#HOWTRUMP?

The United States of Twitter / The People's Republic of Weibo

A comparative analysis of social media data from China and the United States on the subject of Donald J. Trump´s presidential campaign

Carl Fredrik Beckius 帅小白
He Zhaoyang 何兆阳
Bai Hui 白慧
Manuel Roßner
Beatrice Rizzi
Catalina Villa Melendez

#HOWTRUMP?

The United States of Twitter / The People's Republic of Weibo

A comparative analysis of social media data from China and the United States on the subject of Donald J. Trump´s presidential campaign

Carl Fredrik Beckius 帅小白

He Zhaoyang 何兆阳

Bai Hui 白慧

Manuel Roßner
Beatrice Rizzi
Catalina Villa Melendez

WHY TRUMP?

Donald J. Trump announced he was running for President of the United States in June 2015 with controversial and racists comments towards Mexicans. At that time Trump`s candidacy was a joke, a wonderful source for memes and other internet jokes. After the Mexicans came the Muslims, women, persons with disabilities, and also China. In May 2016 Trump was elected the Republican Party`s candidate to the presidential elections, proving that many people in the United States are moved by his racist and hateful speech.

He argues that China is raping the United States and accuses her of manipulating the Yuan to make its exports more competitive on the global market. In order to “Make America Great Again” Trump is proposing the following strategies among others:

- Declare China A Currency Manipulator
- End China’s Intellectual Property Violations
- Eliminate China’s Illegal Export Subsidies And Other Unfair Advantages
- Strengthen the U.S. military and deploying it appropriately in the East and South China Seas.

Check his position towards China here

Those are strong affirmations that usually would cause diplomatic conflicts between countries, or at least reject from the attacked country population. However, people in China seems to admire Trump. According to the BBC article “Donald Trump and China: A complex relationship”, by Vincent Ni and published on the 26th March 2016, Trump is a well-known name in China. He is admired for being a wealthy businessman. Many companies even use the name Trump in order to imply that they are trustful and successful as he is. The article affirms that “*There is no reliable opinion poll on how Mr Trump's comments are received in China, but state media are watching him with a degree of schadenfreude*.”

The purpose of this data gathering, analysis and visualization exercise is to gain a better understanding of Chinese people`s attitude towards Trump, and how it has evolved with time. Social media is therefore the best way to collect data from, since it gives a place to reflect the people`s opinion and a certain degree of anonymity that allows freedom of speech. The following hypotheses will serve as a guiding line through the process:

- Hypothesis 1: ***People in China are interested in Donald Trump´s presidential campaign and react to it on social media.***
- Hypothesis 2**: *The conversation around Donald Trump`s campaign revolves around some controversial ideas (like the Mexican border wall, expelling the Muslims, etc.) while “boring” but non less important points of his campaign are ignored.***

- Hypothesis 3: ***Trump`s statements regarding China influence the perception that Chinese people have of him.***

DATA MINING

Weibo and Baidu

Why chose Weibo as a representative social media in china?

Weibo is one of the most commonly used social software in china, maintaining a tremendous amount (around 200 million) of active users. Compares with similar products like RENREN or QQ, it involves the most popular topics (update every 10 min) happened in this country and with its users most between the ages of 17 of 33, Weibo has dynamic and powerful social influence. So, collecting data from Weibo can make sure at least 2 points:

1. **Every latest news about Donald Trump can be seen**
2. **A lot of different voices can be heard.**

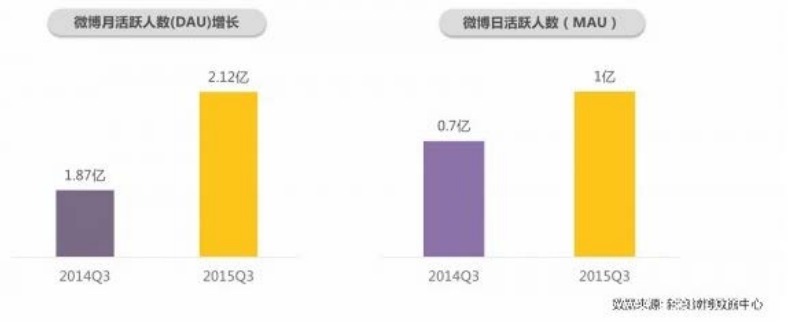
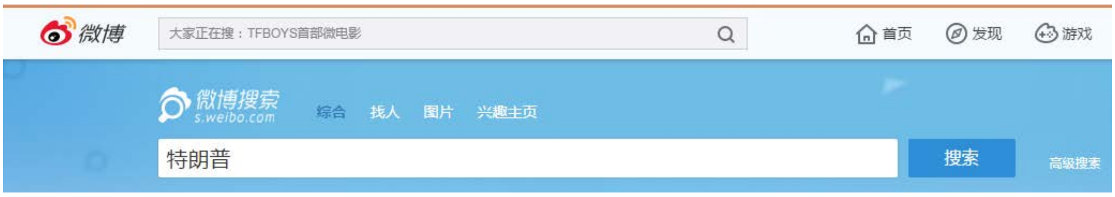


Figure 1：Weibo users
http://www.useit.com.cn/forum.php?mod=viewthread&tid=10921

Which keyword to search？(#特朗普)

There are generally 2 options when you search something about Trump on Weibo, keywords or topic titles. Since getting all the topics related to Trump is relatively difficult for people who is not so familiar with him, keyword became our best choice. Different from twitter, Donald Trump has many versions of translation, official account like Sina News use“特朗普”, while some people call him “川普” as a nickname. In order to get data from both individual and official accounts, “特朗普”is our final decision for keyword.



特朗普

Figure 2: Selected keyword

From when to when?

On June16, 2015, Trump announced his candidacy for president. This day can be considered as a milestone, before that not so many people in China knew this billionaire. So the beginning of our searching period was selected at the particular date. Since the process can be applied to other social topic researches, it will be better if we can get warm data (it might require more coding skills). Due to technical and time constraints, the end of searching period is 2016.5.4 (the end of data collection part of the class)

How to get the data?

Aimed to see the pattern of social media's respond to Trump, a well-structured database need to be established, including user's name, content of microblog, update time, numbers of thumb-ups, followed and comments (all of them can be seen in a single Weibo). A web spider named Gooseeker maybe the best choice if you want to get data from Weibo without coding. The rule of Gooseeker is to get what you can see. It just like you copy the information from Weibo advanced search, so the limit is only hottest and newest tweets can be seen. The simple and efficient way to fix it is to separate searching period into months, then all information will be available.

