Principles of Social Media and Data Mining (CIS – 600)

How Reddit influencers establish a network of linked subreddits

TEAM MEMBERS

Amrith Sai Bharath Kumar	821231347
Midhuneshwar Kandasamy Kanivalavan	351564889
Rajaraajeshwaran Ramasamy Kannan	925189708

TABLE OF CONTENTS

ABSTRACT

The intricacies of how Reddit influencers create interconnected networks across numerous subreddits are the difficulty of this mission. Users of the well-known on-line network website Reddit are capable of starting and participating in a variety of topic-particular subreddits. Influencers are people who have a huge following and have an effect on in those agencies, and they regularly have a prime impact on conversations and the creation of material. The Study makes use of network assessment gear and facts mining methods to look at how influencers assemble networks. The examine targets to map the go-subreddit interactions of important influencers inside unique subreddits. Influence is measured with the aid of looking at metrics like person interaction, submit popularity, and comment threads.

Furthermore, techniques for herbal language processing are used to recognize the context and content of conversations. The effects are presupposed to shed moderate on community improvement, data alternate, and cooperation dispositions in the Reddit surroundings. Gaining knowledge of those network dynamics can assist one understand how on-line communities are organized and the way influencers have an effect on the verbal exchange in well known.

INTRODUCTION

Fig 1. Reddit Introduction

Reddit is an influential and adaptable online community that lets in customers to explore a wide range of topics, take part in debates, and alternate fabric. Reddit, was founded in 2005 through Steve Huffman and Alexis Ohanian, has collected hundreds of thousands of active customers and is now many of the maximum famous websites within the international circuit. The platform is split up into subreddits that are groups devoted to specific subjects, hobbies, or debates. Individuals, generally referred to as Redditors have the option to subscribe to these subreddits in an effort to interact in discussions, offer content, or simply maintain an eye out. Every subreddit runs on its very own and promotes a experience of community around specialized pursuits, ranging from technological know-how and era to pop culture, hobbies, and more.

Because of its distinctive layout, Reddit prioritizes user-generated content and uses a voting system. The community has the ability to upvote and downvote posts and comments, so affecting their exposure. The most well-liked and pertinent conversations rise to the top thanks to this democratic approach to content curation, giving users a dynamic and constantly changing home page. The platform's depth is higher via its numerous person base, which gives a dialogue board for human beings to engage with numerous viewpoints, backgrounds, and testimonies.

Reddit is a powerful social and informative media useful resource because it has evolved into a center for understanding alternate, aid structures, and cultural trends. Reddit promotes lively, welcoming communities, but it also has problems with content moderation—keeping free speech in check while halting the spread of dangerous or false material. Reddit has changed its features and standards throughout time to address these problems, demonstrating its dedication to fostering a welcoming and interesting online community.

