Reddit User Network Analyses

ISOM 673 Social Network Analytics

Yutao (Eric) Gu
Emory University
Atlanta, GA, United States
eric.gu@emory.edu

Prajwal Kuchangi
Emory University
Atlanta, GA, United States
pkuchan@emory.edu

Prarthana Neotia
Emory University
Atlanta, GA, United States
prar.neotia@emory.edu

Jingwen (Kivi) Zuo
Emory University
Atlanta, GA, United States
jingwen.zuo@emory.edu

December 10, 2018

Introduction

Reddit, founded by Alexis Kerry Ohanian, Steve Huffman - college roommates at the University of Virginia, is the product of entrepreneurial spirit sparking creativity through sophisticated tools found in a technical developer's toolkit. Reddit has been and aims to continue being a virtual bulletin board with billions of users. These users are privy to ongoing discussion about relevant topics of the past, present, and imminent future. The forum like environment curated on this platform allows for open discussion and free circulation of opinions and insights. The quality of the content generated and diffused here, and the wit and humor that accompany it, are what make analyzing the existing network interesting. Combining the data found on user comment frequencies on Subreddits, the skills developed in this course have allowed our team to extrapolate the trends and patterns of user behaviors within this network that might not be visible to the naked eyes.

The original data includes 1,576 users and 10,641 unique subreddits. Each row of 14 million observations represents a relation between a user and a commented subreddit with a timestamp. The data contains the commenting relationships from 2006 to 2016, so there is a clear opportunity here to analyze the interest and activity patterns of the platform's readers and commenters.

Analysis Approach

Understanding the wealth of data found on this topic and shaping it so that it contains the fundamental pillars of our venture is the first phase, after which the analysis is delved into using three different lenses.

1. Frequency of Interaction - User Activeness
2. Frequency of Interaction - Popularity of Subreddits
3. Diversity of Interaction - User-Subreddit Network

Analysis Goal

The goal is to combine the insights developed from the above 3 paths of exploration, and leverage the same to build three models that

a. Predicts user activeness
b. Predicts the likelihood of commenting relationships
c. Effectively recommends Subreddits of interest to inactive users

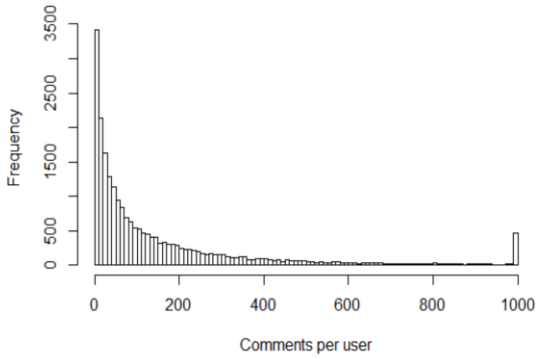
Analysis Process

To accomplish the above, data manipulation using RStudio packages learnt in this class is done, and visuals are built to graphically display findings like network densities and user activity trends.

Parallelly, a neural net model is developed and trained on time - stamped data that is composed of unique sequences of user interactions. Users are then matched on behavior pattern similarity, and the predictions and recommendations are generated. This is based on the assumption that similar past user activity is indicative of behavior emulation going forward.

User Activeness

First, variance of user activity across users is examined. A histogram of user activity is plotted for the month of December 2016.



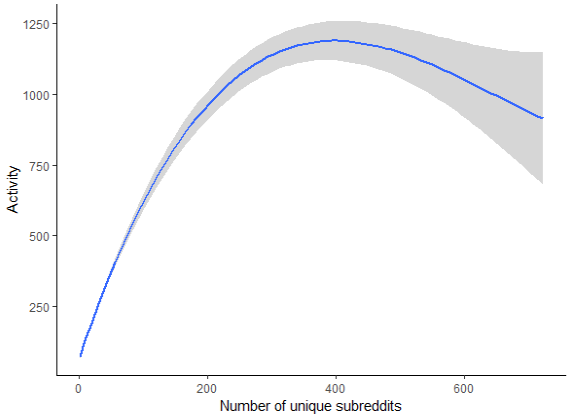
The histogram suggests an exponentially decaying distribution. A large number of users (~30%) make less than a comment per day on average. The median number of comments made per month per user is 70.

Percentile distribution of activity per user								
10%	20%	30%	40%	50%	60%	70%	80%	90%
6	15	28	45	70	107	160	246	430
Inactive	Less active		Active		Very active			

Based on this distribution, users are classified into Inactive, Less active, Active and Very Active groups.