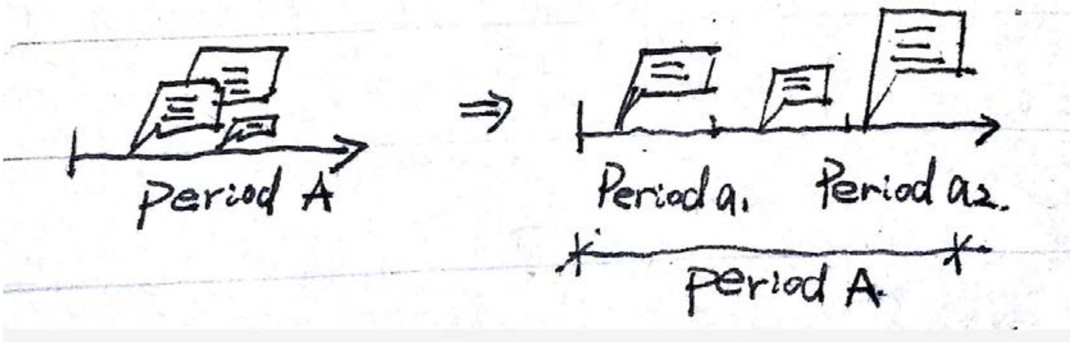
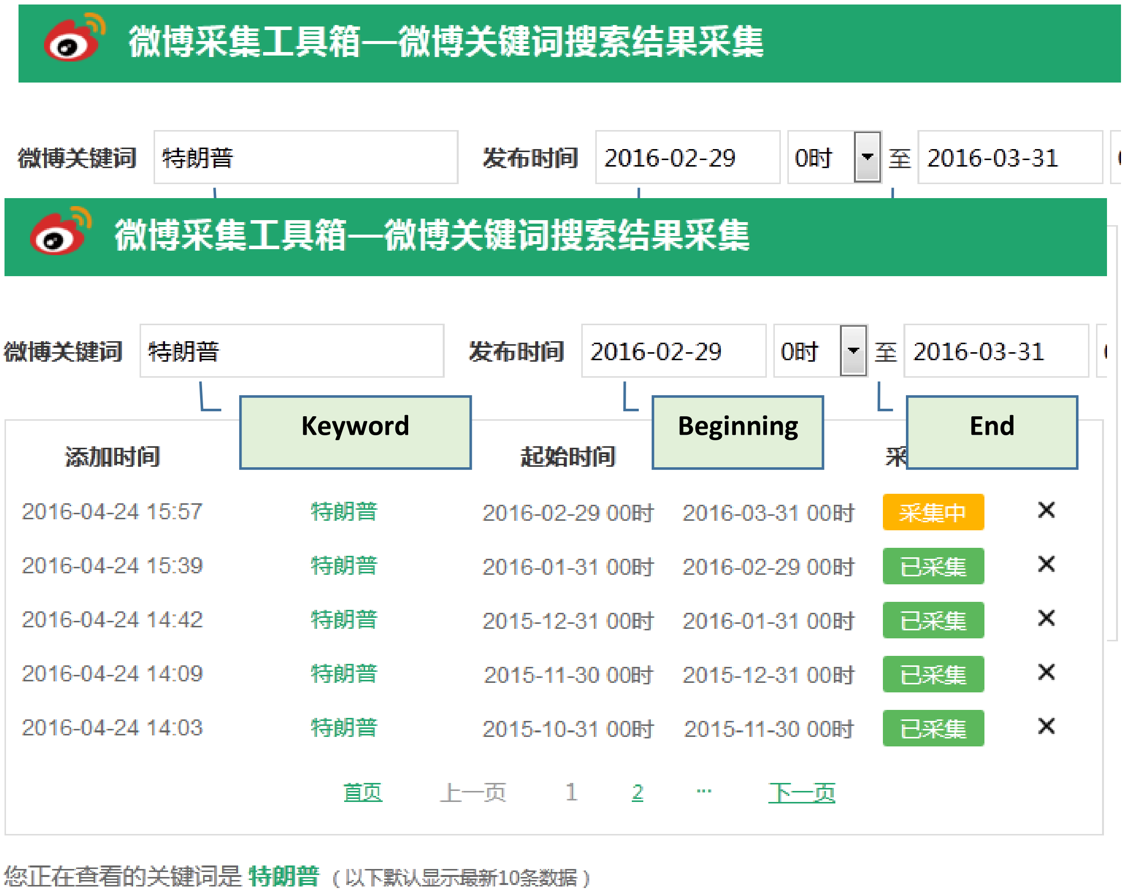


Figure 3:the way to fix problems（draw by the writer）



微博采集工具箱—微博关键词搜索结果采集

微博关键词 特朗普 发布时间 2016-02-29 0时 至 2016-03-31

微博采集工具箱—微博关键词搜索结果采集

微博关键词 特朗普 发布时间 2016-02-29 0时 至 2016-03-31

添加时间		起始时间		采	
2016-04-24 15:57	特朗普	2016-02-29 00时	2016-03-31 00时	采集中	✕
2016-04-24 15:39	特朗普	2016-01-31 00时	2016-02-29 00时	已采集	✕
2016-04-24 14:42	特朗普	2015-12-31 00时	2016-01-31 00时	已采集	✕
2016-04-24 14:09	特朗普	2015-11-30 00时	2015-12-31 00时	已采集	✕
2016-04-24 14:03	特朗普	2015-10-31 00时	2015-11-30 00时	已采集	✕

首页 上一页 1 2 ··· 下一页

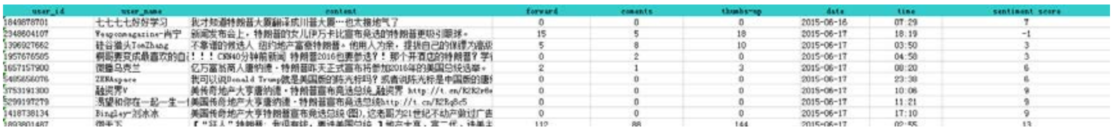
您正在查看的关键词是 特朗普（以下默认显示最新10条数据）

Figure 4:gooseeker
http://www.gooseeker.com/Weibo/secure/key

http://www.gooseeker.com/Weibo/secure/key

Results

Gooseeker provided a highly structured data format, all you need is to change the string data into other forms and delete some unnecessary information. After that, you can have a neat database.



