Ishan Handa
140905248
ML Group 1

Raamish Malik
140905194
ML Group 1

Predicting the Scope of StackOverflow Questions

3rd February 2017

Abstract

Millions of programmers use StackOverflow to get high quality answers to their programming questions every day. There has evolved an effective culture of moderation to safe-guard it. More than six thousand new questions is asked on StackOverflow every weekday. StackOverflow’s core mission is to create an online encyclopedia for all programming knowledge. In order to ensure quality content in the face of rapid growth, community moderators frequently close low quality questions, often asked by newcomers. Currently about 6% of all new questions end up ”closed”.

In other cases, questions that reflect a strong effort but a lack of familiarity with StackOverflow’s guidelines are still met with the same level of hostility. A 2013 study found that 77 percent of users only ask one question and 65 percent of users only answer one question, a direct consequence of this hostile culture and lack of community engagement for newcomers. Ultimately, this cultural epidemic originates from a conflict of interest between older users trying to prevent the dilution of quality content, trying to enforce the rules and unfamiliar newcomers trying to get their programming questions answered. Thus, our goal is to bridge this community divide and remedy the user experience for new users.

The goal of this paper is to build a classifier that predicts whether or not a question will be closed given the question as submitted, along with the reason that the question was closed.

Introduction

In recent time question-answer services like StackOverflow are becoming more popular. Knowledge of such services has been steadily growing so it requires more resources to moderate. Some automation of this process would ease this task. The problem solved in this paper is a small step in this direction. In order to alleviate moderator burden and ease newcomers' transition, we devise two classifiers to predict

1) whether a question will be closed and if close
2) its reason for closure.

We train our models using logistic regression, K-Nearest Neighbours, SVMs, neural networks and boosting before selecting the optimal classifier. Questions on StackOverflow can be closed as off topic (OT), not constructive (NC), not a real question (NRQ), too localized (TL) or exact duplicate. Exact duplicate reason was excluded from our work because it depends on posts history. Posts history actually is present in StackOverflow database dump but its size is about 6GB in xml format, which requires many resources to analyze.

Off topic is a question that is not on-topic of the site or is related to another site in Stack Exchange Network.

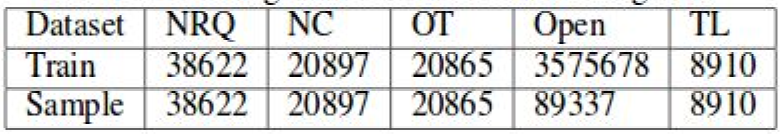
Too localized is a question that is unlikely to be helpful for anyone in the future; it is only relevant to a small geographic area, a specific moment in a time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet.

Not constructive is a question that is not a good fit to Q&A format. It is expected that the answers generally involve facts, references, or specific expertise; this question will likely solicit opinion, debate, arguments, polling, or extended discussion. Not a real question is a question when it's difficult to tell what is being asked here. This question is ambiguous, vague, incomplete, overly broad or rhetorical and cannot be reasonably answered in its current form.

Dataset

For this task the data, with very basic features, was provided by kaggle and it includes train data which contains 3664927 posts and train sample data consisting of 178 351 posts. Full train data and sample train data distribution on closed reasons is shown in table 1.

Table 1: Training data distribution over categories



Dataset	NRQ	NC	OT	Open	TL
Train	38622	20897	20865	3575678	8910
Sample	38622	20897	20865	89337	8910

Our dataset extracted from StackOverflow DataExplorer contains features like -

Badges- Badges is a user activity rewarding system on Stack Overflow.

Comments- This file contains information about who and when commented on the post.

Post History- Contains information about posts' history.

Posts- Contains information about all questions and answers such as creation date, scores, best answer and so on.

Users- Contains user characteristics such as ag e, reputation, personal profile information and so on.

Votes- Contains information about all user vote s – who voted and why.

The final data frame will have questions-answers along with the tags aforementioned which will give us an insight into each record. This will help us to visualise data better and achieve better accuracy with our models.