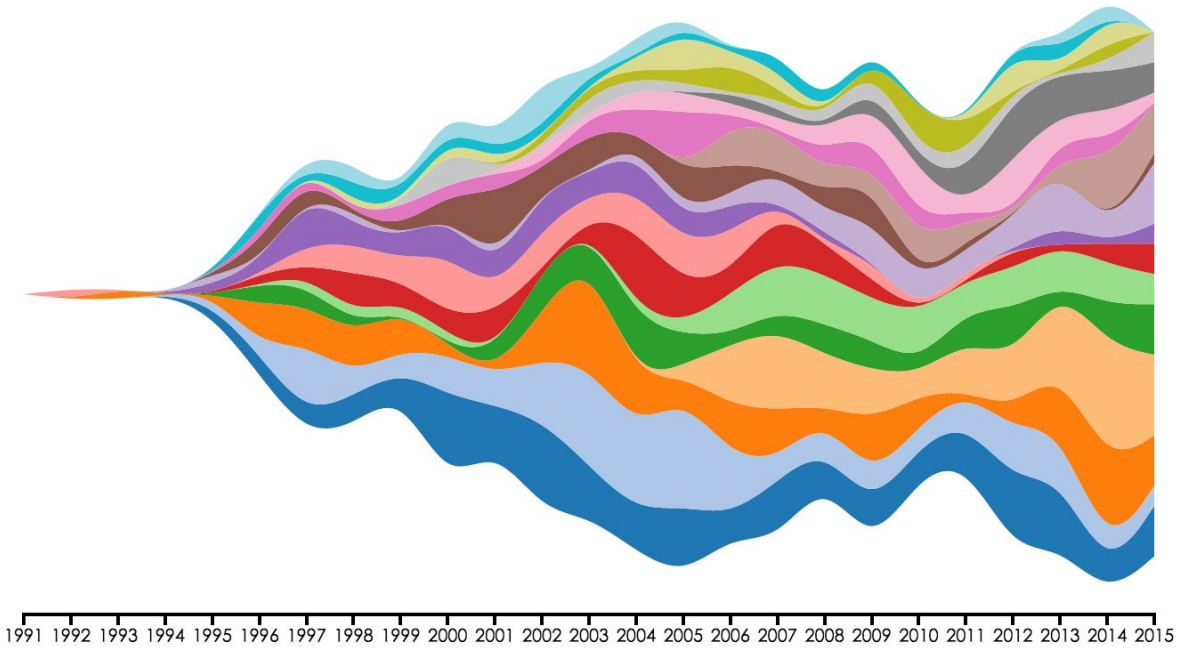
And remove the following views:

- Streamgraph: Streamgraph and line chart are a bit redundant. Most of the information they describe are similar. Thus we decide to discard this one and replace the streamgraph to a calendar.
- Chord chart: The chord chart is not very related to our questions. Because we are mainly focused on the timeline. Thus this chord chart is unnecessary in our visualization.

2. Project Update

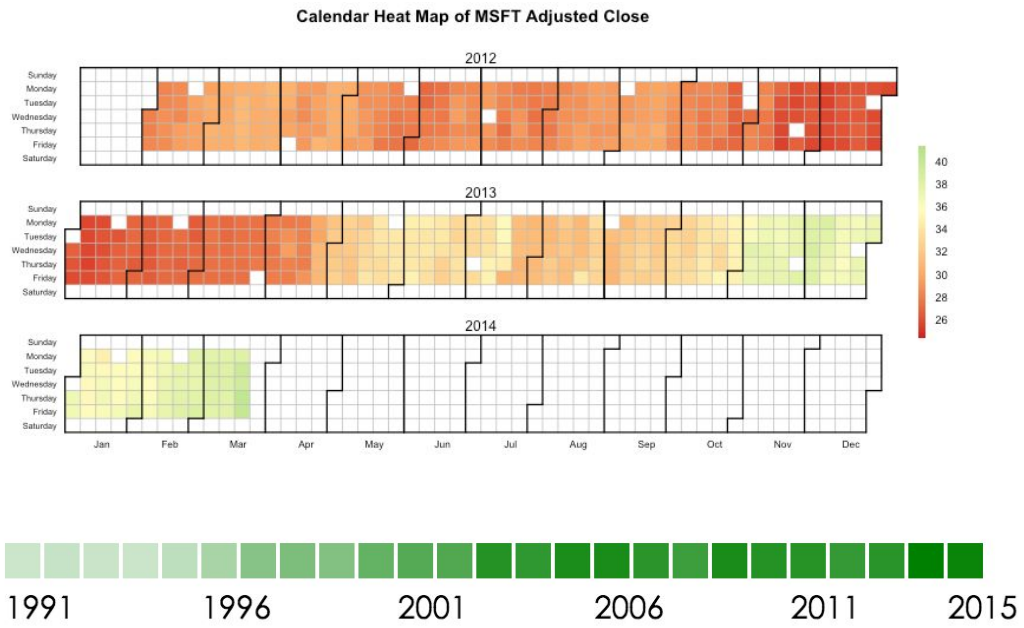
While the developing the technique, we implemented different chart view for experiments, and carefully compared the chart views. We will go through these views in details.

