Date: March 13, 2016
To: Dr. Kwan-Liu Ma, Chris Bryan
From: James Tam 998233811, Christopher Chan 998005560
Subject: Final Report for analyzing and visualizing top Reddit posts
Dataset found at https://github.com/umbrae/reddit-top-2.5-million

### *How to Run Our Visualization*

Requirements: must have python installed

1) Start a server (In Windows we used "python -m http.server")
2) Open up index.html and make sure to have the data folder in the root.

### *Summary*

Our topic will be on what constitutes the top ranking posts on the web 2.0 phenom Reddit.com, a aggregation-- self-proclaimed "Front Page" of the Internet. We have chosen a dataset which contains the top 1,000 all-time posts from the top 2,500 subreddits, or 2.5 million posts in total. The dataset is structured by multiple .csv files named after their respective subreddit. A subreddit is a specific forum moderated for its own topic of discussion and/or agenda. An example of a subreddit is /r/3Dprinting, a community dedicated to topics related 3D printing. Users post content and other users have the ability to vote up or down the post (upvote and downvote), causing content to be listed further up the page or buried under the massive amount of content. This particular dataset does not contain comment data, which would increase the dataset's size exponentially. We will visualize the data using **d3.js** to show what topics are the most highly rated. Since the data contains information we can use to categorize, we will display the data with interactive elements such as limiting the vis to a specific subreddit. We will discuss more on our visualization further into this document.

### *Structure of Data*

For each .csv file that pertains to a subreddit, we are given: **created_utc, score, domain, id, title, author, ups(upvotes), downs(downvotes), num_comments, permalink, selftext, link_flair_text, over_18(can disregard since all our data is "safe for work"), thumbnail, subreddit_id, edited, link_flair_css_class, author_flair_css_class, is_self, name, url, distinguished**.

### *What and Why*

This dataset was polled in late August, 2013, so in a subreddit like /r/worldnews, we should see the NSA as a recent large topic of discussion, likely grabbing most of the attention due to the contents' votes. Reddit was founded in 2005 but has garnered many more users as the website matured, so we should expect to see mostly recent posts to be highly ranked. Although we may

not know the content of the post themselves, the title will contain key words that would describe the particular topic, e.g. “NSA” and “Snowden.” We can ask: what words or phrases allow a post to reach the top lists? Perhaps we may be able to see whether users bother to read the article that is posted. We are given a timestamp, so we can ask: When is the best time to post on Reddit to acquire the most upvotes? From this we may see a correlation between time and popular voting.

