Understanding important factors to identify good and bad questions

A case study of StackOverflow

Jirayus Jiarpakdee, Akinori Ihara, Ken-ichi Matsumoto
E-mail : (jirayus.jiarpakdee.jc6, akinori-i, matumoto)@is.naist.jp

Problem Statement

People read, answer and discuss in **good** question

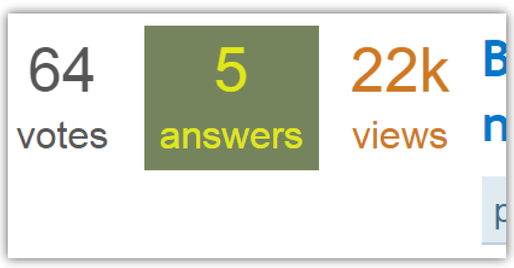


Figure 1 : **Good question**

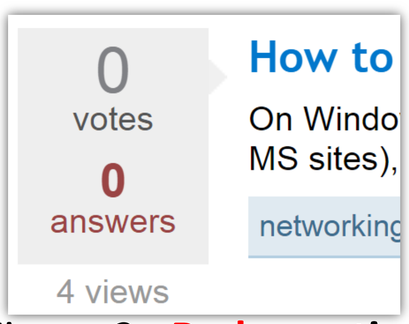


Figure 2 : **Bad question**

Nobody cares about **bad** question

What makes good or bad questions ?

Contribution

Goal : To understand the important factors of good and bad questions

Preprocessing Stage

Questions with Java Tag
1000 **Good** Questions
1000 **Bad** Questions

stackoverflow

① Feature Extraction

Common Features [1]

- **Textual**
- **Community-related**

② Feature Analysis

New Features

- **Textual**
- **Community-related**
- **…**

③ Create new feature from highly correlated features

④ Add new feature

Clustering Stage

Hierarchical Clustering

⑤

Result	None		nSentence/ContentLength	
	C1	C2	C1	C2
Good	569	475	542	458
Bad	855	101	899	101
Recall Good	0.45498		0.458	
Recall Bad	0.89435		0.899	

	nTag	lContent	lTitle	nURL	ARI	Coleman	FKincaid	FReading	GFox	SMOG	nSentence	nWord	LOC	Reputation
nTag	1.000	0.223	0.120	0.126	0.142	0.117	0.150	-0.137	0.123	0.139	0.247	0.257	0.082	0.035
lContent	0.223	1.000	0.158	0.333	0.280	0.245	0.271	-0.251	0.234	0.270	0.728	0.843	0.750	0.088
lTitle	0.120	0.158	1.000	0.053	0.104	0.128	0.095	-0.098	0.080	0.100	0.119	0.157	0.057	0.017
nURL	0.126	0.333	0.053	1.000	0.139	0.211	0.126	-0.155	0.111	0.149	0.288	0.326	0.035	0.052
ARI	0.142	0.280	0.104	0.139	1.000	0.643	0.981	-0.855	0.949	0.824	0.004	0.318	0.117	0.026
Coleman	0.117	0.245	0.128	0.211	0.643	1.000	0.626	-0.835	0.624	0.733	0.137	0.249	0.067	0.059
FKincaid	0.150	0.271	0.095	0.126	0.981	0.626	1.000	-0.900	0.968	0.858	0.000	0.311	0.109	0.028
FReading	-0.137	-0.251	-0.098	-0.155	-0.855	-0.835	-0.900	1.000	-0.888	-0.889	-0.046	-0.268	-0.086	-0.041
GFox	0.123	0.234	0.080	0.111	0.949	0.624	0.968	-0.888	1.000	0.915	-0.021	0.275	0.084	0.032
SMOG	0.139	0.270	0.100	0.149	0.824	0.733	0.858	-0.889	0.915	1.000	0.077	0.316	0.080	0.056
nSentence	0.247	0.728	0.119	0.288	0.004	0.137	0.000	-0.046	-0.021	0.077	1.000	0.879	0.294	0.074
nWord	0.257	0.843	0.157	0.326	0.318	0.249	0.311	-0.268	0.275	0.316	0.879	1.000	0.363	0.080
LOC	0.082	0.750	0.057	0.035	0.117	0.067	0.109	-0.086	0.084	0.080	0.294	0.363	1.000	0.039
Reputation	0.035	0.088	0.017	0.052	0.026	0.059	0.028	-0.041	0.032	0.056	0.074	0.080	0.039	1.000

