

Prediction of 'Closed Questions' on Stack Overflow

Submitted by:

Priya Goyal (10535)
Ayush Mittal (11183)
IIT Kanpur

Advisor:

Prof. Amitabha Mukerjee
IIT Kanpur

Problem and Motivation:

Stack Overflow is widely regarded as the most popular Community driven Question Answering (CQA) website for *programmers*. Stack Overflow strongly emphasises question-answer based format instead of chit-chat/ discussions. Questions on the topics which contain specific algorithms, coding techniques are recommended whereas opinion-based questions/ questions which have tendency to generate discussions rather than answers (e.g. polls) are discouraged. Questions which don't follow the pre-defined set of guidelines are marked '*closed*' (*duplicate, off-topic, not a real question, subjective, too localized*) via a community based-voting system.

Questions are an integral part of Stack Overflow system and hence, quality control of questions plays a significant role in its functioning and popularity. However, with the exponential increasing user base and more than six thousand new questions asked on Stack Overflow every weekday, only 6% of the questions end up being closed. Due to *decrease* in the community voting in closing questions, the workload on the moderators has increased. So in this project we aim to build a classifier which predicts whether or not a question will be closed given the question as submitted, along with the reason that the question was closed.

Previous Work:

Shah et al. propose a classification model with features based on human assessed aspects and question answer meta information to predict answer quality on Yahoo! Answers CQA [3]. Sakai et al. propose evaluation methods based on graded-relevance IR metrics to find the best answers on Yahoo! Chiebukuro (Japanese Yahoo! Answers) [4]. [5] Li et al. analyse factors affecting question quality and propose a Mutual Reinforcement-based Label Propagation approach to predict question quality in Yahoo! Answers. However these approaches focus on the answer quality on CQA sites but it has been proved that the answer quality directly depends on the question quality. The focus of this project is on question quality rather than answer quality.

Dataset:

We obtain the dataset from an online competition site Kaggle.com. The data is very huge with size of around 4 GB. The train data contains both the closed questions and open questions of roughly same size. The data consists of post text and associated metadata at the time of post creation. Metadata is Tags, title, owner creation date, owner user id etc. This will serve as our input.

Approach:

We treat this problem as a binary classification problem and it falls in the category of information retrieval and search. We built upon the initial approach provided in [6]. An important part of this problem is to identify which features to use for model building out of the given user features, Tag features and post features. For this we will look at the data and try to come up with the most important features. We will then use the machine learning methods to build a classifier: SVM, Random forest, ensembling algorithms etc. are some that we intend to use. There is also a new

library for feature learning 'Vowpal Wabbit' [7] which was developed at Yahoo! by John Langford. This library is handy in dealing with large data, important features and it is a *fast* learner.

REFERENCES

- [1] Why are some questions closed, and what does "closed" mean?
<http://stackoverflow.com/help/closed-questions>
- [2] What is a day in life of a stack overflow moderator?
<http://meta.stackoverflow.com/a/166630/214223> February 2013.
- [3] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 411–418. ACM, 2010.
- [4] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community qa answer selection. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 187–196. ACM, 2011.
- [5] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest q&a site in the west. In Proceedings of the 2011 annual conference on Human factors in computing systems, pages 2857–2866. ACM, 2011.
- [6] <http://fastml.com/predicting-closed-questions-on-stack-overflow/>
- [7] https://github.com/JohnLangford/vowpal_wabbit/wiki
- [8] <https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/data>