user_id	user_name	content	forward	coments	thumbs-up	data	time	sentiment score
1849878701	七七七七好好学习	我才知道特朗普大厦翻译成川普大厦…也太接地气了	0	0	0	2015-06-16	07:29	7
2348604107	Vesgomagazine-尚宁	新闻发布会上，特朗普的女儿伊万卡比宣布竞选的特朗普更吸引眼球。	15	5	18	2015-06-17	18:19	-1
1396927662	硅谷猎头TomZhang	不靠谱的候选人 纽约地产富豪特朗普。他用人为亲，提拔自己的保镖为高级	5	8	10	2015-06-17	03:50	3
1957676565	桐聪要变成最喜欢的自	！！！CNN40分钟前新闻 特朗普2016也要参选？！那个开酒店的特朗普？学	0	2	0	2015-06-17	04:50	3
1657157900	微瞳乌克兰	亿万富翁商人唐纳德·特朗普昨天正式宣布将参加2016年的美国总统选举。	2	1	3	2015-06-17	00:20	6
5405656076	ZERAspace	我可以说Donald Trump就是美国版的陈光标吗？或者说陈光标是中国版的唐	0	0	0	2015-06-17	23:38	6
3753191300	融资界V	美传奇地产大亨唐纳德·特朗普宣布竞选总统_融资界 http://t.cn/R2R2r6e	0	0	0	2015-06-17	10:08	9
5299197279	渴望和你在一起一生一	美国传奇地产大亨唐纳德·特朗普宣布竞选总统http://t.cn/R2Rq8c5	0	0	0	2015-06-17	11:21	9
1418738134	Bingley-刘冰冰	美国传奇地产大亨特朗普宣布竞选总统(图),这老哥为21世纪不动产做过广告	0	0	0	2015-06-17	17:10	9
1893801487	微天下	【"狂人"特朗普：我很有钱，要选美国总统】地产大亨，富二代，选美主	112	88	144	2015-06-17	02:55	13

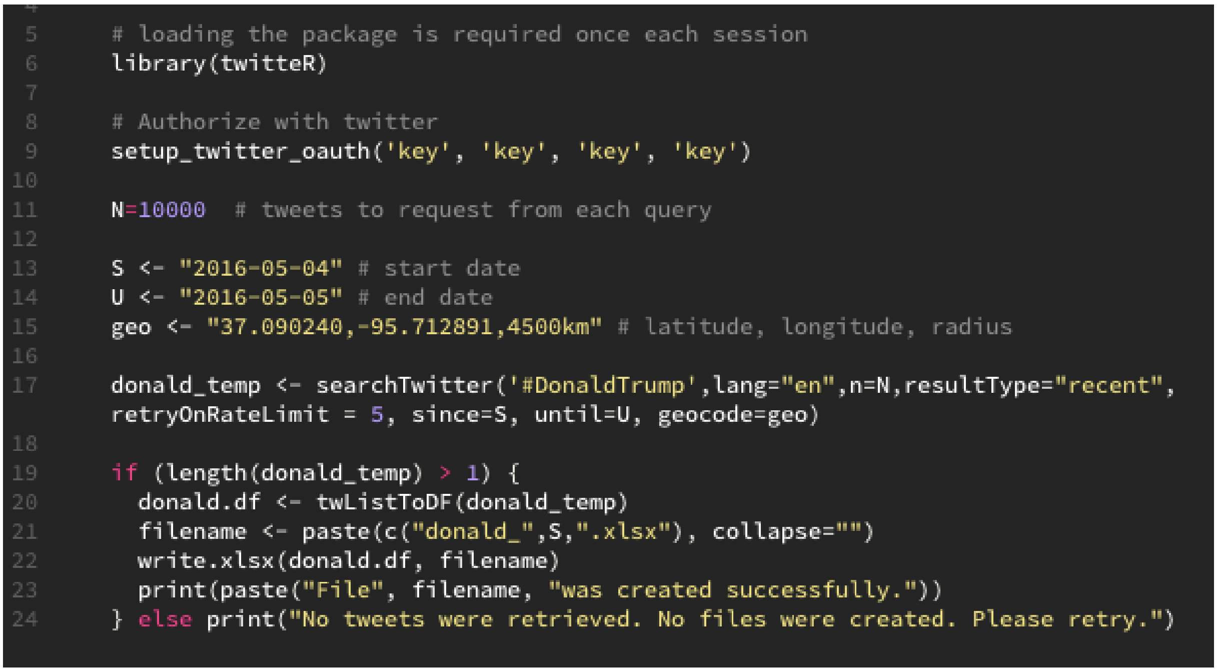
Twitter

For covering US opinions and reactions about Trump, the micro-blogging platform Twitter was chosen. Since its launch in 2006, Twitter has changed from mainly covering personal events to being a worldwide platform for sharing and discussing news, including politics. Twitter's user base in the US is roughly 60 million[1] monthly users. Several nations-wide news companies utilize the platform for spreading their stories, resulting in Twitter collecting both personal and professional opinions.

Twitter was searched using the search query "#DonaldTrump". Through qualitative text analysis this was concluded to be the classification most commonly utilized for tweets (a post on twitter) discussing topics related to Trump. The choice of omitting similar search queries, such as "#trump" and "Donald Trump", was made to get better control over the data. Manual sampling of our data suggested that "#DonaldTrump" covered both democrats and republicans and a wide range of topic.

The public API of Twitter sets certain limitations on data queries. Only data from the last 6-8 days is available. By periodically running searches, data from 10 days was collected. Due to this limitation, comparisons to Sina Weibo were focused to this period of time. Other means of accessing Twitter data are not publicly available and thus not suitable for this study.

