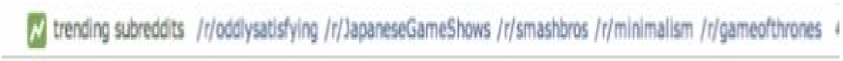
Predicting Trending Subreddits

By Nick Chao, Azam Abdulkadir, Kaval Patel

Introduction:

Reddit, a social media website dedicated to news and discussion is one of the most popular sites on the internet and has an open api. This makes it interesting and easy for data collection and analysis. Before explaining the problem and approach, one must understand the basics of reddit. The website is broken into categories or "subreddits" where a subreddit is dedicated to a specific topic. For example, the subreddit Politics is a subreddit where users post content relating to politics. Other users can see, comment, upvote(a feature that allows a user to promote the post), and downvote(a feature that allows a user to demote the post). However, a post is limited to the subreddit specified by the user who created the content. As of February 2016 there were 853,824 subreddits a user could post to.

One feature of reddit is a trending subreddit bar on the front page **(Figure 1)**. However, according to an article by techcrunch, for the trending feature shown here reddit selects "a half dozen or so non-default subreddits that have seen a particularly high amount of activity lately, and list them at the top of the front page". We believe this to be an inaccurate way to measure trending subreddits and a more accurate representation would include the default subreddits.



trending subreddits /r/oddlysatisfying /r/JapaneseGameShows /r/smashbros /r/minimalism /r/gameofthrones

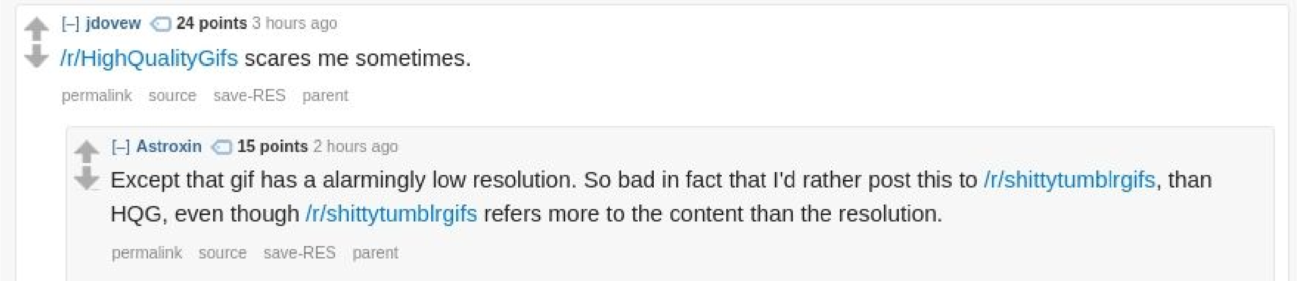
Figure 1

Getting an accurate representation of trending subreddits is important because it provides the user with subreddits that could have interesting content currently posted to them.

In addition, reddit provides a list of recently 'popular' subreddits which does include the default subreddits. However, it is not shown on the front page but we will use it as our gold standard/evaluation metric.

Approach:

Before providing a detailed overview of the approach, some description of reddit's community culture is needed. A common action taken by users is to comment on a post with a link(s) to other subreddit(s) as seen in **Figure 2**.



[–] jdovew 24 points 3 hours ago

/r/HighQualityGifs scares me sometimes.

permalink source save-RES parent

[–] Astroxin 15 points 2 hours ago

Except that gif has a alarmingly low resolution. So bad in fact that I'd rather post this to /r/shittytumblrgifs, than HQG, even though /r/shittytumblrgifs refers more to the content than the resolution.

permalink source save-RES parent

Figure 2

Therefore, we believed a good model of calculating trending subreddits to consist of running PageRank on a graph where each node is a subreddit and a directed edge (n1,n2) exists if in a post to subreddit n1 there exists a comment that links to subreddit n2. We then sort the resulting subreddits by Page Rank. The subreddits with the highest page rank will be considered to be trending.

Our approach begins by selecting 100 subreddits to collect data from. These 100 subreddits are selected by using reddit's api to retrieve all of the posts top posts across all of subreddits in the past 24 hours. The algorithm to select the subreddits is as follows:

```
For each post, P
    If the subreddit, S that P is posted too is new
        Add S to the list of subreddits
    If 100 subreddits have been collected,  Stop
```

```
Then, For each Subreddit, S
    Gather three of the top posts in the past 24 hours and all of their comments
```

Then for each comment collected, construct a link between node n1 to n2 if the comment is in a post to subreddit n1 and the comment contains a link to subreddit n2.

After examining each comment and creating the appropriate edge if necessary, run the Pagerank algorithm on the resulting graph. Sort the subreddits by Page Rank and the top posts will be the trending subreddits as described previously.

Data Collection:

We decided to use recent data and in order to do so we had to create our own dataset. To do so we had to interface with reddit's api. The replies from reddit's api consisted of JSON objects hence the dataset used is in JSON. The stats of the dataset are as follows:

- 132,678 comments
- 2500 comments linked to a subreddit

- Of the comments with links, 147 were unique(didn’t create a edge that already existed in the graph), did not point to itself and pointed to one of the 100 subreddits we were collecting from

Shown below is a image of the constructed graph

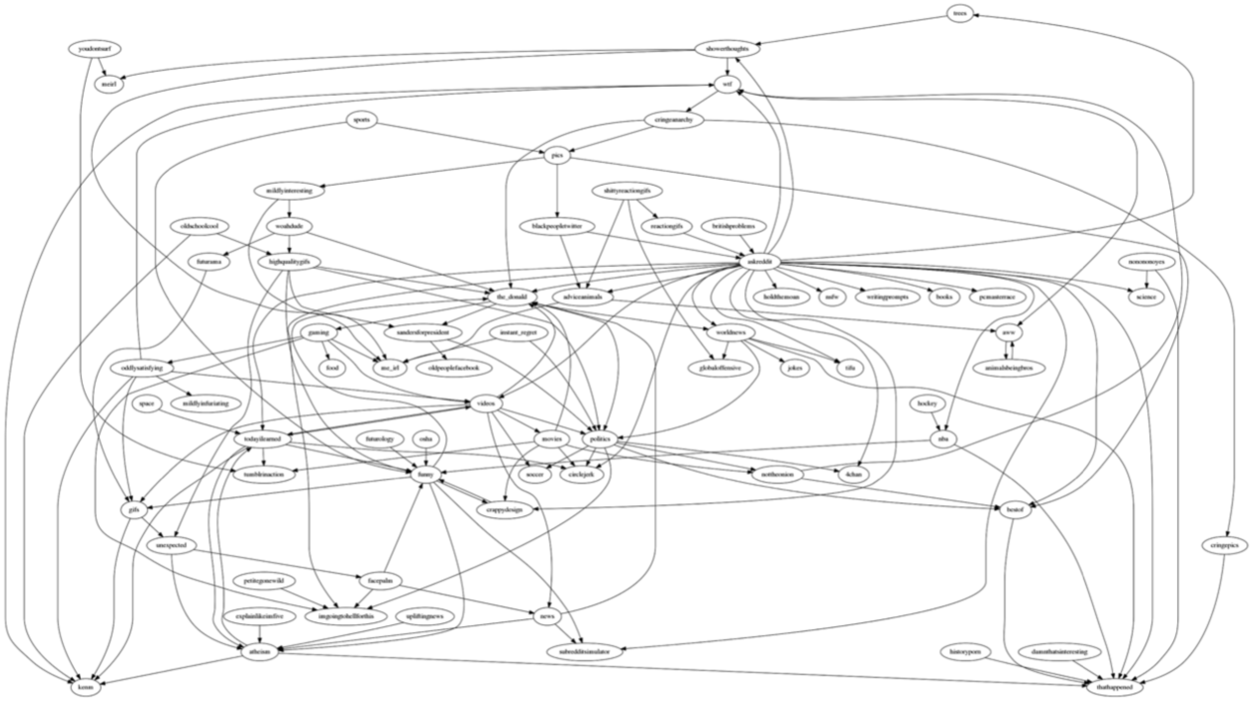


Figure 3

Evaluation Method:

In order to predict the trending subreddits, we used the scaled PageRank algorithm. The algorithm considers every time a subreddit is referenced in a comment from another subreddit and gives an appropriate score for each subreddit. To put it simply, the more references a subreddit has corresponds to a higher PageRank score.