- StreamGraph



How to read it:
The x axis is the timestamp. In this chart, we are using the dataset of IEEE publications, thus it contains the year from 1991 - 2015.
A color represent a unique keyword. The width of the graph in one time unit represent that how many articles contain the keyword.
From the stream graph we can see how the appearance time of different keyword change overtime.
Discuss:
As we discussed before, we the stream graph may be redundant, but it is still a good choice to visualize how the keyword change over time in the main view. However, it does not contain the detail information of keyword including exact appearance number, and it is hard to compare different keywords because the y axis doesn't have specific meaning. In addition, the streamgraph becomes very complicated when the dataset becomes large. Thus finally we decide not to use the stream graph.

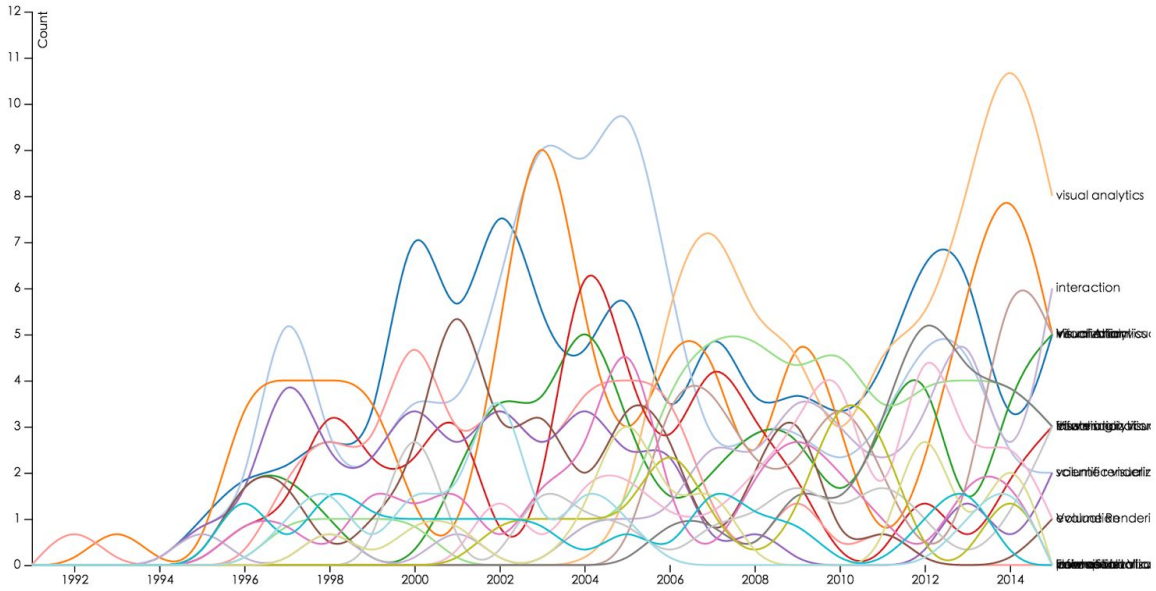
- Calendar



How to read it:
In our design, a view is needed for users to select the time range. After the discussion, we decided to use calendar instead of stream graph. In the stream graph, each grid represents a time unit. The color saturation represent the size of the data in the time unit. For example, in the first calendar, each unit represents a day. The color saturation describes the number between 26-40. The second one is related to the publication dataset. Each grid represents a year, and the color represents the total number of keywords.

Discuss:
Calendar is a good choice for selecting the time range, with providing limited information of the dataset to users. However, we want to provide a little more information in the calendar. Thus finally we decide to keep the calendar, but need to figure out a way to improve it.