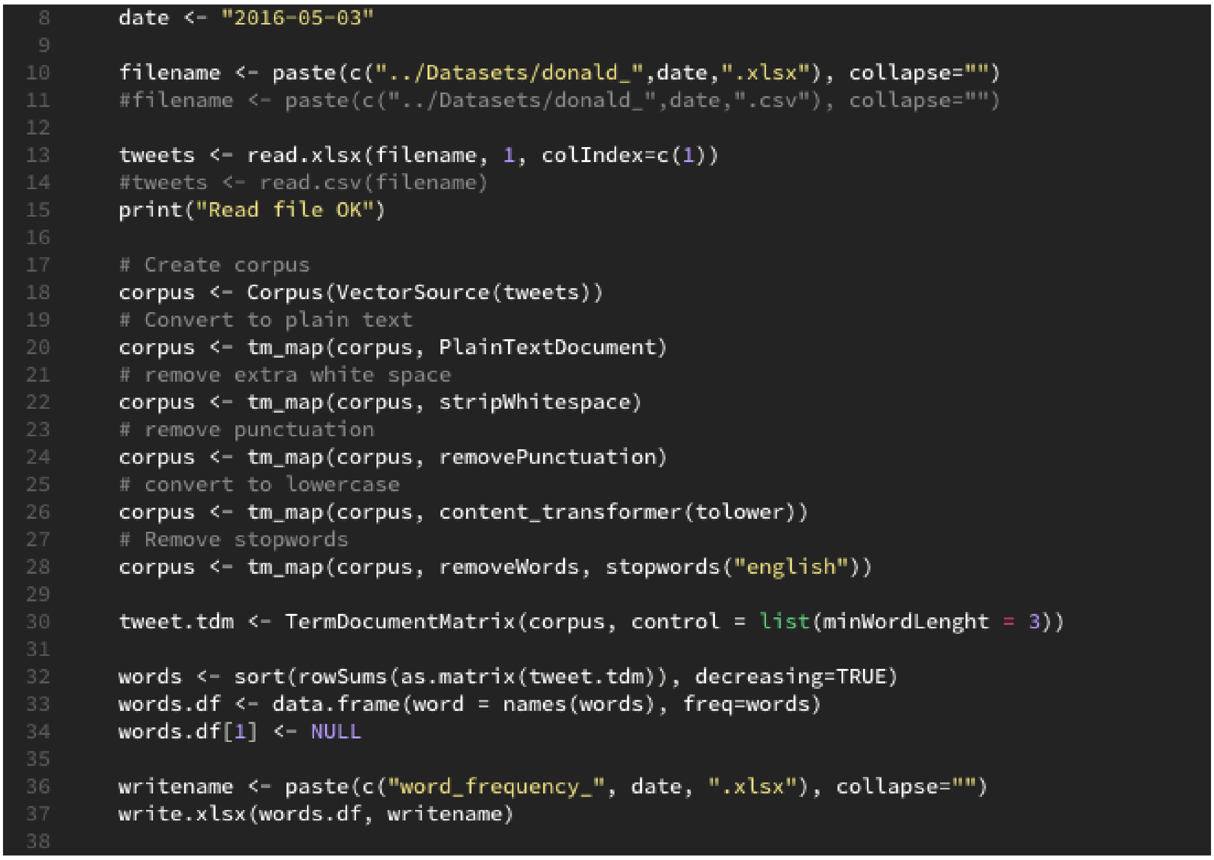
[1] https://about.twitter.com/company



```
# loading the package is required once each session
library(twitteR)

# Authorize with twitter
setup_twitter_oauth('key', 'key', 'key', 'key')

N=10000  # tweets to request from each query

S <- "2016-05-04" # start date
U <- "2016-05-05" # end date
geo <- "37.090240,-95.712891,4500km" # latitude, longitude, radius

donald_temp <- searchTwitter('#DonaldTrump',lang="en",n=N,resultType="recent",
retryOnRateLimit = 5, since=S, until=U, geocode=geo)

if (length(donald_temp) > 1) {
  donald.df <- twListToDF(donald_temp)
  filename <- paste(c("donald_",S,".xlsx"), collapse="")
  write.xlsx(donald.df, filename)
  print(paste("File", filename, "was created successfully."))
} else print("No tweets were retrieved. No files were created. Please retry.")
```

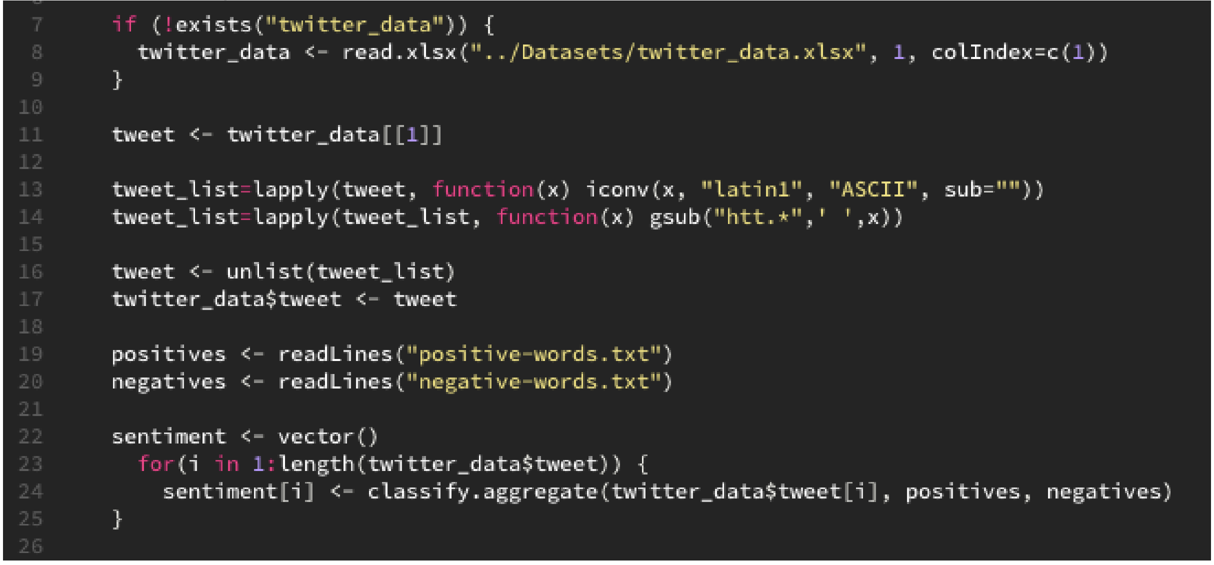
Figure 6 Collecting data from Twitter public API.

