StackOverflow question status prediction

Made by Ihar Shulhan & Kevin Khanda

What is the data

Data consists of 12 columns:

- PostCreationDate - date/time field, which shows when the post was created
- OwnerUserId - int64 field, StackOverflow user ID
- OwnerCreationDate - date/time field, shows when asker account was registered
- ReputationAtPostCreation - int field, stands for asker reputation on a moment, when question was submitted (min value: -15, max value: 171723)
- OwnerUndeletedAnswerCountAtPostTime - int field, number of asker undeleted answers on questions of other users
- Title - string field, matches the question title
- BodyMarkdown - string field; extracted body of a question, may contain code samples
- Tag1-Tag5 - string fields, may contain empty values; define tags, which were assigned by asker or community to the question; tags increase the question popularity (more tags -> more shows -> more chances that question might be visited and answered)

What is the data

The data is a dump of 63462 StackOverflow questions. Some questions are still opened, some are closed by community.

Train data includes both the question and the “open status” of the question, test data includes only question related data.

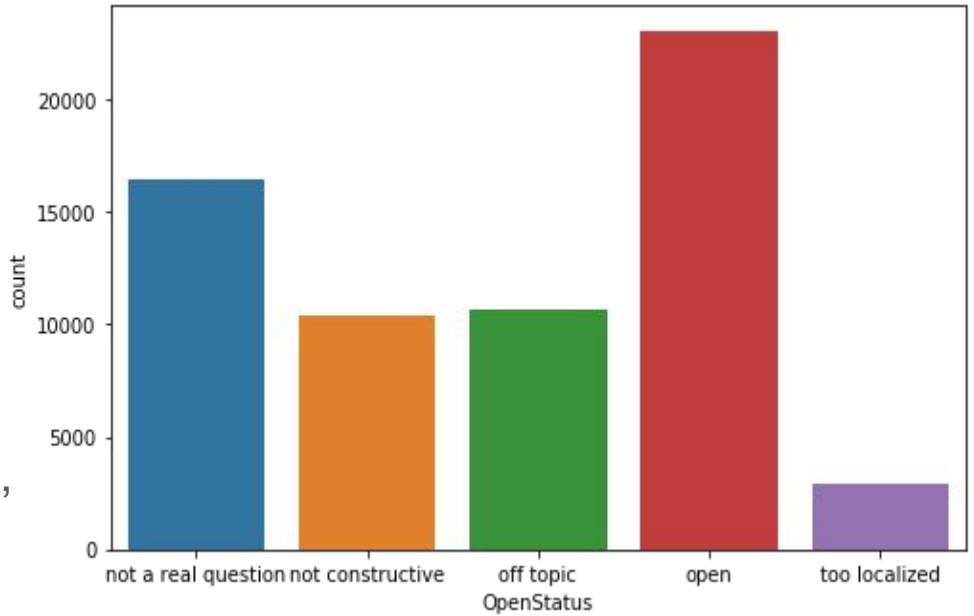
```
Data columns (total 14 columns):
PostCreationDate                        63462 non-null object
OwnerUserId                             63462 non-null int64
OwnerCreationDate                       63462 non-null object
ReputationAtPostCreation                63462 non-null int64
OwnerUndeletedAnswerCountAtPostTime     63462 non-null int64
Title                                   63462 non-null object
BodyMarkdown                            63462 non-null object
Tag1                                    63459 non-null object
Tag2                                    49428 non-null object
Tag3                                    32412 non-null object
Tag4                                    16568 non-null object
Tag5                                    6382 non-null object
PostClosedDate                          40445 non-null object
OpenStatus                              63462 non-null object
```

What should be predicted

OpenStatus - the status of a

Based on the data from previous slide we should provide the most accurate model which will predict the OpenStatus.

Possible statuses are: “not a real question”, “not constructive”, “off topic”, “open”, “too localized”



Features

The most interesting feature is the question inself. Many other features can be generated out of it.

Another valuable features are question tags and user reputation. The popularity of a tags could be determined from the corpus. Higher user reputation tends that the question is more likely to be in status “open” or “not a real question”.