METHODOLOGY

A. Algorithm

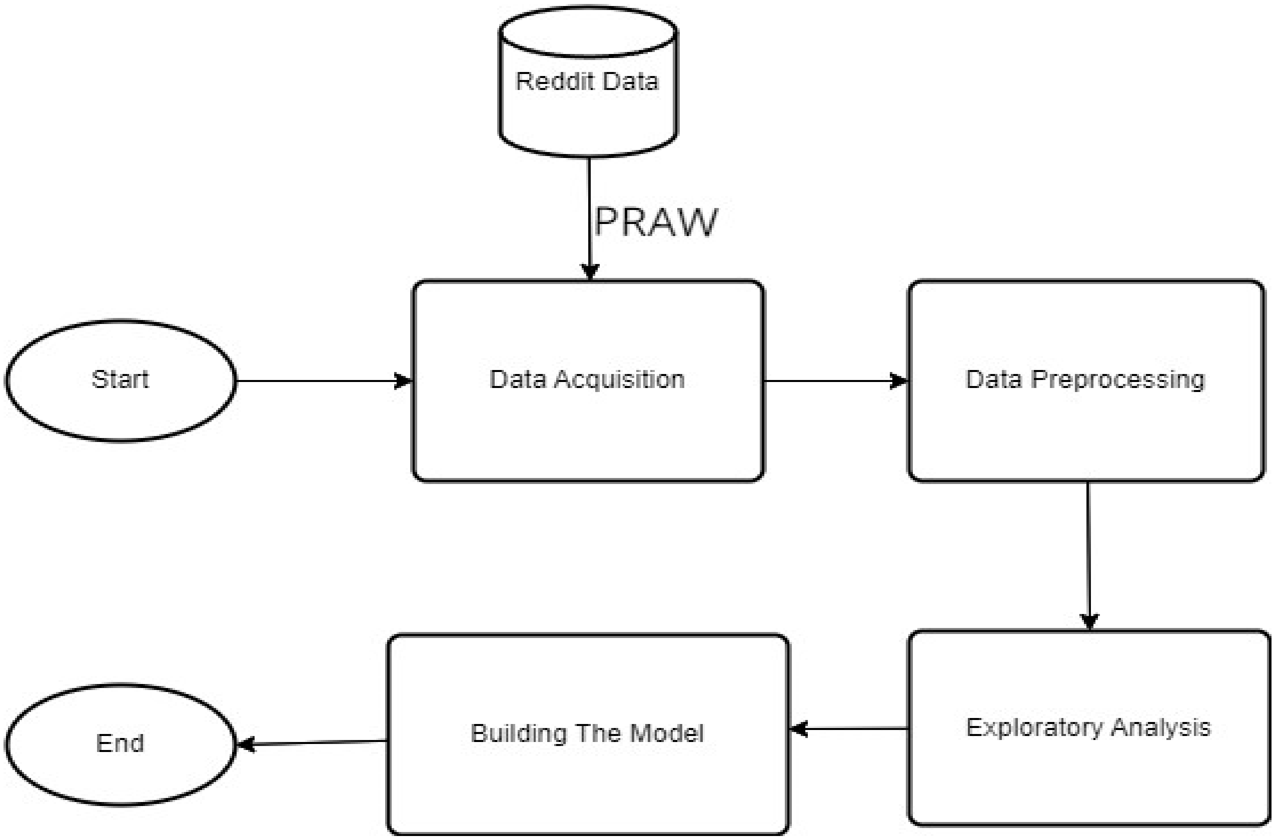


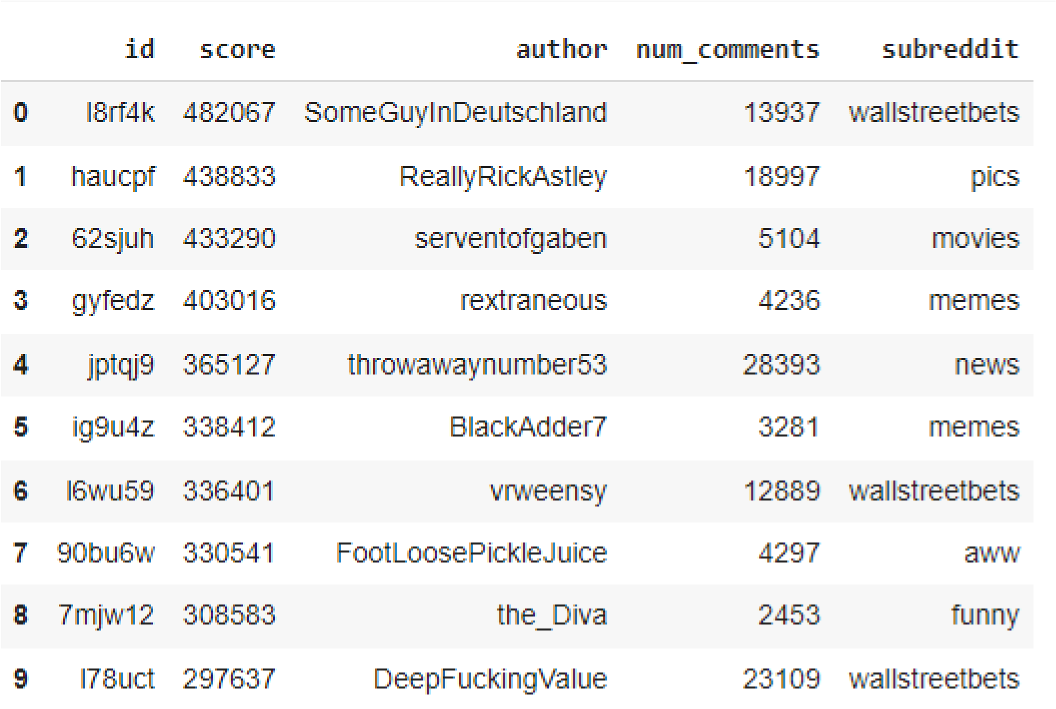
Fig 2. Flow Diagram

From the above flow diagram, first we start with the data acquisition to collect the data from reddit using praw which is dedicated reddit API. Next, we move to the data preprocessing where we clean the data and remove irrelevant data from the data sets. After we get the relevant data, we do exploratory analysis to find out the influential user. Finally we build the product model in form of a network graph using NetworkX python library that contains nodes and edges to represent users and subreddits respectively.

B. Data Acquisition

The first step before we can draw any conclusions from a social media platform is to collect the data. To suit our demands for obtaining user data from active subreddits, we decided to use PRAW - The Python Reddit API Wrapper[1]. To use Reddit's API, you must have a Reddit account. To access Reddit's API as a script application, we needed the Client ID and Client Secret. The instructions on Reddit's First Steps Guide to Creating Keys leads you through the procedure.

