Constructing Better Questions

A Study of Closed Questions on Stack Overflow

Team Discover: Robert Baker, Khalida Hendricks,

Aniket Shah, Jessica Valenti

Mentor: Andrew Castillo

What is Stack Overflow?

- Stack Overflow is a popular website where users can ask and answer questions on a variety of coding topics
- Once questions are answered they are closed
- Unanswered questions remain open for responses
- In some cases, questions are never answered

https://stackoverflow.com/

stack overflow

Importance and Objective

- Importance: Understanding what makes a question likely to be answered would be tremendously helpful knowledge for Stack Overflow users that ask questions

- Objective: Use Stack Overflow question and answer data and Natural Language Processing (NLP) to predict whether questions would be open or closed based on the text provided in the questions

Our data

- Used data from
  https://www.kaggle.com/datasets/stackoverflow/stacksample

- Key data features:
  - Question text
  - Answer text
  - Scores
  - Tags
  - …and many more…

The raw text

- Before editing, the text was messy:

```
'<p>I\'ve written a database generation script in <a href="http://en.wikipedia.org
ute it in my <a href="http://en.wikipedia.org/wiki/Adobe_Integrated_Runtime">Adobe
code>Create Table tRole (\n      roleID integer Primary Key\n      ,roleName varch
fileID integer Primary Key\n    ,fileName varchar(50)\n    ,fileDescription varcha
,fileFormatID integer\n    ,categoryID integer\n    ,isFavorite boolean\n    ,date
integer\n    ,lastAccessTime date\n    ,downloadComplete boolean\n    ,isNew boole
duration varchar(30)\n);\nCreate Table tCategory (\n    categoryID integer Primary
\n    ,parent_categoryID integer\n);\n...\n</code></pre>\n\n<p>I execute this in A
s:</p>\n\n<pre><code>public static function RunSqlFromFile(fileName:String):void {
ionDirectory.resolvePath(fileName);\n    var stream:FileStream = new FileStream();
```

Processing

- We used BeautifulSoup to remove html tags (<p>, <code>, <a href =....)
- After using spaCy to strip out stopwords and punctuation, we obtained lists of words, e.g.:
  - [write, database, generation, script, want, execute, Adobe, application, create, Table, tRole, roleID, integer, Primary, rolename, varchar, Create, Table, tFile, fileID, integer, Primary, filename, varchar, filedescription, varchar,...]

Methods - Bag of words

With our list of words for each question, we implemented a bag of words model to count the number of times each word occurred throughout the entire dataset.

