Reddit Interactive Network Analysis Report

Course：Web Science
Submission Date：07/03/2025
Author：Jiaxin Cheng
GUID:2973117C

1. Data Processing

1.1Introduction:

In this Analysis, we implemented Network Analysis to the data of InvestmentClub subreddit by the interactive model of of exploring Reddit users. We used Reddit forum posts(submissions_cleaned.csv) and comments data (comments_cleaned.csv) and built a social network based on user analysis. This report will introduce in details the methods of data processing and network construction, and provide basic network analysis.

1.2Data Sources

I used two main data files:

1.2.1 submission cleaned.csv(cleaned posts data)

It recorded id、author(poster)、created_utc(post time)、num_commenrs(the number of comments) and so on.

1.2.2comments_cleaned.csv(cleaned comments data)

It recorded author(commentor)、link_id()、body、created_utc（commented time）and so on.

These two datasets provide Reddit users' posting and comment interaction data that can be used to build user-user social networks

1.3 Data Aggregation

1.3.1 Goal

I want to build a user-user network where:

Nodes: Users (including posters and commenters)

Edges: User A commented on User B’s posts, forming a relationship between A → B

1.3.2 Data processing

In order to build a user-user relationship, we performed the following data processing steps:

1.Create a post ID to poster mapping

2.Map commented post ID to post author

3.Remove NaN value.

4.Generate source → target structure.(network_edges.csv)

1.3.3Data Organization Rationale

The reason why I want to build User-User Network (rather than Post-User Network) is that I want to research user interaction model, find crucial influencer and accomplish visualization.

Pseudocode：

```
INPUT:
    - Submissions dataset ("submissions.json")
    - Comments dataset ("comments.json")

PROCESS:
    1. Read both JSON files into dataframes.
    2. Select key columns:
       - From submissions: "id" (post ID), "author" (post author)
       - From comments: "author" (comment author), "link_id" (post ID)
    3. Convert "created_utc" timestamps to human-readable format.
    4. Create a mapping: (post ID -> post author) from the submissions dataset.
    5. Merge this mapping with the comments dataset to identify interactions.
    6. Generate an edge list (source: comment author, target: post author).
    7. Save the cleaned dataset as "network_edges.csv".

OUTPUT:
     - "network_edges.csv" containing the directed interaction network.
```

1.4 Summary

After processing the data, we successfully generated a network edge list (network_edges.csv) that captures user interactions.
The key steps in this process were:

1.Cleaning and filtering the submissions and comments datasets.
2.Mapping comments to their corresponding post authors.
3.Creating a directed network, where an edge exists from the comment author to the post author.
4.Storing the edge list in CSV format for network analysis.

2.Graph construction and Visualizaion

2.1 Introduction

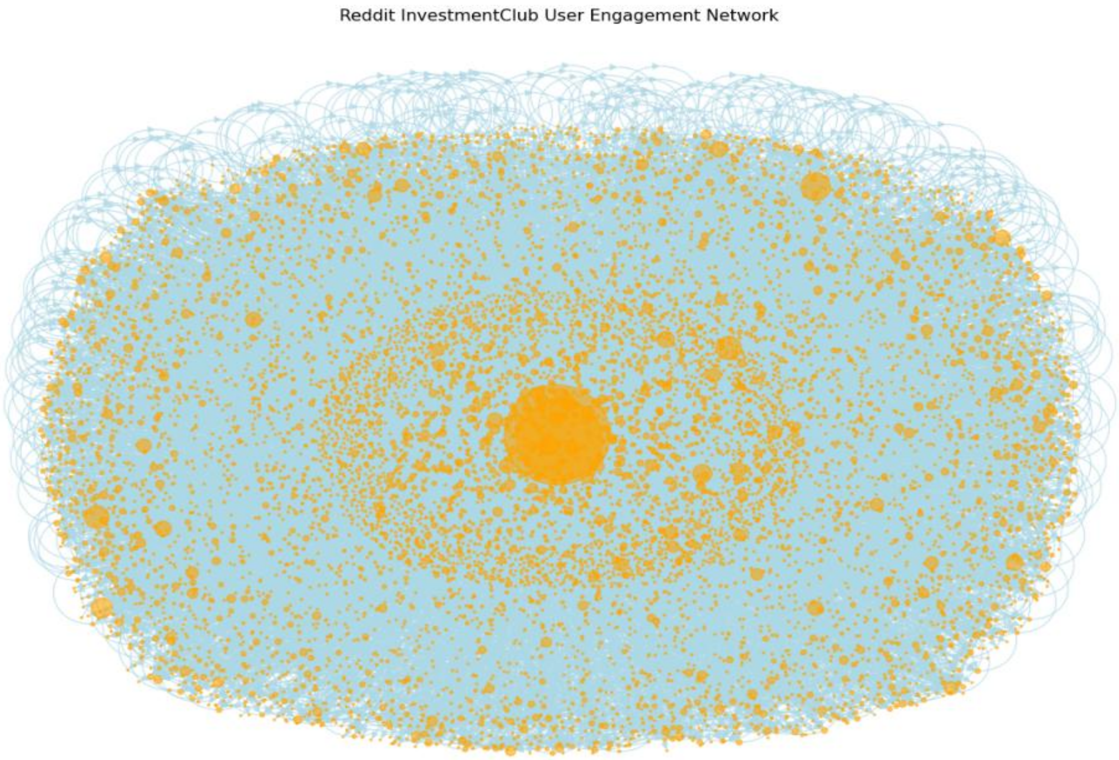
In order to analyse the network data directly and deeply, I created a directed network graph where:

