Reddit Social Analysis Project Proposal

Carissa Bleker[1], Ashely Cliff[1], Sean Whalen[2], and Saeed Beztchi[2]

I. Introduction

Reddit (reddit.com) is a social website where users congregate online to share, vote on and discuss content. Currently Reddit is the 3rd most popular social website in the US, and the 6th most worldwide [1]. Reddit provides numerous forums or communities, termed subreddits, that cover many unique and overlapping topics. These topics range from news, sports, politics, hobbies, entertainment, etc. Each of these subreddits may have tens to millions of users submitting topic specific content, creating discussion threads on these submission, and either "up-voting" or "down-voting" specific submissions or comments.

As a popular website, Reddit has an impact on how information is shared, new phrases are generated, and how perspectives on issues are formed and adopted by users [2]. Analysis of submissions, comments and voting results should help expose these patterns. With the data and methods discussed below, we aim to provide insight into how themes included in subreddits evolve over time, how Reddit users perceive certain themes, and the social networks of Reddit users within and across communities.

II. Data

The hierarchy of Reddit is shown in Fig. 1. The website Reddit contains subreddits focusing on specific topics. Users can submit content to these subreddits, and comment on each others submissions. The data we aim to use will include text collections from submissions, consisting of the text from the comment threads. The comments for each submission will be combined to create a list of strings, and from these strings a list of word frequencies. Meta-data associated with submissions is also available, such as the net number of votes (upvotes - downvotes + fudge factor), submission content, and the date and time the submission was created.

Reddit has a dedicated API for gathering information. PRAW[4] (Python Reddit API Wrapper) works with the Reddit API to provide a Python wrapper and environment. PRAW will be used to gather the information from Reddit for analysis.

III. Methods

Themes

We aim to extract themes from the agglomerated comment text on a particular submission, and analyze how these themes appear and are then replaced in a subreddit over time.

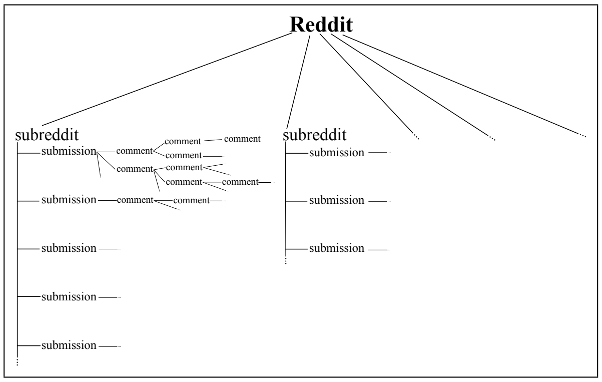

Fig. 1. Reddit submissions hierarchy.

From [3], a "theme" in a text collection is probability distribution of words characterizing a specific topic. These themes can be modeled by a mixture model of a unigram language, estimated by an Expectation-Maximization algorithm.

Controversial themes

Reddit uses tags to describe user interactions with a submission. These tags are Hot, New, Rising, Controversial, Top, and Gilded. By selecting a tag, the submissions within a subreddit will be listed based on these tags. Submissions that are heavily up-voted and interacted with are tagged as Hot. Submissions that receive a high amount of both positive and negative interactions are tagged as Controversial.

By doing text analysis of a submission and its comments, insight can be gained into what themes, users, time of submission result with a specific tag. Specifically, we will focus on what text patterns are highly associated with the Controversial tag.

Users

We can hypothesize that users will tend participate in communities based on their similarities. Given this assumption, there is a probability of a certain user posting in a subreddit based on which other users have posted in that subreddit.

For this analysis, we will build a cohesive network of a number subreddits and the most common users for each subreddit. By doing clustering on this network, we can quantify how users group together around different topics.

IV. Potential Problems

While using an API and an API wrapper makes programming and data management more user friendly, it may include some problems. There are numerous factors specified for the proper scraping and analysis of Reddit data, so it is possible that PRAW may limit our process. Such factors

[1] The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville

[2] The Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville

may include obtaining a shorter history of a subreddit's comments, or possible syntactical errors during processing. Another possible issue that may occur is insufficient data due to the deletion of a user, a user's comments, or an entire subreddit.

V. Timeline

1) Set up OAuth, the authentication method used by Reddit to allow third party software (i.e. scrapers) to access and use Reddit data. This requires having an account with Reddit and creating a client ID and secret that Reddit can use to authenticate the software.
2) Determine which subreddits to focus on. Subreddits will need to be chosen based on how often posts are added, as well as how often the posts are read and commented on, as comments are necessary for analysis.
3) Using one subreddit as a starter, set up PRAW (Python Reddit API Wrapper) to collect information about the subreddit, including a list of submissions going back as far as the API will allow.
4) For each submission within the subreddit, collect all comments into a list for analysis.
5) Collect the names of the top contributors for each submission. Combine top contributors for each submission to create list of top contributors for the subreddit.
6) Analyze comments.
    a) Analyze word patterns and themes within comments.
    b) Analyze differences in word frequencies for submissions with and without the Controversial tag.
7) Compare themes across submissions, using the time the submission was created as the time stamp, showing the evolution of themes within a subreddit.
8) Compare lists of top contributors across subreddits.

TABLE I
TIMELINE

Step	Date
1	10/5
2	10/12
3	10/15
4	10/22
5	10/22
6	11/17
7	11/22
8	11/22

References

[1] SimilarWeb.com. 2017, August. Top Social Network Websites in United States. [online] Available at: https://www.similarweb.com/top-websites/category/internet-and-telecom/social-network [Accessed 27 Sep. 2017].

[2] Weninger, T., Zhu, X.A. and Han, J., 2013, August. An exploration of discussion threads in social news sites: A case study of the reddit community. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 579-583). ACM. Vancouver

[3] Mei, Q. and Zhai, C., 2005, August. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 198-207). ACM.

[4] Boe B. PRAW: The Python Reddit API Wrapper. 2012-, https://github.com/praw-dev/praw/ [Online; accessed 2017-09-29].