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

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

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

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

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.
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 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]**

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?

---

*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
Closed Questions on Stack Overflow

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

(Diamond User/Moderator)
One vote can close the question
5 votes can close the question

(>3000 reputation)
Who can close a question?
How are questions closed?

(<250 reputation)

Exact Duplicate
OffTopic
Not Constructive
Not a Real Question
Too Localized

Why are questions closed?
What happens to closed questions?

Reopened
Deleted
Remain Closed

Denzil Correa, Ashish Sureka (http://correa.in/)
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

<table>
<thead>
<tr>
<th>Users</th>
<th>1.29M (625k askers, 443k answerers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>3.4M (62.21% with accepted answers)</td>
</tr>
<tr>
<td>Answers</td>
<td>6.8M (31.33% marked as accepted)</td>
</tr>
<tr>
<td>Votes</td>
<td>27.5M (72.35% positive, 6.81% favorites)</td>
</tr>
<tr>
<td>Ratio of Answers to Questions</td>
<td>2.16</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed Questions</td>
<td>3.8%</td>
<td>1.52%</td>
<td>1.77%</td>
<td>3.33%</td>
<td>3.82%</td>
<td><strong>102,993</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2.98%)</td>
</tr>
<tr>
<td>Closed Votes</td>
<td>0.03%</td>
<td>0.25%</td>
<td>0.75%</td>
<td>2.21%</td>
<td>3.9%</td>
<td><strong>570,418</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.2%)</td>
</tr>
<tr>
<td>Ratio of Answers to Questions</td>
<td>8.0</td>
<td>5.93</td>
<td>3.11</td>
<td>1.92</td>
<td>1.55</td>
<td>1.92</td>
</tr>
</tbody>
</table>
Closed Question Sub-Category Distribution

- Duplicate: 29.3%
- Not a Real Question: 31.0%
- Subjective: 15.9%
- Too Localized: 6.1%
- Off Topic: 17.7%
Ratio of Closed Questions – Temporal Distribution Plot

- Exact Duplicate
- Off-Topic
- Subjective
- Not a real question
- Too Localized

[Graph showing temporal distribution of question ratios]
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

![Graph showing the effect of new registered users on closed question distribution]

- Pearson Correlation Coefficient = 0.95, p-value < 0.01
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

The diagram shows the distribution of close vote percentages over time, categorized by the number of votes needed to close a question. The categories include 1-vote, 2-votes, 3-votes, 4-votes, and 5-votes. The x-axis represents the months from August 08 to July 12, and the y-axis represents the percentage of closed questions.

- **1-vote**: Represented by blue dots, showing a sharp increase in early 2009 and then a gradual decline.
- **2-votes**: Represented by cyan dots, showing a lower peak compared to 1-vote and a more consistent trend.
- **3-votes**: Represented by green dots, indicating a lower percentage compared to the higher vote categories.
- **4-votes**: Represented by purple dots, showing a relatively low but steady trend.
- **5-votes**: Represented by red dots, showing the highest percentage, with peaks and troughs throughout the year.

Overall, the graph illustrates the distribution of close votes over time, with 1-vote questions having the highest percentage and 5-vote questions having the lowest, indicating a trend of increasing difficulties in closing questions as the vote requirement increases.
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

![Graph showing close vote distribution over time](image-url)
Close Vote Sub-Category Distribution

- Duplicate
- Off-Topic
- Subjective
- Not a Real Question
- Too Localized

Percentage of 'Close' Votes:

- 1-vote
- 2-votes
- 3-votes
- 4-votes
- 5-votes
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

![Graphs and charts showing distribution of title length, number of tags, and body length for different categories (ED, OT, ST, NRQ, TL).]

Denzil Correa, Ashish Sureka (http://correa.in/)  
ACM COSN-2013  
October 8, 2013 25 / 47
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

<table>
<thead>
<tr>
<th>Type</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Languages</td>
<td>java, c++, python, c, perl, r, ...</td>
</tr>
<tr>
<td>Web2.0</td>
<td>php, html5, html, css, apache, javascript, ...</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>ios, unix, linux, android, ubuntu, osx, windows, ...</td>
</tr>
<tr>
<td>Social</td>
<td>Facebook, wordpress, google, ...</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>books, interview-questions, fun, homework, ...</td>
</tr>
</tbody>
</table>
Question Topics

Popular Tags in Closed Questions

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

Normalized Tag Ratio (NTR)

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

\[
\therefore \forall t_i \in T_{CQ} \quad NTR_{t_i} = \frac{R^i_{CQ}}{R^i_{NCQ} + \epsilon}, \quad (T_{CQ} \cap T_{NCQ} \neq \emptyset)
\]
Discriminative Closed Question Tags

![Bar chart showing normalized ratio of tags in closed questions to total questions.](image)
Favorite Vote Distribution

![Favorite Vote Distribution Graph]

Legend:
- Yellow: Subjective
- Blue: Duplicate
- Cyan: Off Topic
- Green: Not a Real Question
- Red: Too Localized

Percentage of Closed Questions

Number of Favorite Votes

- $\geq 5$
- $\geq 10$
- $\geq 25$
- $\geq 50$
- $\geq 100$
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: ≥ 100 ‘favorite votes’ on closed questions in subjective category.