We wanted a Read-Only instance of Reddit utilizing PRAW for our data scraping needs. We also needed a user agent to construct a read-only Reddit instance, which is a unique identifier that Reddit uses to determine the source of network requests.

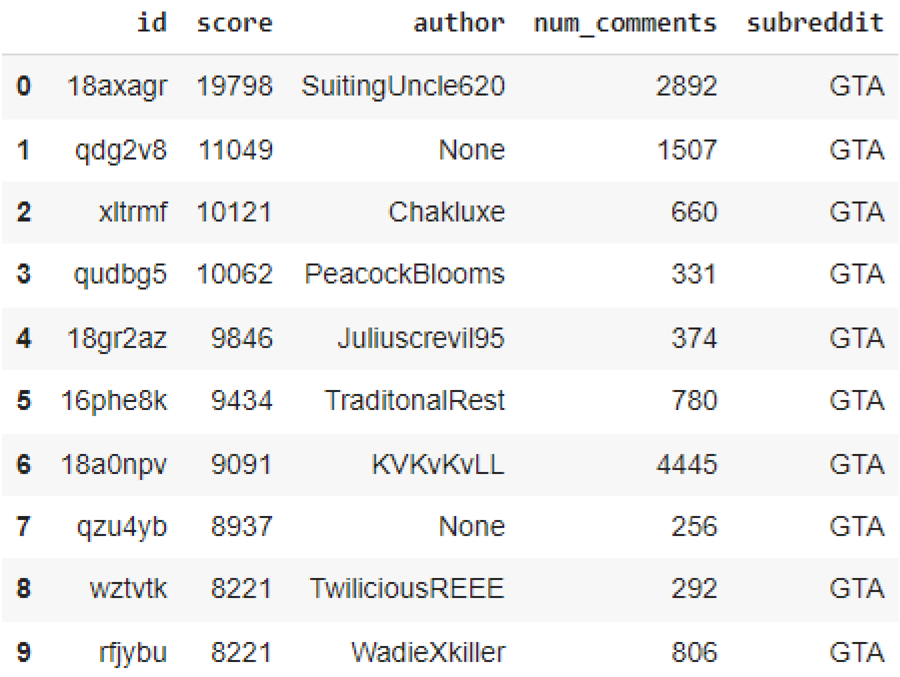


	id	score	author	num_comments	subreddit
0	l8rf4k	482067	SomeGuyInDeutschland	13937	wallstreetbets
1	haucpf	438833	ReallyRickAstley	18997	pics
2	62sjuh	433290	serventofgaben	5104	movies
3	gyfedz	403016	rextraneous	4236	memes
4	jptqj9	365127	throwawaynumber53	28393	news
5	ig9u4z	338412	BlackAdder7	3281	memes
6	l6wu59	336401	vrweensy	12889	wallstreetbets
7	90bu6w	330541	FootLoosePickleJuice	4297	aww
8	7mjw12	308583	the_Diva	2453	funny
9	l78uct	297637	DeepFuckingValue	23109	wallstreetbets

Fig 3. Top 10 Users in r/all

The above figure 3 displays the comprehensive analysis of the Reddit community encompassing the entirety of the platform's content, we have identified the top 10 subreddits featured in the "r/all" section. The associated metrics include the number of comments per subreddit and an examination of the foremost influencers within each community. The influence of users is quantified through their karma scores, which represent the cumulative sum of both upvotes and downvotes received.