Each node represent a user. Origin nodes are the users who write comments (i.e., those who interact with others by responding to posts). Destination Nodes are the users who create the original posts (i.e., those who receive interactions through comments on their posts). Each directed edge represents a comment made by one user on another user’s posts.

Pseudocode for graph visualization:

```
BEGIN
    LOAD network edges dataset
    CREATE a directed graph using NetworkX
    SET node size proportional to degree
    COLOR nodes orange and edges light blue
    USE a spring layout for positioning
    DISPLAY the graph
END
```

Result:



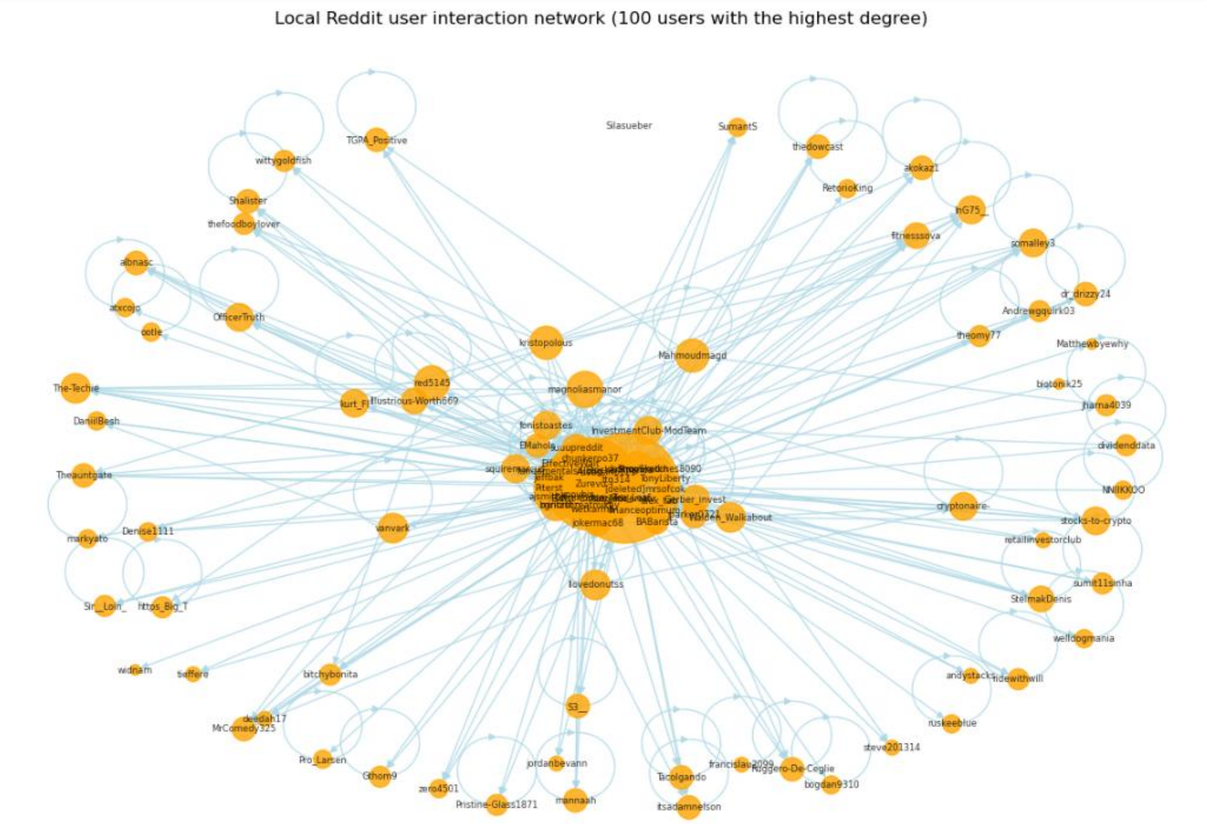
In order to improve clarity, I created Subgraph of Top 100 Most Active Users and Top 20 Most Active Users.

