Fit or Unfit : Analysis and Prediction of Closed Questions on Stack Overflow

Denzil Correa, Ashish Sureka

http://correa.in/

October 8, 2013

Table of Contents

1. Introduction
2. Closed Questions on Stack Overflow
3. Characterization Study of Closed Questions
4. Closed Question Prediction
5. Conclusion

Table of Contents

1. Introduction
2. Closed Questions on Stack Overflow
3. Characterization Study of Closed Questions
4. Closed Question Prediction
5. Conclusion

Research Motivation and Aim

- Stack Exchange – A platform to deploy topic-based community powered Q&A websites
- 103 Q&A sites on diverse topics like code review, parenting, bicycles and audio-video production
- Stack Overflow is the first and most popular Stack Exchange website which caters to professional programmers and programming enthusiasts
- Users earn *reputation* points which is a reflection of their contribution worth to the Stack Overflow community
- Community based voting process to reward(or penalize) *reputation* of users based on question and answer quality

Research Motivation and Aim

- Stack Exchange – A platform to deploy topic-based community powered Q&A websites
- 103 Q&A sites on diverse topics like code review, parenting, bicycles and audio-video production
- Stack Overflow is the first and most popular Stack Exchange website which caters to professional programmers and programming enthusiasts
- Users earn *reputation* points which is a reflection of their contribution worth to the Stack Overflow community
- Community based voting process to reward(or penalize) *reputation* of users based on question and answer quality

Research Motivation and Aim

- Stack Exchange – A platform to deploy topic-based community powered Q&A websites
- 103 Q&A sites on diverse topics like code review, parenting, bicycles and audio-video production
- Stack Overflow is the first and most popular Stack Exchange website which caters to professional programmers and programming enthusiasts
- Users earn *reputation* points which is a reflection of their contribution worth to the Stack Overflow community
- Community based voting process to reward(or penalize) *reputation* of users based on question and answer quality

Research Motivation and Aim

- Stack Exchange – A platform to deploy topic-based community powered Q&A websites
- 103 Q&A sites on diverse topics like code review, parenting, bicycles and audio-video production
- Stack Overflow is the first and most popular Stack Exchange website which caters to professional programmers and programming enthusiasts
- Users earn *reputation* points which is a reflection of their contribution worth to the Stack Overflow community
- Community based voting process to reward(or penalize) *reputation* of users based on question and answer quality

Research Motivation and Aim

- Stack Exchange – A platform to deploy topic-based community powered Q&A websites
- 103 Q&A sites on diverse topics like code review, parenting, bicycles and audio-video production
- Stack Overflow is the first and most popular Stack Exchange website which caters to professional programmers and programming enthusiasts
- Users earn *reputation* points which is a reflection of their contribution worth to the Stack Overflow community
- Community based voting process to reward(or penalize) *reputation* of users based on question and answer quality

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized*.
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized.*
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized.*
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized.*
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized*.
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Stack Overflow is a free, open website to all Internet users
- Maintains strong emphasis on question-answer based format and strongly discourages discussion or *chit-chat*
- Homework, product or service recommendations, non- programming related questions and polls are frowned upon
- Questions which do not follow guidelines viz. are low quality or irrelevant are marked as *closed*
- A question can be marked as *closed* for five reasons – *duplicate, off-topic, subjective, not a real question* and *too localized*.
- Decision to *close* a question is taken by experienced users and community moderators via a systematic voting process.

Research Motivation and Aim

- Despite the existence of vibrant experienced users and self-motivated community moderators, Stack Overflow faces a continuous challenge to maintain content quality
- Important to analyze and study the phenomena of 'closed' questions in order to gain insights on question quality
- The goal of Stack Overflow is to have a knowledge base of question-answers on programming related topics
- A *closed* question is a direct feedback to the question asker that her question may be unfit or needs improvement in its current form
- A system to predict *closed* questions would help both the users and community moderators

Research Motivation and Aim

- Despite the existence of vibrant experienced users and self-motivated community moderators, Stack Overflow faces a continuous challenge to maintain content quality
- Important to analyze and study the phenomena of ‘closed’ questions in order to gain insights on question quality
- The goal of Stack Overflow is to have a knowledge base of question-answers on programming related topics
- A *closed* question is a direct feedback to the question asker that her question may be unfit or needs improvement in its current form
- A system to predict *closed* questions would help both the users and community moderators

Research Motivation and Aim

- Despite the existence of vibrant experienced users and self-motivated community moderators, Stack Overflow faces a continuous challenge to maintain content quality
- Important to analyze and study the phenomena of ‘closed’ questions in order to gain insights on question quality
- The goal of Stack Overflow is to have a knowledge base of question-answers on programming related topics
- A *closed* question is a direct feedback to the question asker that her question may be unfit or needs improvement in its current form
- A system to predict *closed* questions would help both the users and community moderators

Research Motivation and Aim

- Despite the existence of vibrant experienced users and self-motivated community moderators, Stack Overflow faces a continuous challenge to maintain content quality
- Important to analyze and study the phenomena of ‘closed’ questions in order to gain insights on question quality
- The goal of Stack Overflow is to have a knowledge base of question-answers on programming related topics
- A *closed* question is a direct feedback to the question asker that her question may be unfit or needs improvement in its current form
- A system to predict *closed* questions would help both the users and community moderators

Research Motivation and Aim

- Despite the existence of vibrant experienced users and self-motivated community moderators, Stack Overflow faces a continuous challenge to maintain content quality
- Important to analyze and study the phenomena of ‘closed’ questions in order to gain insights on question quality
- The goal of Stack Overflow is to have a knowledge base of question-answers on programming related topics
- A *closed* question is a direct feedback to the question asker that her question may be unfit or needs improvement in its current form
- A system to predict *closed* questions would help both the users and community moderators

Outline of Contributions, Results

1. We present the first characterization study of *closed* questions on Stack Overflow.
2. Question content, answer patterns and temporal trend analysis of 'closed' question.
3. Observations on community participation trends towards 'closed' questions as well as analyze information quality indicators on *closed* questions.
4. We build a predictive model for *closed* question prediction
5. We perform feature analysis and report discriminative features to differentiate *closed* questions from non-*closed* questions

Outline of Contributions, Results

1. We present the first characterization study of *closed* questions on Stack Overflow.

2. Question content, answer patterns and temporal trend analysis of ‘closed‘ question.

3. Observations on community participation trends towards ‘closed’ questions as well as analyze information quality indicators on *closed* questions.