Figure 3 : **Good** Questions Correlation Coefficient

	nTag	lContent	lTitle	nURL	ARI	Coleman	FKincaid	FReading	GFox	SMOG	nSentence	nWord	LOC	Reputation
nTag	1.000	0.010	0.071	0.080	0.024	0.067	0.040	-0.066	0.041	0.120	0.153	0.160	-0.017	-0.055
lContent	0.010	1.000	-0.020	0.031	0.074	0.019	0.063	-0.033	0.054	0.007	0.184	0.228	0.913	-0.034
lTitle	0.071	-0.020	1.000	0.011	0.043	0.057	0.053	-0.081	0.061	0.100	-0.013	0.013	-0.022	0.015
nURL	0.080	0.031	0.011	1.000	0.051	0.116	0.046	-0.068	0.053	0.107	0.133	0.153	-0.021	-0.022
ARI	0.024	0.074	0.043	0.051	1.000	0.535	0.981	-0.883	0.938	0.685	-0.256	0.282	0.028	-0.044
Coleman	0.067	0.019	0.057	0.116	0.535	1.000	0.471	-0.714	0.399	0.550	-0.036	0.130	-0.029	-0.030
FKincaid	0.040	0.063	0.053	0.046	0.981	0.471	1.000	-0.904	0.969	0.728	-0.252	0.294	0.016	-0.049
FReading	-0.066	-0.033	-0.081	-0.068	-0.883	-0.714	-0.904	1.000	-0.853	-0.791	0.183	-0.229	0.018	0.048
GFox	0.041	0.054	0.061	0.053	0.938	0.399	0.969	-0.853	1.000	0.795	-0.247	0.292	0.008	-0.041
SMOG	0.120	0.007	0.100	0.107	0.685	0.550	0.728	-0.791	0.795	1.000	-0.058	0.307	-0.057	-0.031
nSentence	0.153	0.184	-0.013	0.133	-0.256	-0.036	-0.252	0.183	-0.247	-0.058	1.000	0.735	0.095	-0.002
nWord	0.160	0.228	0.013	0.153	0.282	0.130	0.294	-0.229	0.292	0.307	0.735	1.000	0.113	-0.010
LOC	-0.017	0.913	-0.022	-0.021	0.028	-0.029	0.016	0.018	0.008	-0.057	0.095	0.113	1.000	-0.033
Reputation	-0.055	-0.034	0.015	-0.022	-0.044	-0.030	-0.049	0.048	-0.041	-0.031	-0.002	-0.010	-0.033	1.000

Figure 4 : **Bad** Questions Correlation Coefficient

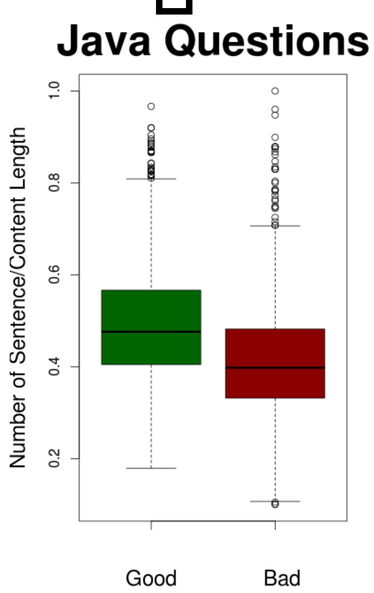


Figure 5 : **New feature in good and bad questions**
p-value < 0.05

After clustering with new feature, we analyze deep into each cluster. Cluster 1 is dominated by good questions, while Cluster 2 is dominated by bad questions. The result shows the same relation with figure 4, good questions tend to have higher number of sentences within the same content length.

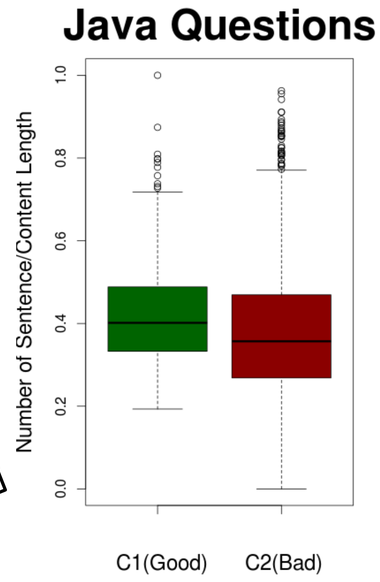
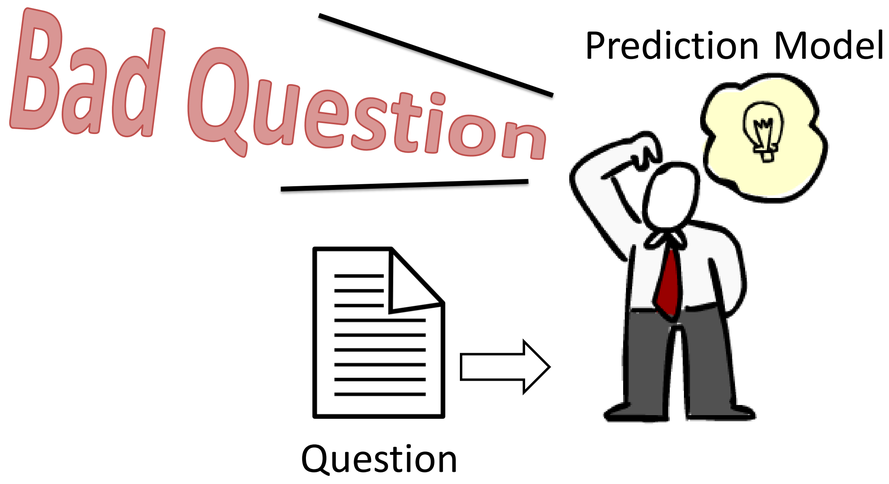


Figure 6 : New features in Cluster

Future Work

To be able to build a prediction model, beside from common features, we want to create new features which will improve the model quality. We are interesting in tag correctness as discussed in [2], one of the important factor which makes question bad is incorrect tagging. Also, meaningfulness of the content, we want to use NLP technique to extract a feature which reflect content's meaningfulness.



[1] L. Ponzanelli, A. Mocci, A. Bacchelli, and M. Lanza, "Understanding and Classifying the Quality of Technical Forum Questions," in In Proceedings of QSIC 2014

[2] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider "Answering Questions about Unanswered Questions of stack Overflow," Proc. MSR, 2013, pp. 97–100.