Firstly, rank users by the activity ranking. I measure user activity by degree centrality which calculate the amount of edges included by each user. Then I classify the users by the degree. Secondly, extract the top 100 active users and order them. Then, I extracted sub graph from main network and only retain the interaction between these users. Thirdly, I use the same visualization's color parameters as the main network. But I ajusted the parameter of size to make it bigger than before.

Pseudocode for graph visualization:

```
BEGIN
    LOAD the full directed graph G

    COMPUTE the degree of each user (number of interactions)
    SORT users by degree in descending order
    SELECT top 100 users with highest degree

    EXTRACT subgraph containing only selected users
    SET node size proportional to user degree
    SET node color as orange and edge color as light blue
    APPLY a spring layout for better positioning
    DRAW the subgraph visualization
    DISPLAY the graph
END
```

Result:

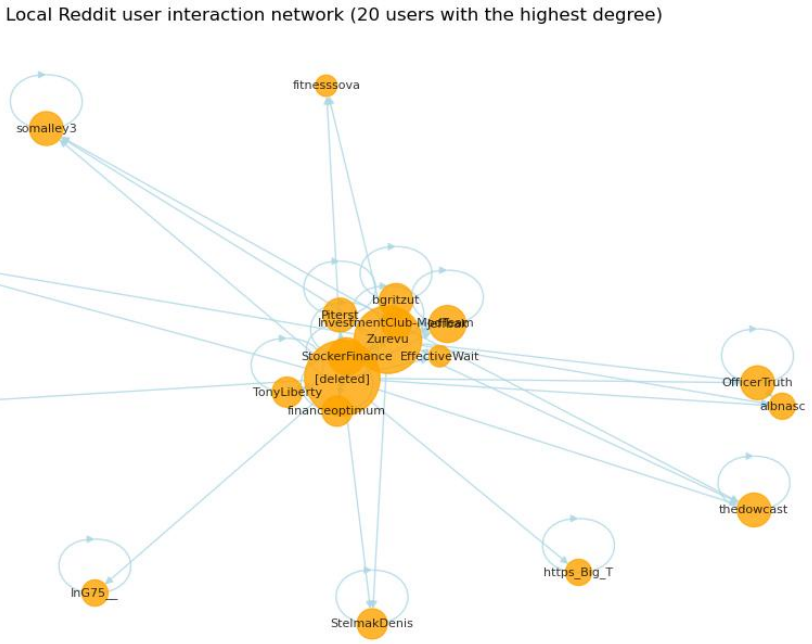


Similarly, I extract Top 20 Most Active Users and create the graph.

Pseudocode for graph visualization:

```
BEGIN
    LOAD the full directed graph G

    COMPUTE the degree of each user (number of interactions)
    SORT users by degree in descending order
    SELECT top 20 users with highest degree

    EXTRACT subgraph containing only selected users
    SET node size proportional to user degree
    SET node color as orange and edge color as light blue
    APPLY a spring layout for better positioning
    DRAW the subgraph visualization
    DISPLAY the graph
END
```

Result:



Local Reddit user interaction network (20 users with the highest degree)

2.2Interpretation of the Data and Observations

2.2.1 Interpretation of full network

The full network visualization displayed interaction structure between many small modes and some very large central nodes. The biggest nodes(high degree users) indicated highly participated users who either often post and receive many reviews or often review to many posts and interact with other users. The existence of light blue directed edge indicate many users interact asymmetrically, which means that some users receive more reviews than they review. The network follows a power law distribution, where little user can generate a lot of discussion but the majority of users contribute least.