4. We build a predictive model for *closed* question prediction

5. We perform feature analysis and report discriminative features to differentiate *closed* questions from non-*closed* questions

Outline of Contributions, Results

1. We present the first characterization study of *closed* questions on Stack Overflow.
2. Question content, answer patterns and temporal trend analysis of ‘closed‘ question.
3. Observations on community participation trends towards ‘closed’ questions as well as analyze information quality indicators on *closed* questions.
4. We build a predictive model for *closed* question prediction
5. We perform feature analysis and report discriminative features to differentiate *closed* questions from non-*closed* questions

Outline of Contributions, Results

1. We present the first characterization study of *closed* questions on Stack Overflow.

2. Question content, answer patterns and temporal trend analysis of 'closed' question.

3. Observations on community participation trends towards 'closed' questions as well as analyze information quality indicators on *closed* questions.

4. We build a predictive model for *closed* question prediction

5. We perform feature analysis and report discriminative features to differentiate *closed* questions from non-*closed* questions

Outline of Contributions, Results

1. We present the first characterization study of *closed* questions on Stack Overflow.
2. Question content, answer patterns and temporal trend analysis of 'closed' question.
3. Observations on community participation trends towards 'closed' questions as well as analyze information quality indicators on *closed* questions.
4. We build a predictive model for *closed* question prediction
5. We perform feature analysis and report discriminative features to differentiate *closed* questions from non-*closed* questions

Table of Contents

1. Introduction
2. Closed Questions on Stack Overflow
3. Characterization Study of Closed Questions
4. Closed Question Prediction
5. Conclusion

What and Who?

What is a Closed Question?

- Deemed unfit for its Q&A format
- No answers but edits on previously posted question-answers and comments are permitted
- Question-answers can also be voted upon and are counted towards reputation points of users

Who can Close a question?

- Experienced users and community moderators can cast a vote
- Users with 3,000+ reputation points and elected community moderators (also called **Diamond** moderators)
- Users with at least 250 reputation points can vote to 'close' their own question

What and Who?

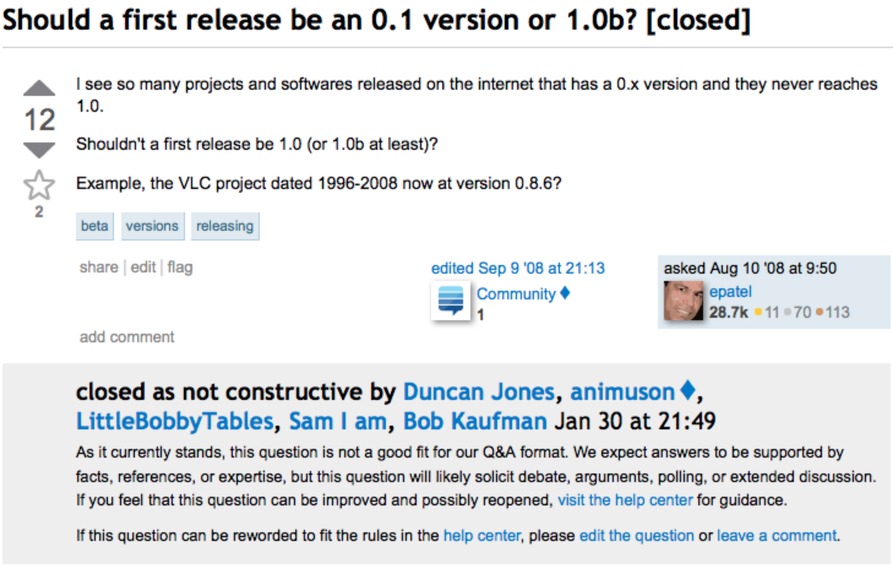
What is a Closed Question?

- Deemed unfit for its Q&A format
- No answers but edits on previously posted question-answers and comments are permitted
- Question-answers can also be voted upon and are counted towards reputation points of users

Who can Close a question?

- Experienced users and community moderators can cast a vote
- Users with 3,000+ reputation points and elected community moderators (also called **Diamond** moderators)
- Users with at least 250 reputation points can vote to ‘close’ their own question

Stack Overflow Closed Question – Example



Should a first release be an 0.1 version or 1.0b? [closed]

12

2

I see so many projects and softwares released on the internet that has a 0.x version and they never reaches 1.0.

Shouldn't a first release be 1.0 (or 1.0b at least)?

Example, the VLC project dated 1996-2008 now at version 0.8.6?

edited Sep 9 '08 at 21:13
Community ♦
1

asked Aug 10 '08 at 9:50
epatel
28.7k ● 11 ● 70 ● 113

closed as not constructive by Duncan Jones, animuson ♦, LittleBobbyTables, Sam I am, Bob Kaufman Jan 30 at 21:49

As it currently stands, this question is not a good fit for our Q&A format. We expect answers to be supported by facts, references, or expertise, but this question will likely solicit debate, arguments, polling, or extended discussion. If you feel that this question can be improved and possibly reopened, visit the help center for guidance.

If this question can be reworded to fit the rules in the help center, please edit the question or leave a comment.

Why are questions Closed?

- **Exact Duplicate** – Question with exactly the same content as earlier questions on this topic
- **Off Topic** – Unrelated to programming or software development within the scope defined
- **Subjective/Not Constructive** – Questions which will likely solicit debate, arguments, polling, or extended discussion
- **Not a real question** – Ambiguous, vague, incomplete, overly broad, or rhetorical and cannot be reasonably answered in its current form
- **Too localized** – Questions unlikely to help any future visitors; it is only relevant to a small geographic area, a specific moment in time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet

Why are questions Closed?

- **Exact Duplicate –** Question with exactly the same content as earlier questions on this topic
- **Off Topic –** Unrelated to programming or software development within the scope defined
- **Subjective/Not Constructive –** Questions which will likely solicit debate, arguments, polling, or extended discussion
- **Not a real question –** Ambiguous, vague, incomplete, overly broad, or rhetorical and cannot be reasonably answered in its current form
- **Too localized –** Questions unlikely to help any future visitors; it is only relevant to a small geographic area, a specific moment in time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet

Why are questions Closed?

- **Exact Duplicate** – Question with exactly the same content as earlier questions on this topic
- **Off Topic** – Unrelated to programming or software development within the scope defined
- **Subjective/Not Constructive** – Questions which will likely solicit debate, arguments, polling, or extended discussion
- **Not a real question** – Ambiguous, vague, incomplete, overly broad, or rhetorical and cannot be reasonably answered in its current form
- **Too localized** – Questions unlikely to help any future visitors; it is only relevant to a small geographic area, a specific moment in time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet

Why are questions Closed?