Following that, we created a function called get_posts to retrieve a specific number of posts on a subreddit. get_posts accepts two parameters, sub_name and n, and returns a Pandas data frame containing a list of the top n posts on a subreddit. We keep track of each post's id, score (the sum of upvotes and downvotes), number of comments, and the subreddit in which it was created. For our initial investigation, we chose r/GTA, a subreddit dedicated to GTA content[3].



	id	score	author	num_comments	subreddit
0	18axagr	19798	SuitingUncle620	2892	GTA
1	qdg2v8	11049	None	1507	GTA
2	xltrmf	10121	Chakluxe	660	GTA
3	qudbg5	10062	PeacockBlooms	331	GTA
4	18gr2az	9846	Juliuscrevil95	374	GTA
5	16phe8k	9434	TraditonalRest	780	GTA
6	18a0npv	9091	KVKvKvLL	4445	GTA
7	qzu4yb	8937	None	256	GTA
8	wztvtk	8221	TwiliciousREEE	292	GTA
9	rfjybu	8221	WadieXkiller	806	GTA

Fig 4. Top 10 Users in r/GTA

D. Data Pre-processing

Meticulous data validation is undertaken to ensure the integrity and reliability of our dataset. Specifically, any instances of null or deleted individuals who appears as the top contributors in subreddits have been excluded from our evaluation This essential step turned into taken to set up the integrity of our data and mitigate the potential impact of erroneous or incomplete information.

Moreover, within a given subreddit, an in-depth examination was conducted to identify recurring contributors who consistently secured top positions in the "all-time" rankings. Subsequently, the resultant data frame was subjected to a comprehensive sanitization process, meticulously addressing, and rectifying any null values present. This rigorous approach towards data cleansing ensures that our analyses and conclusions are founded on a robust and error-free dataset, thereby enhancing the overall validity of our research findings[2].

E. Exploratory Analysis

1.Exploratory Analysis-Part 1

Analyzing the top 500 posts on r/GTA over time reveals some intriguing observations regarding Reddit posts. The scatter plot in figure 5 indicates that the number of comments does not necessarily rise with higher net score articles[4].

The quantity of comments is relatively constant between post scores of 20,000 and 60,000. Our first notion was that the quantity of comments and score would be closely proportional. On a bigger scale, we saw a similar scatter figure for r/all.

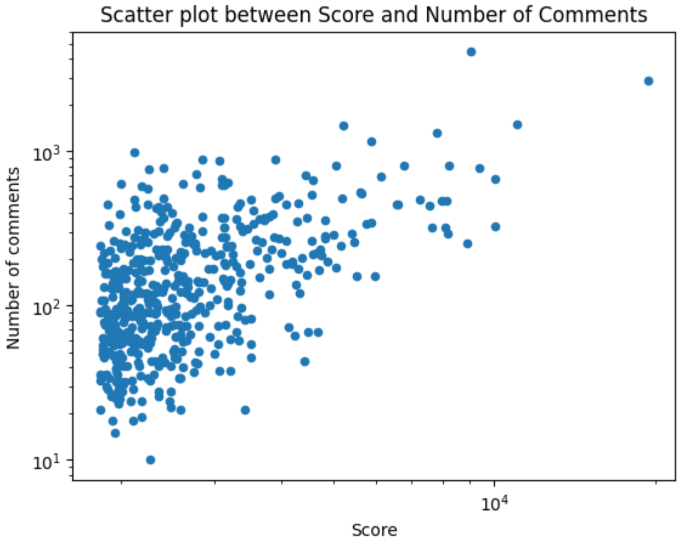


Fig 5. Scatter plot of number of comments and net score of top 500 posts

2. Exploratory Analysis-Part 2

The next step was to extract the top ten posts per user for a large dataset, which would serve as the foundation for detailed bar graphs illustrating post distribution on r/GTA. In this situation, a "Influencer" cut-off was used, which required individuals to have two or more top posts to be considered[6]. However, the subjective character of this criterion became clear, with factors such as subreddit size and submission frequency having an impact.