2.2.2 Interpretation of Subgraph of Top 100 Most Active Users

The aubgrapg of top 100 most active users highlighted core community members and indicated a strongly interacted group. This indicated influenced users often interact with each other. Some users act as hubs which facilitate the discussion between different groups. Cluster formation might represent topic-specific discussions. High in-degree nodes might be discussion leaders or authoritative person.

2.2.3 Interpretation of Subgraph of Top 20 Most Active Users

Some central users(InvestmentClub-Modteam，Zurevu，StockerFinance and effected Wait) dominated network. These users participate and might be toastmaster, frequent contributors or influencer in Subreddit。Some users mainly interacted with center digits and formed small Peripheral cluster. Some users such as somalley3，Thetechie and denise111 who are on the border of edge with less direct interaction.

2.3 Observation

Highly participated users can conduct discussion. A little users boost the majority of interaction who might be community leaders, a small number of users or highly righteous participant. The majority of users participate lowly. And many users participate least via reviewing one or two posts.

2.4 Justification of the approach:

2.4.1 It represent real interaction flow:

The comment system in Reddit is naturally directed and worked on the post, which indicated the clarity of interaction. The directed edge exactly reflected this kind of relationship and therefore ensure that network remained participated structure information.

2.4.2 Make sure key influencer:

I can make sure the attendance of users by analyzing in-degree the node (number of incoming edges).

Super analysis is helpful to make sure the most active reviewer.

2.4.3 Capture community structure:

The direction allow further network analysis such as community identification, test influenced users and measure attendance level.

The approach make sure that network visualization keep insistent with real world dynamics and is useful to analysis and explaination.

3. Analysis of Network Properties

3.1 Introduction

This part aims to analyse the network property in Reddit InvestmentClub and specifically focus on Rich-Club phenomenon. This effect research whether high in-degree users(super users) tend to form closely connected sub network. By researching this phenomenon, we could learn about how influential users in community interact as time going by.

3.2 Research question

How users on support communities, as a group behave over time?

In order to tackle this question, we analysed Rich-Club ratio, which measures the tendency of high-degree nodes to be more interconnected compared to the entire network.

