

Understanding Reddit Communities Through Shared Language

John Flanigan

September 16, 2019

Project Summary

Reddit is a content aggregation site where users can submit links, images, and text posts and then other users can vote and comment on these posts. Every submission is made to a specific group, known as a subreddit. By analyzing the text of the posts and comments within subreddits for similarities, comparisons between subreddits can be made. The results of this analysis can be visualized in an interactive graph and be used to generate subreddit recommendations.

Merits

This project has value because it can lead to new insights about how communities on reddit relate to each other, even if they don't directly interact. It would be simpler to find subreddit relationships by looking at which subreddits a user interacts with but by taking a natural language processing approach, similarities between subreddits can be identified, even if they don't have any users in common.

The analysis done in this project can be used to create a visualization of the way subreddits relate to one another. It could also be used to generate a program to recommend similar subreddits based on either text input or a list of subreddits. This recommendation would be valuable to users who are interested in discovering new subreddits they might not be aware of. Understanding connections between subreddits could also be useful to advertisers who would be looking to expand their advertising reach.

Challenges

Depending on the size of the data and the difficulty of gathering data, it might be necessary to limit this project to a subset of subreddits. This subset could be all English speaking subreddits, a certain number of subreddits based on popularity, or a group of subreddits that already have some aspect in common.