Collection, filtering and sorting and classification of Twitter data was done using R, with “TwitteR” (public API access), “SentR” (sentiment classification) and “TM” (text mining operations) packages. “Twitter” returns a list of tweet information, including: text, user id, whether or not it is a reply, favorite count, retweet count, whether or not it is a retweet, geolocation and creation date. To do text analysis the “TM” packaged was used to remove stop words, links and emoji in order to normalize the text. Sentiment classification was done using the “SentR”. Finally, all data was put together using Excel.



```r
date <- "2016-05-03"

filename <- paste(c("../Datasets/donald_",date,".xlsx"), collapse="")
#filename <- paste(c("../Datasets/donald_",date,".csv"), collapse="")

tweets <- read.xlsx(filename, 1, colIndex=c(1))
#tweets <- read.csv(filename)
print("Read file OK")

# Create corpus
corpus <- Corpus(VectorSource(tweets))
# Convert to plain text
corpus <- tm_map(corpus, PlainTextDocument)
# remove extra white space
corpus <- tm_map(corpus, stripWhitespace)
# remove punctuation
corpus <- tm_map(corpus, removePunctuation)
# convert to lowercase
corpus <- tm_map(corpus, content_transformer(tolower))
# Remove stopwords
corpus <- tm_map(corpus, removeWords, stopwords("english"))

tweet.tdm <- TermDocumentMatrix(corpus, control = list(minWordLenght = 3))

words <- sort(rowSums(as.matrix(tweet.tdm)), decreasing=TRUE)
words.df <- data.frame(word = names(words), freq=words)
words.df[1] <- NULL

writename <- paste(c("word_frequency_", date, ".xlsx"), collapse="")
write.xlsx(words.df, writename)

```

Figure 7 Text manipulation to normalize words.



```r
if (!exists("twitter_data")) {
  twitter_data <- read.xlsx("../Datasets/twitter_data.xlsx", 1, colIndex=c(1))
}

tweet <- twitter_data[[1]]

tweet_list=lapply(tweet, function(x) iconv(x, "latin1", "ASCII", sub=""))
tweet_list=lapply(tweet_list, function(x) gsub("htt.*",' ',x))

tweet <- unlist(tweet_list)
twitter_data$tweet <- tweet

positives <- readLines("positive-words.txt")
negatives <- readLines("negative-words.txt")

sentiment <- vector()
  for(i in 1:length(twitter_data$tweet)) {
    sentiment[i] <- classify.aggregate(twitter_data$tweet[i], positives, negatives)
}

```

Figure 8 Sentiment matching with sentiment lexicon.

The resulting dataset consists of statuses for 10 days, 2016-04-25 to 2016-05-04 with an average of 1800 per day (including retweets). As retweets cannot be counted as individual opinions, these were removed from the dataset. The resulting dataset consists of 2846 statuses.

DATA ANALYSIS

Sentiment Analysis from Weibo

It´s always so hard to describe a feeling. There are several methods on internet telling how to do it with coding. Basically, it working with a dictionary which store positive words and negative word separately. Then, go through the text to find words that matching in the dictionary. The last step is scoring the text based on counting numbers of words of each type. In terms of Chinese, things get more complicate, unlike English Chinese words stick to each other in one sentence, which means the coding can't work with Weibo content. Luckily, professor Shen from Qing Hua university created a software named as ROST, it can separate words can give sentiment scores. (Since the code of ROST is not open to the public, the principle of scoring still remain as a black box to us)

Something about Weibo

Q1：What makes people talk?

In order to find the event that triggered the heat talk on Weibo, number of posts and searching index from Baidu (biggest searching engine in China) are compared, attached with big news about Trump.

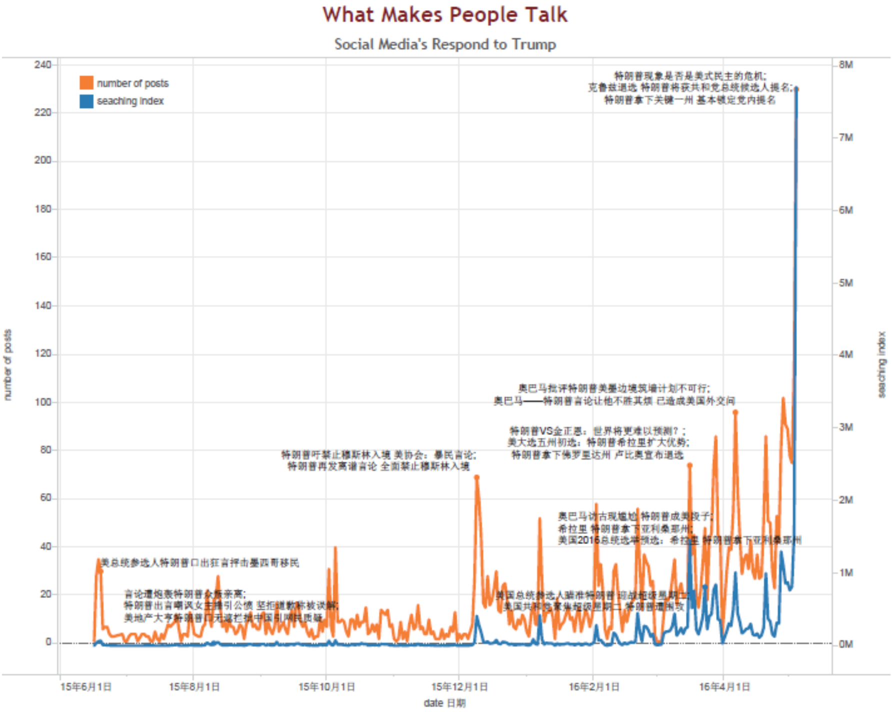


Figure 6: what makes people talk (using Tableau)

We can easily find the common tendency that people show an increasing focus on trump and both curves have sharp fluctuations. We just compare the peak points of every month between the two curves. According to the day of events, we check the specific contents of Weibo posts and the news of Baidu to see what happened on that day. We found that if they both have peak points. The events are always the same and mainly involve speech about China or some interview of trump or the results of primary elections. But Weibo curve shows some funny video of trump can also cause peak points, just like what happened on 10th July and 10th September. This phenomenon can show some kind of difference between public media and social media.

Q2：How do people feel?

The picture on the top is using the median of sentiment score each day (to avoid the extreme cases) , red is for negative feelings while blue for positive, thickness of the line is according to the amount of total posts that day. Second picture is based on number of 2 type of posts.