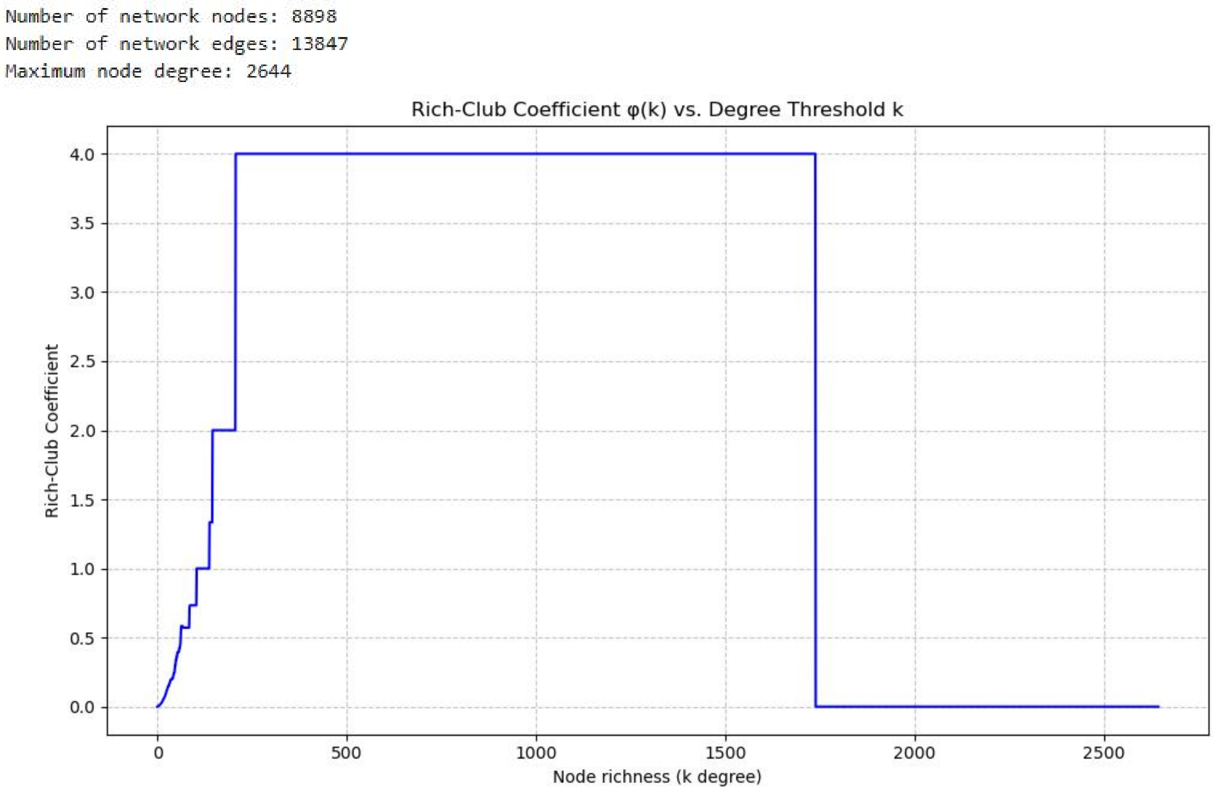
3.3 Methodology

Firstly, I calculate the Rich-Club ratio φ(k) of different degrees of threshhold and plot the relationship between k and φ(k). This helps to ensure whether degree nodes (more active users) form stronger inner network.

Pseudocode for Rich-Club Analysis:

```
BEGIN
    LOAD Reddit interaction graph G
    COMPUTE degree for each node
    FOR k from 0 to maximum node degree:
        FIND nodes with degree greater than k
        COMPUTE number of such nodes N>k
        IF N>k < 2:
            CONTINUE
        CREATE subgraph with N>k nodes
        COMPUTE number of edges E>k
        CALCULATE rich-club coefficient  $\Phi(k) = 2E_{>k} / (N_{>k} * (N_{>k} - 1))$
    PLOT  $\Phi(k)$ vs k
END
```

Result:



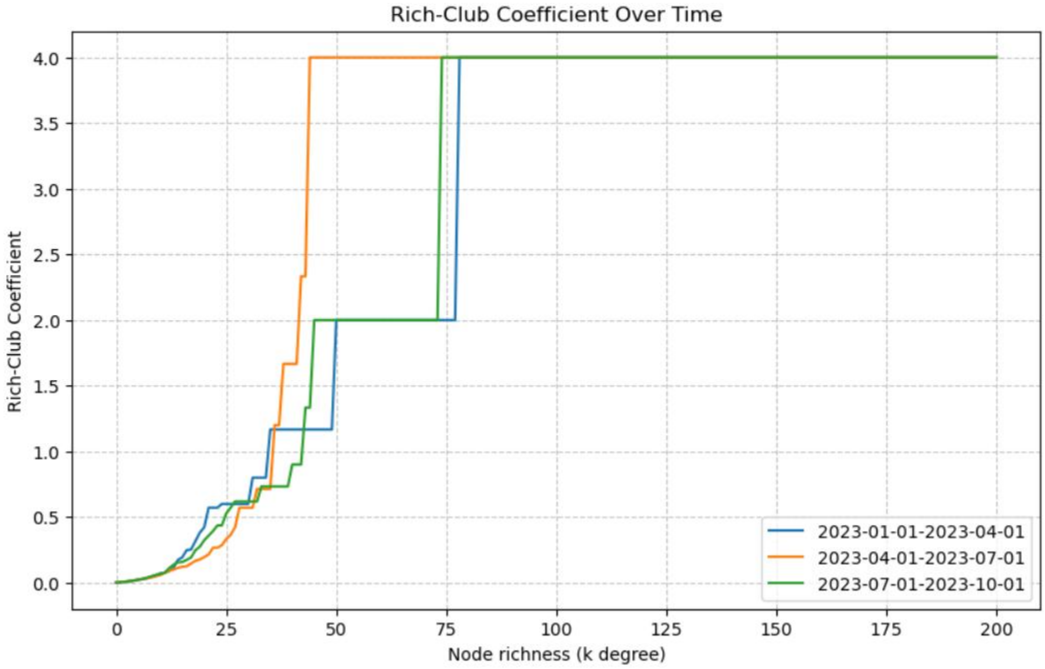
```
Number of network nodes: 8898
Number of network edges: 13847
Maximum node degree: 2644
```

Secondly, I segmented interactive data into different periods and calculate Rich-Club coefficient in each period. This helps me to observe how the interconnectivity among highly active users changes over time.

