Reddit Threads Monitoring

This Project will help investigator explore what's going on Reddit more easily and find out the attitude of people towards threads.

Yi Qin	He Huang	Chenglong Sun	Jianan Xiang
yq468@nyu.edu	hh1443@nyu.edu	cs4543@nyu.edu	jx696@nyu.edu
N18068309	N13422894	N17840926	N17143210

Project page (on Github): https://github.com/NYU-CS6313-Fall16/Reddit-Threads-17
Video: https://vimeo.com/196916683
Working demo: http://redditmonitor17.herokuapp.com/

Problem Statement

One sentence summary: our project aims to help investigators explore what's going on Reddit, what top threads are and the attitude of people on these threads and what time people contribute to these threads.

Exploring threads activity over time and people's attitude on those threads is difficult for investigators if they look at the Reddit page directly. The raw thread page has so much information that investigator can't extract valuable information they needed, especially the activity of thread over time and the attitudes of people towards the thread.

Thus we want to help investigators filter and visualize the useful thread data so that they can understand the activity of thread and people's attribute on threads.

Visualization Questions

Given the nature of our data, we assume few things about investigator: the investigator has the intention to explore more about the dataset and perhaps want to pinpoint a certain entity for investigation.
According to our data from Reddit api, Reddit threads/posts could be used for this entity.
Questions about the threads:

- What top 30 threads are, what are their title keywords and Reddit score ?
- What subreddit do they belong to ?
- What attitudes do people have on these threads, positive or negative ?
- What time slot does certain thread attract users' focus?

Data Attributes

Our original data is from Reddit API and because we use Node.js to build backend framework, we use snoowrap package to extract thread data. We only extract some useful attributes through the API and transfer these attributes into new form so that we can use them at front-end.

The derived attributes are as below:

Title keyword: We used "keyword-extractor" to extract keyword from title, the reason we do this is that original title sometimes is too long and can't be understood by the investigator quickly.. Extract only meaningful keywords from title can provide user concise and direct information about thread title.

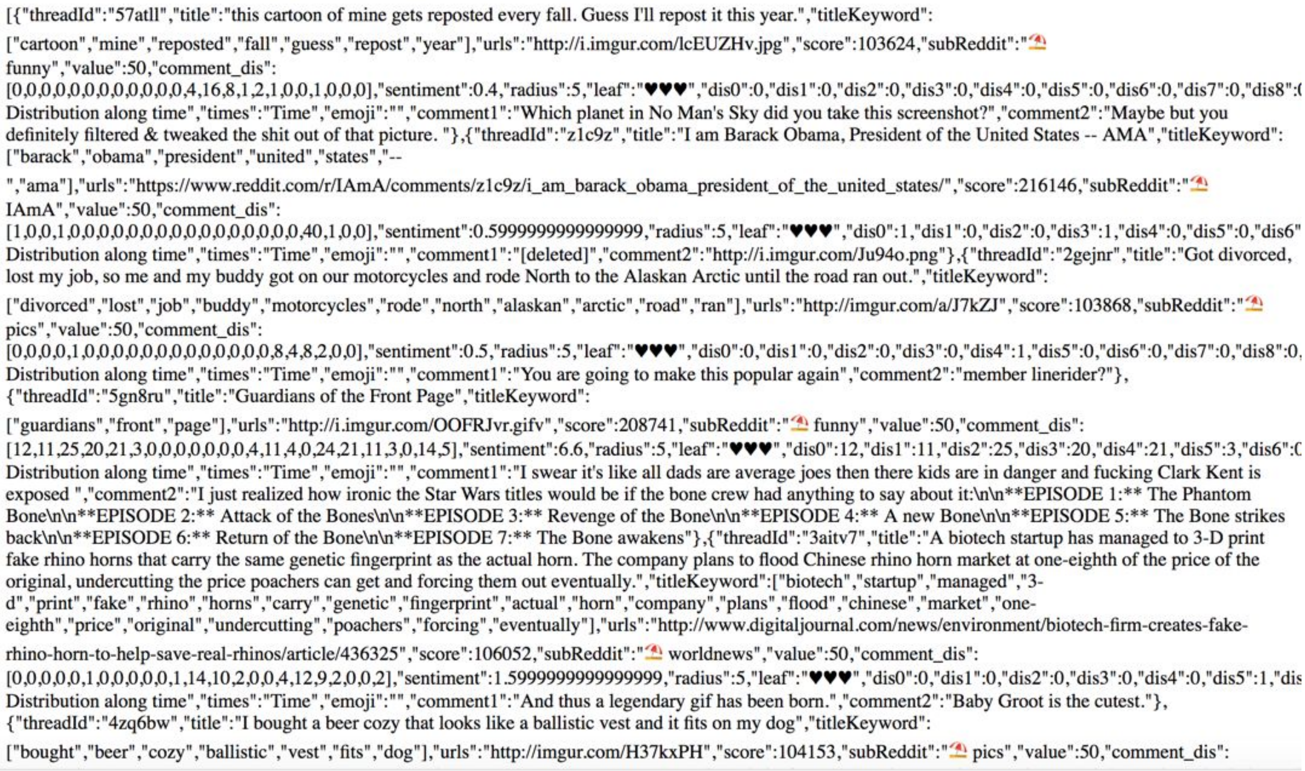
Comment_distrubution: we want to analyze the comment time distribution of thread, so we first extract time of each comment and do summation of each time range, finally we build an array to record these summations, index of array represents time range.

Sentiment: we analyzed the top 30 comments in each thread.Then calculate a value that can represent the overall sentiment of these comments in each thread . In this way, user can know the attitude of people towards this thread.

Here is form that contains all the attributes:

Attribute Name	Attribute Type	Attribute Meaning	Value Range/Categories	Derived or not
threadId	categorical	the id of a particular forum	All threads	Not
title	unstructured text	Thread title	All threads	Not
titleKeyword	Array of string	Thread title keyword	Depends	Yes
urls	categorical	Thread content url	All threads	Not
score	Quantitative	Score of thread		Not
subReddit	categorical	subReddit thread belongs to	All threads	Not
value	Numerical		50	Yes
comment_dis	Numerical Array	The distribution of comment across 24 hour, each slot represents 1 hour range.	Depends.	Yes
sentiment	Numerical	Sentiment from thread comments	(-1,1)	Yes
radius	unstructured text	Control	50	No
max	Numerical	The maximum number of comment_dis	Depends	yes

