Introduction

Stack Overflow Question Quality

Ryan Tobin, James Westbrook

The problem we are interested in is "can we infer the qualities of a good Stack Overflow question?" We will attempt to answer this question by classifying existing questions into two categories: 'good', and 'bad'. For our purposes, we will define a 'good' question as one which has positive votes, and a 'bad' question as one which does not. With this simplistic labeling approach, we can construct a model to *classify* questions as good or bad by analyzing important features. Though such a model could be used to predict whether unpublished questions will be deemed good or bad, this is not what we are interested in. Instead, this is an *inference* problem - we want to understand the features of a good Stack Overflow question so that we as humans can ask better questions. Though there are general guidelines provided by Stack Overflow about what constitutes a good question, such as providing a minimal reproducible example and explaining steps already taken to attempt to solve the problem, it would be interesting to both verify that these features actually do make for 'good' questions, and perhaps we could find new features which make a question 'good'. We expect that a successful model can predict whether a Stack Overflow question is good or bad at or above a 90% success rate (arbitrarily set).

We consider this problem to be important because it is grasping at the idea of using computers to enhance human communication. In particular, we want to model proper question-asking. If a model for good Stack Overflow questions exists, the most obvious implication is that it could be analyzed to help Stack Overflow users ask better questions and get better answers to questions, but there is no reason to believe that this model couldn't generalize beyond Stack Overflow. For example, perhaps we could also model good search engine queries. Regardless, this is a potential step towards making learning more efficient.

As for how we will collect our data, we can see a table of all Stack Overflow question IDs at Stack Overflow Data (kaggle.com). Unfortunately, this table doesn't contain much information on the posts themselves, so obtaining usable data requires significant (but reasonable) web scraping using the Stack Overflow API.