Pseudocode for Rich-Club Analysis Over Time:

```
BEGIN
    LOAD Reddit interaction data with timestamps
    DEFINE time windows (e.g., quarterly segmentation)
    FOR each time window:
        EXTRACT interactions within the time range
        CONSTRUCT network graph G_t
        COMPUTE degree for each node in G_t
        FOR k from 0 to maximum node degree:
            FIND nodes with degree greater than k
            COMPUTE number of such nodes N>k
            IF N>k < 2:
                CONTINUE
            CREATE subgraph with N>k nodes
            COMPUTE number of edges E>k
            CALCULATE rich-club coefficient   $\Phi(k) = 2E>k / (N>k * (N>k - 1))$
        PLOT   $\Phi(k)$ vs k for each time period
END
```

Result:



3.4 Observation and interpretation

3.4.1 Rich-Club Coefficient Tendency

Low Level Users（k < 100）： Rich-Cub coefficient is too low， Which means the majority of users do not form interconnected group preferentially.

Intermediate User（100 < k < 500）： Rich-club coefficient rise largely，which means more participated users begin to form dense sub network.

High degree users（k < 500）： Rich-club coefficient stabilizes before falling down, which means highly connected users’ center possess limited amounted extremely high node, forming isolated interaction.

3.4.2community structure

The existence of Rich-Club effect indicate that a part of users dominate the interaction in community.This kind of structure is very common in supporting community, electric users or toastmaster would drive discussion.

The decrease in φ(k) at very high degrees may suggest that top contributors interact more with the general community rather than forming exclusive elite circles.

3.4.3 Rich-Club Coefficient Trends Over Time

1.First Time Period (Jan 2023 - Apr 2023):

The rich-club coefficient displays a steady increase, indicating that high-degree users were forming more closely connected clusters over time.

2.Second Time Period (Apr 2023 - Jul 2023):

The coefficient increased significantly at lower degrees, suggesting that moderately active users started forming more interconnected subgroups.

3.Third Time Period (Jul 2023 - Oct 2023):

The coefficient stabilized at a high level, suggesting a well-established rich-club structure among highly active users.

3.5 Conclusion:

This analysis shows that highly active users formed a more and more closed sub network over time, indicating that the effect of Rich-Club in Reddit InvestmentClub become bigger and bigger. These influential users might have significant influence in maintaining discussion flow, instructing new comers and shaping community practices.

4. Sentiment Analysis

4.1 Introduction

This part aims to analyse the users' emotion in Reddit InvestmentClub and specifically focus on the research question that how to analyze the emotional changes of users in the community, especially the comparison between high-scoring users and low-scoring users.

4.2 Mthedology

4.2.1. Sentiment analysis

Apply Vader sentiment Analysis to extract semtiment scores from users' posts.

Compare the distribution of emotion between high score and low score users.

4.2.2 Topic Modelling

Apply potential LDA to confirm the crucial topic

Track how topic change indifferent windows

4.2.3 Comparison between high score users and low score users

Split users based on their submission and comment scores.

Compare the sentiment score and distribution.

4.3 Implementation

Pseudocode for Rich-Club Analysis:

```
1. Read data:
    Read post data from "submissions_cleaned.csv" (df_submissions)
    Read comment data from "comments_cleaned.csv" (df_comments)

2. Handle missing values:
    If the post 'title' is empty, fill in "No Title"
    If the comment 'body' is empty, fill in "No Content"
```

3. Initialize the sentiment analyzer:

Create a VADER SentimentIntensityAnalyzer instance

4. Calculate emotional scores:

For each post 'title', calculate the sentiment score and deposit it into the 'sentiment' column

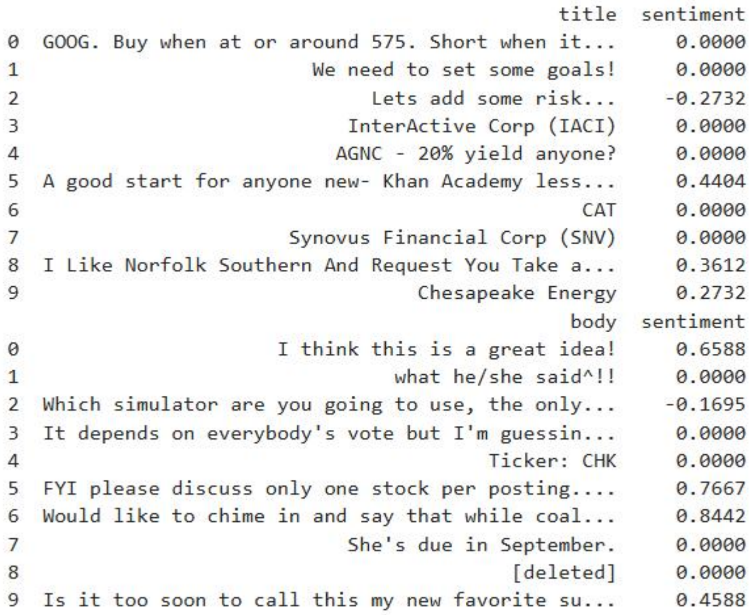
For each comment 'body', calculate the sentiment score and deposit it into the 'sentiment' column