Data(JSON) exported to front-end is like below:



```
[{"threadId":"57atll","title":"this cartoon of mine gets reposted every fall. Guess I'll repost it this year.","titleKeyword":
["cartoon","mine","reposted","fall","guess","repost","year"],"urls":"http://i.imgur.com/lcEUZHv.jpg","score":103624,"subReddit":"⛱
funny","value":50,"comment_dis":
[0,0,0,0,0,0,0,0,0,0,0,0,4,16,8,1,2,1,0,0,1,0,0,0],"sentiment":0.4,"radius":5,"leaf":"♥♥♥","dis0":0,"dis1":0,"dis2":0,"dis3":0,"dis4":0,"dis5":0,"dis6":0,"dis7":0,"dis8":
Distribution along time","times":"Time","emoji":"","comment1":"Which planet in No Man's Sky did you take this screenshot?","comment2":"Maybe but you
definitely filtered & tweaked the shit out of that picture. "},{"threadId":"z1c9z","title":"I am Barack Obama, President of the United States -- AMA","titleKeyword":
["barack","obama","president","united","states","--
","ama"],"urls":"https://www.reddit.com/r/IAmA/comments/z1c9z/i_am_barack_obama_president_of_the_united_states/","score":216146,"subReddit":"⛱
IAmA","value":50,"comment_dis":
[1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,40,1,0,0],"sentiment":0.5999999999999999,"radius":5,"leaf":"♥♥♥","dis0":1,"dis1":0,"dis2":0,"dis3":1,"dis4":0,"dis5":0,"dis6"
Distribution along time","times":"Time","emoji":"","comment1":"[deleted]","comment2":"http://i.imgur.com/Ju94o.png"},{"threadId":"2gejnr","title":"Got divorced,
lost my job, so me and my buddy got on our motorcycles and rode North to the Alaskan Arctic until the road ran out.","titleKeyword":
["divorced","lost","job","buddy","motorcycles","rode","north","alaskan","arctic","road","ran"],"urls":"http://imgur.com/a/J7kZJ","score":103868,"subReddit":"⛱
pics","value":50,"comment_dis":
[0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,8,4,8,2,0,0],"sentiment":0.5,"radius":5,"leaf":"♥♥♥","dis0":0,"dis1":0,"dis2":0,"dis3":0,"dis4":1,"dis5":0,"dis6":0,"dis7":0,"dis8":0,
Distribution along time","times":"Time","emoji":"","comment1":"You are going to make this popular again","comment2":"member linerider?"},
{"threadId":"5gn8ru","title":"Guardians of the Front Page","titleKeyword":
["guardians","front","page"],"urls":"http://i.imgur.com/OOFRJvr.gifv","score":208741,"subReddit":"⛱ funny","value":50,"comment_dis":
[12,11,25,20,21,3,0,0,0,0,0,0,0,0,4,11,4,0,24,21,11,3,0,14,5],"sentiment":6.6,"radius":5,"leaf":"♥♥♥","dis0":12,"dis1":11,"dis2":25,"dis3":20,"dis4":21,"dis5":3,"dis6":
Distribution along time","times":"Time","emoji":"","comment1":"I swear it's like all dads are average joes then there kids are in danger and fucking Clark Kent is
exposed ","comment2":"I just realized how ironic the Star Wars titles would be if the bone crew had anything to say about it:\n\n**EPISODE 1:** The Phantom
Bone\n\n**EPISODE 2:** Attack of the Bones\n\n**EPISODE 3:** Revenge of the Bone\n\n**EPISODE 4:** A new Bone\n\n**EPISODE 5:** The Bone strikes
back\n\n**EPISODE 6:** Return of the Bone\n\n**EPISODE 7:** The Bone awakens"},{"threadId":"3aitv7","title":"A biotech startup has managed to 3-D print
fake rhino horns that carry the same genetic fingerprint as the actual horn. The company plans to flood Chinese rhino horn market at one-eighth of the price of the
original, undercutting the price poachers can get and forcing them out eventually.","titleKeyword":["biotech","startup","managed","3-
d","print","fake","rhino","horns","carry","genetic","fingerprint","actual","horn","company","plans","flood","chinese","market","one-
eighth","price","original","undercutting","poachers","forcing","eventually"],"urls":"http://www.digitaljournal.com/news/environment/biotech-firm-creates-fake-
rhino-horn-to-help-save-real-rhinos/article/436325","score":106052,"subReddit":"⛱ worldnews","value":50,"comment_dis":
[0,0,0,0,0,1,0,0,0,0,0,1,14,10,2,0,0,4,12,9,2,0,0,2],"sentiment":1.5999999999999999,"radius":5,"leaf":"♥♥♥","dis0":0,"dis1":0,"dis2":0,"dis3":0,"dis4":0,"dis5":1,"dis
Distribution along time","times":"Time","emoji":"","comment1":"And thus a legendary gif has been born.","comment2":"Baby Groot is the cutest."},
{"threadId":"4zq6bw","title":"I bought a beer cozy that looks like a ballistic vest and it fits on my dog","titleKeyword":
["bought","beer","cozy","ballistic","vest","fits","dog"],"urls":"http://imgur.com/H37kxPH","score":104153,"subReddit":"⛱ pics","value":50,"comment_dis":
```