PageRank Initialization:

-Each node represents a subreddit and a directed edge from Node A to Node B represents A referencing B in a comment.

-The score for each node in the graph is initialized to a PageRank score of 1/N, where N is the number of nodes in the graph.

-We chose a K such that after running the PageRank Update Rule K times, the graph reaches equilibrium. This means that no matter how many more times the Update Rule is

applied, the PageRank score for all nodes remains the same. For this algorithm, K = 10,000.
-A scaling factor S is chosen in order to account for graph properties that lead to inaccurate scores. For this algorithm, S = 0.8
-After every iteration of the PageRank Update rule, the sum of every node's PageRank should be equal to 1.

-The PageRank Update Rule:
-Each node will give its current PageRank/Number of outgoing edges to each node that is the recipient of the outgoing link.
-One complete iteration consists of the above rule running for every node once.
-After each iteration, the PageRank score for every node will be multiplied by the scaling factor S, in this case 0.8, and the result of that will be increased by (1 - S)/N.
-The above steps will happen K times.

Once the PageRank algorithm was completed, we had a list of size N corresponding to each subreddit's PageRank score. In order to make our results more accurate, we used a scaling factor as mentioned above. This was necessary because of issues with some of the graph's characteristics. One of the major issues was a cycle found between two nodes. This led to the score being trapped between the two nodes instead of being further passed around in the graph. There were nodes that did not have any out-links, which also led to score being trapped.

Experimental Results:

After refining our total dataset and creating the graph, the scaled PageRank algorithm was used to generate a PageRank score for each of the subreddits. Once each subreddit was assigned a PageRank score, the subreddits were ordered in descending order. Once the ordering was done, based on the hypothesis, the subreddit with the highest score was ranked as the most active/trending subreddit and the subreddit with the lowest score was ranked as the least active/trending. Once all one hundred subreddits were ranked, we were able to compare the calculated PageRank list with the list provided by Reddit. The list provided by *Reddit* is updated daily so the list used for comparison is from the twenty-four hour time period when the dataset was collected.

Based on the comparison of the top twenty of both lists, the results were not very promising. Of the twenty subreddits on both lists, only two appeared in the top twenty for both lists. However, neither of these subreddits were ranked correctly on the PageRank list. After comparing the remaining subreddits, we constructed a graph to show the standard deviation of the subreddits that intersect both lists. The standard deviation was computed with the difference in position between our calculated position and actual position. In order to create the standard deviation graph, five buckets were created and each bucket contained eight sequential subreddits that intersect both position lists. Although the standard deviation is relatively high, averages around twelve, the subreddits ranked 17-24, bucket 3, had the lowest standard deviation. This shows that these subreddits were the closest to their positioning on the active/trending subreddit list acquired from *Reddit*.