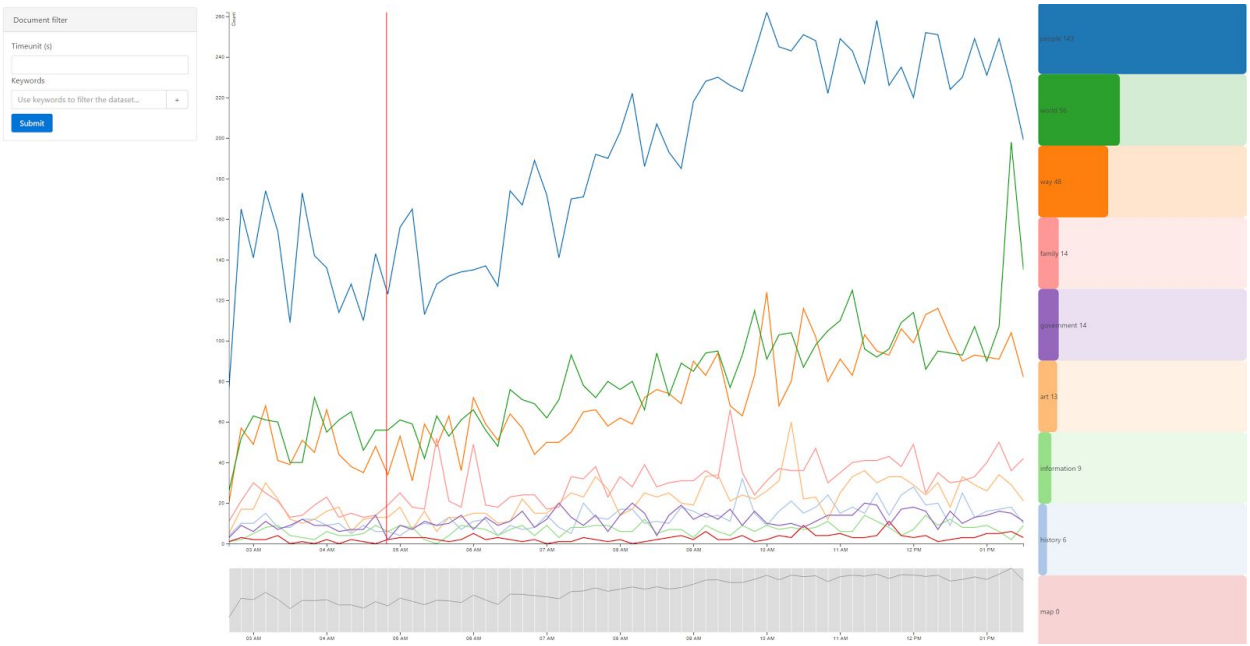
- Line Chart



How to read it:
In the above line chart, x axis represents the time unit, y axis represents the frequency. Each line(color) represents a keyword. From the chart, we can find how these keyword change over time.

Discuss:
Line chart is always a straightforward solution to show the changing pattern of keywords, and it can also answer lots of interesting questions. But as we see, when the number of keywords increases, the figure becomes complicated. And It is difficult to compare different keywords, and difficult to see how their ranking changes over time. Thus we decide to keep this chart, but need to improve it.