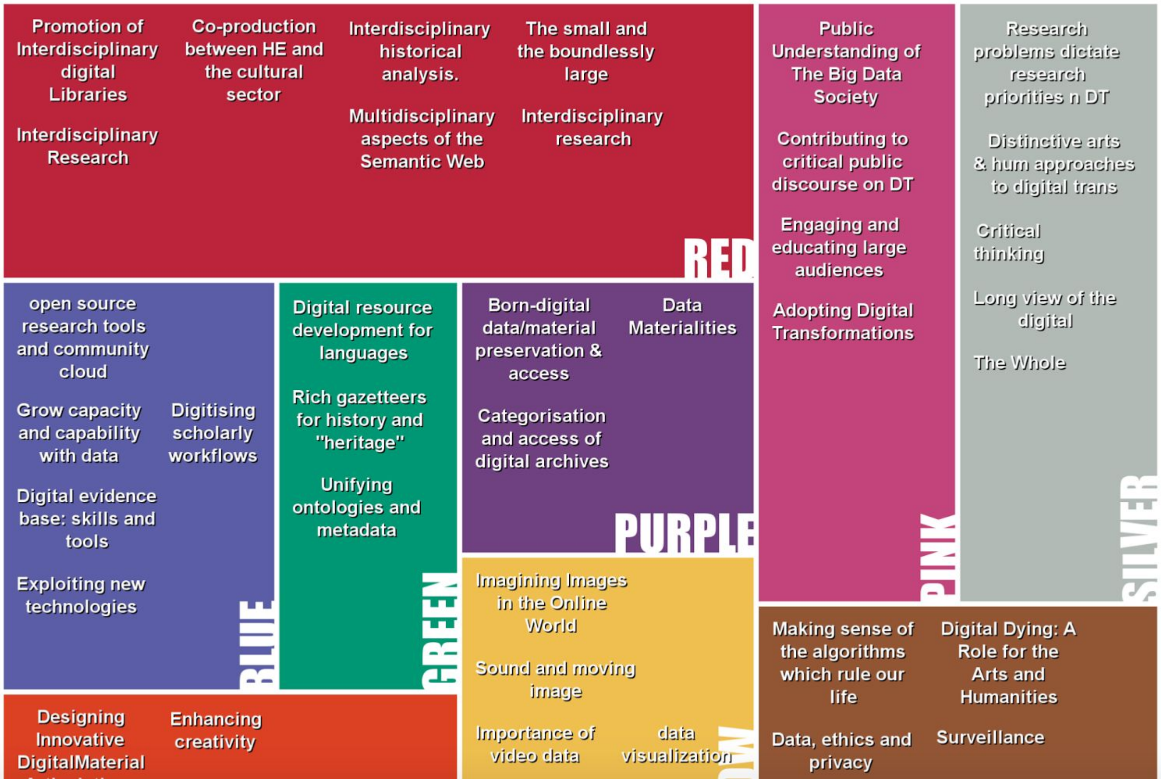
Related Works

We take some ideas from following projects

Previous works has been done on other kinds of events monitoring or recording.

Digital Transformations Event Tree Map

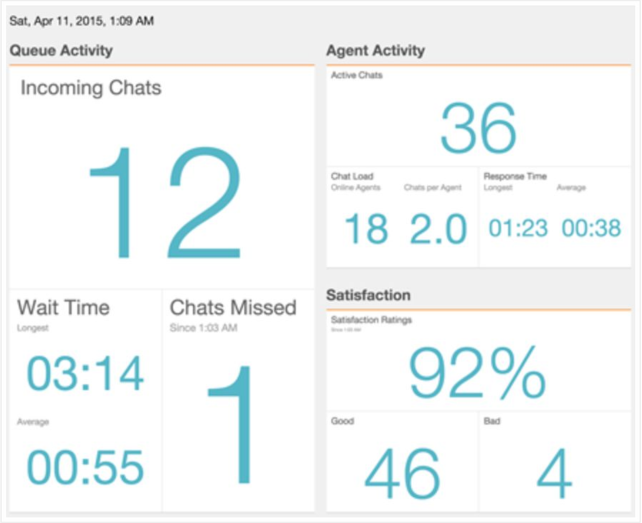
By Wendy Matcham



- Each section represents groups of challenges which the attendees thought were similar.
- Hover Over or Tap any challenge to see the longer description submitted by the attendees
- Click or Tap a colour name to see a video describing the area challenges
- **Relation:** This project groupby the similar answers together and uses different colors to identify different groups, that means we could group our thread object data by subReddit section and each section has unique color hue as identifier. Another hint we get from this treemap is that they just show short description of contents, in our project, we only show the title keyword, not title itself.

Monitoring real-time chat metrics

By Nora Mullen



Sat, Apr 11, 2015, 1:09 AM

Queue Activity

Incoming Chats 12

Wait Time
Longest 03:14
Average 00:55

Chats Missed
Since 1:03 AM
1

Agent Activity

Active Chats 36

Chat Load
Online Agents 18
Chats per Agent 2.0

Response Time
Longest 01:23
Average 00:38

Satisfaction

Satisfaction Ratings 92%

Good 46

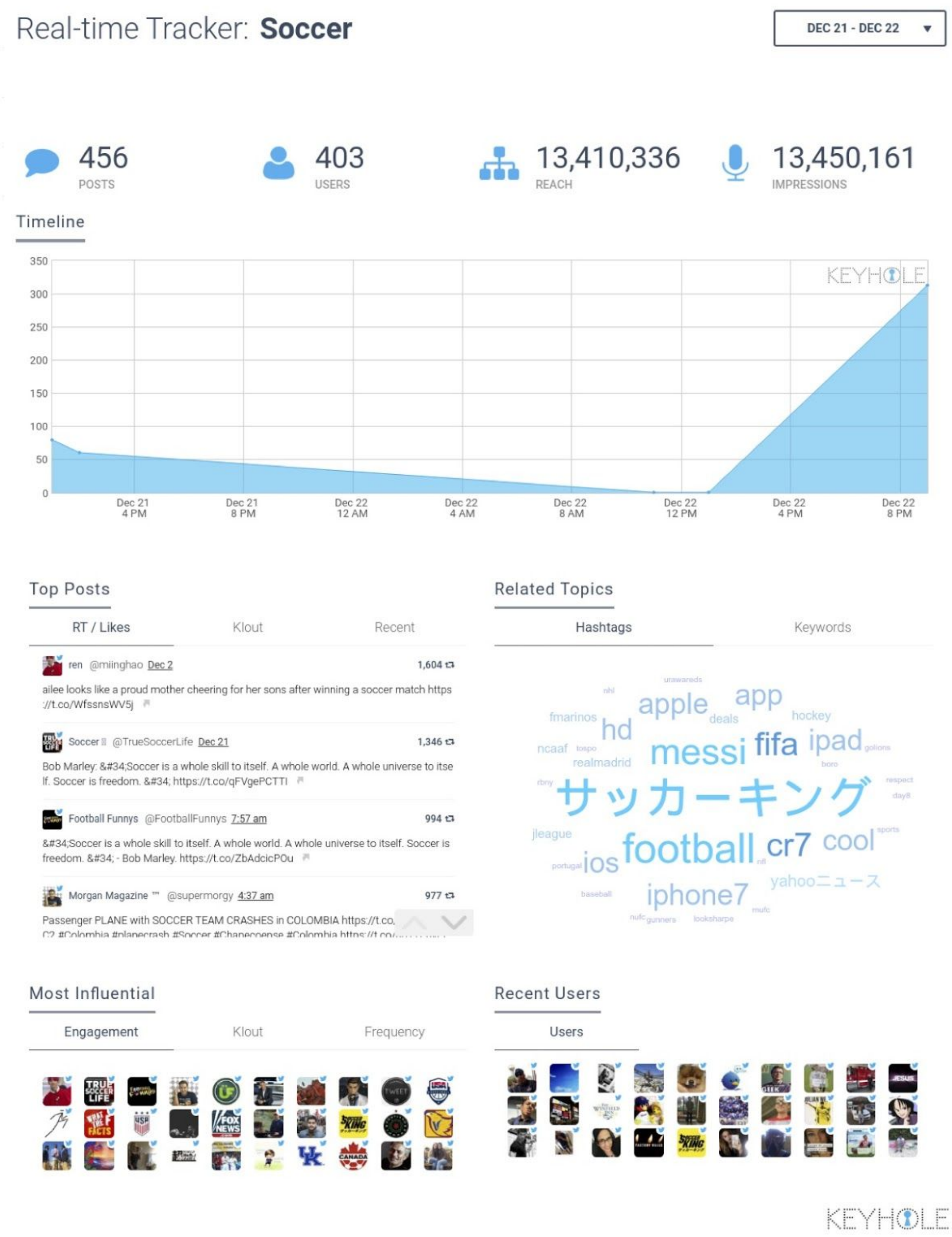
Bad 4

Description: an overview of key chat metrics, including queue size, customer wait times, and chat satisfaction, on a single screen.

