Final Project Proposal
Group: Abdelmalek Hajjam, Monu Chacko

Table of Contents

Introduction

For the final project we will use the knowledge gained in the class to analyze a dataset of comments for a given discussion. We will Take a discussion thread from reddit and analyze its comments. Reddit is a social news aggregation site, rating contents, and discussions. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. This network can be analyzed to determine its sentiment and network itself.

Kaggle provides us many datasets that are collected from many places. It also includes comments data from reddit. We can use this dataset to do our analysis.

Reddit comments dataset https://www.kaggle.com/reddit/reddit-comments-may-2015 contains about 1.7 billion records. To perform

analysis, kaggle provides us a smaller subset of data. We will use that data for this project.

This dataset has the following columns:

created_utc, ups, subreddit_id, link_id, name, score_hidden, author_flair_css_class, author_flair_text, subreddit, id, removal_reason, gilded, downs, archived, author, score, retrieved_on, body, distinguished, edited, controversiality, parent_id

The body field in this dataset is the text comment.

Network Analysis

Comments in a discussion happens between two or more persons, or directly to the topic. This forms a social network. Each person can be treated as a node. The edge between 2 nodes is the comment response from one node(person) to the other. Hence an undirected graph. The weight of the edge is the number of comment responses between the 2 persons. There can also be other influencing data that can be introduced to determine the weight.

Sentiment Analysis

We will perform text processing to determine the sentiment of the comment. This data can be added to the above network to perform more analysis.

Analysis

For our analysis we will use degree centrality. The importance of the node in relation to its neighbor will be analyzed. The sentiment analysis along with degree centrality will give us interesting insights about the relationship in a network. For example a person could be important to the network if the person has negative sentiment and is controversial.

Issues

There are number of problems that we might face in this analysis. For example if people do not respond to each other but just respond to the discussion then it might not form a network like we would want. The presents of bots or auto responders might also be a problem.