Discovering structure in Reddit Comment Networks

SHRIYAK SRIDHAR

Summary

In this project, I look at different Reddit communities and explore the structure that emerge as members of these communities interact with each other. The evolution of these communities across time is described and various discussions on these forums are studied to group them based on content. The discussions that emerge from these communities are summarized with a topic model that is a statistical model using a collection of words. The communities are clustered together using Mixture Models.

Introduction

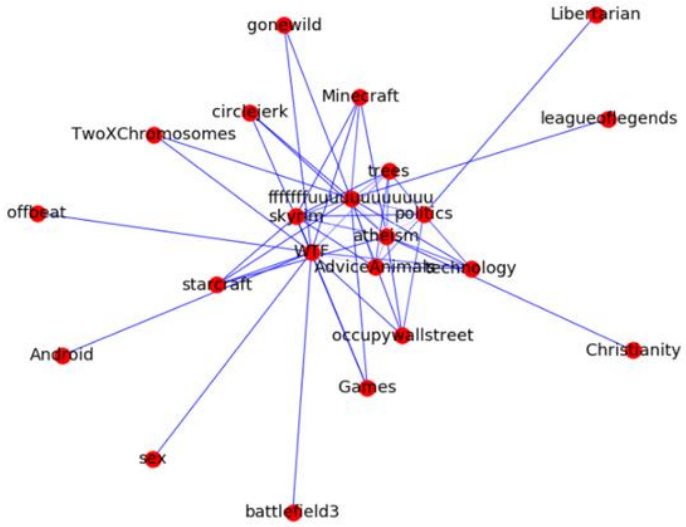
Reddit is the fourth most popular website in the world [1] and with an average time of 16 minutes per day spent by a user on the site, it is the most addictive of the top twenty websites. Since 2009, the site has been growing exponentially with over two million comments posted on the site every day. A repository of every single Reddit comment posted is available at [2] and this project describes the structure of the comment networks that emerge. Reddit contains individual subreddits, niche forums with a dedicated topic. Users often comment on several subreddits and similar topics of discussion are present on the site. With over ten thousand new subreddits added every month, finding an appropriate subreddit is challenging.

Using the comment network, I draw a map of reddit and describe what the different parts of the site are discussing. Attempts to cluster Reddit communities have been performed by Bustain and Golbeck [3] but they focus solely on user interactions. Weninger et al [4] look at the discussions on the site to understand if it can be useful for search engines and Singer et al [5] describe the evolution of Reddit submissions with time. None of these methods look at the content of the comments for their analysis and my project reinforces the findings from interactions by looking at the comment contents and various topics that arise by exploring the content.

Methodology

Interactions between different sections of a member graph is often studied in social media but comment networks are a relatively unstudied aspect. With reddit, since there is no formal social circle for every user, we analyze the interactions of different users on the various subreddits that are created. To analyze the interactions between two forums, we first look at the set of distinct users each of them contain. Once we have their user base, we look at how many members are common to both the bases. This forms the foundation for the measure of interactions between the two communities. In graph networks, this is called the pairwise intergroup distance.

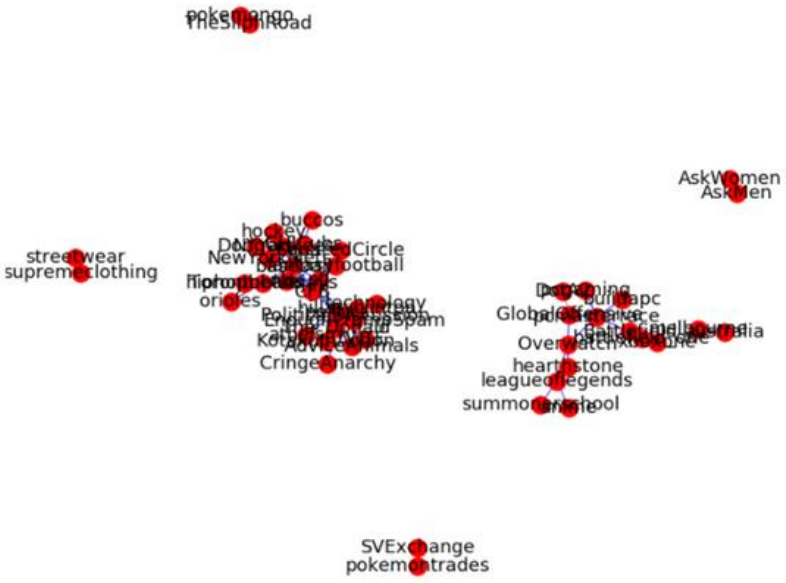
To obtain the final score we can use to map the distances between these groups, we also study the frequency of comments made by these set of users on their respective communities. This gives us the pairwise distances that we use to plot the communities on the map. Using a Spring layout, we obtain the proximity of these different communities and allow some groups to form naturally.