- Real-time dashboard for recording important chat metrics.
- Three main blocks, queue activity, agent activity and satisfaction
- Each main block is divided by specific chat metrics and they are shown real-time.
- **Relation:** In our project, we groupby data according to their subreddit. In our original thought, when we step into thread level, we still use this kind of block to show different statistics of thread.

Keyhole

One of top 25 social media monitoring tools



Real-time Tracker: **Soccer**

DEC 21 - DEC 22 ▼

456 POSTS

403 USERS

13,410,336 REACH

13,450,161 IMPRESSIONS

Timeline

Top Posts

RT / Likes | Klout | Recent

ren @miinghao Dec 2 1,604

ailee looks like a proud mother cheering for her sons after winning a soccer match https://t.co/WfssnsWV5j

Soccer @TrueSoccerLife Dec 21 1,346

Bob Marley: &#34;Soccer is a whole skill to itself. A whole world. A whole universe to itself. Soccer is freedom. &#34; https://t.co/qFVgePCTTI

Football Funnys @FootballFunnys 7:57 am 994

&#34;Soccer is a whole skill to itself. A whole world. A whole universe to itself. Soccer is freedom. &#34; - Bob Marley. https://t.co/ZbAdcicPOu

Morgan Magazine ™ @supermorgy 4:37 am 977

Passenger PLANE with SOCCER TEAM CRASHES in COLOMBIA https://t.co,
C2 #Colombia #planecrash #Soccer #Chapecoense #Colombia https://t.co/

Related Topics

Hashtags | Keywords

Most Influential

Engagement | Klout | Frequency

Recent Users

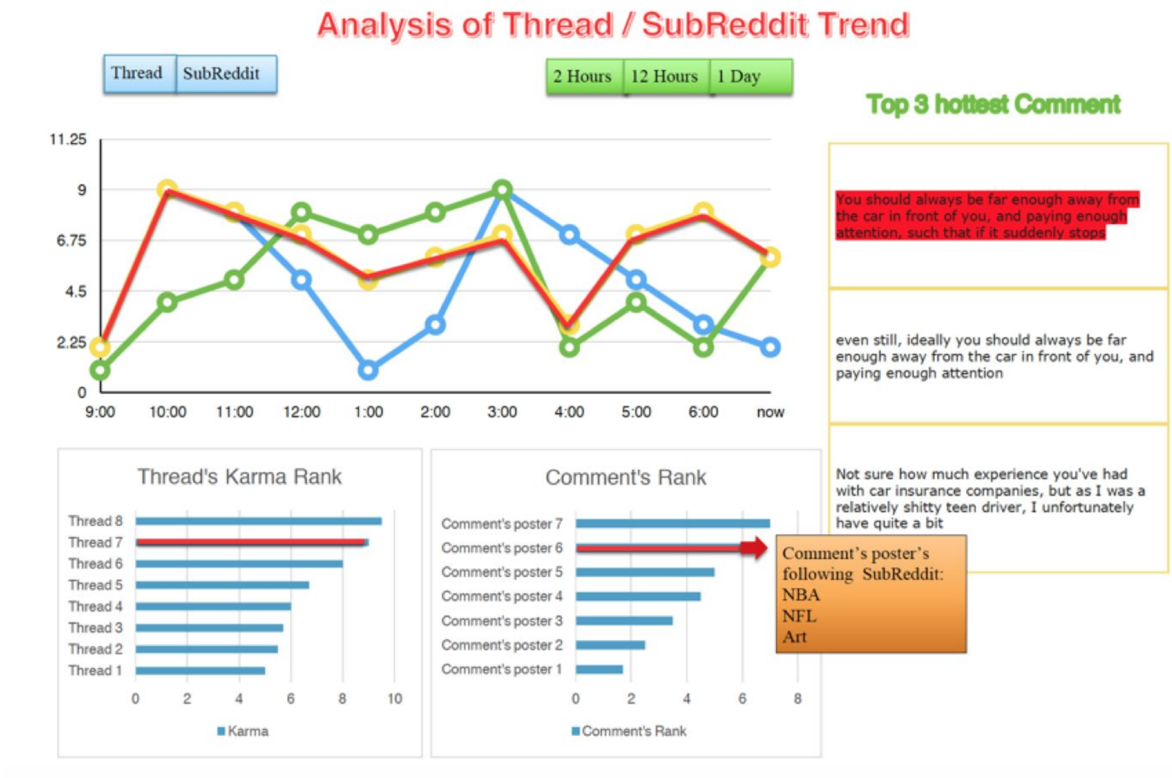
Users

Description: Real-time media monitoring visualization and analysis

- This dashboard tool shows many information about certain topic
- It shows timeline, top posts, related topics, keywords and user information
- **Relation**: our project shows the comment distribution across time and we also extract keywords from title.

Design Iterations

Initial Mockup:



You should always be far enough away from the car in front of you, and paying enough attention, such that if it suddenly stops

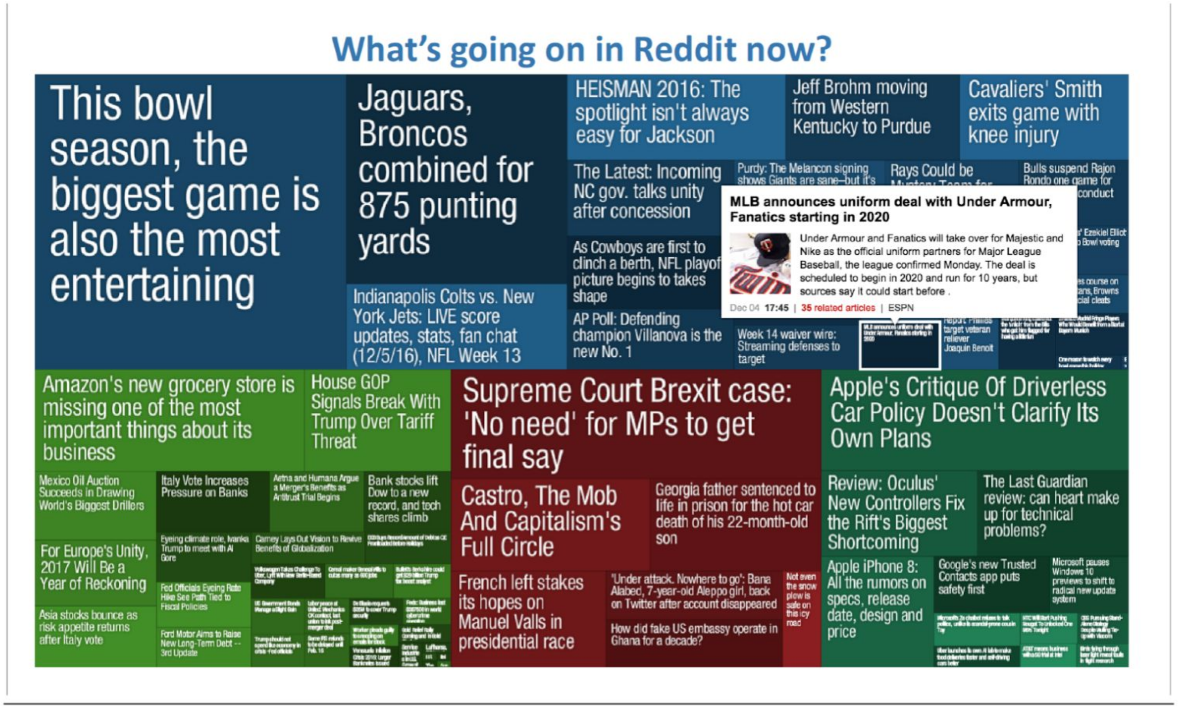
even still, ideally you should always be far enough away from the car in front of you, and paying enough attention

