What StackOverflow Tells Us
About Programming Languages

Rahul Thankachan
University of Southern California
Computer Science Department
Los Angeles, CA 90089
thankach@usc.edu

Nada Aldarrab
University of Southern California
Computer Science Department
Los Angeles, CA 90089
naldarra@usc.edu

Prathmesh Gat
University of Southern California
Computer Science Department
Los Angeles, CA 90089
gat@usc.edu

Abstract

StackOverflow is a huge collaborative environment for developers. Many people benefit from direct answers they get to their questions. In this project, we investigate the potential indirect benefits that we might get by analyzing StackOverflow data to unfold trends, help improve programming language design, and predict future advancements. Our analysis shows that scripting languages have gained more popularity in the last five years. Our response time predictor achieves an f-score of 0.81 using the J48 classifier.

Programmers and coding enthusiasts often face issues and challenging problems while carrying out development activities. Over the years, a vast number of public forums and websites have provided common platforms where programmers could share knowledge and ask questions freely. The advent of large number of platforms have resulted in creation of huge social Knowledge Markets. These knowledge markets provide rich tools for discovering and sharing knowledge resources. StackOverflow has become a prime leader in this space and boasted a large and vibrant community of users who collaborate on large number of specific programming questions. The large social data and crowdsourced knowledge make it a perfect fit for our data analysis project. It has so many interesting features, such as:

- Multiple programming languages discussed: Unlike many forums which are language specific, StckOverflow is language independent. Many programming languages are discussed.
- Social Data: StackOverflow allows users to share posts, comments, upvotes, etc, which creates large and rich social data.
- Programmer-Specific: Unlike other social sites, StackOverflow is a programmer-specific website where questions are restricted to programming exclusively.

Related Work

Some research has been done on StackOverflow data. Allamanis and Sutton (2013) performed topic modeling analysis to model Java topics of questions. They used Latent Dirichlet Allocation (LDA) to categorize the type and topics of questions. They suggested using these topic models to evaluate the orthogonality of different programming languages. They also found out that StackOverflow questions are about the code and are not application domain specific. They also performed some temporal analysis and found out that people post less questions on the weekends than on the weekdays.

Another interesting study by Bhat, et al. (2014) was performed to predict response time on StackOverflow. They modeled the problem as a binary classification problem (with response time being less than an hour or more than an hour). They extracted some tag based and some non-tag based question features. They fed those features into four different classifiers. They report their best prediction results to be around 70 percent in accuracy and f-score.

Data

We used the Stack Overflow Data Dump, published August 18, 2015 by Stack Exchange, Inc. This data set includes more than 26.5 million posts by more than 4.5 million users.

Popular Programming Languages

We performed basic analysis on the whole Stack Overflow dataset, and found out the top ten programming languages discussed on StackOverflow. Figure 1 shows the top ten programming languages that people discuss on StackOverflow. For each of the top ten programming languages, we found the occurrences of the programming language tags in posts and listed the top ten tags (i.e Programming Languages). Javascript ranks first while C ranks tenth.

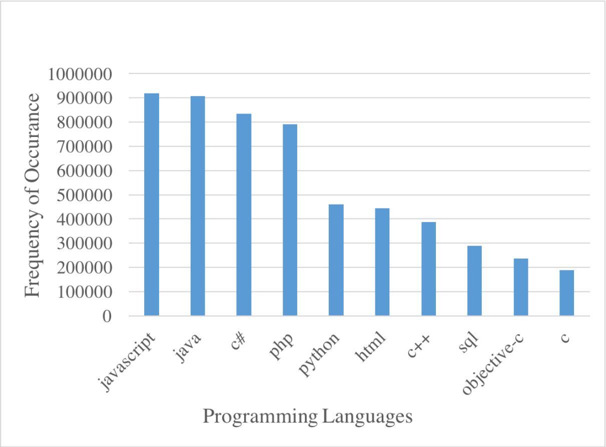


Figure 1: Top Ten Popular Programming Languages

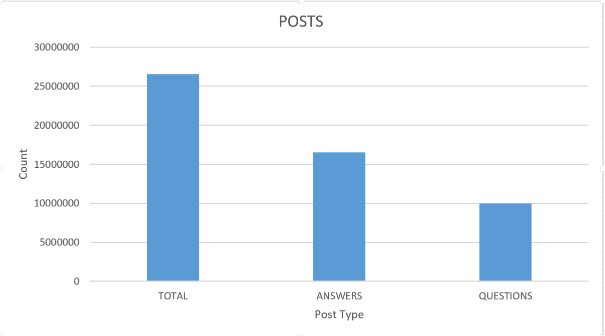


Figure 2: Number of Posts/ Questions / Answers

Figure 2 shows the basic statistics on the posts. Posts that are answers are more than questions. Answers are around 16 million while questions are around 10 million.

- Number Of Answers = Posts that are Answers
- Number Of Questions = Posts that are Questions
- Total = Number Of Answers + Number Of Questions

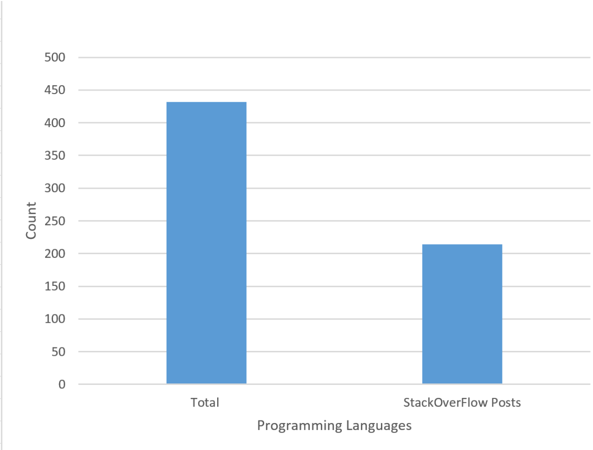


Figure 3: Programming languages and Programming languages Discussed

Figure 3 shows the basic statistics on the programming languages discussed on StackOverflow. We scraped some popular websites which listed programming languages and used a regex filter to extract the programming languages. Approximately 45 percent of all programming languages in the world are discussed on StackOverflow.

Temporal Trends

Here we performed temporal analysis on the whole Stack Overflow dataset to find out the time period during which a particular programming language was popular in terms of Question count, