Chianson Siu, Wei Liao

CSE160: Final Project Part 1

12 February 2019

Analyzing Stack Exchange’s Interpersonal Skills

**Summary of research questions.**

Our project is interested in studying the prevalent interpersonal issues among people in 2018. Using data from Stack Exchange’s Interpersonal Skills website, our study will seek to address 3 different research questions about the website’s posts and users.

1. What type of questions are more likely to be have an answer?

Given a post has an accepted answer, we hypothesize that either longer posts or posts about relationships will most likely have an accepted answer. This may be because longer posts have spent more time detailing their question, so other users are able to clearly understand what they are asking whereas as short post may leave out key details. Additionally, a post about a relationship, such as boyfriend or girlfriend, may likely have an answer because it is old and extremely common topic. Therefore, more people are likely to have experience or knowledge about relationship issues, allowing for a greater chance of receiving an answer.

1. What type of users have high reputation scores?

Given an equal number of approved responses between users, we hypothesize that users with longer answers will have higher reputation scores. This may be because a longer answer requires more effort and as a result is more thought out. If it completely answers the question, then it is more likely to be upvoted for other users, giving the answerer more points than just the 15 points given for an accepted answer.

1. Given a user a low number of reputation points, what could be a contributing factor to the low score?

User with low reputation points are those that are negative towards others (low upvote high downvote, compute ratio). Two Pearson correlation graph.

We hypothesize that users who use words categorized as racist, sexist, or otherwise is considered harmful, will be overrepresented among users with low reputation scores. This is because a post marked as abusive by 6 or more other users results in a loss of 100 points, the largest reputation loss possible influenced by other users (a user can lose more points if they offer over 100 points as bounty for one of their questions).

**Motivation**

These research questions are worth investigating because it can be used as proxy to discover what types of interpersonal issues were prevalent in 2018, while also informing people about the user base of Interpersonal Skills website. Understanding prevalent social topics is useful for understanding how society may be influencing the social culture over the years. This may be of interest to psychologist or mental health organizations, where the researchers in the organizations may need to understand contemporary problems to provide meaningful and effective solutions.Additionally, Interpersonal Skills is only 1 of 173 online communities under Stack Exchange. Therefore, the company may be interested in this analysis because it will inform them more about their users, which could lead to future decisions on how to improve their websites. This could range from simple aesthetics to more functional changes like adding or removing privileges based on knowing what types of users have high reputation scores.

**Methodology**

For the first research question, this study will analyze post titles, as well as the body of the post. By using text analysis to conduct word counts for each post, we will be able to reoccuring themes in all questions. Then we will compare the top reoccurring words with questions that have accepted answers, showing the statistics for numbered of answered questions per keyword or group of keywords. After that, we will repeat the process using the tags of the post and length of post.

For the second research question, this study will first filter out only for users with high reputation scores. Then we will analyze the posts, questions, answers, and comments made by these uses to see if there hidden reason why one user may have a higher reputation than another.

For the third research question, this study will first filter out only for users with low reputation scores. Then we will analyze the posts, questions, answers, and comments made by these users to see if there is a hidden reason why these users have low reputation scores.

**Work Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Task | Estimated Completion Time | Assigned Group Member |
| 1 | Organize data into a dataframe using the pandas library. This will make it able to be queried. | 2 hours | Chianson |
| 2 | Perform text analysis on posts and titles to answer research question 1. | 5 hours | Wei |
| 3 | Perform tag analysis on posts and titles to answer research question 1 | 3 hours | Chianson |
| 4. | Filter for high reputation users and analyze their posts, comments, answers, and questions for research topic 2. | 5 hours | Wei |
| 5. | Filter for low reputation users and analyze their posts, comments, answers, and questions for research topic 3. | 4 hours | Chianson |
| 6. | Visualize data for steps 2-5 using matplotlib and include them in a written report. | 6 hours | Chianson & Wei |

Our team will collaborate by programming parts individually and sharing our code through github. For more difficult steps, we will also meet in person to do pair programming on one person’s computer. The report portion of the project will also be divided evenly, with each member (Chianson and Wei) writing separate sections and then checking each other’s work.

**Questions**

We are not quite sure what analysis or methodology to implement. Our plan was just to aggregate the data in various ways and look for patterns or relationships, but the assignment description sounds like it wants actual statistical formulas. What formulas should we consider using given the approach we have decided to take? Neither of us have a background in mathematics or statistics.