BodyMarkdown
I am following a tutorial on Function x <http://www.functionx.com/vcnet/xml/readwrite.htm>. I h...
I just couldnt find a decent review of this book http://www.careercup.com/ \r\nHas anyone in thi...
I am planning to develop a web application in java using struts/servlets/jsp technology.\r\n\r\n...
If there is a multi select control and you press Ctrl+A in Firefox to select all values then it ...
I have a array value bellow\r\n\r\n[[[6,'SPAIN LA LIGA',0],[[843188,'RCD Espanyol','Real Madrid...
please tell me examples of rotateLeft method of Integer class in java
I have looked for the 'patch' system but I don't think this is what I want. I've also done a lit...
/SXML5 is the only version of MSXML that supports XML digital signatures. \r\n\r\nDoes anyone h...
[Phishing][1] is a very serious problem that we face. However, banks are the biggest targets. W...

Generated features

First of all, BodyMarkdown and Title features were tokenized with Keras Tokenizer based on dictionary.

Sequences of word counts were generated for both Title and BodyMarkdown.

Stop-words were not removed by a reason - it matters how the question is asked and articles can tell more about this.

title_sequences
[1078, 70, 101560, 176, 5, 782, 6, 247, 118, 85, 32, 8, 1905, 11, 70]
[8, 2, 41282, 416, 1919, 2, 1850]
[5967, 19, 45]
[2265, 5, 271, 118, 7, 923, 10, 1289, 178]

Title
Compiler error 3921? Following a tutorial and can't work out what is causing this error
Is the CareerCup Book Worth the money?
Mailing with Java
Ctrl+A doesn't work in Firefox for multi select
How to detect this type of array value use php