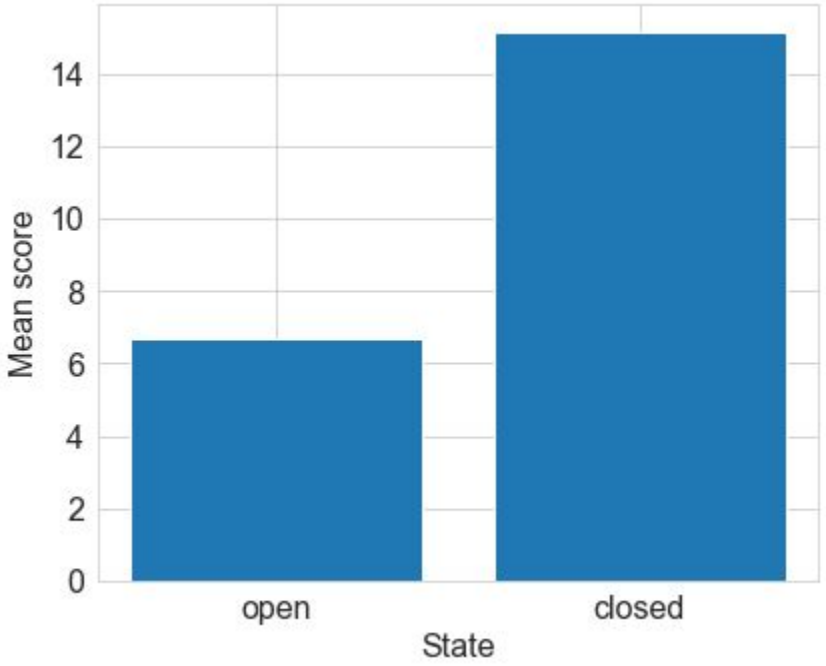
Methods - Logistic regression

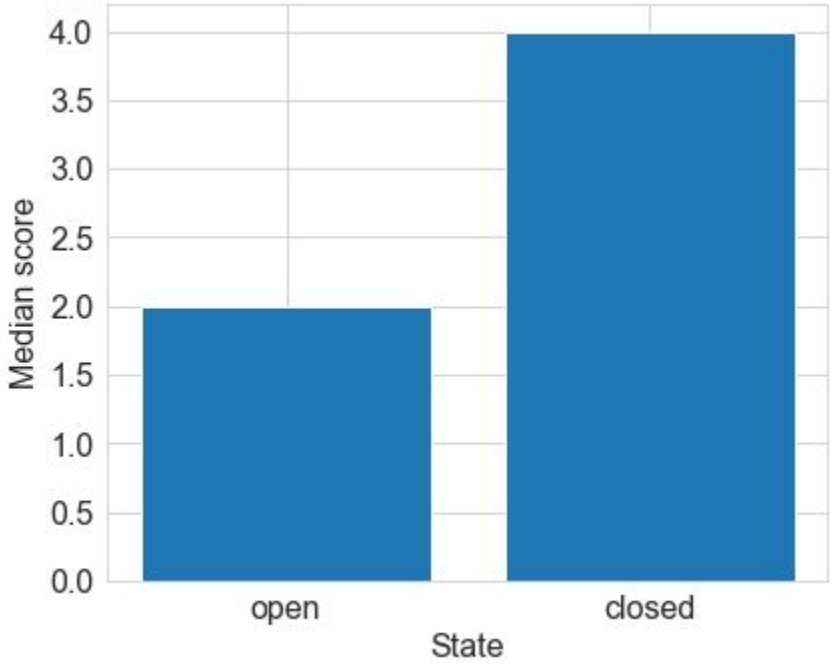
We then implemented a logistic regression model to compare new questions against our existing bag of words and predict if that new question would be closed or open.

In addition to testing the accuracy of our predictions, we examined various quantities you can get from the confusion matrix and explored different diagnostic curves.

