Improving Crowdsourced Documentation:

Examining Answers on Stack Overflow

Scott Crain, Niranjana Gajjelli, and Marc Hudak

I. Problem Definition

Our problem focuses specifically on the answers that have been deemed to be high quality. Specifically, we will be examining answers above a certain, as-yet-undetermined score threshold 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. Our primary goal is to find trends in these metrics across the data, so that we can identify aspects of the "good" answers and, perhaps, be able to apply these trends to the general problem of writing better documentation, both in crowdsourced form and otherwise.

II. 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.

2) Accepted Answers

Accepted answers are answers to questions which the asker felt most adequately answered the prompt. 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.

III. Input and Output/Methodology

The metrics we will be evaluating are "Response Time", "Presence of code snippets in the answer", “Length of the answer (in words)", "Reputation of the answerer", 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 put into an Excel sheet. Excel will aid us in running statistics on the data. 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.