Final Visualization



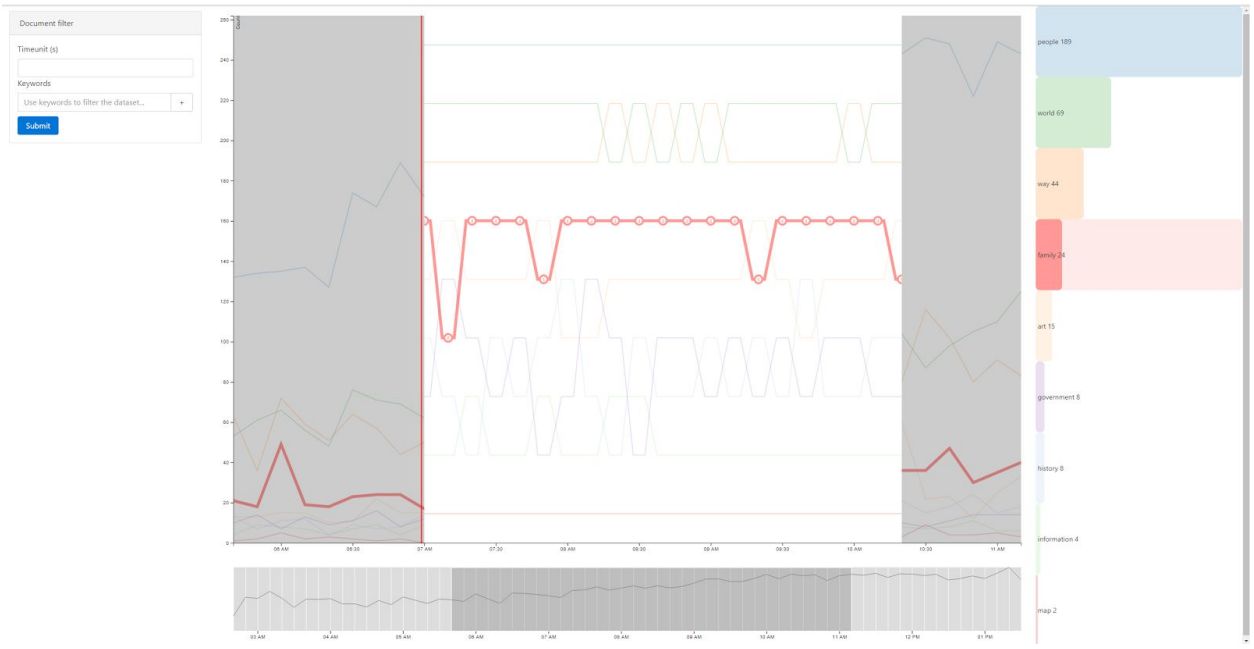
How to read it:

There are 4 different views in the final visualization. It contains document filter, main line chart, ranking bar chart and calendar.

In the document filter, user can set the a timeunit. Different dataset may have different timeunit. For example, the IEEE publication dataset has a timeunit of one year. But the twitter dataset only has a timeunit of one second. For the twitter dataset, users may want to know how the keyword change in every one minute or one hour. Thus they can set the the timeunit in the document filter.

In addition, users can also type in the keywords they want to check. In our current visualization, we only show the top 10 popular keywords in the dataset. However, users may not be interested in these keywords. Thus they are allowed to input the keywords which they are interested.

The main line chart is quite similar to our initial mockup. X axis represents the time unit, y axis represents the frequency. Each line represents a keyword with a unique color. In order to show the ranking of keyword list, we developed an interactive function by selecting a range in this line chart, as shown in the below figure. The selected area is replaced by the ranking mode. In his mode, y axis presents the ranking(1-10). And when selecting a keyword, a number which represents the ranking will show up in the line. Users can easily understand how the ranking of a keyword changes over time. By doing this, users do not need any button to switch between two different modes. The same color represents the same keyword all the time. It is easy for user to check the ranking of any keyword they want.