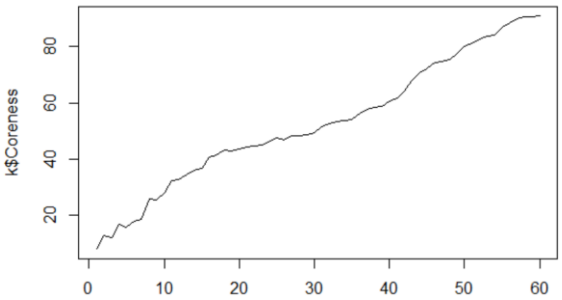
The segmentation is also based on judgement. A large number of users are prone to "lurking". Users may be on the social media site for a long time before making a comment. Further data is needed to ascertain that, but any user who makes almost a comment per day should be considered active. Since there are many users who are more active, only the users who fall below the 10th percentile are classified as inactive.

More importantly, the behavior of networks of active and inactive users is explored. Also, whether active users tend to follow more Subreddits is studied.



A clear pattern seems to emerge. Activity increases as the number of Subreddits increase but only up to a point. This may be due to excessive noise on the user's feed, if the user is subscribed to too many Subreddits.

Next, whether coreness is important to maintain high user activity in a Subreddit is explored. To analyze this, a popular Subreddit "AskTrumpSupporters" is selected. The Subreddit becomes largely popular towards the end of 2016. Two users are assumed to be connected together if they comment on the Subreddit on the same day. If user remains inactive for more than 7 days, the tie is considered to have decayed and is deleted. Coreness is calculated across 60 days from November to December 2016. It is confirmed that coreness of the user network is extremely important to maintain activity in the Subreddit.



Next, the general pattern of active users is dug deeper into. While analyzing each type of user, it is unveiled that most users are intermittently active. However, some users are perennially active (make more than 50 comments each month in 2016) and some users are perennially inactive (make less than 10 comments every month in 2016). The characteristics that allow for a user to be perennially active are further explored. Do they add new Subreddits to their feed on a regular basis? Are they active only in few Subreddits while remaining in a small group? Also, what makes users inactive? Are they not following new and relevant Subreddits?

The behavior of a perennially inactive user- "ancientfutureguy" was analysed and contrasted with the behavior of a perennially active user "Goonboo".

Below is a list of each Subreddit that the inactive user has commented from January to June 2016.

Month					
1	2	3	4	5	6
skateboarding	todayilearned	battlefront	reallifedoodles	AskReddit	gaming
aww	reallifedoodles	TameImpala	Unexpected		Games
	facepalm	SandersForPresident	macdemarco		indieheads
	movies	WTF			JusticePorn
	AskReddit				
	macdemarco				

It was identified that the perennially inactive user tends to comment on very few numbers of Subreddits but also that they tend to move onto new Subreddits each month while almost never going back to the previous Subreddits that they commented on.

Next, this is contrasted with the activity of a perennially active user. It was found that the perennially active user tends to remain active in his core Subreddits and also branches out to new Subreddits (Figure on next page).

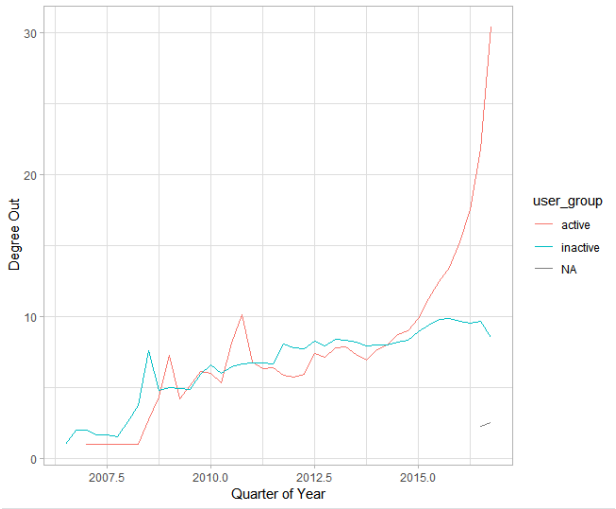
To summarize, user activity depends on the number of Subreddits that the user subscribes to and also the coreness of the Subreddit. A perennially active user tends to remain active in his core Subreddits while also branching out to new Subreddits. Thus, it is vital for our recommender algorithm to suggest relevant and active Subreddits to ensure high and maintain levels of user activity.