- **Exact Duplicate –** Question with exactly the same content as earlier questions on this topic
- **Off Topic –** Unrelated to programming or software development within the scope defined
- **Subjective/Not Constructive –** Questions which will likely solicit debate, arguments, polling, or extended discussion
- **Not a real question –** Ambiguous, vague, incomplete, overly broad, or rhetorical and cannot be reasonably answered in its current form
- **Too localized –** Questions unlikely to help any future visitors; it is only relevant to a small geographic area, a specific moment in time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet

Why are questions Closed?

- **Exact Duplicate –** Question with exactly the same content as earlier questions on this topic
- **Off Topic –** Unrelated to programming or software development within the scope defined
- **Subjective/Not Constructive –** Questions which will likely solicit debate, arguments, polling, or extended discussion
- **Not a real question –** Ambiguous, vague, incomplete, overly broad, or rhetorical and cannot be reasonably answered in its current form
- **Too localized –** Questions unlikely to help any future visitors; it is only relevant to a small geographic area, a specific moment in time, or an extraordinarily narrow situation that is not generally applicable to the worldwide audience of the internet

How and What Happens Next?

How are questions Closed?

- Automatically marked *closed* if it receives 5 *close* votes
- **Diamond** moderator votes are final and binding
- One can only vote once to ‘close’ a question

What happens to a Closed question?

- *Reopened* if improved from its current form
- Similar community based vote procedure as closing
- If questions are very poor in quality, they are deleted

How and What Happens Next?

How are questions Closed?

- Automatically marked *closed* if it receives 5 *close* votes
- **Diamond** moderator votes are final and binding
- One can only vote once to 'close' a question

What happens to a Closed question?

- *Reopened* if improved from its current form
- Similar community based vote procedure as closing
- If questions are very poor in quality, they are deleted

Stack Overflow Closed Question Lifecycle

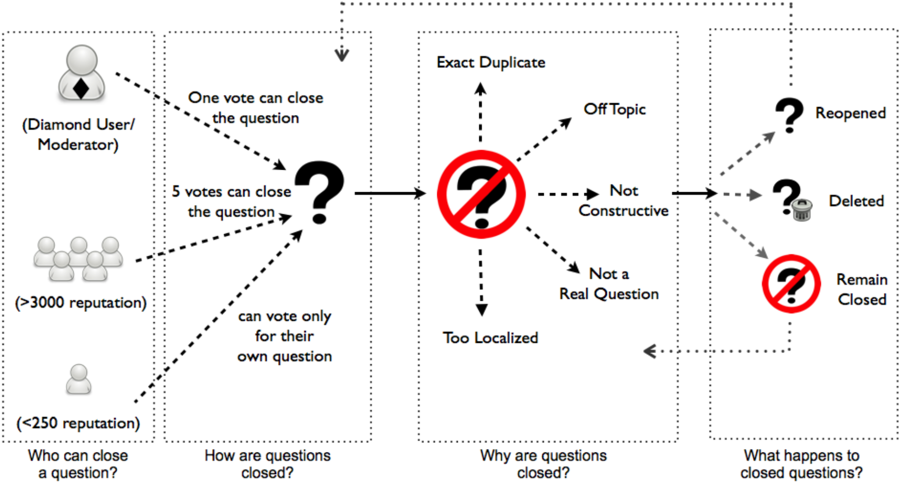


Table of Contents

1 Introduction

2 Closed Questions on Stack Overflow

3 Characterization Study of Closed Questions

4 Closed Question Prediction

5 Conclusion

Stack Overflow Dataset Details

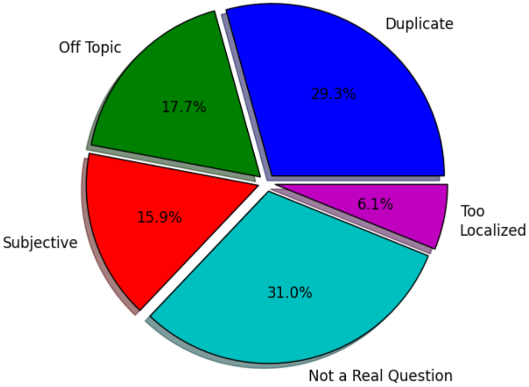
Stack Overflow August 2012 dataset statistics

Users	1.29M (625k askers, 443k answerers)
Questions	3.4M (62.21% with accepted answers)
Answers	6.8M (31.33% marked as accepted)
Votes	27.5M (72.35% positive, 6.81% favorites)
Ratio of Answers to Questions	2.16

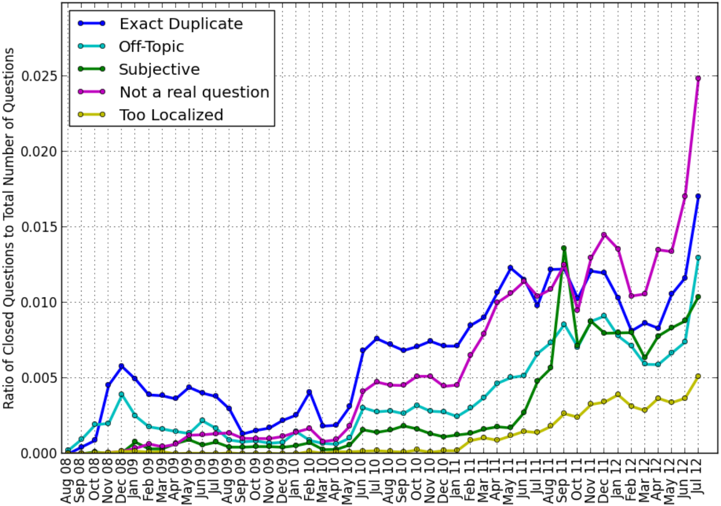
Statistics of ‘Closed Questions’ in Stack Overflow from August 2008 to August 2012.

	2008	2009	2010	2011	2012	Total
Closed Questions	3.8%	1.52%	1.77%	3.33%	3.82%	**102, 993** (2.98%)
Closed Votes	0.03%	0.25%	0.75%	2.21%	3.9%	570,418 (0.2%)
Ratio of Answers to Questions	8.0	5.93	3.11	1.92	1.55	1.92

Closed Question Sub-Category Distribution



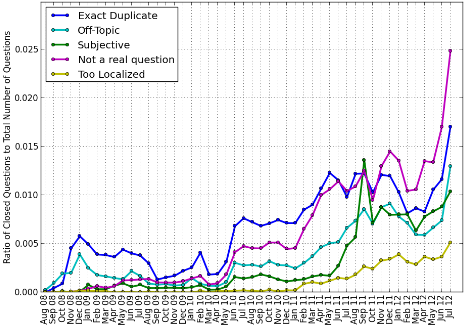
Ratio of Closed Questions – Temporal Distribution Plot



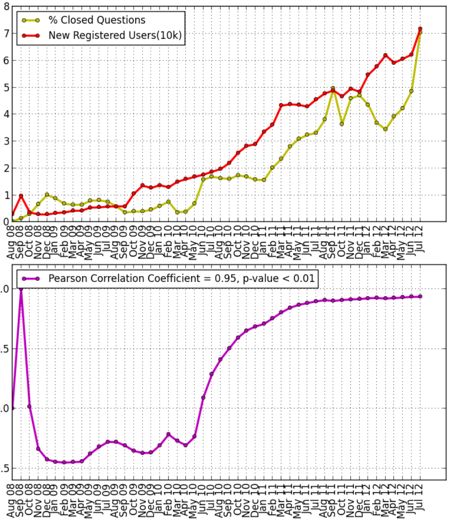
Temporal Distribution Analysis

Key Findings

1. Increase in ratio of *closed* questions over time
2. *Duplicate* and *Not a Real Question* most common categories
3. Sharp increase in ratio after January 2011



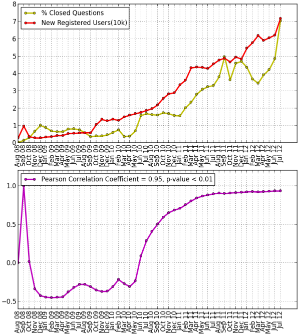
New Registered Users Effect on Closed Question Distribution



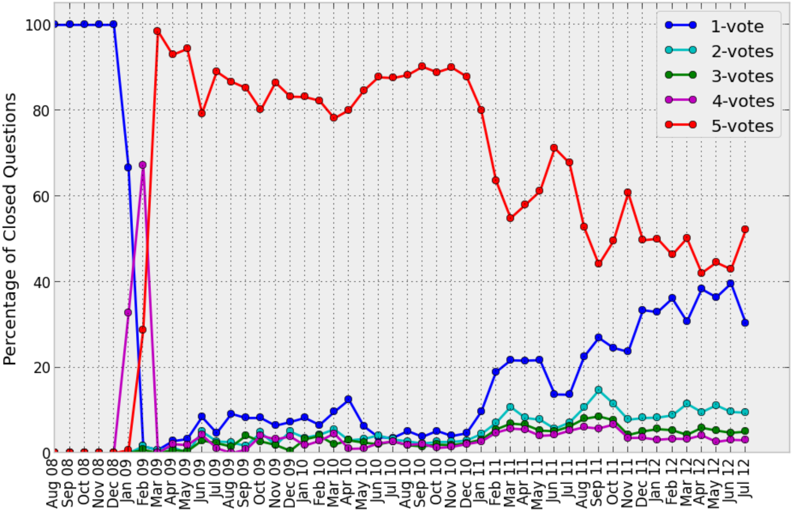
New Registered Users Effect on Closed Question Distribution

Observations

1. High correlation
2. Pearson Correlation Coefficient $=$ 0.95, p-value $<$ 0.01
3. Sharp increase in correlation after January 2011



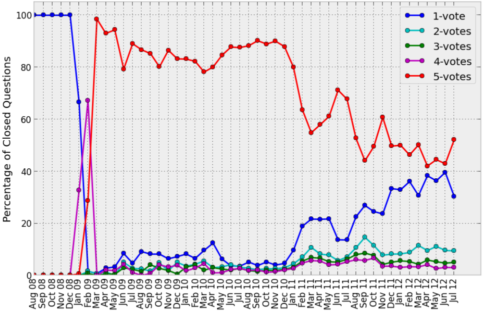
Close Vote Distribution



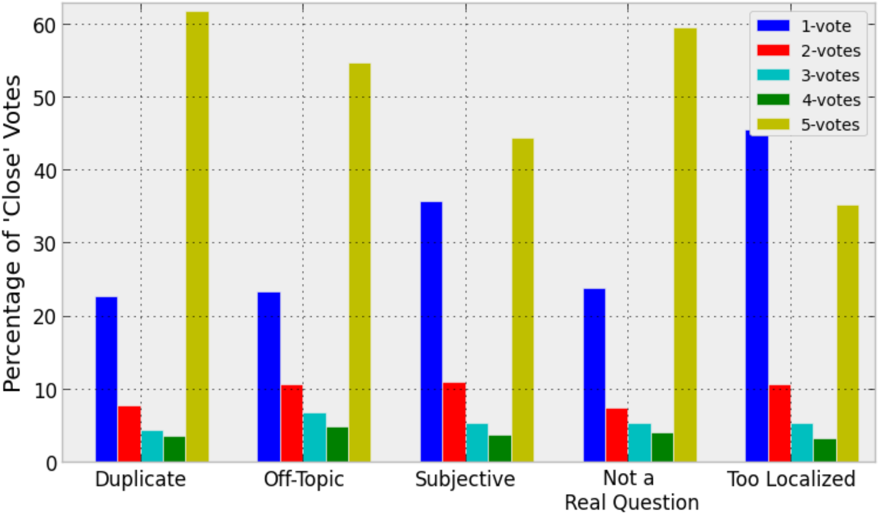
Close Vote Distribution

Observations

1. 26.5% of questions are closed by ONLY **Diamond Moderator**
2. 45% of questions have at least one **Diamond Moderator** vote
3. Increase in **Diamond Moderator** intervention, Decline in community participation



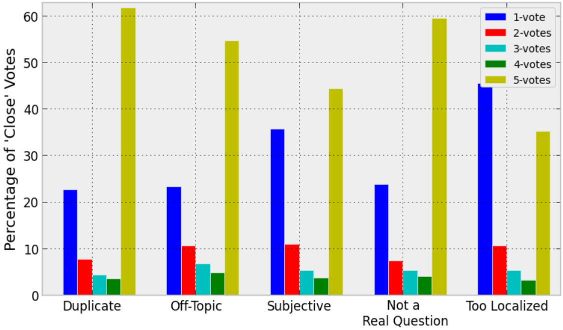
Close Vote Sub-Category Distribution



Close Vote Sub-Category Distribution

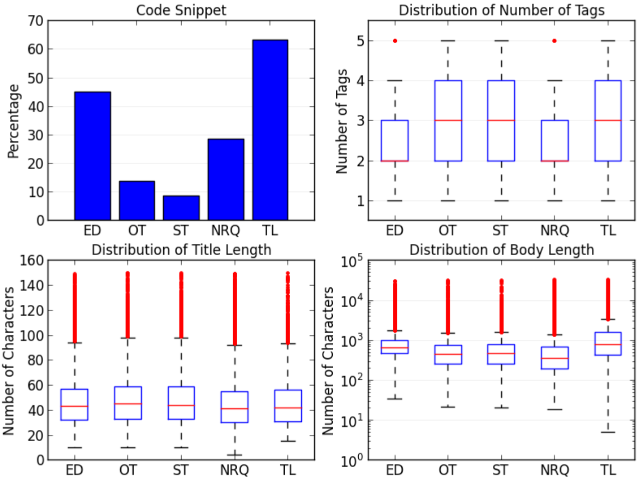
Observations

1. Community participation on *Duplicate*, *Off Topic* and *Not a Real Question*
2. *Subjective* sees equal community-moderator participation
3. *Too Localized* has highest **Diamond moderator** intervention