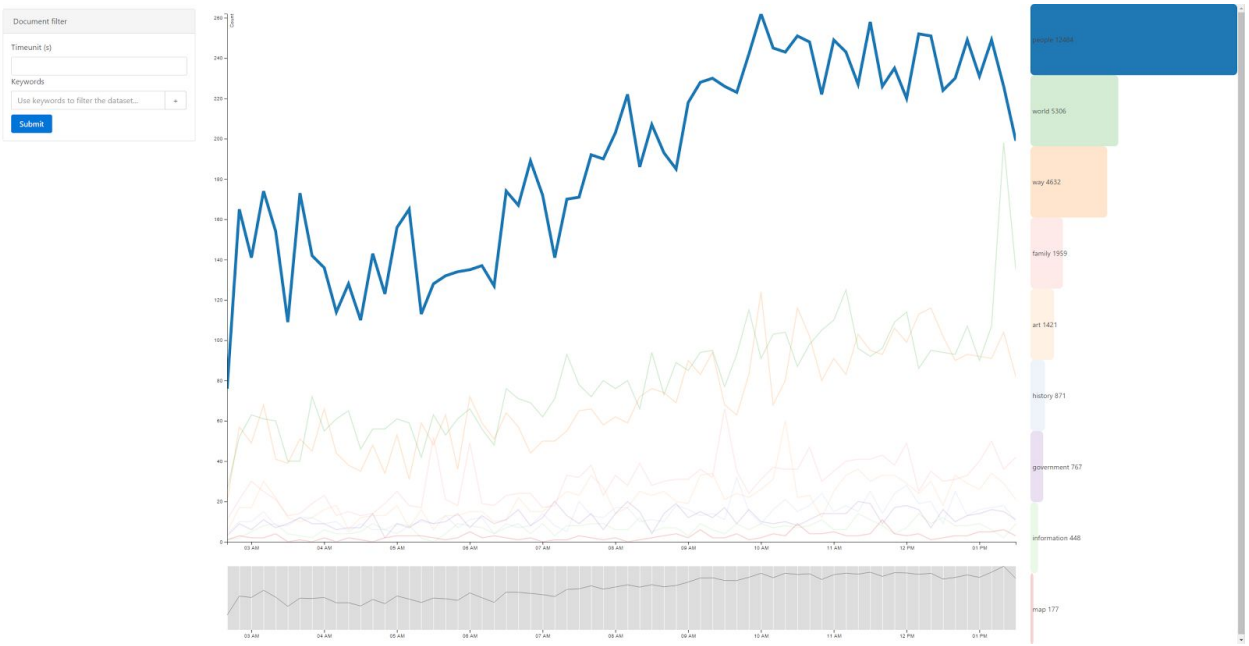


Ranking bar shows the details information of the top K keywords. The keywords are shown in descending order from top to bottom, with its exact appearance number. The length of the bar represents its relative value to the most popular keywords. User can also select the keywords in the ranking bar, and all the related information in the bar chart and line chart will be highlighted.

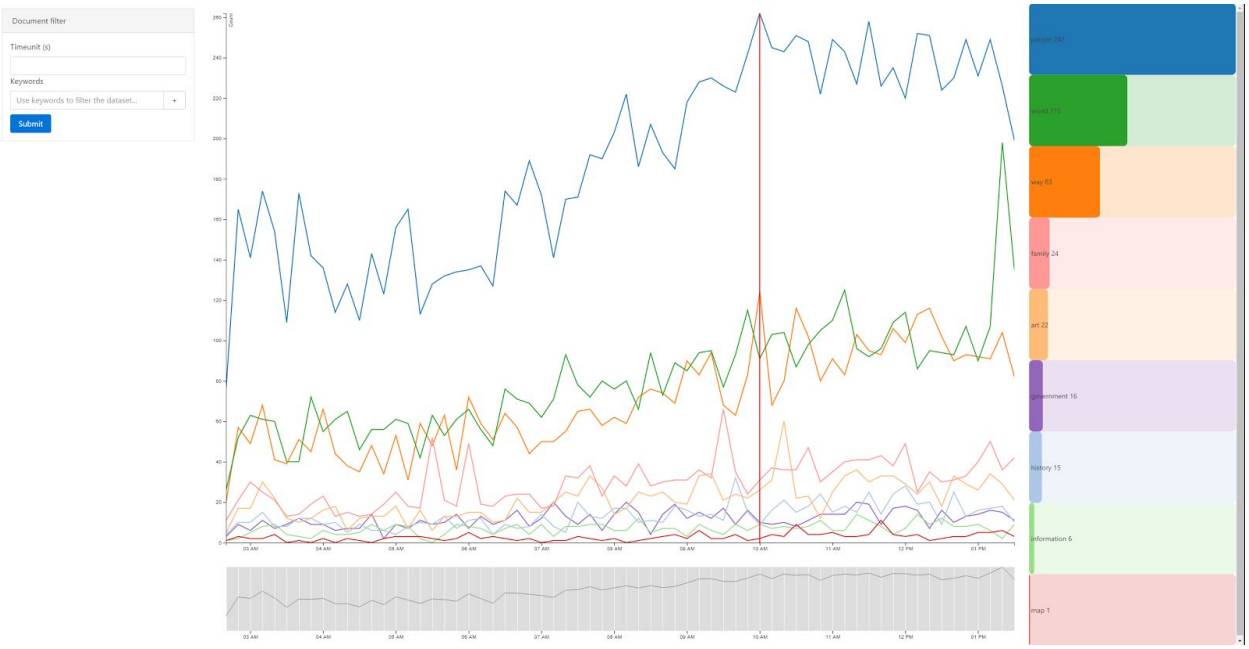
The calendar shows the time range of the dataset. Users can easily select the time range by selecting an area in the calendar. There is also a line showing the total number of keywords. We use the line instead of color saturation mainly because it provides a better way to compare the numbers between different time unit.