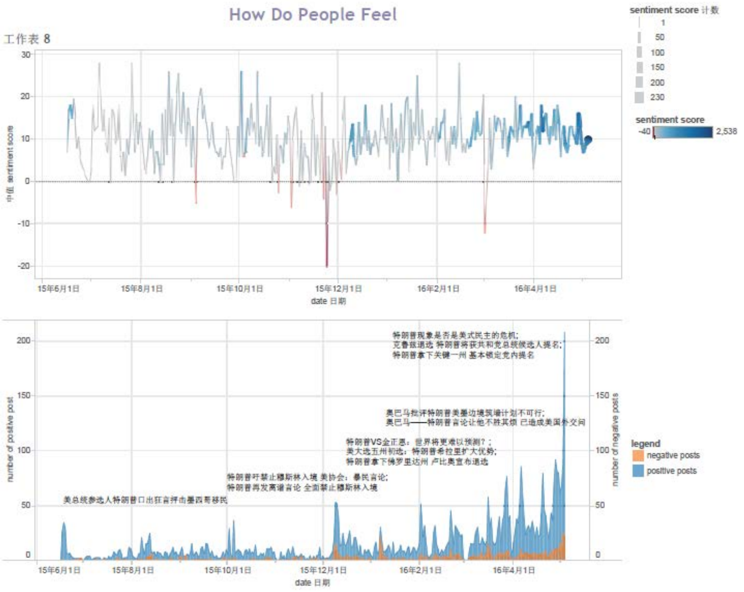


Figure 9:how do people feel

According to the picture, it seems obvious that generally people tend to like Trump as he won states one after another, but when people feel bad, they do say something extreme on Weibo.

Q3：Who is concerned?

In this part, Gantt chart has been used to see the distribution of each ID’s posts according to time. Top 10 official accounts and individual accounts were selected out. Maybe more information can be dug out like what the official account mainly about.

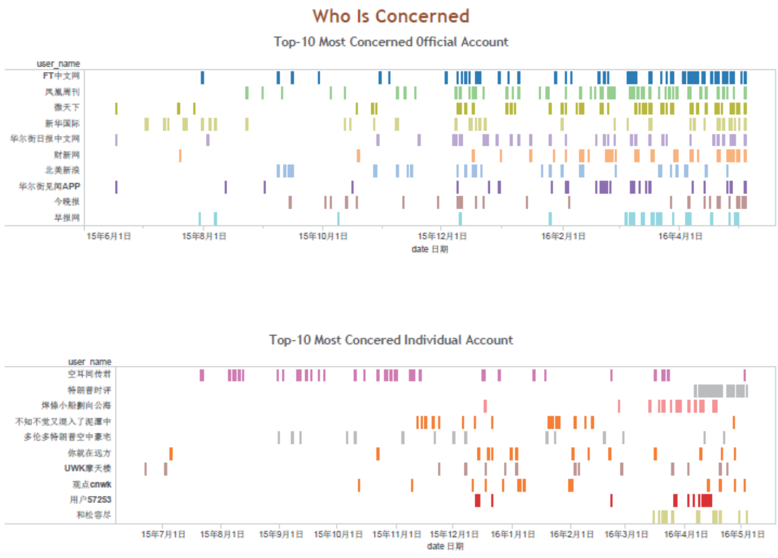


Figure 10:who is concerned

Q3：Who speaks for public

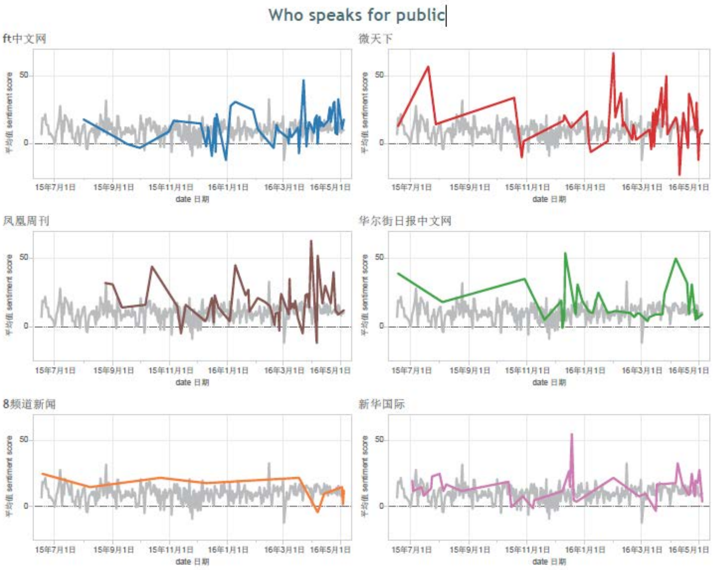
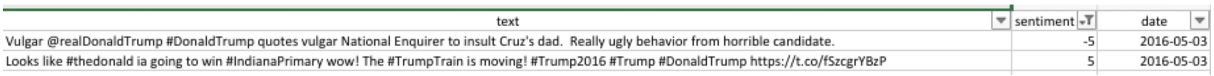


Figure 11: Who speaks for public?

There are actually 2 steps of doing this. Cross-search most concerned official account with most influent ones (using the number of total likes, sum up with follows and comments), then you can get 6 accounts. Then use the mean sentiment score of all individual accounts as comparison, you can see International News from Xinhua News Agency is the most representative official account.

Sentiment Analysis from Twitter:

For sentiment analysis, a lexicon of 6,800 words[2] was used. This lexicon separates words in positive in negative, making it possible to quantify sentiment. The algorithm used for analysis is very simple, the difference of the positive and negative words used in a text. As tweets are capped at 140 characters the number of words used don't vary too much and this simple model makes for easy comparison. Grouping of the sentiment scores showed that there was a huge emphasis to neutral scoring tweets. Manual analysis lead to the conclusion that lots of tweets are written with implicit sentiment, without using words that carry meaning in themselves but rather how they are used in the context. Despite this, the outliers do show a difference between people's opinions regarding Trump.



text	sentiment	date
Vulgar @realDonaldTrump #DonaldTrump quotes vulgar National Enquirer to insult Cruz's dad. Really ugly behavior from horrible candidate.	-5	2016-05-03
Looks like #thedonald ia going to win #IndianaPrimary wow! The #TrumpTrain is moving! #Trump2016 #Trump #DonaldTrump https://t.co/fSzcgrYBzP	5	2016-05-03

Figure 10 Strong opinions regarding Trump.

Error sources:

On of the limitations of the Twitter public API is the geographical position, which can not be set to a state, but rather to a point and a radius. In order to cover as much of US as possible while maintaining a low level of noise, both states of Alaska and Hawaii were omitted. The final search radius was set to 4500 km. This results in some tweets

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

originating from Mexico, Canada and possible other nations possibly being included, though because of the significantly lower usage-frequency in these countries the overall result should not be affected significantly.

