TRINITY COLLEGE DUBLIN

Abstract

Integrated Computer Science
Computer Science and Statistics

Masters of Computer Science

Classifying the Quality of Questions and Answers From Stack Overflow

by Geoffrey Hodgins

This dissertation aims to discover indicators of quality, and to use this knowledge to correctly classify the quality of questions and answers from Stack Overflow. The proliferation of technical questions and answers on Q&A websites such as Stack Overflow means there is more information available than ever. However, the ease of publishing such information also tends to mean the quality varies significantly. The job of moderating Stack Overflow is left to the community. Stack Overflow performs some basic quality analysis, but this is an area where improvement would have many benefits to not only Stack Overflow, but many other domains where the quality of text is important.

Machine Learning techniques are leveraged to discover empirical evidence and anecdotal insight. Specifically, Random Forests are used due to their white-box nature, and features such as readability indexes and similarity measures were engineered from raw text data. Promising quality classification performance is found, along with interesting insights from analyzing the Random Forest classifier. Based on these results, a number of areas for further work are outlined that would greatly benefit this research area. These suggestions include improving tools for technical text data, and content analysis.