Findings

What is current most popular keyword? How does it change over time? What makes the keyword change over time?



The data between 03:00 AM to 01:00 PM is shown as above.
From the chart we can find that the most popular keyword is “People”. In 10:00 AM, it arrives the peak with 242 times of appearance. In most of tweets on Dec 18, users are talking about “People”.

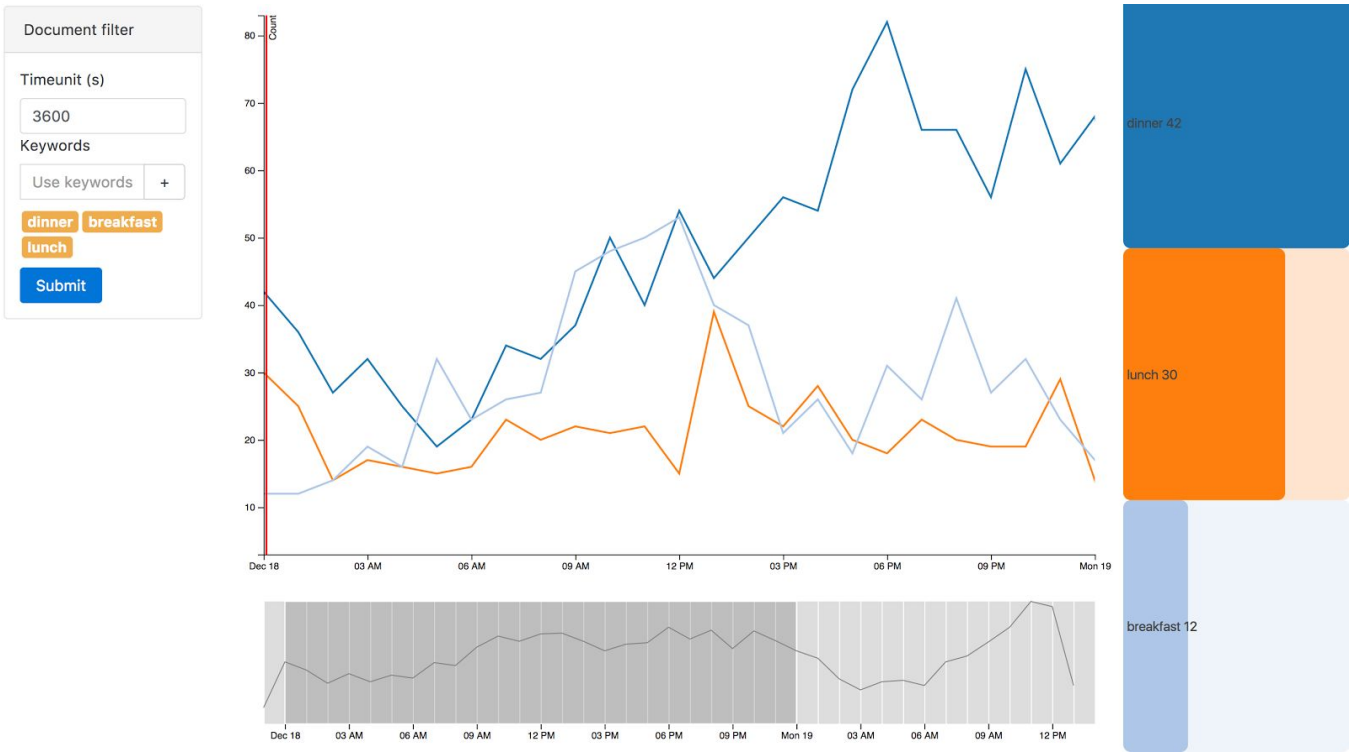


In addition, the appearance of “People” keeps increasing until 10:00 am, mainly due to more users woke up and posted tweets.

From the calendar we can see that the amount of tweets increases. It is the main reason that keyword ”People” increased.

How does the keyword “breakfast”, “lunch” and “dinner” change over time? What can you find from the changing pattern?

The changing pattern of keywords “breakfast”, “lunch” and “dinner” on the day Dec 18 is shown below.
First, it clearly indicates that individuals actually pay attention much more to “dinner” in general.



12am Dec 18 -- 12am Dec 19

In details,
a.) At midnight, “dinner” appeared 42 times and “lunch” appears 30 times. And they both experienced a general decline until morning 8AM. This is mainly due to that people talk about having lunch or dinner together tomorrow.
b.) Then, focus on time range 12am -- 8am and have a look the exact changing in the night.

