# Project Proposal

Group members: Mahmoud Hamza, Jingting Xu

## Project Overview

The aim of this project is to analyze social network data from Reddit and predict trends in a particular subreddit ‘ChatGPT’ and try to predict its growth over time based on various metrics such as the number of comments, upvotes, and downvotes. We will be using machine learning algorithms and network analysis techniques to uncover patterns in the data and make predictions about future trends.

## Method Innovations

Basic steps:

* connecting to reddit API
* get data using the subreddit ‘chatGPT’
* create the graph, where nodes are users and edges are connections between users
* get summary statistics (eg. transitivity, centrality, density)
* network visualization
* community detection

More advanced:

* Build a recommendation system to recommend reddit users based on how they perceive chatGPT (i.e. opportunity, threat, nonsense, etc..)
* compare different methods of recommendation based on performance and speed.
* predict the subreddit growth

## Dataset

We will be using data from Reddit's API to collect information about chatGPT subreddit. We will be collecting data on the number of comments, upvotes, and downvotes for each post, as well as information about the users who posted and commented on the content. We will also be collecting data on the number of subscribers to the subreddit and the growth rate over time. In addition, we may analyse the posts’ texts to aid in the recommendation based on natural language processing

## Updated Weekly Timeline

> We are on track. A single note added for week1 and we will begin week2

Week 1: ***All done (data\_preparation notebook)***

* Collect data from Reddit's API
* Clean and preprocess the data
* summary stats using networkx
* Added: Add more data to graph and optimize the graph used.

Week 2:

* Build machine learning models for the recommendations
* compare models based on performance and speed
* Fine tune the models

Week 3:

* Evaluate the performance of our machine learning models on test data
* consider predicting future trends in the subreddit chatGPT
* Write up our findings and conclusions in a final report

## Preliminary results

In the notebook `data\_preparation\_stats.ipynb` , we present multiple descriptive statistics

1- Top 10 posts for the subreddit “ChatGPT” along with their score

2- We created a graph based on the top 20 hot posts (as an initial trial)

3- we calculated the global coefficient (transitivity) of our graph

3- We calculated different centrality measures:

a- betweeness centrality

b- degree centrality

c- closeness centrality

4- we added centrality attribute to our nodes based on the betweeness centrality

5- We visualized the graph with all nodes labelled

6- we revisualized the graph using labels for the top 10 nodes based on closeness centrality

7- we visualized the graphs based on the communities detected by louvain algorithm

References

<https://medium.com/social-media-theories-ethics-and-analytics/network-analysis-from-social-media-data-with-networkx-13605d711590#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjFhYWU4ZDdjOTIwNThiNWVlYTQ1Njg5NWJmODkwODQ1NzFlMzA2ZjMiLCJ0eXAiOiJKV1QifQ..cPWVTr5ZLQFoiHR3f3Tt0iObh2jzyj4XGZXxHBsfhJ1dri0aToTjGBriG5Mzv61-QacxnuBRRTMFYgEUtqd6F20-EOb5j_9nnOH6bk7l2Fvqlg8dkUeqC0S8HDpeOlklBLaHjv2cCbz7Ex6opnDeN_jh6lQBPfzdjYVGjna9o66d4UD7cudUQ_8apE5KEXCnvAVWSNgBDJLQiO9-ZHwQ-joOOpUFnTSCGbvxMwskgHuaEYYkq7agllwvrYuKfwyJlUdVnU6wi307oRi2FtTyjuBLzdNXOi0Apt9Kx333cLvQphk5_PR09MC-xl0tCAgc3GcLE2a1aa_p_ekI5-g_IQ>

<https://www.kaggle.com/code/anshulmehtakaggl/recommending-you-github-users-viz>

<https://github.com/topics/social-recommendation>