Generated features

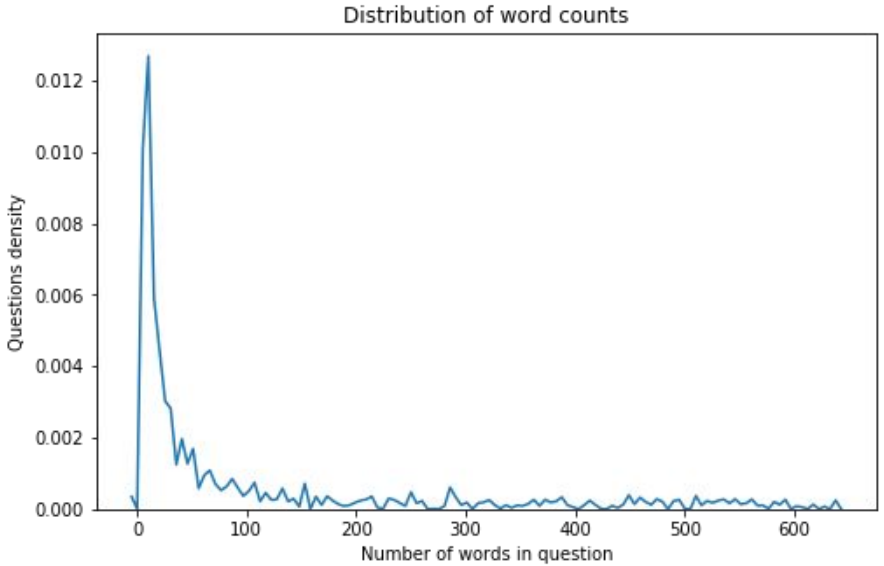
Number of words in a question

Here is the distribution on word counts in BodyMarkdown

And the crosstable presents the distribution of statuses for questions

1. True row - words count <= 75

2. False row - words count > 75)

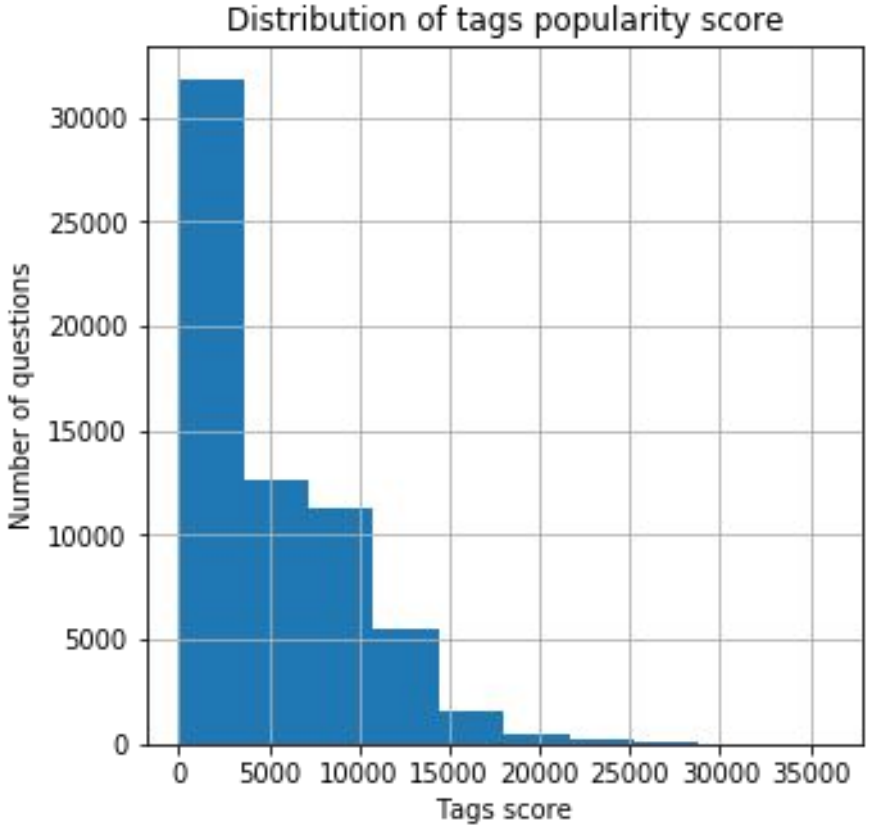


OpenStatus	not a real question	not constructive	off topic	open	too localized	All
body_word_count						
False	5094	4492	4432	14634	1789	30441
True	11365	5929	6202	8383	1142	33021
All	16459	10421	10634	23017	2931	63462

Generated features

Tags popularity

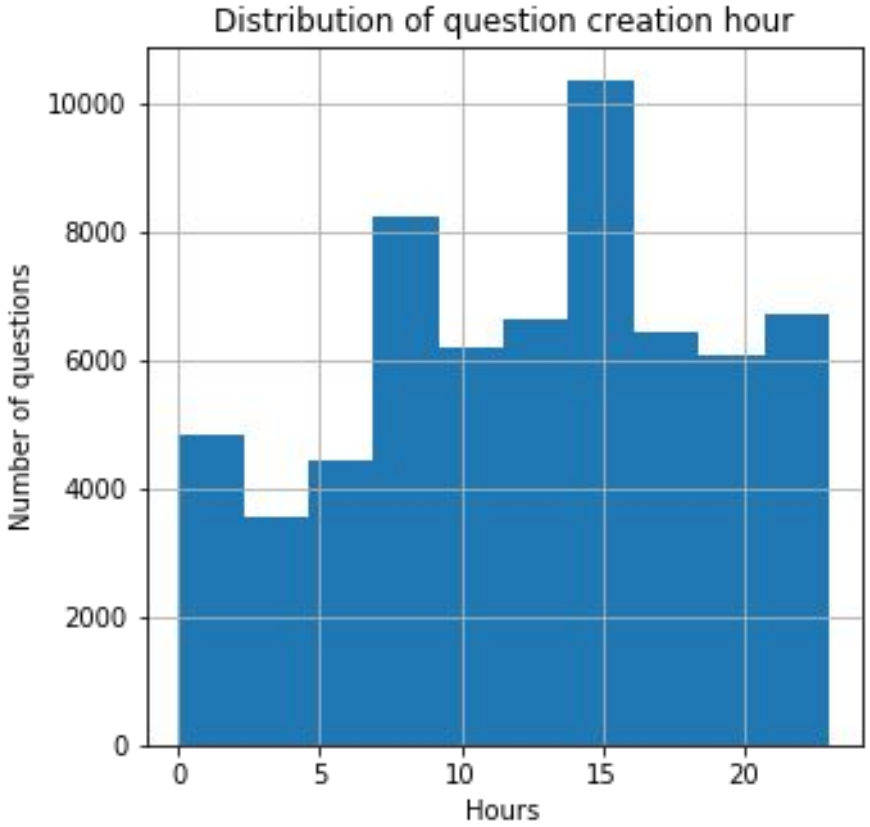
We defined feature named tags_weight, which is a sum of occurensies of each tags from columns Tag1 to Tag5, ignoring the empty tags.