Month					
1	2	3	4	5	6
AdviceAnimals	AdviceAnimals	AdviceAnimals	AdviceAnimals	AdviceAnimals	AdviceAnimals
AskReddit	AskReddit	AskReddit	AskReddit	AskReddit	AskReddit
awesome	aww	daschund	aww	blackdesertonline	blackdesertonline
Diablo	creepy	Edmonton	blackdesertonline	CrappyDesign	darksouls3
ffxiv	Diablo	ffxiv	creepy	darksouls3	Edmonton
funny	Edmonton	funny	darksouls3	DkS3Builds	funny
Games	ffxiv	gaming	Edmonton	Edmonton	gifs
gifs	funny	keto	ffxiv	ffxiv	pics
history	Games	mildlyinteresting	Games	funny	technology
JusticePorn	gifs	MonsterHunter	gaming	Games	todayilearned
LifeProTips	islam	technology	IdiotsFightingThings	gaming	videos
MonsterHunter	MonsterHunter	videos	keto	popping	watchpeopledie
pics	pics	watchpeopledie	pics	technology	worldnews
todayilearned	science	woahdude	science	todayilearned	WTF
UnexpectedThugLife	space	worldnews	SWORDS	videos	
videos	technology		technology	watchpeopledie	
watchpeopledie	todayilearned		todayilearned	woahdude	
worldnews	videos		videos	worldnews	
	watchpeopledie		watchpeopledie		
	worldnews		worldnews		
	WTF		WTF		

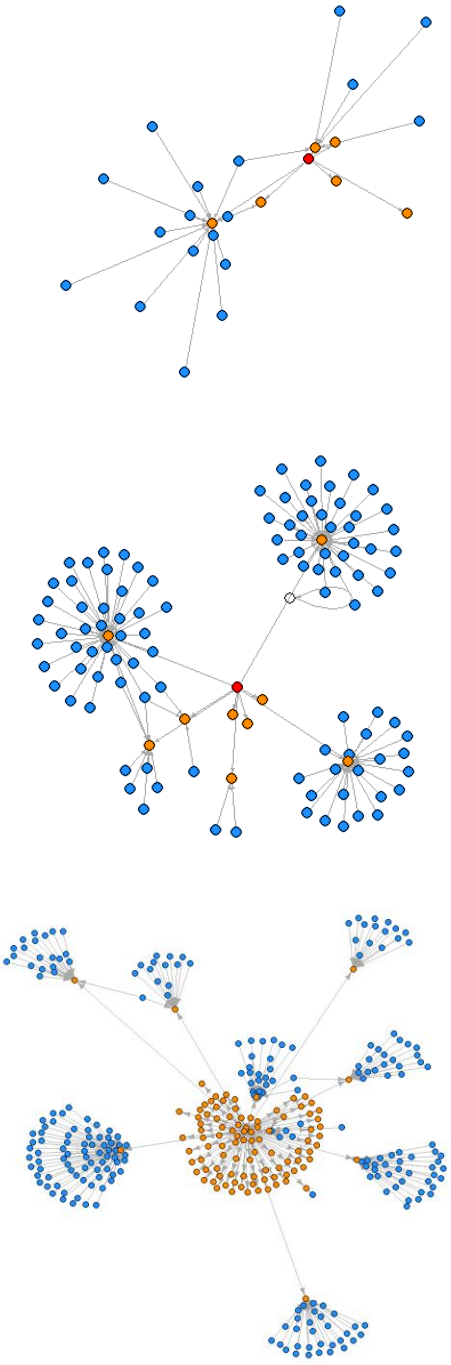
User-Subreddit Network

A user-Subreddit network is built from the edgelist in the original dataset and weighted on the frequency of visits within the time period. The network has directed relationships between users and Subreddits, as it is based on users' comments on Subreddits. Due to the nature of rapid changes in the digital world, three months are implemented as the decay time for Subreddits that users engage with. Therefore, the timestamp is converted to quarters from 2006 to 2016. User-Subreddit networks are constructed using edgelist within the same quarter. Since our analyses are focused on user behaviors, the assessment of user-Subreddit network will be analyzed from the perspective of users.