1. 2009 Reddit comment network

In the 2009 network, we can see the presence of one central cluster with densely connected nodes. There is no single theme to these nodes but they are connected radially to several other solitary nodes. While this was still the early stages of the social networking site, we can still see the popular topics of discussion containing:

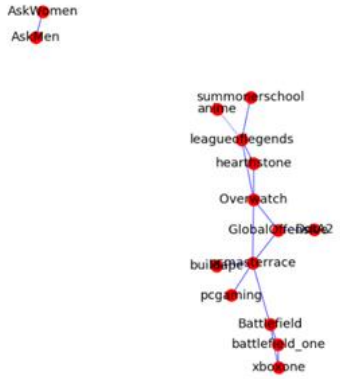
- Gaming: Minecraft, skyrim, leagueoflegends, Games, starcraft, battlefield3
- Technology: technology, Android
- Religion: atheism, Christianity
- Politics: politics, Libertarian, occupywallstreet



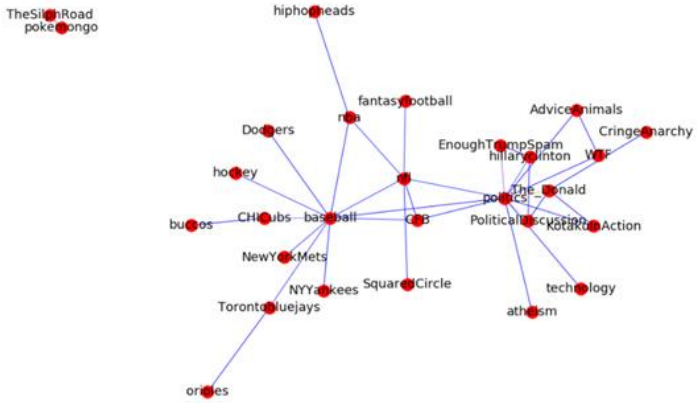
2. 2016 Reddit comment network

We then study the 2016 network to get a glimpse of site when it is more mature. This network is very different from the 2009 network with the emergence of several new cliques. Cliques are informal groups that are present in a social network with high intra group interactions. When the cliques are examined manually, we can see the emergence of several topics around which these cliques are centered. The outer nodes formed are also completely isolated. Unlike the 2009 network, they are no longer connected to the central clusters and they are self-sustained. In addition to the earlier topics, we also see the growth of memes, fashion and seeking advice.

A look at the central nodes shows the topics clearly and important roles played the major sports communities in forming the bridge between the sports and politics clusters. We then look at the content of the comments within these communities to automatically identify the abstract topics. This is a method called topic modelling and helps us find the groups of words that best represents the information for these groups. These groups have a different set of infrequent words that best represent them along with some common words across the site.