It was highlighted in Figure 2, a graphic depiction of post distribution, how hard it is to identify influencers in the r/GTA setting. These results demand more analysis of the variables influencing user involvement and content creation in the vibrant Reddit community. They bring up more general issues regarding how online communities are evolving and the need to modify methods for recognizing key contributors[5].

Understanding the subtleties of user-generated content is becoming more and more important as digital platforms expand in order to comprehend these vibrant online communities as a whole. As the research develops, it not only offers details on the state of user dynamics on r/GTA at the moment, but it also advances a progressive perspective on the dynamic landscape of online interactions and community influence.

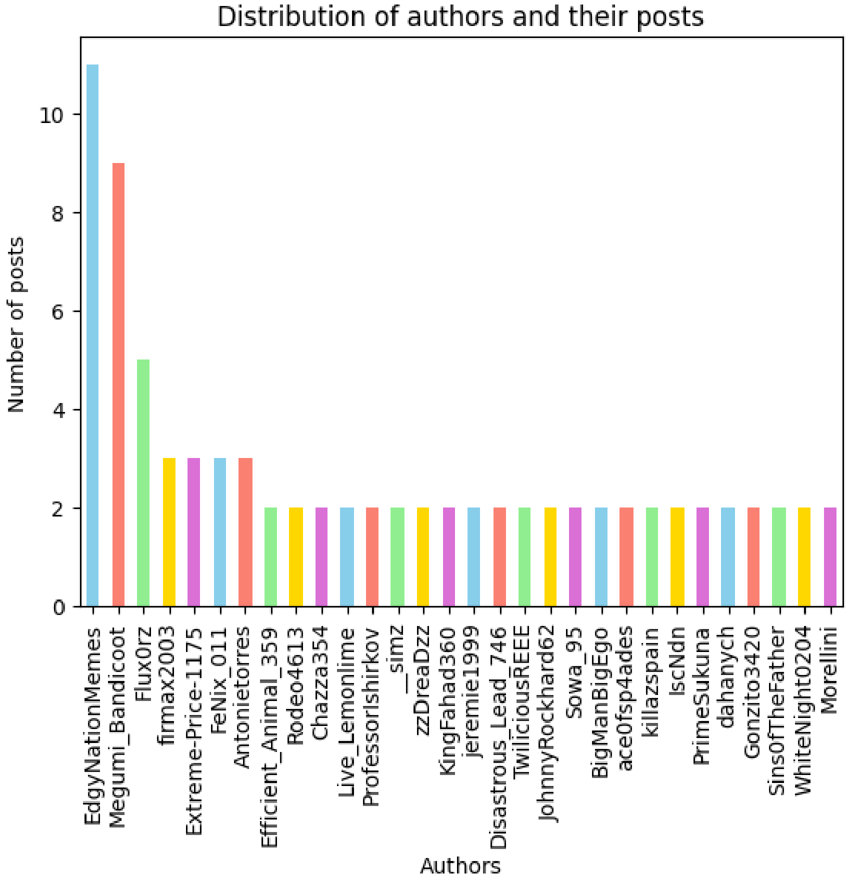


Fig 6. Bar Graph of Distribution of Top 30 Users and their Posts

3. Exploratory Analysis-Part 3

By analyzing the extensive impact of prominent Reddit users, we aimed to pinpoint authors who had swayed over many subreddits. To visualize this network of contributors, we employed data to generate a network graph that showed which specific subreddits these influencers were active in. For simplicity's sake, we concentrated on subreddits where influencers have contributed at least twice, using the X-axis to indicate the particular subreddits and the Y-axis to show submission numbers. Figure 3 illustrates the distribution of subreddits where influencers have contributed significantly.

Figure 3 illustrates graphically the proliferation of other subreddits where these influential posts have had an impact. The Y-axis effectively transmits submission numbers, highlighting the regularity and effectiveness of their efforts, while the X-axis displays the range of subreddits they have connected with. By emphasizing both the scope of these influencers' reach and the depth of their contributions inside certain communities, this chart paints a complete picture of these users' Reddit activity.