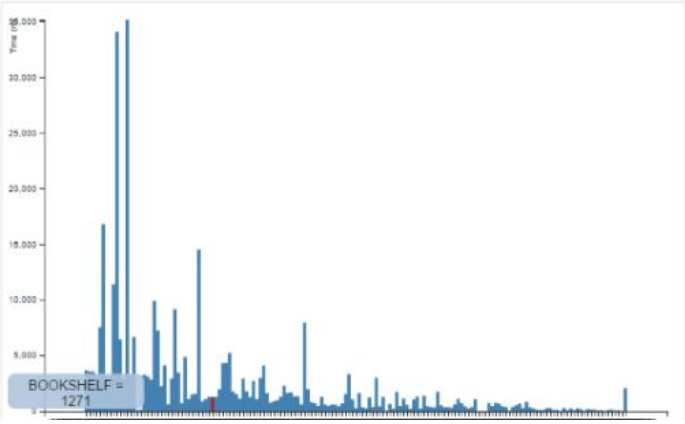
***Initial Proposal***

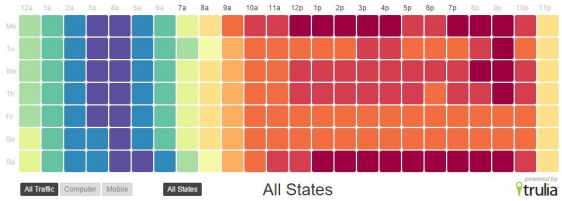
We want to design a way to see what sorts of topics are the most discussed, generally agreed upon, or are the “meta” at the time. We’ll be leveraging web programming, **jQuery,** and **d3.js** in order to display our visualizations. We will have several visualizations linked together in order for users to interact with the data. One visualization may be a overview bar chart to list the popular subreddits and its posts. We plan on creating an interface to select the present subreddits through a search feature and/or list, then showing a pie graph visualizing the percentage of the domains used, and also a heat map to visualize highly voted posts in correlation to time. We’ll allow the user to sift through these views in order to find answers and gain insights to the questions they initially had, or even create new questions after seeing our visualizations. We’ll utilize many visual channels, such as color (for the heat map), and size (for the graphs and charts such as a pie chart/bar graph).

*Bar chart*: utilize size(length) to show highest voted posts, most populated(in terms of content) subreddit.
*Heat map*: utilize color in order to display the upvote amount based on time
*Pie chart*: utilize color to differentiate the slices and size to display the percentages of upvotes based on either keyword or subreddit.

How our works will be divided:

James: Bar chart
Chris: Heat map
Both: Parsing and processing data, pie chart, and linking the interfaces together.
Example visualization to use:




***Process Report***

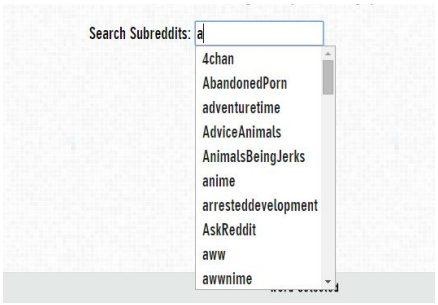
We decided to do a few things differently as we went through the process of parsing data for our project. In doing so, we focused on several key attributes from the original data: title, score, and date. Using these, we created more attributes such as an individual word, average score from a word, the TF-IDF score of a word, and the number of times a word has been mentioned.

To clarify, here is the breakdown of the non-obvious attributes:

*Average score of a word*: given a word's score is the score of the title it came from, take the sum of the word scores and divide by the number of time the word was mentioned in the subreddit.
*TF-IDF*: frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. (Wikipedia)

We designed our interface and usage pattern as the following:

Type in search bar to select a desired subreddit to see visualization for.


Then three visualizations-- a bubble cloud, a bar chart, and a heat map-- will display. An additional visualization-- a table-- will display when the user clicks either a bubble or a bar.

We ended up changing our word cloud for a bubble chart with increased interactivity. We disregarded the pie chart as what we envisioned it to do did not provide any useful or insightful information. Instead of having an overall view, the bar chart complements our focus in this visualization-- words. We believe words give the user enough information to achieve the goal of the visualization, to figure out what's popular amongst a subreddit. We let the heatmap serve as a overall indicator of a particular subreddit, showing when the most scores get allocated in a week.

We had challenges in parsing the dataset-- it was very large. We could not just use a single subreddit's csv file. We wrote many python scripts to offset the load the browser would process and as a result we have several folders with csv's of different attribute sets.

We learned techniques in dealing with large datasets (python scripting) and many design cue usage details (color scheme, consistency) to make the website beautiful and simple to use.

***Final Visualization Details***

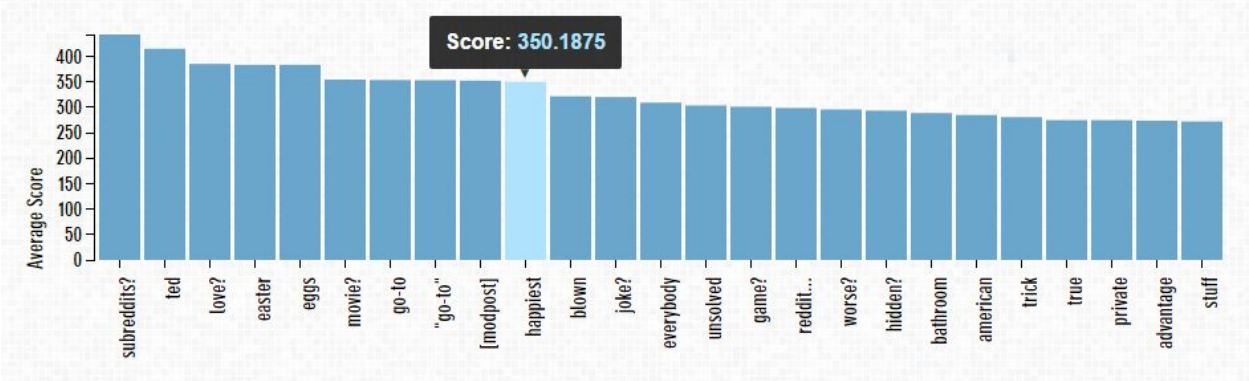
The topic of our visualization now focuses on the words pertaining to a subreddit.