3. Gaming network - 2016



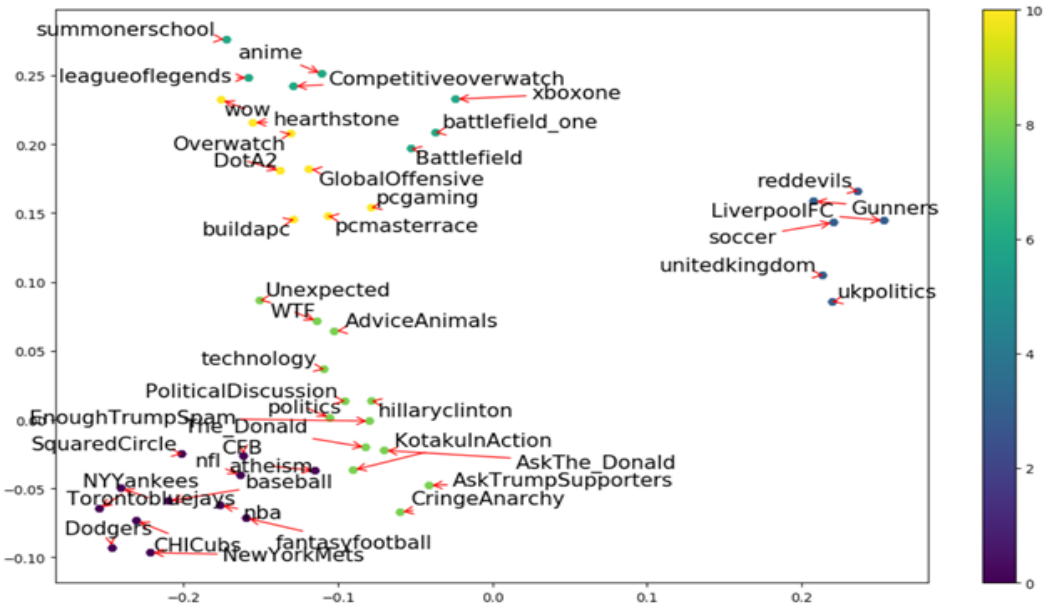
4. Sports and Politics network - 2016

Before we can analyze the comments, we first have to perform some text preprocessing. The comments are all broken up into words, non alphanumeric characters are removed and all words converted to lowercase. We then remove a set of stop words. There are some common words that appear very often across all topics and add very little value in discriminating the comments from each other. These are excluded from the vocabulary and are called stop words. In addition to the stop words provided by [6], we add a custom set of stop words specific to reddit.

We assume that a comment is a document and we aim to compute the frequency of words within the document and across the set of all documents, ie the corpus. We use Tf-Idf (term frequency, inverse document frequency) to determine how important a word is to the document and we use this statistics to then compute the set of words for each topic. We use Latent Direchlet Allocation to obtain our topic and this method assumes that every document is a mixture of multiple topics in some proportion. The collection of words forming the topic can have different probabilistics presence in the comments belonging to a topic. We study the ideal number of topics using the information gain while being aggressive with the topics we discard.

Finally, since we see a large overlap in the topics provided by topic modelling and topics of the subreddits communities, we attempt to cluster these subreddits based on their user base and content. We only perform this on the 2016 dataset and using the elbow method, we are able to identify the

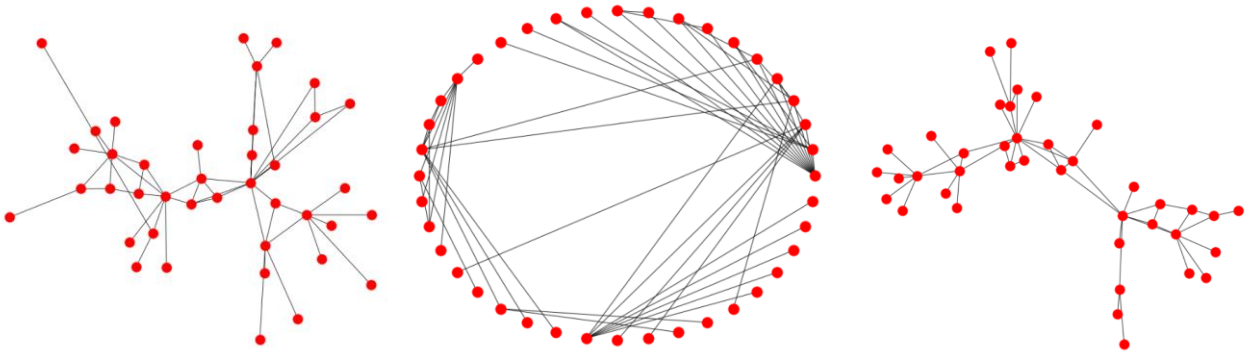
ideal number of clusters for the Gaussian Mixture Model. We can see a clear emergence of some clusters and the topics that they are focussed on. Reddit is a predominantly American site so the emergence of the European cluster with discussion on their local politics, sports teams and regions is especially interesting. This shows the nature of Reddit and how the cliques are formed in the different parts of the site.