5. Output result:

Show the first 10 posts (title and emotional score)

Show the first 10 comments (body and emotional score)

Result:



```
                                             title  sentiment
0  GOOG. Buy when at or around 575. Short when it...     0.0000
1                           We need to set some goals!     0.0000
2                                Lets add some risk...    -0.2732
3                              InterActive Corp (IACI)     0.0000
4                              AGNC - 20% yield anyone?     0.0000
5  A good start for anyone new- Khan Academy less...     0.4404
6                                                  CAT     0.0000
7                          Synovus Financial Corp (SNV)     0.0000
8  I Like Norfolk Southern And Request You Take a...     0.3612
9                                    Chesapeake Energy     0.2732
                                              body  sentiment
0                           I think this is a great idea!     0.6588
1                                  what he/she said^!!     0.0000
2  Which simulator are you going to use, the only...    -0.1695
3  It depends on everybody's vote but I'm guessin...     0.0000
4                                          Ticker: CHK     0.0000
5  FYI please discuss only one stock per posting....     0.7667
6  Would like to chime in and say that while coal...     0.8442
7                               She's due in September.     0.0000
8                                            [deleted]     0.0000
9  Is it too soon to call this my new favorite su...     0.4588
```

1. Create a canvas of size (12,5)

2.

2. Plot the distribution of sentiment scores for Submissions:

a. Select the first subgraph (position 1,2,1)

b. Plot a histogram (bins=30) and enable kernel density estimation (KDE=True)

c. Set the title "Submissions Sentiment Distribution"

d. Set the x-axis label "Sentiment Score"

3. Plot the distribution of sentiment scores for Comments:

a. Select the second subgraph (position 1,2,2)

b. Plot a histogram (bins=30) and enable kernel density estimation (KDE=True)

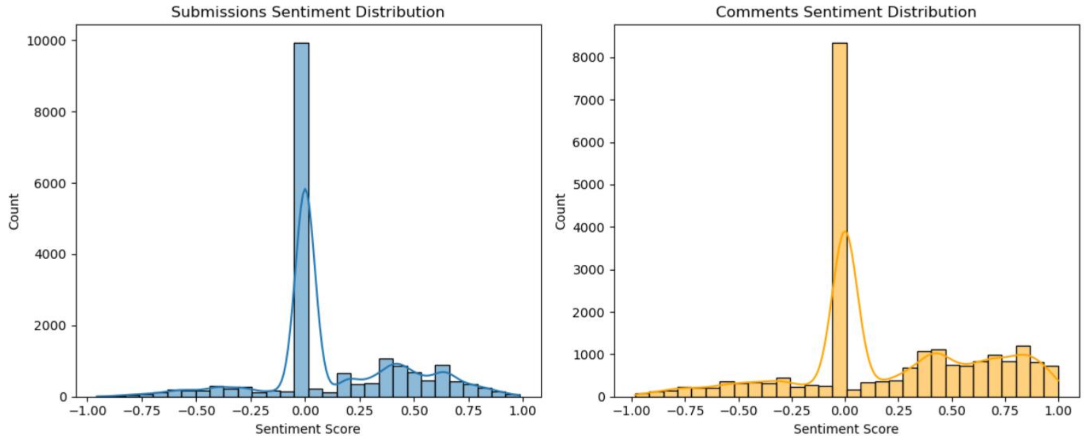
c. Set the color to orange `color="orange"`

d. Set the title "Comments Sentiment Distribution"

e. Set the x-axis label "Sentiment Score"