Not sure how much experience you've had with car insurance companies, but as I was a relatively shitty teen driver, I unfortunately have quite a bit

Our initial mockup wants to show the thread activity, its relative rank and most popular comment as well as hottest comment, the top line graph shows the timeline of thread across time. The left graph on the bottom shows thread ranks according to their score. The right graph on the bottom shows comment rank according to comment karma score, its vertical axis shows the comment poster name. The right graph shows the specific hottest comment of thread. If we click a comment, the thread containing this comment will be highlighted, in addition, comment content will also be highlighted.

Our initial mockup is denied by professor, so we change our idea about project in the progress report.

Progress Report:


What's going on in Reddit now?
This bowl season, the biggest game is also the most entertaining
Jaguars, Broncos combined for 875 punting yards
Indianapolis Colts vs. New York Jets: LIVE score updates, stats, fan chat (12/5/16), NFL Week 13
HEISMAN 2016: The spotlight isn't always easy for Jackson
Jeff Brohm moving from Western Kentucky to Purdue
Cavaliers' Smith exits game with knee injury
The Latest: Incoming NC gov. talks unity after concession
Purdy: The Melancon signing shows Giants are sane—but it's
Rays Could be
Bulls suspend Rajon Rondo one game for conduct
As Cowboys are first to clinch a berth, NFL playof picture begins to takes shape
AP Poll: Defending champion Villanova is the new No. 1
Week 14 waiver wire: Streaming defenses to target
target veteran reliever Joaquin Benoit
Amazon's new grocery store is missing one of the most important things about its business
House GOP Signals Break With Trump Over Tariff Threat
Mexico Oil Auction Succeeds in Drawing World's Biggest Drillers
Italy Vote Increases Pressure on Banks
Aetna and Humana Argue a Merger's Benefits as Antitrust Trial Begins
Bank stocks lift Dow to a new record, and tech shares climb
For Europe's Unity, 2017 Will Be a Year of Reckoning
Eyeing climate role, Ivanka Trump to meet with Al Gore
Carney Lays Out Vision to Revive Benefits of Globalization
Fed Officials Eyeing Rate Hike See Path Tied to Fiscal Policies
Asia stocks bounce as risk appetite returns after Italy vote
Ford Motor Aims to Raise New Long-Term Debt -- 3rd Update
Supreme Court Brexit case: 'No need' for MPs to get final say
Castro, The Mob And Capitalism's Full Circle
Georgia father sentenced to life in prison for the hot car death of his 22-month-old son
French left stakes its hopes on Manuel Valls in presidential race
'Under attack. Nowhere to go': Bana Alabed, 7-year-old Aleppo girl, back on Twitter after account disappeared
How did fake US embassy operate in Ghana for a decade?
Not even the snow plow is safe on this icy road
Apple's Critique Of Driverless Car Policy Doesn't Clarify Its Own Plans
Review: Oculus' New Controllers Fix the Rift's Biggest Shortcoming
The Last Guardian review: can heart make up for technical problems?
Apple iPhone 8: All the rumors on specs, release date, design and price
Google's new Trusted Contacts app puts safety first
Microsoft pauses Windows 10 previews to shift to radical new update system

MLB announces uniform deal with Under Armour, Fanatics starting in 2020

Under Armour and Fanatics will take over for Majestic and Nike as the official uniform partners for Major League Baseball, the league confirmed Monday. The deal is scheduled to begin in 2020 and run for 10 years, but sources say it could start before .

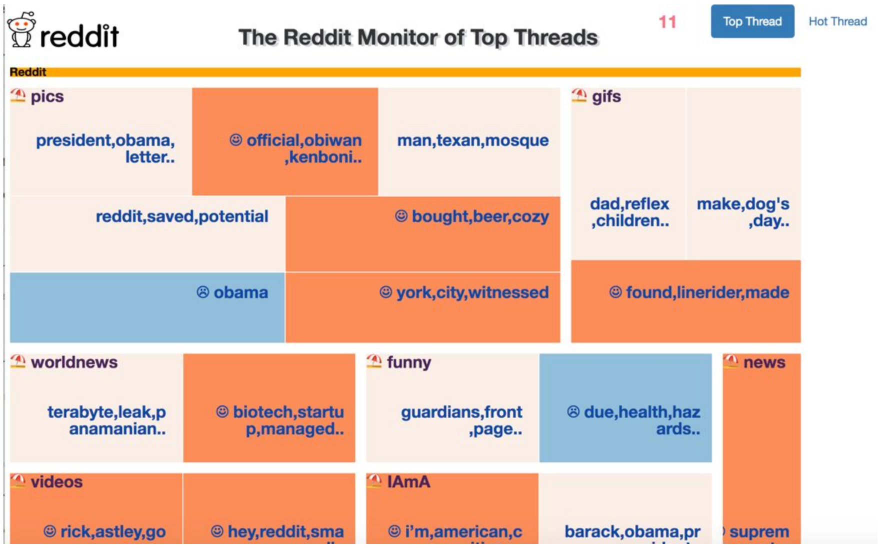
Dec 04 17:45 | 35 related articles | ESPN

This time, we changed our project visualization structure into treemap, each small block represents a thread and shows the thread title contents. If threads belong to same subreddit, then we groupby them together, different color hue represents different subreddit. Here, the color saturation in subreddit represents the attitude of people towards thread, the color is more darker, corresponding attitude is more negative. As for interaction, if we click one thread, a small window will show us the details of thread, like url, contents and keyword of the thread.