5. Clusters of Reddit communities - 2016

Numerical Studies

Since the project involved creating a network, finding a topic model and clustering the nodes; different studies were done to tune the parameters for each of these steps.

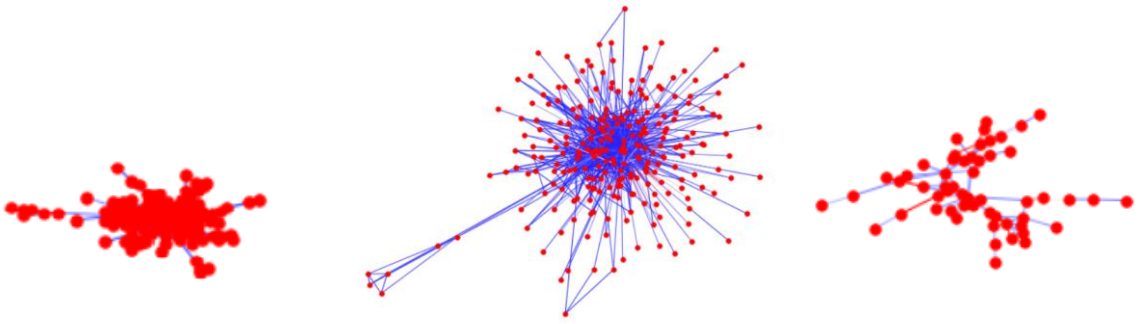


i.Kamada Kawai ii.Circular iii.Spring

6. Networks plotted with different layouts

When creating the network for Reddit, we have to analyze the structure and choose between circular, spectral and spring. As the network in 2009 consists of a few central nodes with several outer, the Spectral layout provides the best effect and for the 2016 network, we use the Spring layout to enable the clusters with fewer interactions to be placed further away from the central nodes. These different layouts are depicted above. In addition to the layouts, we also have to finetune the level of

interactions to observe. I decided to use a weighted graph with the weights denoting the pairwise level of interaction. As we reduce the threshold, more nodes appear and form densely connected network. If we instead increase the threshold very high, we only observe the large communities and get a very fragmented network. Various values for the parameters along with the networks are shown below.



i.Threshold = 20 ii.Threshold = 1 iii.Threshold = 40

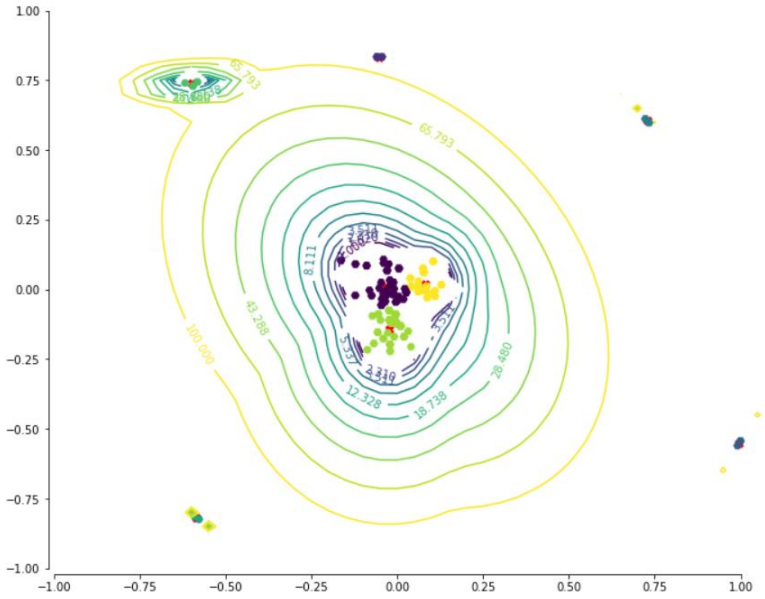
7. 2016 network using different thresholds

The initial topic model consisted several stop words that are not recorded in popular sources. I added the frequently appearing words across the topics and this helped the discriminating power of the model. There are no fixed methods to identify the the exact number of topics and I used the information gain to obtain the final collection of words that forms the topic.