Question Title, Body and Code Snippet

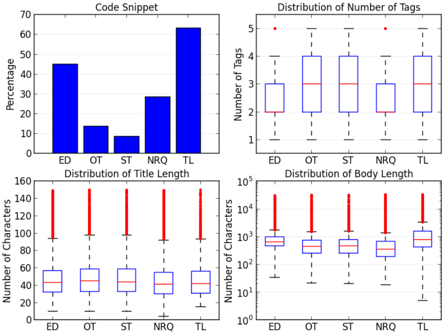
ED = Exact Duplicate, OT = Off-Topic, ST = Subjective, NRQ = Not a Real Question, TL = Too Localized



Question Title, Body and Code Snippet

Observations

1. **31%** of closed questions contain code snippets
2. *Subjective* lowest percentage of source code
3. *Too Localized* has good percentage of source code but are unfit
4. *Not a Real Question* is a clear misfit



Question Topics

Popular Tags in Closed Questions

Type	Tags
Languages	java, c++, python, c, perl, r, . . .
Web2.0	php, html5, html, css, apache, javascript, . . .
Operating Systems	iOS, unix, linux, android, ubuntu, osx, windows, . . .
Social	Facebook, wordpress, google, . . .
Miscellaneous	books, interview-questions, fun, homework, . . .

Question Topics

Popular Tags in Closed Questions

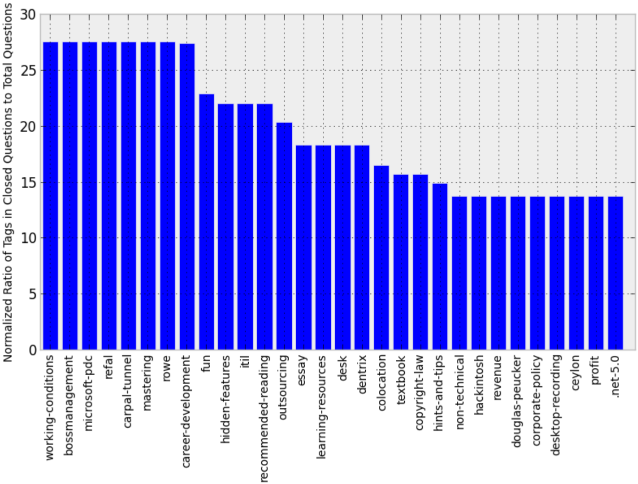
Type	**Tags**
Languages	java, c++, python, c, perl, r, . . .
Web2.0	php, html5, html, css, apache, javascript, . . .
Operating Systems	iOS, unix, linux, android, ubuntu, osx, windows, . . .
Social	Facebook, wordpress, google, . . .
Miscellaneous	books, interview-questions, fun, homework, . . .

Normalized Tag Ratio (NTR)

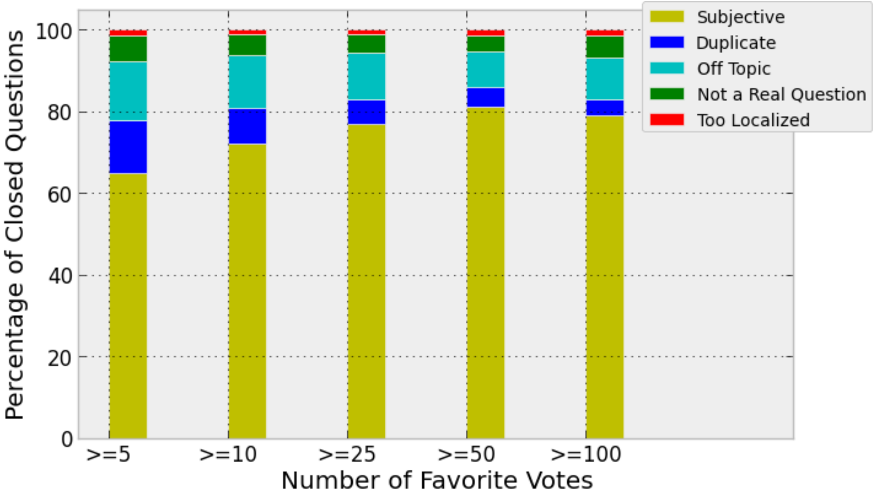
$\forall t_i \in T_{CQ} \quad where \quad t_i \in \{t_1 \dots t_n\}, \quad R^i_{CQ} = \frac{count(t_i)}{\sum_{i=1}^{n} count(t_i)} \forall t_j \in T_{NCQ} \quad where \quad t_j \in \{t_1 \dots t_m\}, \quad R^j_{NCQ} = \frac{count(t_j)}{\sum_{j=1}^{m} count(t_j)}$

$$\therefore \forall t_i \in T_{CQ} \quad NTR_{t_i} = \frac{R^i_{CQ}}{R^i_{NCQ} + \epsilon}, \quad (T_{CQ} \cap T_{NCQ} \neq \emptyset)$$

Discriminative Closed Question Tags



Favorite Vote Distribution



Favorite Vote Distribution

Observations

1. 19% of *closed* questions have at least 1, 3% have at least 5 favorite votes
2. *Subjective* category attracts a very high number of favorite votes from users
3. *What is right is not always popular and what is popular is not always right.* Einstein

Examples : $\geq$ 100 'favorite votes' on closed questions in *subjective* category.