It is possible to find out that certain users have considerable influence that extends beyond the confines of particular subreddits by navigating this intricate network structure. The interconnectivity of their interaction demonstrates how barriers between different online communities may be broken down and ideas and information can be shared. This analysis broadens our understanding of notable contributors and illuminates the dynamic interactions that occur at the crossroads of numerous subreddits, adding a range of viewpoints and contributions to the Reddit community.

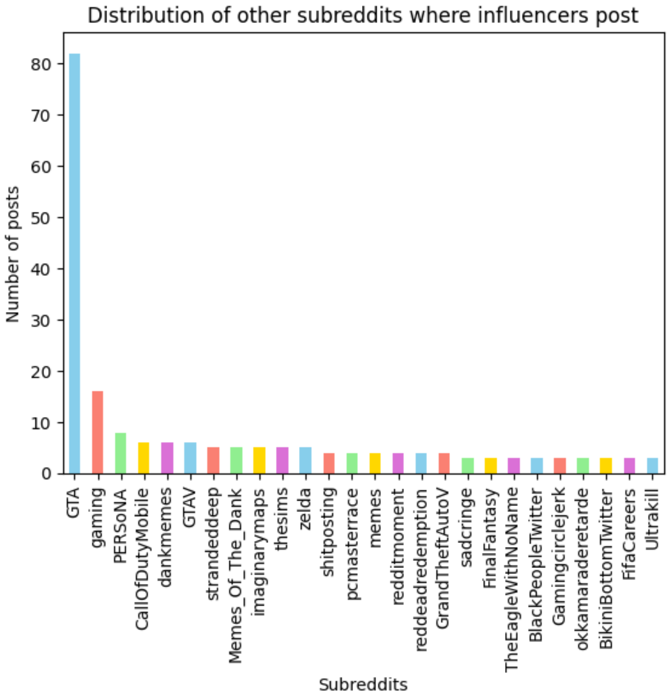


Fig 7. Bar Graph of Top 30 Users in r/GTA's Distribution of Other Subreddits the Users Post

F. Building the Model

We utilized NetworkX's spring layout because we needed to take a more uniform approach. Based on a physical concept, the Spring Layout technique is frequently applied to force-directed graph sketching. It mimics a real-world system in which edges serve as springs and nodes represent charged particles. Finding an equilibrium configuration—where the pressures are balanced and the layout is aesthetically beautiful while revealing the graph's underlying structure—is the aim.

It's crucial to remember that although Spring Layout works well for small to medium-sized graphs, it is less appropriate for extremely big networks due to its computational complexity. Furthermore, the graph's properties and the particular algorithmic parameters selected may have an impact on the layout's efficacy. The final layout can be affected by changing parameters like the number of iterations or the optimal distance (k).