Topic 1 - *politics*		Topic 2-*gaming*		Topic 3-*sports*		Topic 4-*other*	
people	0.005711	good	0.005944	games	0.003602	lol	0.008317
trump	0.00455	play	0.005336	post	0.003126	f*ck	0.005966
make	0.002577	game	0.005128	play	0.002927	god	0.004031
money	0.002179	team	0.004651	time	0.002909	awesome	0.003942
clinton	0.001757	wow	0.003218	yeah	0.002899	man	0.003672
country	0.001668	kill	0.002631	work	0.002608	damn	0.003447
vote	0.001668	damage	0.002561	years	0.002349	guy	0.003331

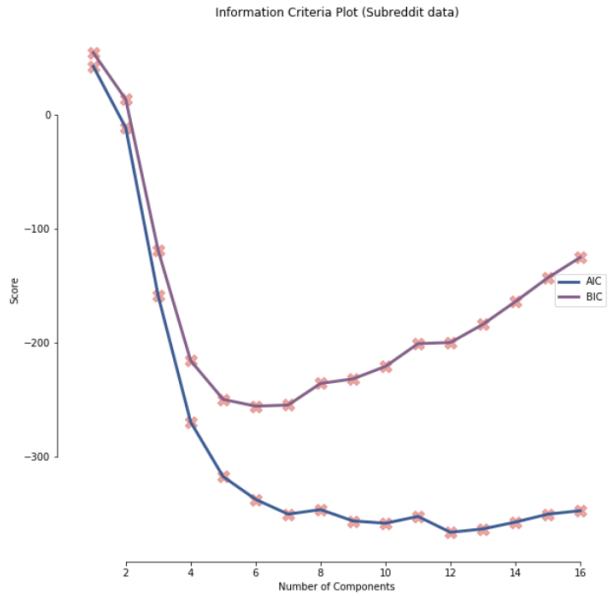
8. Topic Model: collection of words

To perform clustering, I initially used K-means but it quickly became evident that the radial boundaries of the method were not going to capture the nature of the Reddit network. I then switched over to Gaussian Mixture models. With mixture models, the data is sampled with the underlying assumption that it is generated from a set of finite Gaussian process with either diagonal, spherical (same as K-means), tied or full covariance matrix.



9. Contour plot for GMM

Since the full covariance structure is the most descriptive, we choose that and to decide on the number of clusters present in the data, we use the elbow method plotting both AIC and BIC. Bayesian Information Criteria has information gain that decays with size and Akeike Information Criteria places a harsher penalty on the models with higher complexity. Since GMMs work on the Expectation Maximization method, BIC seems to work well for our given dataset.



10. Elbow plot for GMM

Conclusion

For this project, we can conclude that as the site matured, it has become more diverse. It has included topics of discussion for many niche areas and even conventional topics that may not be welcome in regular social media are present on Reddit. This however creates clear cliques that appear, each with a different in topic within that community. Given a particular comment, we can classify which cluster a comment belongs using the topic models we generate. We can suggest other subreddits in this cluster at a minimum distance from the users post for the user to cross post to as well.

From the studies done during the project, it is evident that there is potential for recommender system. The content based collaboration method can be used to suggest different subreddits that have content similar to the users interest and current comments. Based on his interactions with other users (both positive and negative), we can also suggest communities that he can be a part of or avoid. During the project I also tried predicting comment score but with the vast variety in comments this task proved very difficult. A richer dataset including the context of comment, its neighbors and location in the comment thread are all necessary.

References

[1] Alexa site rankings claiming reddit is 4th
[2] data source
[3] https://dl.acm.org/citation.cfm?id=2579231
[4] https://ieeexplore.ieee.org/abstract/document/6785761/authors
[5] https://dl.acm.org/citation.cfm?id=2576943
[6] https://www.ranks.nl/stopwords