Favorites	Title	Answers	Views
5894	List of freely available programming books	112	569,199
2228	Hidden features of Python	100	212,589
1685	What is the best comment in source code you have ever encountered?	519	1,051,784
421	Worst security hole you've seen?	163	32,840
140	What is the most useful R trick?	34	13,197

Favorite Vote Distribution

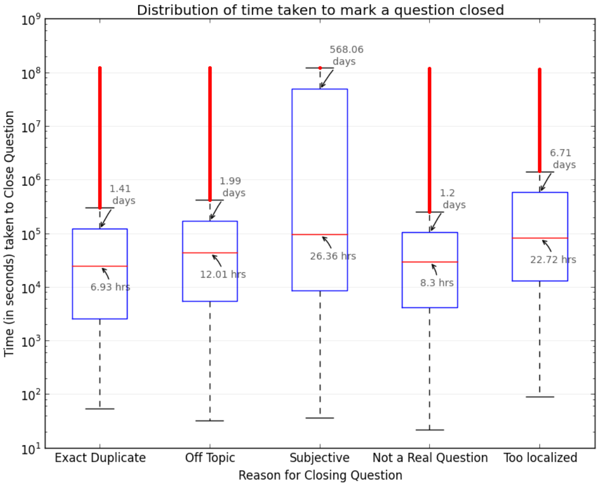
Observations

1. 19% of *closed* questions have at least 1, 3% have at least 5 favorite votes
2. *Subjective* category attracts a very high number of favorite votes from users
3. *What is right is not always popular and what is popular is not always right.* Einstein

Examples : $\geq$ 100 'favorite votes' on closed questions in *subjective* category.

Favorites	Title	Answers	Views
5894	List of freely available programming books	112	569,199
2228	Hidden features of Python	100	212,589
1685	What is the best comment in source code you have ever encountered?	519	1,051,784
421	Worst security hole you've seen?	163	32,840
140	What is the most useful R trick?	34	13,197

Closure Time Analysis



Closure Time Outliers

Category	1-vote	2-vote	3-vote	4-vote	5-vote
Duplicate	**55.44%**	11.68%	4.25%	2.18%	26.45%
Off-Topic	**42.06%**	16.21%	6.31%	3.47%	31.96%
Subjective	**64.64%**	16.66%	4.9%	2.26%	11.54%
Not a Real Question	**46.97%**	9.52%	6.28%	3.5%	33.74%
Too Localized	**68.22%**	11.85%	3.62%	1.86%	14.45%

Observations

1. High percentage of *Diamond moderator* intervention
2. Content value takes time to reach maximum community potential

Closure Time Outliers

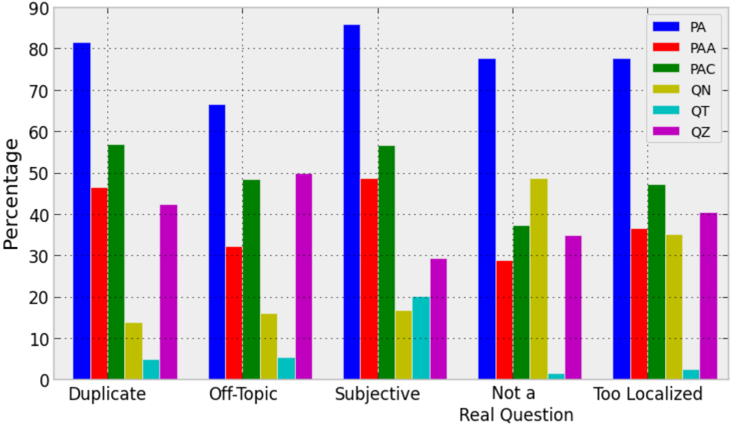
Category	1-vote	2-vote	3-vote	4-vote	5-vote
Duplicate	**55.44%**	11.68%	4.25%	2.18%	26.45%
Off-Topic	**42.06%**	16.21%	6.31%	3.47%	31.96%
Subjective	**64.64%**	16.66%	4.9%	2.26%	11.54%
Not a Real Question	**46.97%**	9.52%	6.28%	3.5%	33.74%
Too Localized	**68.22%**	11.85%	3.62%	1.86%	14.45%

Observations

1. High percentage of *Diamond moderator* intervention
2. Content value takes time to reach maximum community potential

Question Scores and Answer Patterns

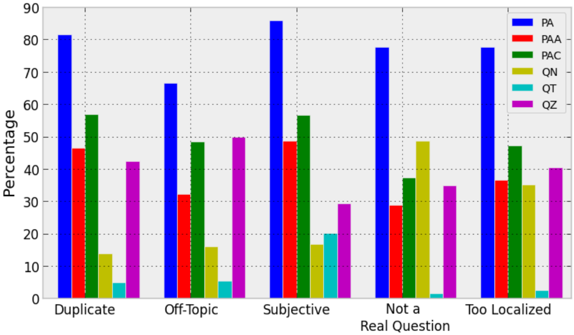
PA = Percentage of Answers, PAA = Percentage of Accepted Answers, PAC = Percentage of Accepted Answers given that a closed question has an answer, QN = Percentage of Questions with Negative Score, QT = Percentage of Questions with Score ≥5, QZ = Percentage of Questions with Zero Score



Question Scores and Answer Patterns

Observations

1. Eagerness to answer or hungry for karma for *Duplicate* questions
2. *Not a Real Question* has a high QN, low QT, high QZ – very poor quality
3. *Subjective* has high QT indicating popularity



Locked, Protected, Community Wiki

Locked

- Can not receive any new answers or any form of votes on question-answers
- Marked by **Diamond Moderator** ONLY to prevent reputation gaming

Protected

- Prevents newly registered users from answering these question
- Prevent noisy answers like “Thank You”, “+1” from new users

Community Wiki

- Donate and transfer ownership of the question to the community
- Helps Stack Overflow to be a knowledge base of quality information

Locked, Protected, Community Wiki

Locked

- Can not receive any new answers or any form of votes on question-answers
- Marked by **Diamond Moderator** ONLY to prevent reputation gaming

Protected

- Prevents newly registered users from answering these question
- Prevent noisy answers like “Thank You”, “+1” from new users

Community Wiki

- Donate and transfer ownership of the question to the community
- Helps Stack Overflow to be a knowledge base of quality information

Locked, Protected, Community Wiki

Locked

- Can not receive any new answers or any form of votes on question-answers
- Marked by **Diamond Moderator** ONLY to prevent reputation gaming

Protected

- Prevents newly registered users from answering these question
- Prevent noisy answers like “Thank You”, “+1” from new users

Community Wiki

- Donate and transfer ownership of the question to the community
- Helps Stack Overflow to be a knowledge base of quality information

Locked, Protected, Community Wiki

	Number of 'Closed' Questions		
Category	**Locked**	**Community Wiki**	**Protected**
Exact Duplicate	732(33.8%)	160(9.9%)	36(10.3%)
Off Topic	**1180(54.5%)**	273(16.8%)	70(20.1%)
Subjective	188(8.7%)	**978(60.3%)**	**202(58%)**
Not Real Question	50(2.3%)	192(11.8%)	28(8%)
Too Localized	114(0.6%)	10(0.6%)	12(3.4%)
Total	2,264	1,613	348