The network is assessed by its centrality measures. Because of the directed network between users and Subreddits, degree-in represents how many users have commented on the Subreddits, and degree-out represents how many Subreddits a user has commented on. Furthermore, the average degree-out centrality has a positive relationship with the time, as users tend to spread out their interests in more Subreddits over time. However, the graph also shows that there is a significant difference between active and inactive users. The average degree-out was very similar for both groups before 2015, then the centrality for active users increases dramatically after but remains at around 10 Subreddits for inactive users. Therefore, the active users are more of generalists since they tend to spread out their interests over time, but inactive users are specialists who would stay with relatively small number of Subreddits.

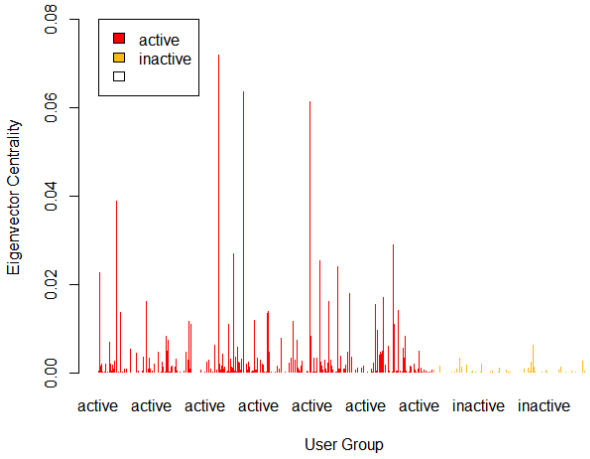


Closeness means the reachability of users and Subreddits within the network. In more details, each Subreddit will form a cluster, and two Subreddit clusters can be connected to another Subreddit cluster through a user who has commented on both Subreddits. Therefore, a node with high closeness will have more exposure to new Subreddits that its neighbors are connected to, especially when the neighbors have high degree centrality. Three examples are shown on the right with increasing value of closeness centrality. Colors of red, orange and blue represent the target node, Subreddits the target node comments on, and other users who comment on the same Subreddits. The examples show that the more users a node's Subreddits have, the higher the node's closeness centrality will be especially when the node itself has many Subreddits.



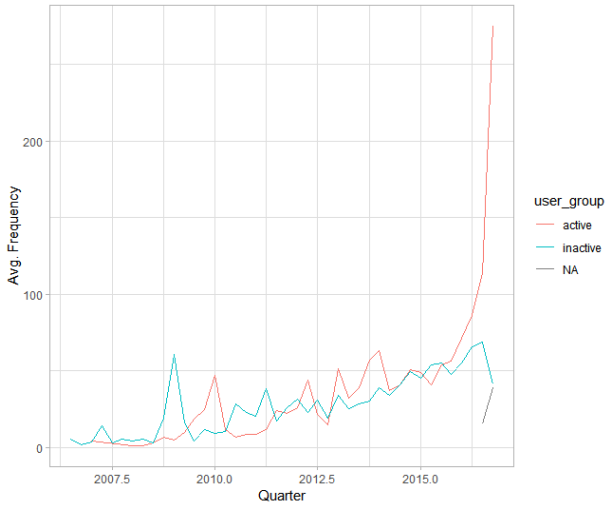
The eigenvector represents the sum of the centrality values of the user or Subreddit that it is connected to. There is an significant difference between the active and inactive user groups. The mean eigenvector is 0.0031 for active users, and 0.00025for

inactive users. It means that actives users are more centered in the network.



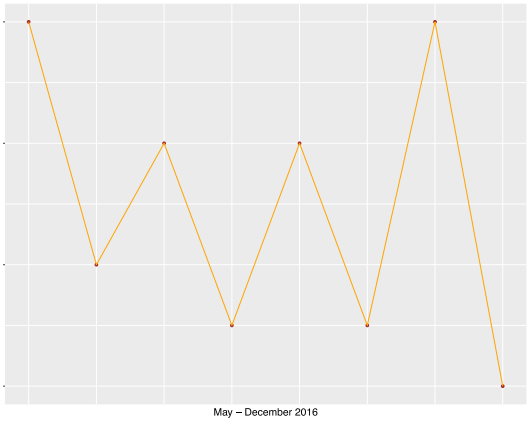
Another perspective is the frequency of a user's comments on the same Subreddits, and an average of frequency per quarter is calculated for both user groups. Inactive users have a much lower mean of 50 visits to the same Subreddits and active users' mean is at around 200. At the same time, the trend of average frequency shows a very similar result to the increase of commented Subreddits over time. The user interaction increased dramatically after the second quarter of 2015.

	mean	s.d.
Active users	199.47	263.51
Inactive users	49.98	91.05



Both line charts above show an important change in the second quarter of 2015, which was the time when Reddit had a new CEO. When Reddit had its interim CEO – Ellen Pao, Reddit shut down five Subreddits that was cited to be related to harassment. Users had argued that the accusations restricted their free expressions. There was also a protest conducted by the users to set some popular Subreddits to private followed with a petition to remove the former CEO. During this time, Ellen Pao was the CEO from the 4th quarter of 2014 to the 2nd quarter of 2015, and she eventually stepped down replaced by Reddit's co-founder Steve Huffman. After that, both average Subreddits and average commenting frequency per user went up.

