**Improving Crowdsourced Documentation:**

**Examining Answers on Stack Overflow**

Scott Crain, Niranjana Gajjelli, and Marc Hudak

Abstract

The goal of our project is to divorce questions from answers and focus on what makes an answer desirable. One assumption that we make is that highly voted answers reflect what people like to see in answers.

1. **Introduction**

As an unofficial source of documentation, Q&A sites such as Stack Overflow are immensely popular among programmers, and that popularity is only increasing as time passes. As of October 19, 2014, Stack Overflow has over 8.2 million questions and 14 million answers. A mere year ago, Stack Overflow had over 5.5 million questions and about 10.5 million answers. Both of these are nearly 50% increases over a single year. Stack Overflow is constantly growing, and shows no signs of slowing down. However, as with all forms of community-driven content, the quality of the material on Stack Overflow is subject to questioning. With the increasing popularity of Stack Overflow as not only a Q&A forum, but as a form of unofficial documentation to those who come after, low-quality content is a very bad thing to have.

The primary forms of quality control on Stack Overflow are the "voting" and "accepted answer" systems. For any given post, be it question or answer, any Stack Overflow user can vote the post either up or down, and these votes are aggregated into a "score" on the answer. The simplest and most obvious use of this is that posts that score higher are more visible. Users are encouraged to participate and vote through the Reputation system. Therefore, generally speaking, a highly-voted post will be a high-quality, useful post as judged by the Stack Overflow community while a low (or negative) scoring post has been judged to be of little or no value. On the other hand, the accepted answer system is only available to question askers. The asker of a question can choose to mark any answer to his question as the "accepted answer", causing it to appear at or near the top of the answer list for his question, along with the highly-voted posts. Practically, this is very similar to the voting system, though there is the key difference that the only person who determines the accepted answer is the question asker, instead of the whole community.

1. **Problem Definition and Research Questions**

Our problem focuses specifically on the answers that have been deemed to be high quality. Specifically, we will be examining answers with a score of at least 14 and the set of accepted answers. These answers will be evaluated against a set of metrics to answer the question of "What makes for a good answer on Stack Overflow?”. The accepted answers and the highly-voted answers will, nominally, comprise two different data sets, but both are being measured in the same way. To this end, our primary goal is to answer the following two questions: “What factors contribute to Accepted Answers?” and “How do the factors that contribute to Accepted Answers differ from those that contribute to Highly Voted Answers?”

1. **Sources of Information**

The information we use in our analysis is derived from two sources – highly voted answers and accepted answers.

1. **Highly Voted Answers**

Highly voted answers are a subset of all submitted Stack Overflow answers which contains only the highest rated answers as voted by the community. To obtain this subset, we first look at the user-voted rating of each answer and determine a rating threshold derived from the average score. The set of highly voted answers is then defined as all answers that exceed the average score.

1. **Accepted Answers**

Accepted answers are answers to questions which the asker felt most adequately answered the question at hand. Contrary to highly voted answers, accepted answers do not have to meet a certain score threshold and, in extraordinary cases, may even be the least preferred answer as rated by the community.

1. **Input and Output/Methodology**

The metrics we will be evaluating are "Response Time", "Presence of code snippets in the answer", “Number of Nouns in the answer”, “Number of Verbs in the answer”, “Number of Adjectives in the answer”, “Number of words in the answer”, and "Presence of links in the answer". We will extract our answer set from a SQL Database and then run the metrics on each answer, generating data that we can run statistics on. Our primary tool for these statistics is the R programming language. The primary output will be in the form of bar charts and graphs to conduct analysis and draw conclusions from the statistics. We can then put forward an analysis and conclusion stating how to analyze the quality of an answer to build a higher reputation and have your answer accepted and highly voted on Stack overflow.

1. **Threats to Validity**
2. We looked at the top rated answers overall. Due to the nature of Stack Overflow, it is likely that the pool of top rated answers is not distributed evenly across questions and many answers could have come from the same question. This means that there could have been a better written answer with lower visibility while our analysis instead took a less quality response because it had higher visibility.
3. **Related Papers**