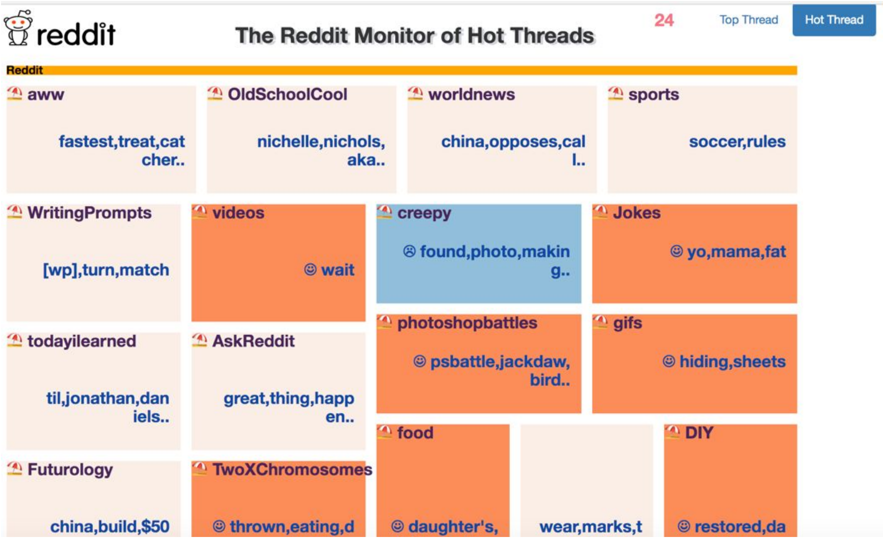
Final Visualization

The Treemap in our final visualization has three levels. Besides the treemap in each level, each level contains a control panel and the directory. In the control panel, there is count down timer, it will shows how many seconds left to refresh the data(We refresh the data every 60 seconds). And there is a pill that we can choose the top threads to visualize or hot threads to visualize. The Directory can show the user which level you are at, and if you click the directory, you can return to the previous level.

1. The first level shows all the general observations of each thread. Different subreddits are divided by larger white gaps and subreddit names are in black font color which are shown on the top left of frame. Each small box represent a thread which are divided by smaller gaps. The blue name shows the title keyword of thread. At this level, the size of small box doesn't represent anything, we make them in same size because in our view, each thread is equally important. If their sizes are decided by other thing like thread score or something else, bigger box has more chance to be clicked than smaller box. The emoji at the front of thread **title keyword** and the **background** of each small box represents the attitudes of people towards thread. Smiling **emoji** and orange background means positive, sadness emoji and dark blue **color** means negative, and all the others illustrates the sentiment is mild.

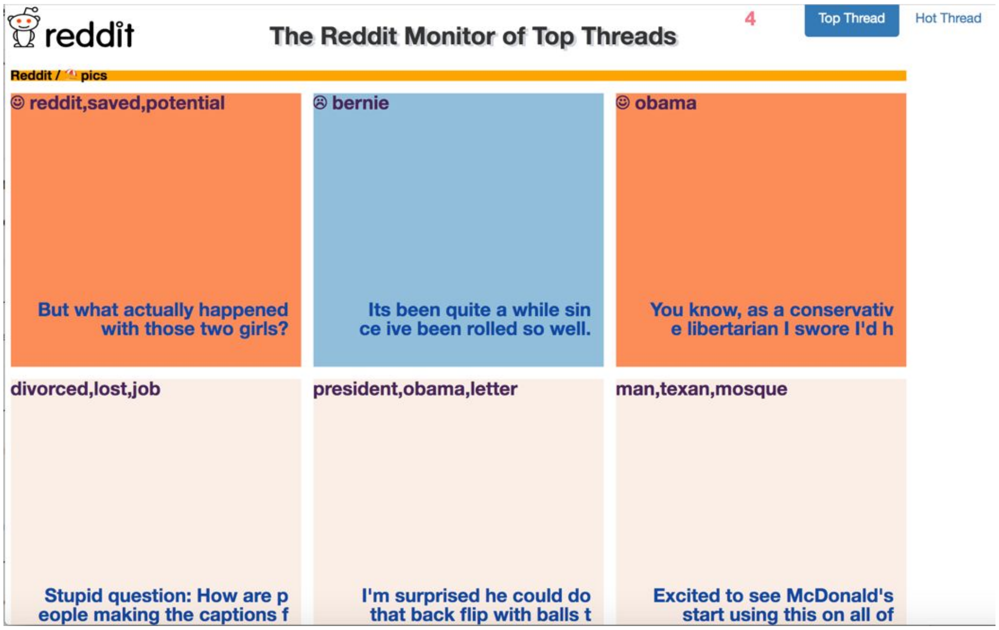


The Top thread monitor first level interface



The Hot thread monitor first level interface

2. The graph below shows the second level, when you click any thread on first level, the visualization will be transferred into graph below with slow transitioning effect. The second level is based on subreddit and subreddit information is shown in the directory, for example, below graph represents the thread in subreddit *pics*. Its attributes title **keywords, background color and the emoji** still represent the same thing in the first level, besides these attributes, we add **hottest comment** in each thread at this level, so the user can see the comments details for the corresponding thread.



☺ reddit,saved,potential

But what actually happened with those two girls?

☹ bernie

Its been quite a while sin ce ive been rolled so well.