Observations

- *Exact Duplicate* and *Off Topic* most prone to reputation gaming
- *Subjective* questions contain a high number of *community wiki* donations as they are philosophical rather than factual
- High percentage of *Subjective* questions are *protected* due to popularity

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- ≈ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- ≈ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- $\approx$ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- $\approx$ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- ≈ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- ≈ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- ≈ 31% questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Observation Summary

- Increasing trend in percentage of *closed* questions over time
- Positive correlation with a high confidence between new registered users and percentage of *closed* questions %
- Decrease in community participation to mark a question *closed*, Increase in **Diamond Moderator** work load
- Topics on *closed* questions are vague and non-programming related
- *Subjective* category questions are popular
- *Not a Real Question* category questions have very low quality
- $\approx 31\%$ questions have code snippets
- *Too Localized* have high code snippets but not popular
- *Duplicate* and *Off Topic* are very attractive to reputation gamers

Table of Contents

1 Introduction

2 Closed Questions on Stack Overflow

3 Characterization Study of Closed Questions

4 Closed Question Prediction

5 Conclusion

Features for Classification

Set	Category	Number	Features
A	**User Profile**	3	*Age of Account* *Badge Score* *Previous Posts with Negative Score*
B	**Community**	3	*Post Score* *Accepted Answer Score* *Favorite Score*
C	**Question Content**	3	*Number of URLs* *Number of Stack Overflow URLs* *Number of Popular Tags*
D	**Textual Style**	9	*Title Length* *Body Length* *Number of Tags* *Number of Punctuation Marks* *Number of Short Words* *Code Snippet Length* *Number of Special Characters* *Number of Lower Case Characters* *Number of Upper Case Characters*

Badge Score (BS)

$$BS = \sum_{i=0}^{n} \frac{1}{\#\text{users who have } b_i}$$

Features for Classification

Set	Category	Number	Features
A	**User Profile**	3	*Age of Account* *Badge Score* *Previous Posts with Negative Score*
B	**Community**	3	*Post Score* *Accepted Answer Score* *Favorite Score*
C	**Question Content**	3	*Number of URLs* *Number of Stack Overflow URLs* *Number of Popular Tags*
D	**Textual Style**	9	*Title Length* *Body Length* *Number of Tags* *Number of Punctuation Marks* *Number of Short Words* *Code Snippet Length* *Number of Special Characters* *Number of Lower Case Characters* *Number of Upper Case Characters*

Post Score (PS)

$$PS = \sum_{i=0}^{n} score(q_i) + \sum_{j=0}^{m} score(a_i)$$

Features for Classification

Set	Category	Number	Features
A	**User Profile**	3	*Age of Account* *Badge Score* *Previous Posts with Negative Score*
B	**Community**	3	*Post Score* *Accepted Answer Score* *Favorite Score*
C	**Question Content**	3	*Number of URLs* *Number of Stack Overflow URLs* *Number of Popular Tags*
D	**Textual Style**	9	*Title Length* *Body Length* *Number of Tags* *Number of Punctuation Marks* *Number of Short Words* *Code Snippet Length* *Number of Special Characters* *Number of Lower Case Characters* *Number of Upper Case Characters*

Favorite Score (FS)

$$FS = \sum_{i=0}^{n} score(fq_i) + \sum_{j=0}^{m} score(fa_i)$$

Features for Classification

Set	Category	Number	Features
A	**User Profile**	3	*Age of Account* *Badge Score* *Previous Posts with Negative Score*
B	**Community**	3	*Post Score* *Accepted Answer Score* *Favorite Score*
C	**Question Content**	3	*Number of URLs* *Number of Stack Overflow URLs* *Number of Popular Tags*
D	**Textual Style**	9	*Title Length* *Body Length* *Number of Tags* *Number of Punctuation Marks* *Number of Short Words* *Code Snippet Length* *Number of Special Characters* *Number of Lower Case Characters* *Number of Upper Case Characters*

Accepted Answer Score (AAS)

$$AAS = \sum_{i=0}^{n} 15$$

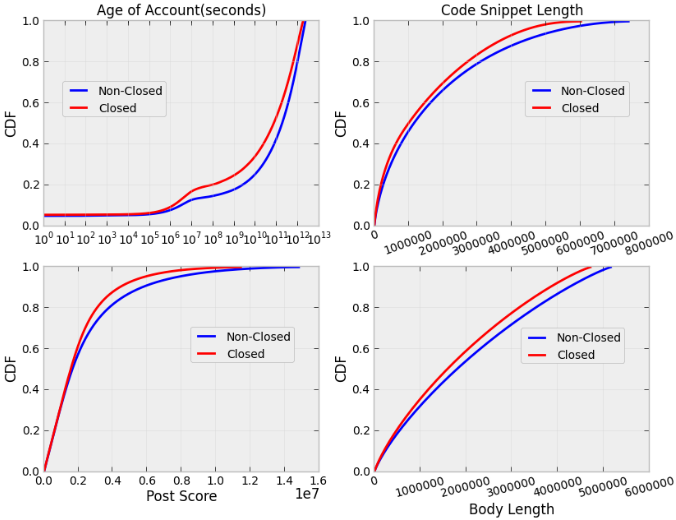
Features for Classification

Set	Category	Number	Features
A	**User Profile**	3	*Age of Account* *Badge Score* *Previous Posts with Negative Score*
B	**Community**	3	*Post Score* *Accepted Answer Score* *Favorite Score*
C	**Question Content**	3	*Number of URLs* *Number of Stack Overflow URLs* *Number of Popular Tags*
D	**Textual Style**	9	*Title Length* *Body Length* *Number of Tags* *Number of Punctuation Marks* *Number of Short Words* *Code Snippet Length* *Number of Special Characters* *Number of Lower Case Characters* *Number of Upper Case Characters*

Number of Popular Tags (#PT)

$$\#PT = \|T \cap PT\|$$

Cumulative Distribution Function Plots



Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Experimental Testbed, Setup and Classifier

Challenges, Issues etc.