An example of the finding in centrality is an analysis conducted on Brexit-related Subreddits. The plot below illustrates user activity in May – December, 2016.



The trend observed here implies hyper activity right after the declaration of Brexit (June, 2016), and a further surge later that year. It could be interesting to explore the reasons behind this surge in a following phase of this project analysis. For instance, was this surge powered by the fact that Donald Trump became president elect in November of the same year?

The most active user on the Brexit multireddit is 'prodmerc', and their user behavior conveys the same message as that delineated in the User-Subreddit Network. 'prodmerc' maintains a high and progressively increasing closeness and outdegree measure through the quarters of 2016, which implies that not only are they active in the Brexit-related subreddits but also branches out a fair amount. This complements the takeaways about user behavior remaining in the same neighborhood of activeness and diversity touched upon later in this section.

*For similar findings in other major multireddits, please refer to the Appendix.

Predictions

A model to predict user activity was built using centrality measures and age of the user as the independent variables. Age of the user on the network is taken as the difference in months between the user's first and most recent comments in the observation window. It was found that both centrality and age of the user were significant. Increased centrality increases activity whereas increased age of the user on the network decreases activity. This can be explained by the fact that users tend to explore the website more and comment more frequently when they are new on the network.

```
Call:
glm.nb(formula = activity ~ age_diff + degree + eigenv, data = regression_merge,
    init.theta = 0.9448289846, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5473  -0.9961  -0.3079   0.3348   5.0437

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.6285955  0.0112924  498.44   <2e-16 ***
age_diff    -0.0333541  0.0004446  -75.01   <2e-16 ***
degree       0.0123586  0.0002316   53.37   <2e-16 ***
eigenv      28.3230571  1.0886981   26.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An ERGM model is built to predict that if network familiarity influences a user choosing to comment on a particular Subreddit. The outcome variable will be the directed relationship from a user to a Subreddit. To limit computational strain, we only consider relationships from October to December of year 2016. For the same reason, we subset the users into two group to be included in the model, including who commented on between 40-60 and 140-200 unique Subreddits. There are three variables in the model: edges, triangle, and user groups.

The result of the model is shown below. All three variables have significant relations with the network, as potential network familiarity affects users commenting on certain subreddits.

```
==========================
Summary of model fit
==========================

Formula:   N ~ edges + triangle + nodematch("usergroup")

Iterations:  19 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC %  z value Pr(>|z|)
edges               -7.249807   0.003684      3 -1967.68   <1e-04 ***
triangle            -2.907001   0.214028      1   -13.58   <1e-04 ***
nodematch.usergroup -3.242147   0.173234      4   -18.71   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 144527162  on 104254310  degrees of freedom
 Residual Deviance:   1187401  on 104254307  degrees of freedom