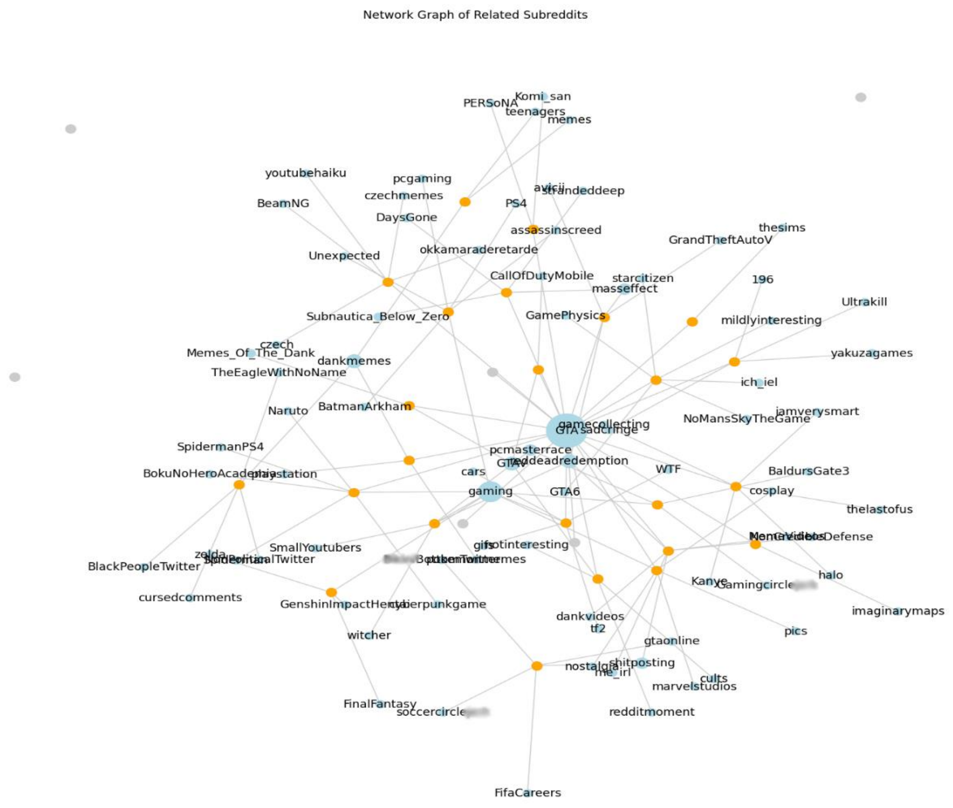


Fig 8. Network Graph Using the GTA Subreddit

In the above network graph, We have used the GTA as the base subreddit in our test run. in the network graph above the orange node represents the influential users and the edges connecting them are the sub reddits which the influential user connects mutually with both the base sub reddit and the other subreddit. The influential user who has the most no of child nodes and connected with the base node will be ranked in the top of the influential user.

CONCLUSION & FUTURE WORKS

In summary, using our procedure, we were able to create a network graph of connected subreddits. A few things to keep in mind are that our notion of what an "influencer" is is rather simplistic. Determining what constitutes an "influencer" threshold varies depending on the subreddit being studied and is influenced by factors like user activity and subreddit size. If an influencer appears too frequently in a subreddit, our sampling of the top 500 posts may yield results that are relatively limited.

In the event that we carried out this experiment again, we would take user feedback into consideration when identifying influencers. Because Reddit threads are hierarchical, processing user comment analysis without optimisation would take an extremely lengthy time. However, we were unable to investigate influence through the comments channel due to time and computational limitations. Furthermore, enhancing our algorithm to discover subreddit similarity based on posts and comments would be an intriguing task to determine if one approach outperforms the other.

To sum up, the study of how Reddit influencers create a network of related subreddits has shed light on how dynamic and interconnected online communities are. We have discovered patterns of cooperation, information exchange, and community development within the Reddit ecosystem by applying data mining, network analysis, and natural language processing techniques.

The identification of influential users across a range of subreddits, the mapping of interactions between subreddits, and the measurement of influence using metrics like user engagement and content popularity are some of the key discoveries. The research has illuminated the ways in which influencers, operating outside the confines of certain subreddits, are crucial in influencing conversations and content development.

The knowledge acquired from this effort lays the groundwork for future investigations into the complex dynamics of online influence, cooperation, and information sharing as Reddit develops. Understanding how influencers bring together diverse communities might help us better manage the opportunities and difficulties posed by the always changing digital social interaction landscape on sites like Reddit.

References:

[1] K. H. Prasad, T. A. Faruquie, S. Joshi, S. Chaturvedi, L. V. Subramaniam and M. Mohania, "Data Cleansing Techniques for Large Enterprise Datasets," 2011 Annual SRII Global Conference, San Jose, CA, USA, 2011, pp. 135-144, doi: 10.1109/SRII.2011.26.

[2] R. Shaji, "Exploratory data analysis on Reddit data: An efficient pipeline for classification of flairs," 2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM), Taichung, Taiwan, 2021, pp. 65-68, doi: 10.1109/BigMM52142.2021.00018.

[3] K. J. Millman and M. Aivazis, "Python for Scientists and Engineers," in Computing in Science & Engineering, vol. 13, no. 2, pp. 9-12, March-April 2011, doi: 10.1109/MCSE.2011.36.

[4] P. Sinthong and M. J. Carey, "Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 3119-3129, doi: 10.1109/BigData52589.2021.9671603.

[5] S. Kazi et al., "Preprocessy: A Customisable Data Preprocessing Framework with High-Level APIs," 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 2022, pp. 206-211, doi: 10.1109/CDMA54072.2022.00039.

[6] J. DSouza and S. Velan S., "Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225621.