☺ obama

You know, as a conservativ e libertarian I swore I'd h

divorced,lost,job

Stupid question: How are p eople making the captions f

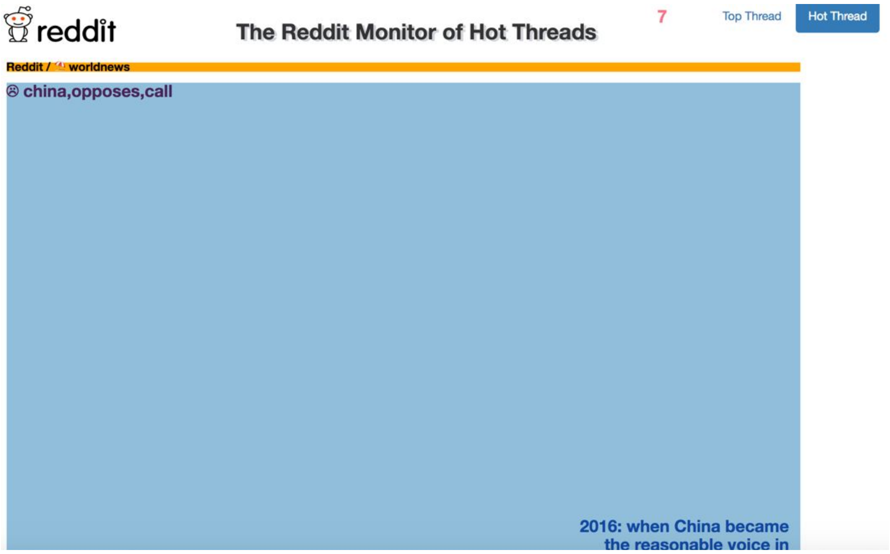
president,obama,letter

I'm surprised he could do that back flip with balls t

man,texan,mosque

Excited to see McDonald's start using this on all of

The Top thread monitor second level interface



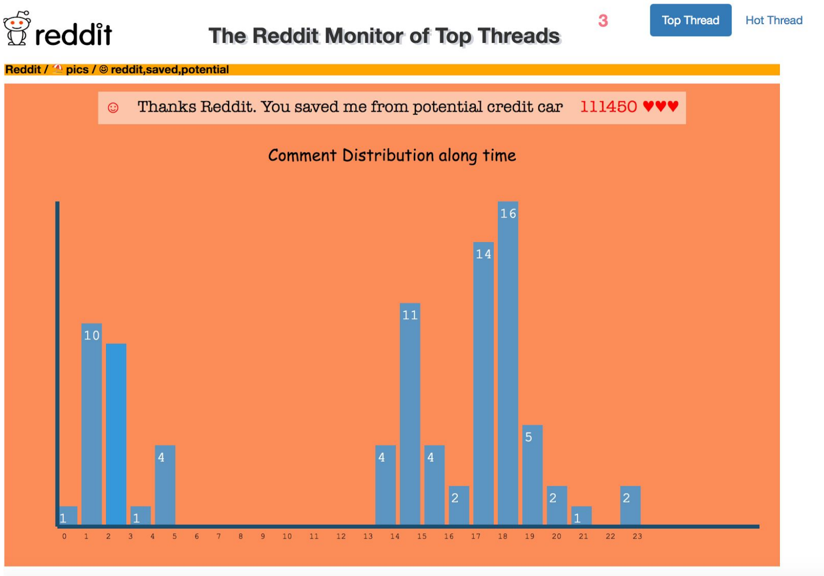
☹ china,opposes,call

2016: when China became the reasonable voice in

The Hot thread monitor second level interface

3. The graph below shows the third level, when you click any thread on second level,

the visualization will be transferred into graph below with slow transitioning effect. The **title keyword and subreddit information** will be displayed in the **directory**. The **background color** represents the attitudes of people towards this thread, this color inherits from second level. The text within white frame is the **actual title** of thread, this title could be clicked which render a page showing thread contents. Moreover, the number following the title represents the **thread score**. The **bar chart** below shows the distribution of comments across time. Because Reddit API rule, we couldn't extract all comments. And we assume those comments could represent whole comments distribution across time.



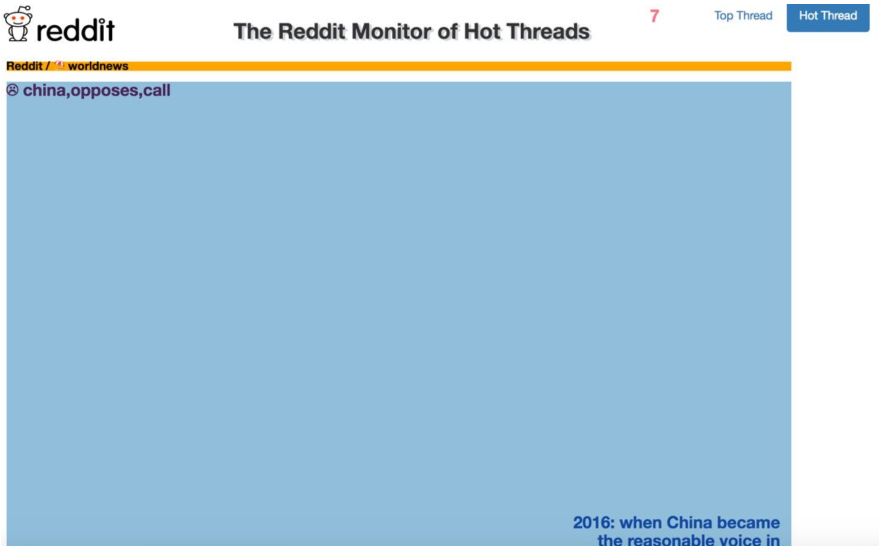
The Reddit Monitor of Top Threads

Reddit / pics / ☺ reddit,saved,potential

☺ Thanks Reddit. You saved me from potential credit car 111450 ♥♥♥

Comment Distribution along time

The Top thread monitor third level interface



The Reddit Monitor of Hot Threads

Reddit / worldnews

☹ china,opposes,call

2016: when China became
the reasonable voice in

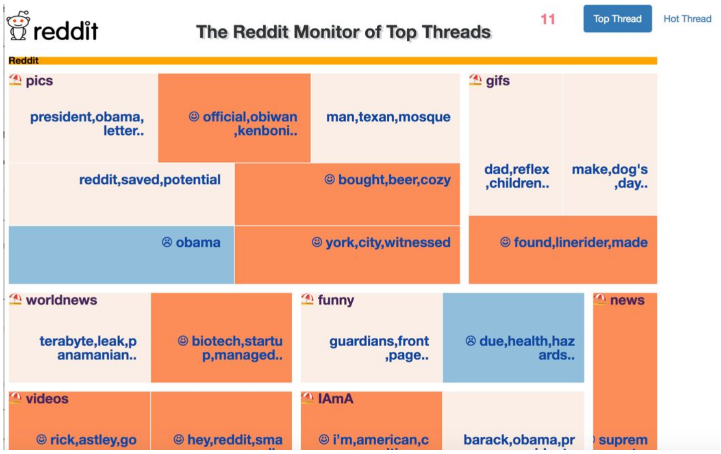
The Hot thread monitor third level interface

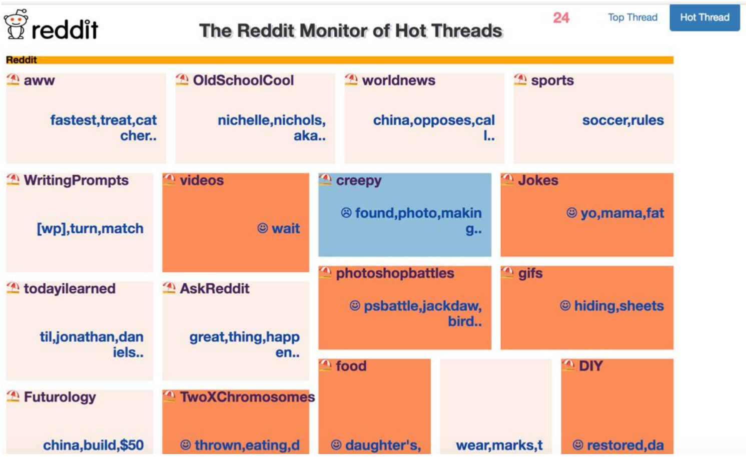
Findings

Our project implements real-time Reddit Monitoring, the investigator could know what's happening on Reddit, including current top threads and current hot threads, so our project don't have much data to do analysis but we can still find some interesting patterns from project. Here, top threads are ranked by total scores of threads since they are posted. Threads become hot threads because they are discussed and viewed most in most recent time.