AIC: 1187407    BIC: 1187456    (Smaller is better.)
```

Recommender System

To improve user inactivity by helping them find communities they might be interested in, a recommender system is built based on users' commenting history.

The most popular method in recommender system, which is collaborative filtering, is not used here because it is already widely implemented in industry and some of Reddit's existing features are built on top of that. More importantly, as what has been proven in previous analyses, most active users are "generalist" who have interests in a variety of topics. Therefore, the focus of the recommender system is set to help inactive users develop new interests instead of just recommending topics similar to what they already like. To do so, a machine learning method based on sequential prediction is used to predict the "next interesting Subreddit" that users might like according to their historical activity. More specifically, users discover new Subreddits over time as they interact with the Reddit ecosystem. However, because of network diffusion effect, their new discoveries are affected by the Subreddit communities they have previously been interacting with. Therefore, a sequence of Subreddit discovery can be generated for each existing user and a model can be trained to "learn" the patterns. When the model "sees" a sequence of Subreddit discovery patterns of a user that matches a similar sequence of interactions of other users, it recommends their Subreddits that the current user has yet to discover.

The algorithm used for the recommender system is a recurrent neural network. It has internal memory so it can memorize information from the input they receive. Unlike a feed-forward neural network, information cycles through a loop within the hidden layer of a recurrent neural network to enable it to process a sequence of inputs. It is widely used for sequential prediction for time series, speech, text, audio, video, etc.

Thirty thousand rows of time-stamped data are sampled out from the raw data for model building. Using a larger subset of data would theoretically improve model performance but would also cost a lot more computational power. About 5,000 unique Subreddits is extracted from the subset of data and this is good enough for the model to capture the underlying pattern behind user activity.

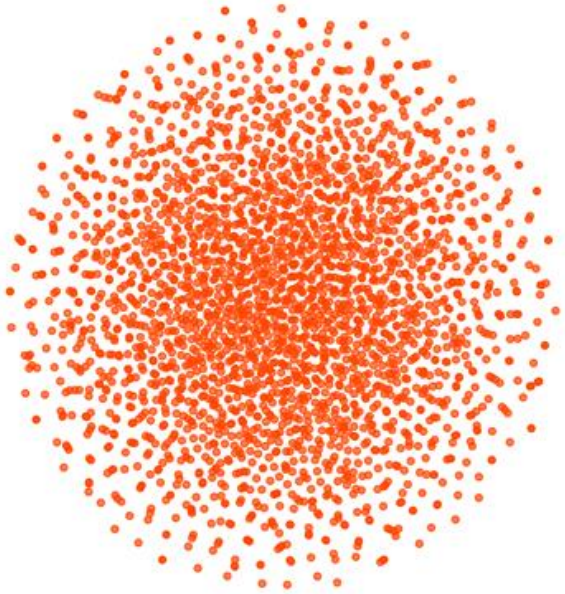
Every extracted Subreddit is indexed to save computational power and a list of Subreddit interaction history of each user is generated. After data cleaning and putting all the sequences together in a data frame, a new dataset is formed to train the model.

The type of model that is built can be classified as a “many to one” sequential model. Given the past interaction, we want to predict which Subreddit the user is most likely to interact in the next step. Using the previous example, given a sequence of [18, 7, 9, 24,15], the following expanding window prediction is expected:
[18] -> 7
[18, 7] -> 9
[18, 7, 9] -> 24
[18, 7, 9, 24] -> 15
The array on the left stores the historical interaction and the index on the right is the prediction of a Subreddit that a user might interact next. Weight, which is measured by the frequency of appearance of a sequence, is also an important factor to help predict the next value when some sequences are very similar. For example, if [18, 7, 9, 24, 15] have more appearances than [18, 7, 9, 24, 72], then our model should recommend 15 when it sees a sequence of [18, 7, 9, 24]. To train such model, a well-implemented python framework called “Spotlight” is used to turn the original sequence into a sequential interaction object.