<table>
<thead>
<tr>
<th>Favorites</th>
<th>Title</th>
<th>Answers</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>5894</td>
<td>List of freely available programming books</td>
<td>112</td>
<td>569,199</td>
</tr>
<tr>
<td>2228</td>
<td>Hidden features of Python</td>
<td>100</td>
<td>212,589</td>
</tr>
<tr>
<td>1685</td>
<td>What is the best comment in source code you have ever encountered?</td>
<td>519</td>
<td>1,051,784</td>
</tr>
<tr>
<td>421</td>
<td>Worst security hole you’ve seen?</td>
<td>163</td>
<td>32,840</td>
</tr>
<tr>
<td>140</td>
<td>What is the most useful R trick?</td>
<td>34</td>
<td>13,197</td>
</tr>
</tbody>
</table>
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: ≥ 100 ‘favorite votes’ on closed questions in subjective category.

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

Distribution of time taken to mark a question closed

- **Exact Duplicate**: 1.41 days
- **Off Topic**: 1.99 days
- **Subjective**: 568.06 days
- **Not a Real Question**: 1.2 days
- **Too localized**: 6.71 days

Time (in seconds) taken to close question

- **Exact Duplicate**: 6.93 hrs
- **Off Topic**: 12.01 hrs
- **Subjective**: 26.36 hrs
- **Not a Real Question**: 8.3 hrs
- **Too localized**: 22.72 hrs
## Closure Time Outliers

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

### Observations

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

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

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 $\geq 5$, 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

<table>
<thead>
<tr>
<th>Category</th>
<th>Locked</th>
<th>Community Wiki</th>
<th>Protected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Duplicate</td>
<td>732(33.8%)</td>
<td>160(9.9%)</td>
<td>36(10.3%)</td>
</tr>
<tr>
<td>Off Topic</td>
<td>1180(54.5%)</td>
<td>273(16.8%)</td>
<td>70(20.1%)</td>
</tr>
<tr>
<td>Subjective</td>
<td>188(8.7%)</td>
<td>978(60.3%)</td>
<td>202(58%)</td>
</tr>
<tr>
<td>Not Real Question</td>
<td>50(2.3%)</td>
<td>192(11.8%)</td>
<td>28(8%)</td>
</tr>
<tr>
<td>Too Localized</td>
<td>114(0.6%)</td>
<td>10(0.6%)</td>
<td>12(3.4%)</td>
</tr>
<tr>
<td>Total</td>
<td>2,264</td>
<td>1,613</td>
<td>348</td>
</tr>
</tbody>
</table>

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

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

**Badge Score (BS)**

\[
BS = \sum_{i=0}^{n} \frac{1}{\#\text{users who have } b_i}
\]
# Features for Classification

<table>
<thead>
<tr>
<th>Set</th>
<th>Category</th>
<th>Number</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>User Profile</td>
<td>3</td>
<td>Age of Account, Badge Score, Previous Posts with Negative Score</td>
</tr>
<tr>
<td>B</td>
<td>Community</td>
<td>3</td>
<td>Post Score, Accepted Answer Score, Favorite Score</td>
</tr>
<tr>
<td>C</td>
<td>Question Content</td>
<td>3</td>
<td>Number of URLs, Number of Stack Overflow URLs, Number of Popular Tags</td>
</tr>
<tr>
<td>D</td>
<td>Textual Style</td>
<td>9</td>
<td>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</td>
</tr>
</tbody>
</table>

**Post Score (PS)**

\[
PS = \sum_{i=0}^{n} \text{score}(q_i) + \sum_{j=0}^{m} \text{score}(a_i)
\]
Features for Classification

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

**Favorite Score (FS)**

\[
FS = \sum_{i=0}^{n} \text{score}(f_{qi}) + \sum_{j=0}^{m} \text{score}(f_{ai})
\]
## Features for Classification

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

### Accepted Answer Score (AAS)

\[
AAS = \sum_{i=0}^{n} 15
\]
# Features for Classification

<table>
<thead>
<tr>
<th>Set</th>
<th>Category</th>
<th>Number</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>User Profile</td>
<td>3</td>
<td>Age of Account, Badge Score, Previous Posts with Negative Score</td>
</tr>
<tr>
<td>B</td>
<td>Community</td>
<td>3</td>
<td>Post Score, Accepted Answer Score, Favorite Score</td>
</tr>
<tr>
<td>C</td>
<td>Question Content</td>
<td>3</td>
<td>Number of URLs, Number of Stack Overflow URLs, Number of Popular Tags</td>
</tr>
<tr>
<td>D</td>
<td>Textual Style</td>
<td>9</td>
<td>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</td>
</tr>
</tbody>
</table>