*Bubble Cloud*:


*Under the word in a node/bubble is the TF-IDF score of the word. This shows how important the word is in relation to the rest of the sources. The higher the score, the larger the bubble. This

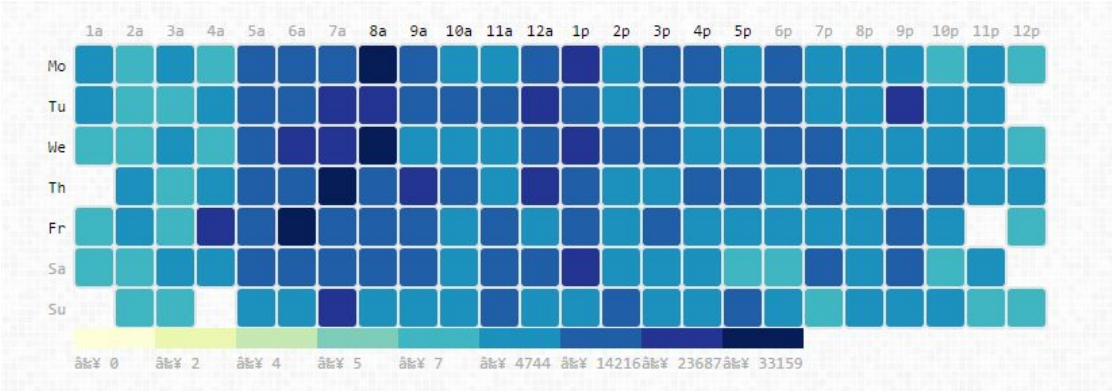
interface is interactive; the user can drag the bubbles around. Clicking the bubble will set the active word (See *Tables*).

*Bar Chart*:




*The chart shows the average score of the top 25 words of the particular subreddit. Hover over the bar to see the exact average score of the word in a tooltip. Clicking the bar will set the active word (See *Table*).

*Heat Map*:




*Shows the high scoring time slots (each square) of the subreddit. The darker the color the higher scoring. This visualization serves as a overall view for the particular subreddit.

*Tables*:

| Score | Title |
|---|---|
| 767 | Hayao Miyazaki upsets some fans and angers conservatives with his latest film |
| 702 | Hayao Miyazaki Fan Art |
| 613 | Like Miyazaki movies? I present the Ghibli Museum in Japan (as requested) |
| 612 | When I chose Hayao Miyazaki as an artist to research and draw a peice of artwork from him 2 1/2 d |
| 604 | New Hayao Miyazaki film, summer 2013! |
| 591 | My girlfriend knows I like Miyazaki.... here is my Anniversary present. |

*After selecting a word by either the bubble cloud or bar chart, this table will be filled with titles that mention the word and the post's score.

| Visited Subreddit | Word Selected |
|---|---|
| anime | miyazaki |
| AskReddit | superpower |
| zelda | skyrule |
| bestof | explains |
| bestof | badass |

*This is a history view which stacks the most recently visited word at the top.

***Work Allocation*:**

James: Scripting to parse dataset, Search Bar, Bar Chart, Styling and fitting elements onto web page, Write-Up

Chris: Scripting to parse dataset, Bubble Cloud. Heat Map, active word Table and History Table