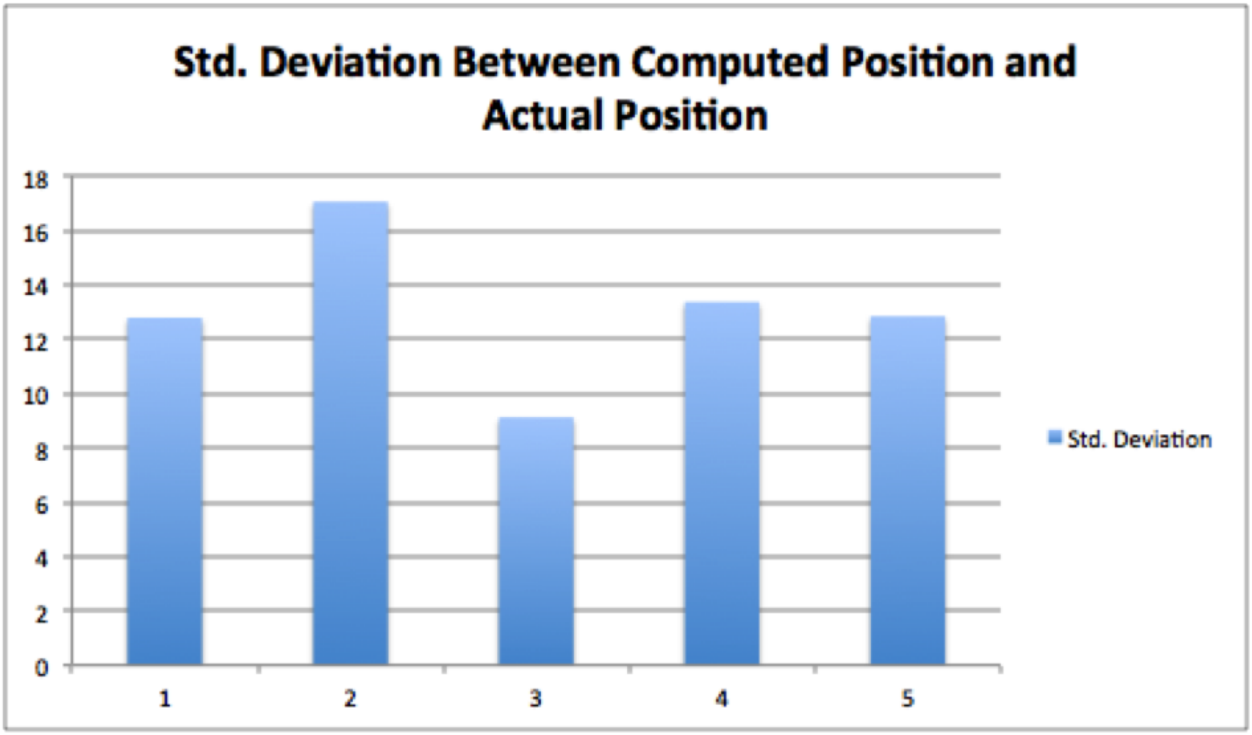


Figure 4

Although the results produced were inaccurate when compared to the golden standard there are several factors that contributed to the results that were produced. Reddit is a large community with nearly nine hundred of thousand subreddits and many more posts and comments. Comments on Reddit are almost over two billion in number and the dataset used only accounted for .0076% (Fisher). And of the .0076% even less was used because of most of the comments did not meet the criteria. When looking at the sheer size of Reddit, collecting data for a twenty-four hour period is not nearly sufficient to accurately predict trending/active subreddits.

In addition, the Reddit API requires many calls to be made to gather the required data and as a result we had to limit our dataset in order for the runtime to be within a reasonable timeframe. As we reviewed the graph in Figure 3 we created based on in and out links, we noticed a lot of places where the PageRank score would get to but had no way to move because the node did not have any out links. This is a major reason why we had skewed results. We also noticed while scouring through the comments that users generally link to smaller more obscure subreddits rather than large, known subreddits. Because users do not tend to link the larger subreddits this also led to a skew in our data from the golden standard because they had less in links therefore they produced a lower PageRank score. These factors led to our results being widely inaccurate when compared to the golden standard we were comparing it with.

References:

Fisher, Jonathan. "The Numbers Show That Reddit Has a Lot to Celebrate on Its 10th Birthday." Business Insider. Business Insider, Inc, 23 June 2015. Web. 09 May 2016.

Kumparak, Greg. "Reddit Starts Listing Trending Subreddits To Get More Users Into Its Smaller Communities." TechCrunch. N.p., 10 Apr. 2014. Web. 09 May 2016.

Reddit. N.p., n.d. Web. https://www.reddit.com/subreddits