The data is then split into a training set and a test set with an 80:20 ratio. However, one problem with a normal recurrent neural network model is they have short memory and have difficulty in memorizing things far away in a sequence. In other words, they tend to make prediction based on the most recent inputs instead of taking the whole sequence into account. In order to solve this problem, a single layer of LSTM (Long short-term memory) is used as our hidden layer because it can extend the model’s memory. And to make the learning process more reliable, a low learning rate is applied even though it would take more time on optimization. The performance of our model is measured by a cross-entropy loss function.

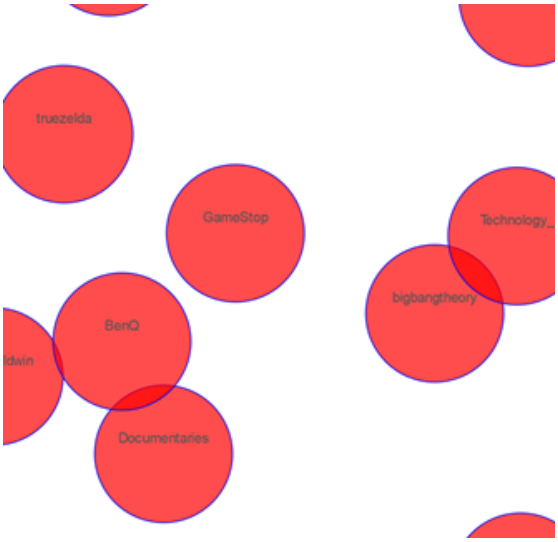
Our training loss is 3.67 and our validation loss is 3.98. The validation loss is a bit larger than our training loss, indicating that our model may over-fit a little, but not by too much.

During the model training process, a high dimension embedding space is learned so that the model captures the similarity between different Subreddits by evaluating their corresponding weights. With dimensionality reduction techniques, the embedding layer can be visualized in a two-dimension space. Figure below is the visualization of the embedding space generated.



Each dot represents a unique Subreddit and the shorter the Euclidean distance between two points, the closer the relationship they have. In other words, they are more frequently brought up together in a sequence during the training process. Even

though it is not technically a graph of network, its structure is very similar to a "Core-Peripheral" network. This suggests that in fact many subreddits are very interconnected to others even though they might be different in nature. This is an illustration of how users can develop new interest through exploring those subreddits in the "core". This visualization can also be used to help us validate the model by inspection. For example, figure below is a "close view" of some points that are very close to each other in the space.



By examining the names of those Subreddits, it suggests that the pattern our model captures is reasonable because those Subreddits are relevant but also have their differences in niche. For example, gamers who are following "truezelda" and "Gamestop" might extend their interest to "technology", technology and game related TV shows such as "bigbangtheory", and electronics brand "BenQ".

The trained model can be deployed by feeding the discovery sequence of inactive users into the model as input and the predictive results will be the recommendations to those users.

Conclusion

To recap, users are classified into active and inactive categories by the analyses on user activeness, and user behavior analysis is conducted on those who comprise both groups. Metrics like degree, closeness and Eigenvector centrality are computed for each user and Subreddit, and the active users are observed to connect to more subreddits with high centrality measures.

These classifications and patterns are then used to help build predictive models on user activeness and relations of user commenting on particular subreddits. Lastly, user discovery sequences of relatively inactive users will be plugged in to the recurrent neural networks. The project is concluded with the generation of Subreddit recommendations for users through this model, thereby enhancing user activity and participation.

For future possible extension, the pattern of user behaviors discovered in this project, as well as the recommender system, can be applicable to websites that function in similar way such as Quora and Pinterest because those websites also have sub-communities. Related studies can be done to compare user behavior of communities on different websites to uncover some universal patterns to gain a deeper understanding of internet ecosystem. In addition, if implementing the dataset of users' timestamped comments on the same reddit, users could be classified into discussing or commenting type and their behavior will be further investigated and embedded in to more comprehensive prediction models and the recommender system.

References