1. Top threads have less subreddits than hot threads.
After clarifying these two concepts, we can find that some interesting phenomenon exists from below two graphs. As we can see, in left graph, there exists only 7 subreddits and *pics* subreddit contains five threads. In right graph, there exists 25 subreddits, each subreddit just hold only one thread. Not like left top threads graph, right hot threads graph changes frequently and each subreddit only holds one thread at most time.

2. ***Pics** and **gifs*** are always the the most popular subReddits in Top Thread
We find that *pics* and *gifs* are always the most popular subReddit, there are always at least 3 top threads ranked in the top 30. We can find that people more like the picture-type threads than other types like videos, text etc. Picture is a very direct and easiest way for user to understand what's happening now or express something.

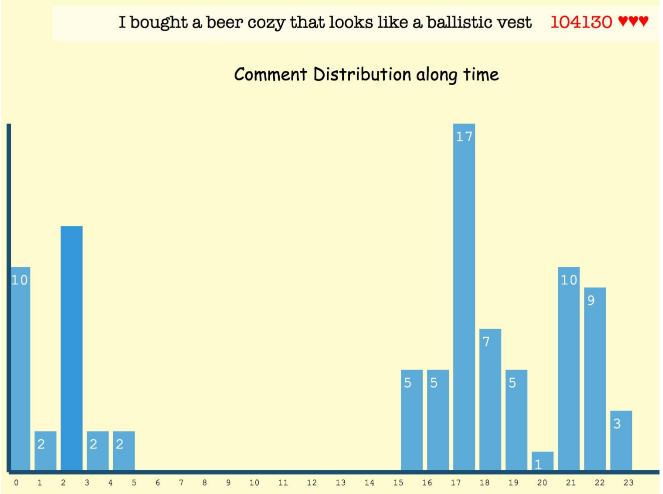


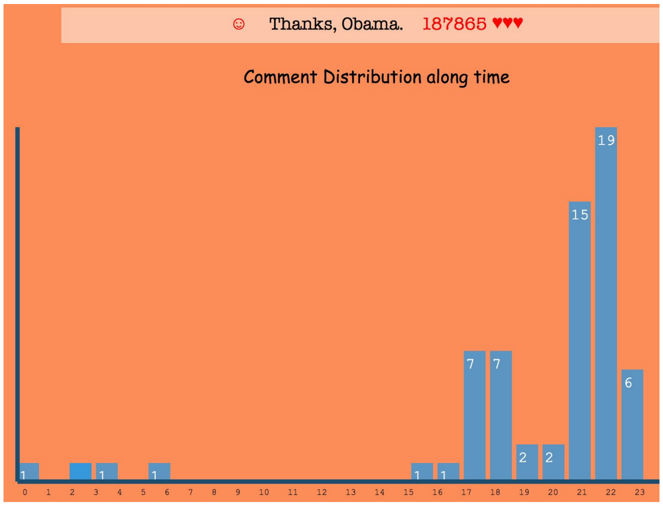


Top thread Hot Thread

3. Users are always active at some time slot.

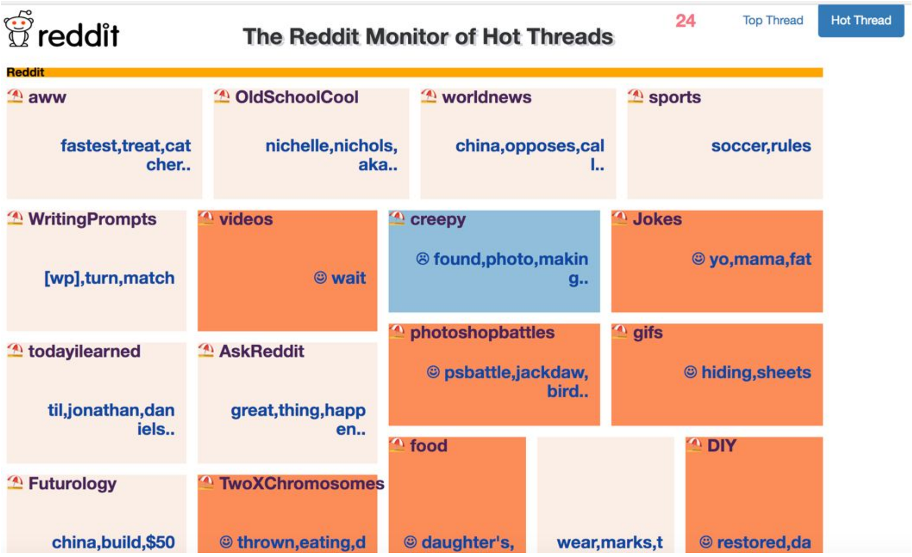
The most active time slots of all threads are from 17:00 to 23:00, even though it has already been proved by tons of surveys from social scientists, we also find the same result. We visualized the comments distribution of top threads in detail, as you can see the graph below, Reddit users' most active time is in the night, maybe the night time is the causal time after work and study.





4. Attitudes of users tend to be neutral

As to most trending threads, people will take a neutral attitude about them, in another word, the amount of positive sides or negative sides are equal, every user will express his real feeling or attitude about threads, they may debate with each other, the final result of people's attitude is neutral. Not too many threads have obvious positive or negative attitude unless they are some hot-button issues.



reddit The Reddit Monitor of Hot Threads 24 Top Thread Hot Thread

Reddit

aww
fastest,treat,catcher..

OldSchoolCool
nichelle,nichols,aka..

worldnews
china,opposes,call..

sports
soccer,rules

WritingPrompts
[wp],turn,match

videos
☺ wait

creepy
☹ found,photo,making..

Jokes
☺ yo,mama,fat

photoshopbattles
☺ psbattle,jackdaw,bird..

gifs
☺ hiding,sheets

todayilearned
til,jonathan,daniels..

AskReddit
great,thing,happen..

food
☺ daughter's,

wear,marks,t

DIY
☺ restored,da

Futurology
china,build,$50

TwoXChromosomes
☺ thrown,eating,d

Limitations and Future Works

The major limitations of our project for now is that we don’t have enough data analysis. If we had more time, we’d like to add user data to find some connections between user activity and threads. For example we could show which user contribute most to this thread, what their comments are, how often their activity in this thread or what time they posted their comments.