Generated features

Date features

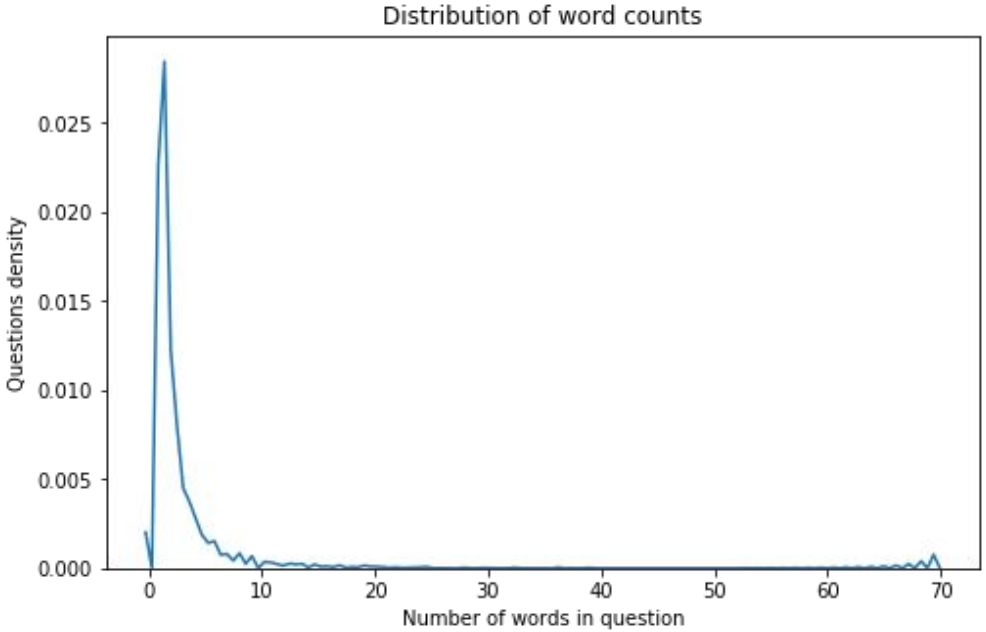
From dates we have extracted days of week, hour of question creation and difference between asked registration and question creation.



Generated features

Number of programming language keywords

For word occurencies from set of keywords from programming languages we assign a score, which is a number of such occurencies in a Title or BodyMarkdown



Feature preparation

All numeric features were converted to float64 type. NaN values were filled with zeros. Scaling has been done with a standard scaler

Title, BodyMarkdown and Tags were tokenized to generate new features as it was mentioned before, and after that these features were vectorized using TF-IDF vectorizer. We also tried LDA, but it has shown itself worse.

Also BodyMarkdown and Title were transformed using word2vec. We have tried different pretrained embeddings and the best performance was achieved with "fasttext-wiki-news-subwords-300".

We also tried to perform augmentation - translated texts on Russian and back to increase the amount of data, but it didn’t perform well.

From date fields we extracted hour of question creation and day of week, which then were transformed to categorical features.

Models & parameters

Model	Parameters	Score
KNeighborsClassifier	n_neighbors=10, weights='distance'	
SGDClassifier	loss="log", penalty="l2", max_iter=60	0.617 +/- 0.005
RandomForestClassifier	n_estimators=190, max_depth=19, class_weight='balanced', min_samples_leaf=7, max_features=275	0.600 +/- 0.001
XGBClassifier	max_depth=2, eta=0.3, alpha=0.01, grow_policy=lossguide, max_leaves=5, gamma=0.02, colsample_bylevel=0.5, subsample=0.8	0.638 +/- 0.004
LGBMClassifier	learning_rate=0.05, num_leaves=10, max_depth=5, feature_fraction=0.9, bagging_fraction=0.5, reg_alpha=0.05, reg_lambda=0.05	0.639 +/- 0.004
NNClassifier	Dropout(0.38) of input, then Dense(23) layer with selu activation, BatchNormalization and final layer with Softmax activation	0.633 +/- 0.005

Final model

Final model is an average between XGBoost, LGBM and Neural Network

Cross validation: 0.6455 (0.0059)
Leaderboard: 0.57442