Word frequency:

Word frequency was calculated by first normalizing tweets by removing stop words, and using stemming. Word frequencies were then calculated for each day. It became obvious that other presidential candidates were often mentioned together with Trump. Twitter's high usage of hashtags to group topics also affect the frequency, as popular hashtags, which often consist of several words together, overtake many regular keywords in the rankings.

Word Cloud

A difficult part of the analysis of social media content is the shift from quantitative to qualitative insights. Because it is not (yet) possible to pin the meaning of complex sentences or to distinguish irony from serious statements it might not even be possible to gather meaningful qualitative insights on a larger scale at all (see sentiment analysis).

One possible path to at least approach a kind of relation between words is to look at the words that are used together with a word of interest (in this case #trump). Still the meaning of the relations can't be clearly defined. But in an appropriate display the viewer can estimate the particular notion quite quickly. First of all the data needs to be cleaned though. Like in the text you're reading now, filling words without a particular meaning have the highest count. After those are erased its worth looking for coincidences that can also lead to misinterpretations.

Compared to other forms of visualization there is a unique connection the cloud makes: It links the quantity of a word to its font size. A word that occurs more often is then rendered bigger than the others. If the initial search word is included it will always be the biggest, because it is in every post of the selection.

In the special case of an animated word cloud the output will also show the development over time. If the words used in each timeframe are very different an animation might be less useful to trace a growth or shrinking, but rather for visualizing the temporary context. If there are two very unique expressions over a longer period of time, a word cloud might really help to see peaks in the relation between these two words or to identify them in the first place. Because the legend of such a chart is somehow included in itself it is also a useful to clarify the overall image, making further explanation unnecessary since it relies on our capacity to find headlines or other kinds of emphasized text.

Because a word cloud can be created easily automatically from p.e. the data of a blog, it has been used quite often lately. Not only in terms of style the word cloud has therefore a somewhat random image, also the content often is not that interesting.

In the Trump-Case on the other hand it fits quite well, because there are different topics he talks about during his campaign.

RESULTS

Hypothesis 1: *People in China are interested in Donald Trump´s presidential campaign and react to it on social media.*

The following diagram shows the number of Weibo posts using #特朗普 as well as the number of Baidu searches related to Donald Trump since June 2015 when he announced his candidacy, until the beginning of May 2016 when he was elected as frontrunner for the Republican Party.

Refer to the Graphic: What makes people talk? Chinese Social Media`s Response to Donald J. Trump

Over the studied time, some high peeks on the number of posts are noticeable, while the Baidu search curve is steadier during the first months, and turns peakier after December 2015.

Hypothesis 2: *The conversation around Donald Trump`s campaign revolves around some controversial ideas (like the Mexican border wall, expelling the Muslims, etc.) while "boring" but non less important points of his campaign are ignored.*

To prove this hypothesis two animated word clouds where developed, one in Chinese for Weibo and one in English for Twitter. They show the changes through a period of 10 days of the most used words:

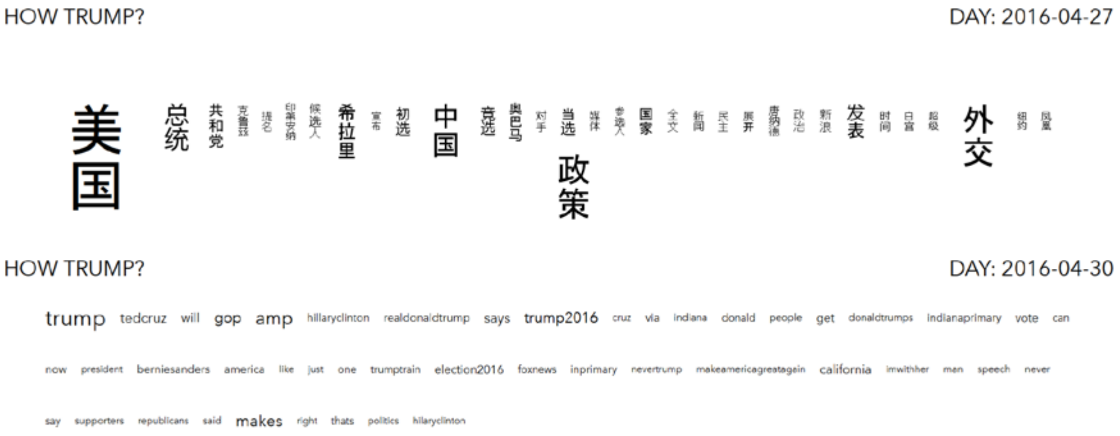

HOW TRUMP?
DAY: 2016-04-27
美国
总统
共和党
克鲁兹
提名
印第安纳
候选人
希拉里
宣布
初选
中国
竞选
奥巴马
对手
当选
政策
媒体
参选人
国家
全文
新闻
民主
展开
唐纳德
政治
新浪
发表
时间
白宫
超级
外交
纽约
凤凰
HOW TRUMP?
DAY: 2016-04-30
trump tedcruz will gop amp hillaryclinton realdonaldtrump says trump2016 cruz via indiana donald people get donaldtrumps indianaprimary vote can
now president berniesanders america like just one trumptrain election2016 foxnews inprimary nevertrump makeamericagreatagein california imwithher men speech never
say supporters republicans said makes right thats politics hilaryclinton

Figure 11

Click here to check the animated word clouds.

Click here to check the GitHub file.

Unfortunately, the word cloud was not a great tool to prove this Hypothesis. Even if it looks appealing (when animated), the results are not reveling or precise. Taking apart the limitations from this method, is possible to see that mostly all the words used related to other candidates or to general events like the elections or the primaries, but words related to his policies do not appear.

Hypothesis 3: *Trump`s statements regarding China influence the perception that Chinese people have of him.*

In order to prove this Hypothesis, collections of sentiments were considered to have information

on the different perception of Chinese people on Trump's sentences.
The sentiments were collected from Twitter and Weibo during a specific period of time from May the 25th to April the 5th, the goal was to visualize the flow of perceptions recorded in the two social networks.
The interesting elements that were providing relevant information for the research were the quantity of posts, the information about time and date in which they were published and the sentiment scale.
The team decided to try different types of diagrams to find the solution that describes in a better way three level of information. The first attempt was about the representation with a Scatter Plot diagram, but for the visualization was not easy because the readable layer of quantity had interferences with the time distribution. The team decided to change the chart's typology.
The Second attempt was to use the a 3d representation of the chart using the isometric view, but the issue that emerged was the representation of different values of data in the front were hiding the representation of data in the back. For this reason the team, after different experiments, decided to use a 3d diagram to have three different level of information for each social network in the same chart.
The idea of the 3D model is to have the value regarding the time on the X axis, the scale of sentiment's score on the Y axis and the quantity of posts on the Z axis. The area in which the reader could read the analysis is composed by a repetition of bar chart elements for each day. For each day there are two different elements that represent the tendency of posts of Weibo and Twitter.