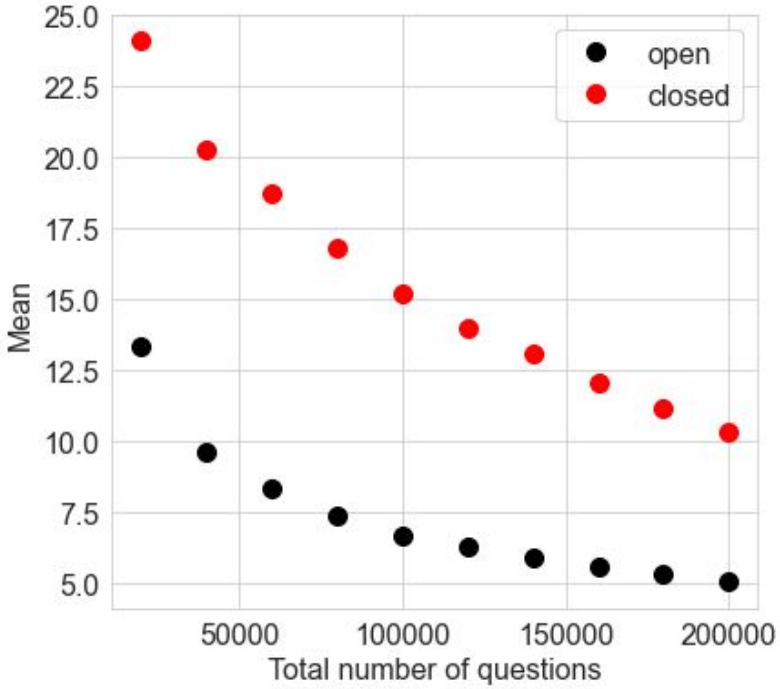
Analysis & Results

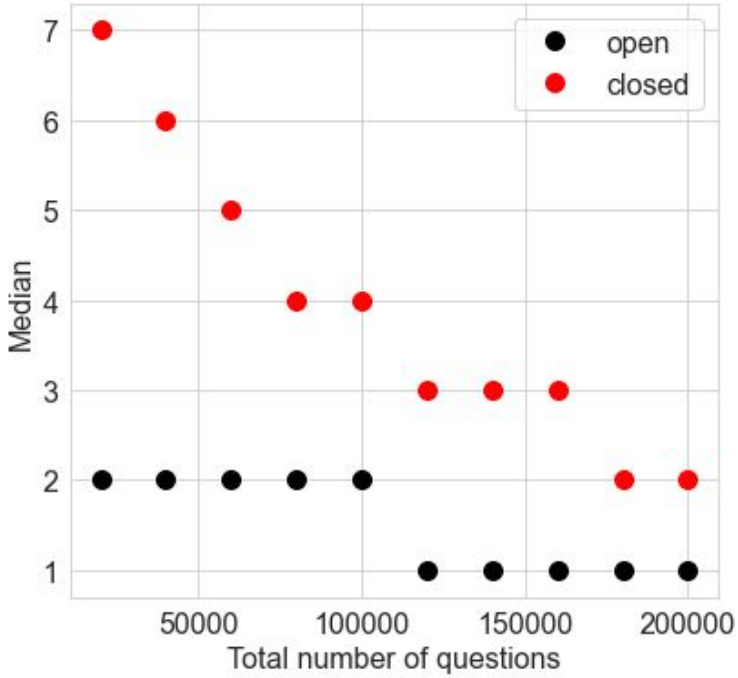
Analysis & Results





Analysis & Results

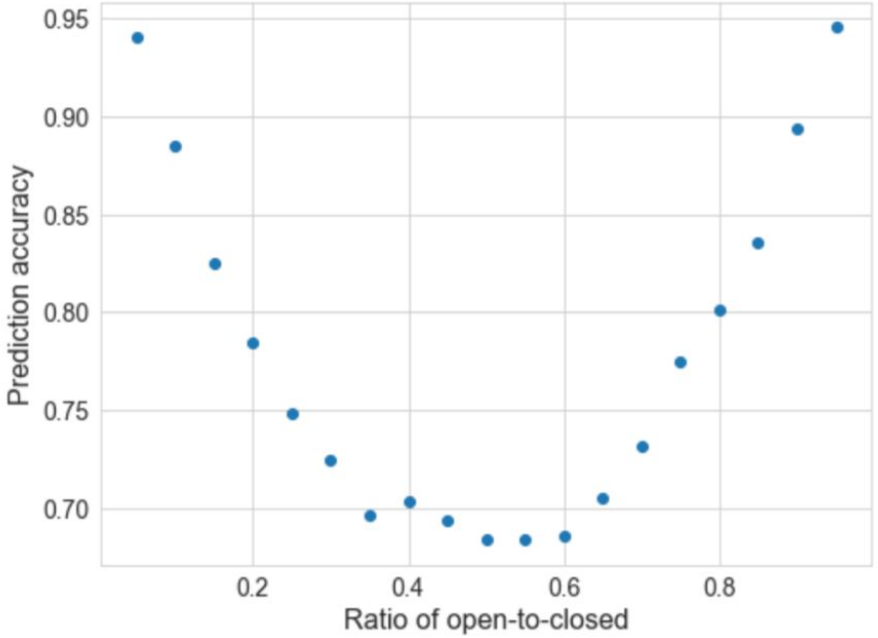




Analysis & Results

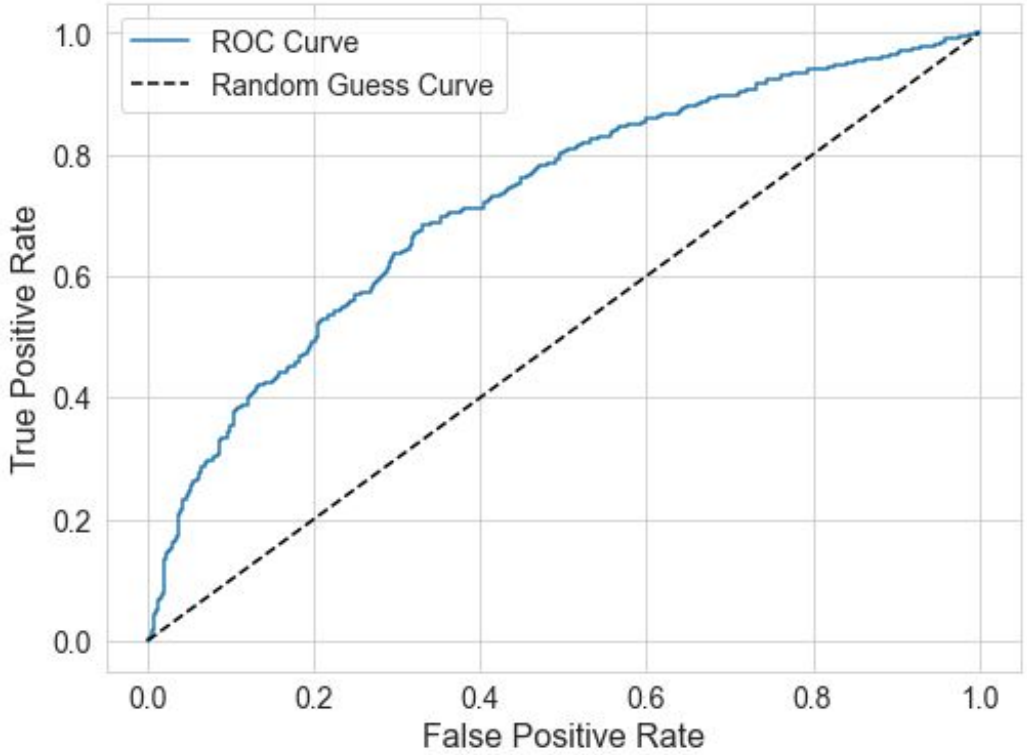
Depending on the characteristics of the train/test datasets, we found an accuracy between 94% and 68%.

When trained on 50-50 data, the model can correctly predict if a question is closed 68% of the time.



Analysis & Results

		PREDICTED	
		CLOSED	OPEN
ACTUAL	CLOSED	408	192
	OPEN	198	402



Conclusion

SUMMARY:

- Open/closed status is correlated to answer quality
- Open/closed status can be predicted 68% of the time with a simple logistic regression model.

FUTURE IMPROVEMENTS:

- Continue to refine data cleaning and processing to better capture keywords
- Utilize some degree of syntax in the model, either through bigrams and trigrams, or though a more sophisticated NLP machine learning technique