[1] McGregor, Jena (July 6, 2015). "More than 200k people have signed a petition calling for Reddit's Ellen Pao to step down". The Washington Post. Retrieved July 7, 2015.

[2] Griffin, Andrew (2015-06-11). "Reddit bans communities including 'Fat People Hate' as users say anti-harassment policies could be 'beginning of the end'". *The Independent*.

[3] Mike, Isaac (July 10, 2015). "Ellen Pao Is Stepping Down as Reddit's Chief". *The New York Times*. The New York Times Company.

[4] Donges, Niklas (Feb 25, 2018), "Recurrent Neural Networks and LSTM" *Towards Data Science*

[5] Huang, Steeve (Feb 16, 2018), "Introduction to Recommender System (Neural Network Approach)" *Towards Data Science*

[6] MacLean, Cole (2016), "A Recurrent Neural Network Based Subreddit Recommendation System"

[7] Akash Kandpal (Jan 1, 2016), "Recurrent Neural Network - LSTM"

[8] Shengxian Wan (2015), "Next Basket Recommendation with Neural Networks"

[9] Donkers, Tim (August 27-31,2017), "Sequential User-based Recurrent Neural Network Recommendations"

[10] Liu, David (Jun 2, 2016), "A Recurrent Neural Network Based Recommendation System"

Appendix

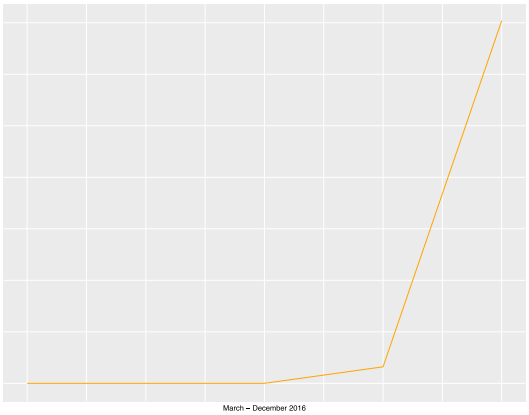
Multireddit analysis*

This section analyzes other examples of user activeness being reflective of diversity. The timelines chosen for the above analyses vary depending on the relevance of these topics, and includes appropriate windows of before and after the 'event' occurred.

1. Donald Trump

a. "Impeach_Trump"

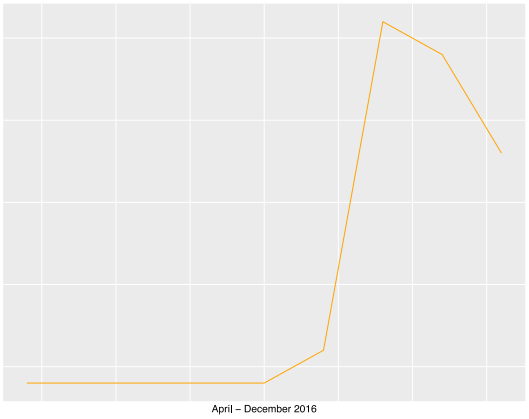
The plot below illustrates user activity in March – December, 2016.



The surge in activity observed towards the end of the specified timeline makes functional sense, since user sentiment for impeachment can only get strong *after* the election (after November).

b. "NeverTrump"

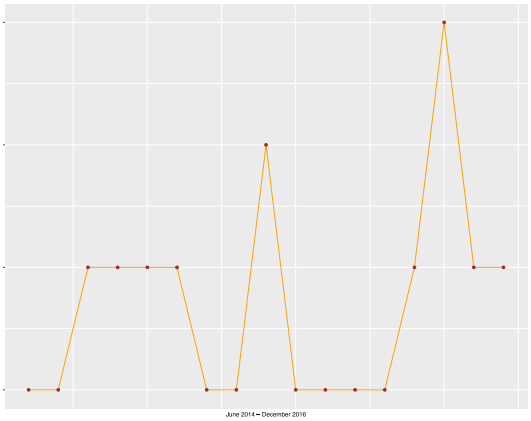
The plot below illustrates user activity in April – December, 2016.



The surge in activity observed here is much earlier, which would mean users were commenting more frequently on this multireddit as they approached the November election. It can also be observed that there is a conspicuous drop after around December, 2016, which would imply a loss in this thread group's attractiveness (after the election's conclusion).

2. FIFA

The plot below illustrates user activity in June 2014 – December 2016.

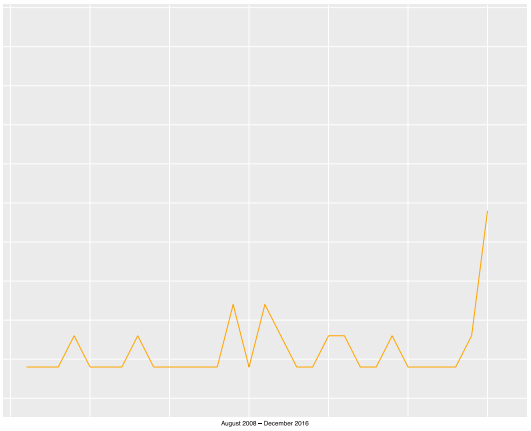


The trend observed here is interesting because the last two FIFA World Cups were in 2014 and then 2018. However, there is a relative surge in activity that can be observed around 2015. It is unlikely that excitement for the upcoming World Cup spiked in the intermission between the two

tournaments, but this surge could potentially be attributed to the crowning of Reddit's new CEO and original founder – Steve Huffman. Overall user activity has been observed to spike upward after this change in executive leadership, and it is possible that the same surge is being reflected in the plot above.

3. Barack Obama

The plot below illustrates user activity in August 2008 – December 2016.



The trend here implies that user activity was somewhat low in the beginning and spiked up halfway through Obama's term. Interestingly, the trend shot upward toward the end of his term and continued slightly after.

It could be interesting to explore whether Obama being succeeded by Trump had much to do with continued activity on Obama's multireddit. For instance, the distress caused amongst anti – Trump supporters has fueled much hue and cry on numerous platforms, potentially including Reddit.

Similarities between Multireddits

It is also interesting to explore the similarities between multireddits. The multireddits of Obama and Donald Trump, for example, share 126 users in common. This overlap makes sense from a partisan perspective, because an Obama supporter is likely not ecstatic about Trump's presidency, and might be more keen on participating in Democratic threads and those resounding dismay about the 2016 election results.