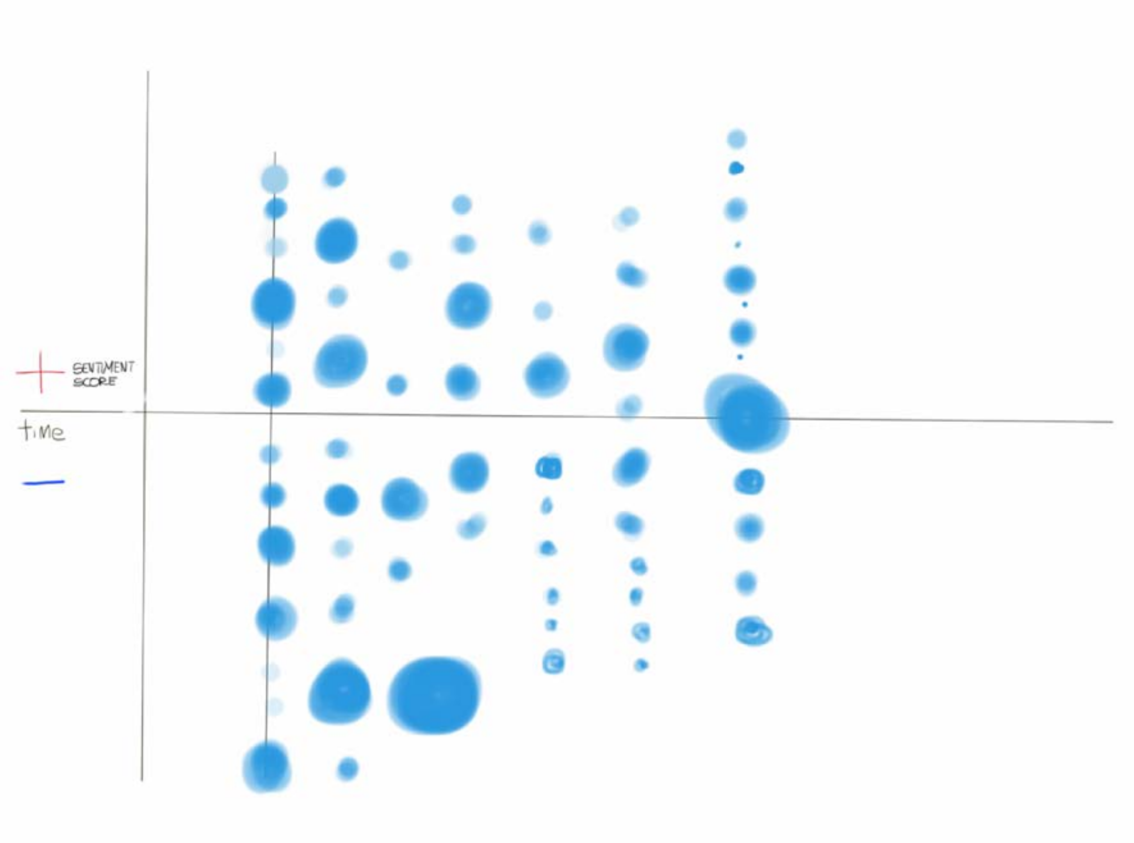


Figure 12. Scatter Plot

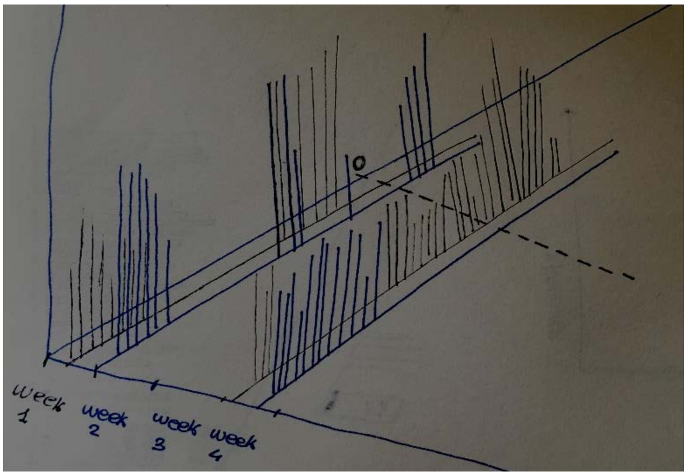


Figure 13

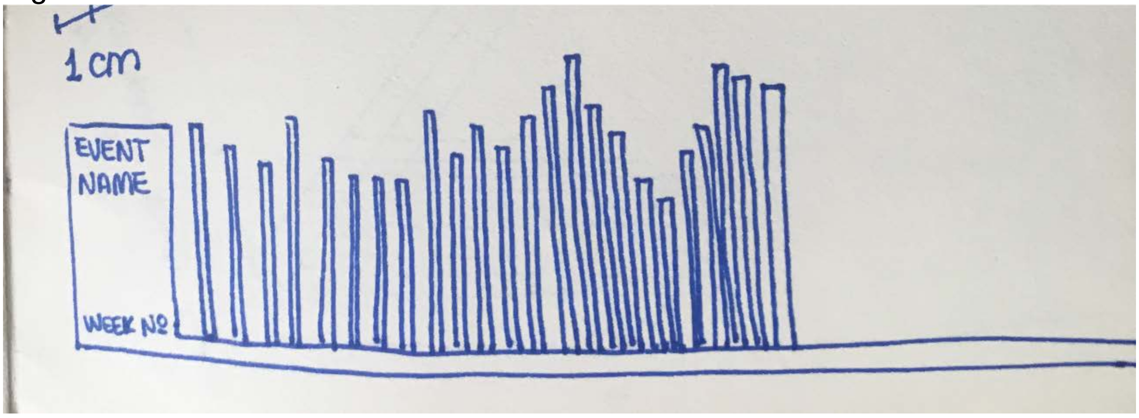


Figure 14

Sadly, due to a lack of time and resources it was not possible to create the physical model and the team decided that the best way to represent it (with the time constraint) was the following:

Refer to Graphic: Sentiment on Twitter / Weibo