Number of Popular Tags (#PT)

\[
#PT = |T \cap PT|
\]
Cumulative Distribution Function Plots

- **Age of Account (seconds)**
- **Code Snippet Length**
- **Post Score**
- **Body Length**

[Denzil Correa, Ashish Sureka](http://correa.in/)
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Closed’ (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-‘Closed’ (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
<tr>
<td>Classifier</td>
<td>Stochastic Gradient Boosted Trees</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Sub-sample size</td>
<td>0.7</td>
</tr>
<tr>
<td>Classification Runs</td>
<td>10 (for each +ve/-ve pair)</td>
</tr>
<tr>
<td>Feature Sets</td>
<td>{A}, {A, B}, {A, B, C}, {A, B, C, D}</td>
</tr>
<tr>
<td>Train-Test Split</td>
<td>70%-30%</td>
</tr>
<tr>
<td>Cross Validaton</td>
<td>10-folds</td>
</tr>
</tbody>
</table>

Denzil Correa, Ashish Sureka (http://correa.in/)
ACM COSN-2013
October 8, 2013
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Closed' (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-'Closed' (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
</tbody>
</table>

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

Denzil Correa, Ashish Sureka (http://correa.in/)
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Closed’ (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-‘Closed’ (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
<tr>
<td>Classifier</td>
<td>Stochastic Gradient Boosted Trees</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Sub-sample size</td>
<td>0.7</td>
</tr>
<tr>
<td>Classification Runs</td>
<td>10 (for each +ve/-ve pair)</td>
</tr>
<tr>
<td>Feature Sets</td>
<td>{A}, {A, B}, {A, B, C}, {A, B, C, D}</td>
</tr>
<tr>
<td>Train-Test Split</td>
<td>70%-30%</td>
</tr>
<tr>
<td>Cross Validation</td>
<td>10-folds</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Closed’ (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-‘Closed’ (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
<tr>
<td>Classifier</td>
<td>Stochastic Gradient Boosted Trees</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Sub-sample size</td>
<td>0.7</td>
</tr>
<tr>
<td>Classification Runs</td>
<td>10 (for each +ve/-ve pair)</td>
</tr>
<tr>
<td>Feature Sets</td>
<td>{A}, {A, B}, {A, B, C}, {A, B, C, D}</td>
</tr>
<tr>
<td>Train-Test Split</td>
<td>70%-30%</td>
</tr>
<tr>
<td>Cross Validation</td>
<td>10-folds</td>
</tr>
</tbody>
</table>
Closed Question Prediction

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Closed’ (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-‘Closed’ (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>203,382 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Closed’ (+ve class)</td>
<td>101,691</td>
</tr>
<tr>
<td>Non-‘Closed’ (-ve class)</td>
<td>101,691 (10-times)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Stochastic Gradient Boosted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Sub-sample size</td>
<td>0.7</td>
</tr>
<tr>
<td>Classification Runs</td>
<td>(for each +ve/-ve pair)</td>
</tr>
<tr>
<td>Feature Sets</td>
<td>{A}, {A, B}, {A, B, C}, {A, B, C, D}</td>
</tr>
<tr>
<td>Train-Test Split</td>
<td>70%-30%</td>
</tr>
<tr>
<td>Cross Validation</td>
<td>10-folds</td>
</tr>
</tbody>
</table>
Classification Performance

Confusion Matrix

<table>
<thead>
<tr>
<th>True</th>
<th>Predicted</th>
<th>Closed</th>
<th>Non-Closed</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Closed</td>
<td>69.6%</td>
<td>29.1%</td>
</tr>
<tr>
<td>True</td>
<td>Non-Closed</td>
<td>30.4%</td>
<td>70.9%</td>
</tr>
</tbody>
</table>

Accuracy, F1 and AUC score

![Graph showing Accuracy, F1 Score, and AUC for different feature sets A, A,B, A,B,C, and A,B,C,D]
Feature Importance

- #PunctMarks
- #SplChars
- Code Snippet Length
- Age Of Account
- #ShortWords
- Prev. Neg. Posts
- Post Score
- #LowerCaseChar
- Title Length
- Body Length
- #UpperCaseChar
- Badge Score
- Accepted Answer Score
- Favorite Score
- #URLs
- #StackOverflowURL
- #Tags
- #Popular Tags

Relative Percentage
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* questions 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* questions 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* questions 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.