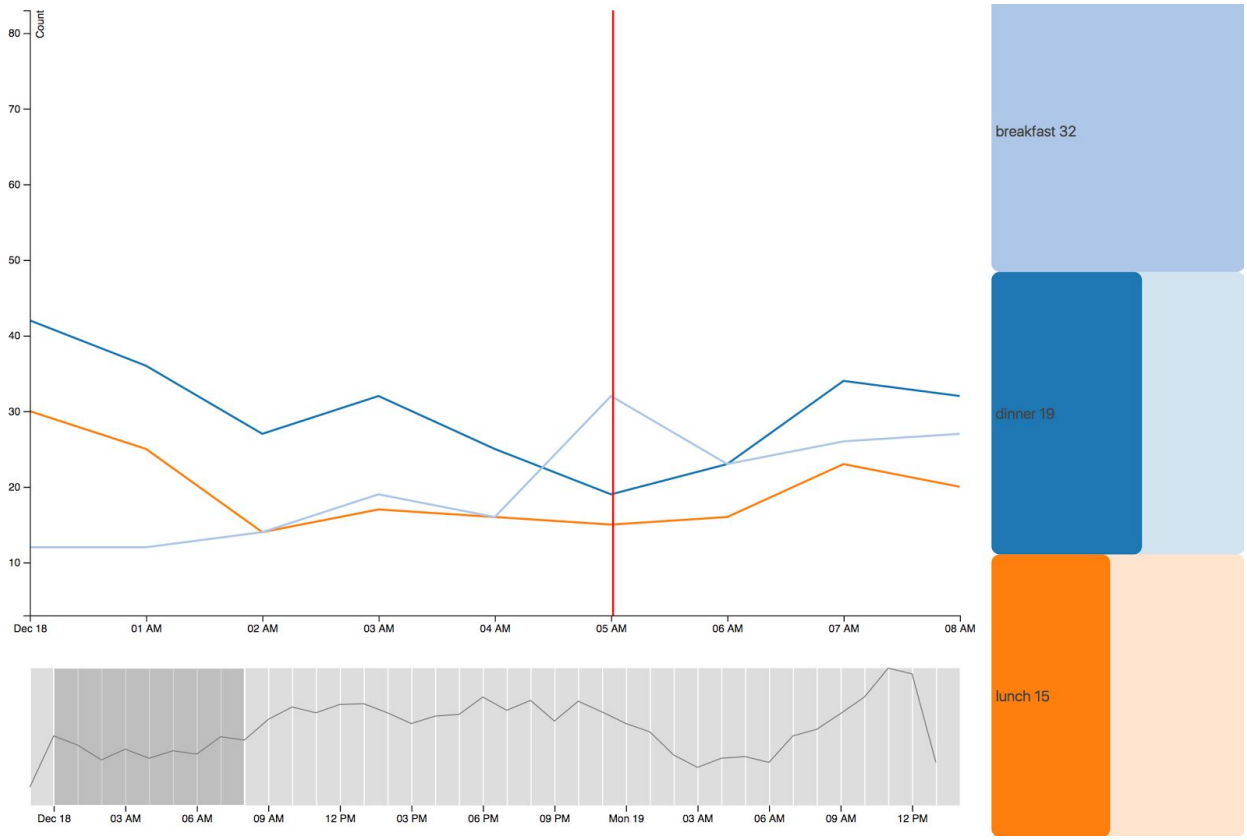


Figure 2. 12am Dec 18 -- 8am Dec 19

There was a booming of “breakfast” at 5am. This is probably caused by “early birds” already woke up and maybe they took a photo and talked about their breakfast, to show they had a nice start of a day.

c.) Then, we have a look of the daytime scenario, 8am -- midnight Dec 18.