3. Adjust the image layout to ensure that the two subgraphs do not overlap

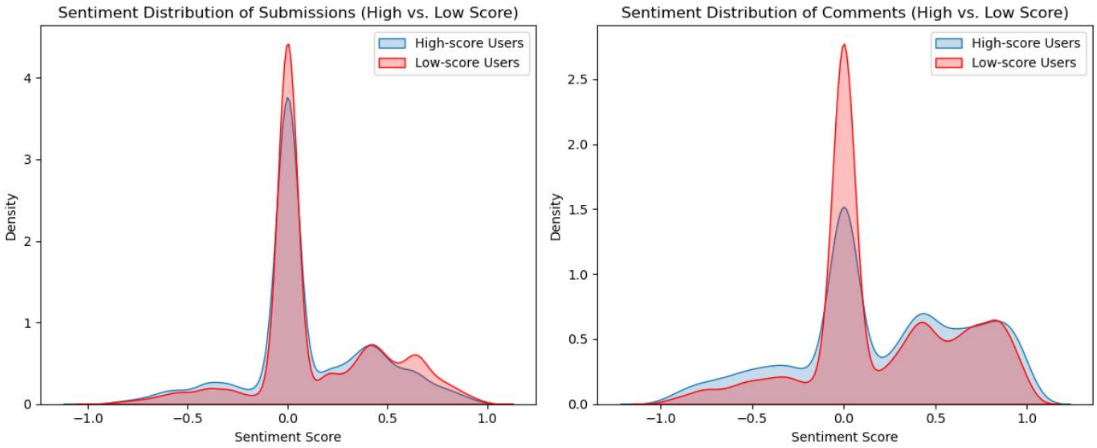
4. Display the image

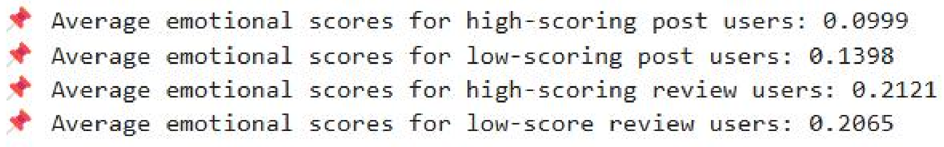


```
# Create a figure with a size of 12x5
Initialize a figure with size (12, 5)

# Plot sentiment distribution for Submissions
Create subplot (1, 2, 1)
    Plot KDE (Kernel Density Estimate) for "sentiment" scores of High-score Users in df_sub_high_low
    Fill the KDE plot and label it as "High-score Users"

    Plot KDE for "sentiment" scores of Low-score Users in df_sub_high_low
    Fill the KDE plot with red color and label it as "Low-score Users"

    Set title: "Sentiment Distribution of Submissions (High vs. Low Score)"
    Set x-axis label: "Sentiment Score"
    Add legend

# Plot sentiment distribution for Comments
Create subplot (1, 2, 2)
    Plot KDE for "sentiment" scores of High-score Users in df_com_high_low
    Fill the KDE plot and label it as "High-score Users"

    Plot KDE for "sentiment" scores of Low-score Users in df_com_high_low
    Fill the KDE plot with red color and label it as "Low-score Users"

    Set title: "Sentiment Distribution of Comments (High vs. Low Score)"
    Set x-axis label: "Sentiment Score"
    Add legend

# Adjust layout to prevent overlap
Apply tight layout
```

Display the plots

Show the figure





```
📌 Average emotional scores for high-scoring post users: 0.0999
📌 Average emotional scores for low-scoring post users: 0.1398
📌 Average emotional scores for high-scoring review users: 0.2121
📌 Average emotional scores for low-score review users: 0.2065
```

4.4 Results and Interpretation

General sentiment trends: The sentiment distribution of Reddit submissions and comments is centered around a neutral value (0), with some highly positive and negative outliers.

High vs. low scoring users:

High-scoring submission users tend to have slightly lower average sentiment scores (0.0999)(0.1398) compared to low-scoring users.

In contrast, high-scoring comment users have slightly higher sentiment scores (0.2121)(0.2121)(0.2065) compared to low-scoring users.

This suggests that popular opinions may not necessarily be positive, but that high-scoring comments tend to be more constructive or engaging.

Topic modeling insights: Future extensions could analyze how different discussion topics are related to user engagement and sentiment trends.

4.5 Conclusion

By combining network structure analysis with content-based insights, we can better understand how user engagement evolves. These additional analyses provide a broader view of how users interact and what topics drive community engagement.