- 1302 questions do not have any information about the question asker
- High imbalance dataset – 3% only in +ve class
- Reputation of the user at question creation time is not available
- Original text of question is not available
- No particular *winner* discriminatory feature

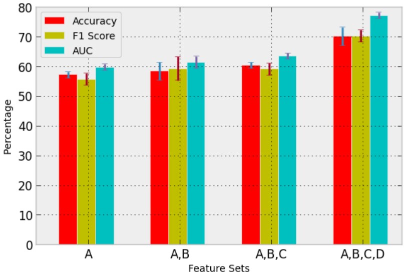
Dataset	203,382 questions
'Closed' (+ve class)	101,691
Non-'Closed' (-ve class)	101,691 (10-times) random sample with replacement
Classifier	Stochastic Gradient Boosted Trees
Learning Rate	0.1
Sub-sample size	0.7
Classification Runs	10 (for each +ve/-ve pair)
Feature Sets	{A}, {A, B}, {A, B, C}, {A, B, C, D}
Train-Test Split	70%-30%
Cross Validation	10-folds

Classification Performance

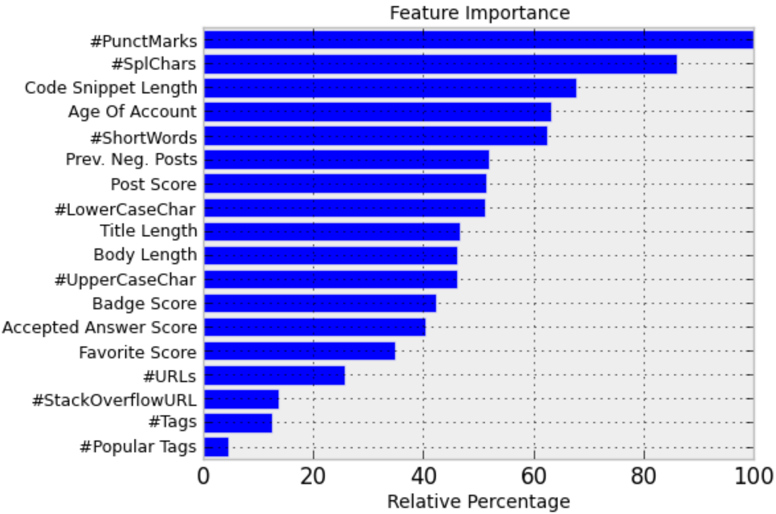
Confusion Matrix

		Predicted	
		Closed	**Non-Closed**
True	**Closed**	**69.6%**	30.4 %
	Non-Closed	29.1%	**70.9%**

Accuracy, F1 and AUC score



Feature Importance



Feature Importance

Observations

- Top five features – *Punctuation Marks, Special Characters, Code Snippet Length, Age of Account, Short Words*
- More *Punctuation Marks* and *Special Characters* in 'closed' than non-'closed questions
- *Closed* question have short *Code Snippet Length* and *Title Length*
- Importance of *Age of Account* indicate new users are more prone to submit a *closed* question
- *Closed* question are less informative and descriptive

Feature Importance

Observations

- Top five features – *Punctuation Marks, Special Characters, Code Snippet Length, Age of Account, Short Words*
- More *Punctuation Marks* and *Special Characters* in 'closed' than non-'closed questions
- *Closed* question have short *Code Snippet Length* and *Title Length*
- Importance of *Age of Account* indicate new users are more prone to submit a *closed* question
- *Closed* question are less informative and descriptive

Feature Importance

Observations

- Top five features – *Punctuation Marks, Special Characters, Code Snippet Length, Age of Account, Short Words*
- More *Punctuation Marks* and *Special Characters* in 'closed' than non-'closed questions
- *Closed* question have short *Code Snippet Length* and *Title Length*
- Importance of *Age of Account* indicate new users are more prone to submit a *closed* question
- *Closed* question are less informative and descriptive

Feature Importance

Observations

- Top five features – *Punctuation Marks, Special Characters, Code Snippet Length, Age of Account, Short Words*
- More *Punctuation Marks* and *Special Characters* in 'closed' than non-'closed questions
- *Closed* question have short *Code Snippet Length* and *Title Length*
- Importance of *Age of Account* indicate new users are more prone to submit a *closed* question
- *Closed* question are less informative and descriptive

Feature Importance

Observations

- Top five features – *Punctuation Marks, Special Characters, Code Snippet Length, Age of Account, Short Words*
- More *Punctuation Marks* and *Special Characters* in 'closed' than non-'closed questions
- *Closed* question have short *Code Snippet Length* and *Title Length*
- Importance of *Age of Account* indicate new users are more prone to submit a *closed* question
- *Closed* question are less informative and descriptive

Table of Contents

1 Introduction

2 Closed Questions on Stack Overflow

3 Characterization Study of Closed Questions

4 Closed Question Prediction

5 Conclusion

Conclusion

1. We investigate the phenomena of *closed* questions on Stack Overflow
2. We present our characterization study and draw multiple key insights based on temporal, community patterns, content analysis and popularity
3. We build a predictive model to classify *closed* questions with 70.3% accuracy
4. We find that *closed* question are less descriptive and informative in nature
5. Our study benefits the Stack Overflow community – both the users as well as the community elected moderators

Conclusion

1. We investigate the phenomena of *closed* questions on Stack Overflow
2. We present our characterization study and draw multiple key insights based on temporal, community patterns, content analysis and popularity
3. We build a predictive model to classify *closed* questions with 70.3% accuracy
4. We find that *closed* question are less descriptive and informative in nature
5. Our study benefits the Stack Overflow community – both the users as well as the community elected moderators

Conclusion

1. We investigate the phenomena of *closed* questions on Stack Overflow
2. We present our characterization study and draw multiple key insights based on temporal, community patterns, content analysis and popularity
3. We build a predictive model to classify *closed* questions with 70.3% accuracy
4. We find that *closed* question are less descriptive and informative in nature
5. Our study benefits the Stack Overflow community – both the users as well as the community elected moderators

Conclusion

1. We investigate the phenomena of *closed* questions on Stack Overflow
2. We present our characterization study and draw multiple key insights based on temporal, community patterns, content analysis and popularity
3. We build a predictive model to classify *closed* questions with 70.3% accuracy
4. We find that *closed* question are less descriptive and informative in nature
5. Our study benefits the Stack Overflow community – both the users as well as the community elected moderators

Conclusion

1. We investigate the phenomena of *closed* questions on Stack Overflow
2. We present our characterization study and draw multiple key insights based on temporal, community patterns, content analysis and popularity
3. We build a predictive model to classify *closed* questions with 70.3% accuracy
4. We find that *closed* question are less descriptive and informative in nature
5. Our study benefits the Stack Overflow community – both the users as well as the community elected moderators

Thank You!

Connect with me

http://correa.in/
@denzil_correa

Thank You!

Connect with me

http://correa.in/
@denzil_correa

One more thing ...

It's Stack Overflow, not Stackoverflow.