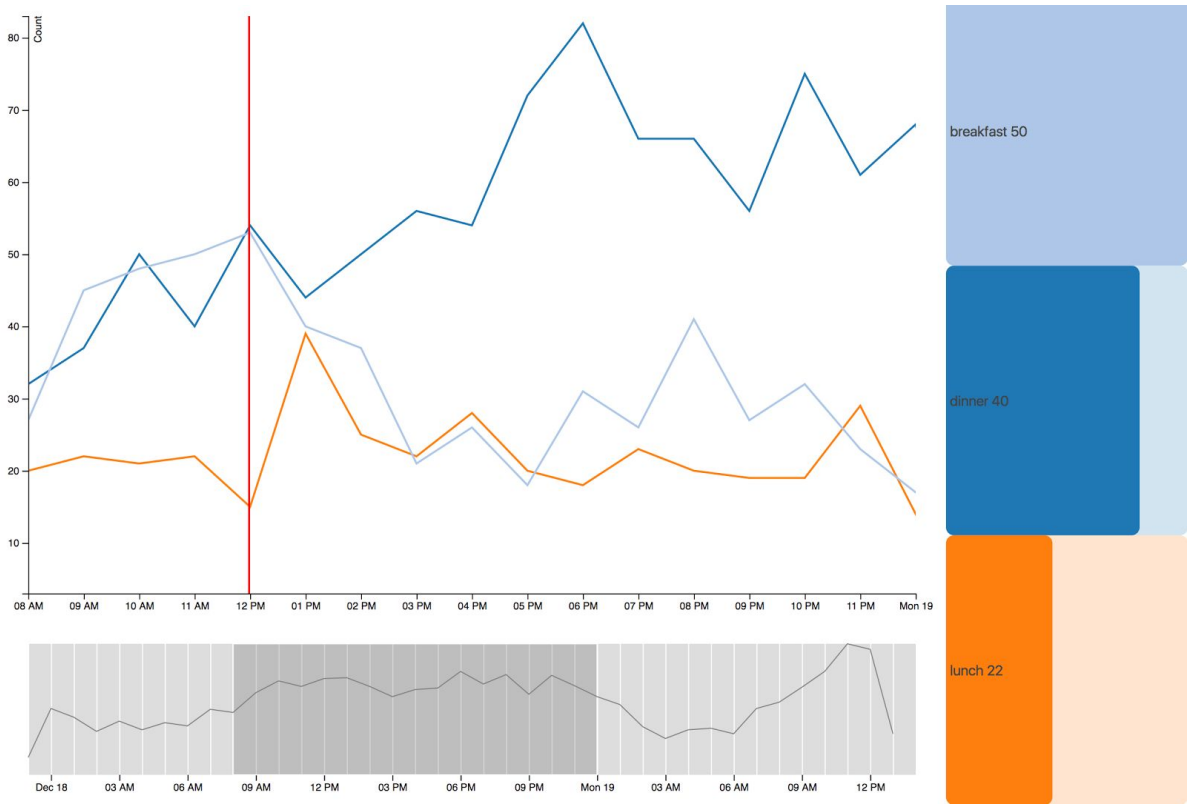


Figure 3. 12am Dec 18 -- 8am Dec 19

i.) Why there is a valley bottom at 12pm noon regarding “lunch”? And why did “breakfast” reach a top value, almost the same as “lunch” which people talking about most frequently.
In personal opinion, firstly, people mainly still tweet about “breakfast” at noon, as they like to take a phone and tweets out after breakfast. This is contrast to most people tweet and talk about “dinner” and “lunch” before these two actually happen. And 12pm (noon) is the exact time for those people who are used to get up late and have a breakfast like bruch and lazily tweet it, probably with the a photo.

ii.) The peaking value of “lunch” appearing at 1pm also indicates that actually most people have lunch at that time. And probably it is most leisure time of office workers, when they can tweets something about their day.

d.) And after the afternoon, people continues to talk about “dinner” more often. It reaches a peak at 6pm.
e.) And there is still a peak value about “breakfast” at 8pm. Why? It is about those people liking to plan for tomorrow, probably they “@” some friends with a beautiful tweet photo, want to have a breakfast together the next day.

Limitations and Future Works

Our project still has limitations. Primary one is the chart becomes complex if we want to visualize large dataset. For example if we want to see top 50 keywords, the chart will become a mass. Our current solution is only showing top 10 keywords. But if user want to check some keywords with lower ranking, they need to type in the keyword manually, which may have a bad user experience.

Another limitation is that the performance of current calendar is still not good when the dataset covers large time range with very small time unit. In the future work, we may design a hierarchical calendar to improve the performance.

Because of the variation of different datasets, the performance of our system also varies. For example, it is really hard to extract meaningful keywords from the twitter dataset. Our current method can only provide a possible solution for our system. A more robust method is still needed to generate keywords with timestamp from the different datasets.