Constructing Better Questions: A Study of Closed Questions on Stack Overflow

Team Discover: Robert Baker, Khalida Hendricks, Aniket Shah, and Jessica Valenti

Stack Overflow is a popular website where users can ask and answer questions on a variety of coding topics. In general, once questions are answered they are closed and no new answers can be provided. Unanswered questions remain open for responses, but in some cases questions are never answered. Understanding what makes a question likely to be answered would be tremendously helpful knowledge for Stack Overflow users that ask questions on this website.

Using Stack Overflow question and answer data (provided via Kaggle) and Natural Language Processing (NLP), we predicted whether questions would be open or closed based on the words used in the questions. To make the raw data usable, we cleaned the text of each post and extracted important words. These important words were fed into a bag-of-words model and we implemented a logistic regression model to predict if a previously unseen question would be open or closed. Regardless of the characteristics of the training data, we found a prediction accuracy of greater than 50%, meaning that even this simple model can help identify which word choices will lead to a question receiving satisfactory answers. We further identified a strong dependence of the prediction accuracy on the characteristics of the underlying dataset. That is, if the training data is heavily weighted toward closed or open questions, the prediction accuracy artificially increases without actually providing any further predictive power.

The results of this analysis lead to two recommendations:

1) Implementing even a simple model can help predict how best to ask a question such that it will receive a satisfactory answer. Doing so will save time so that Stack Overflow users can make progress on other projects without having to devote more time to debugging or troubleshooting their code.
2) Analyzing the characteristics of the training data can highlight overconfidence in the model predictions. As such, to maximize the reliability of our predictions, adequate time and resources should be devoted to ensure we have a good sample of training data.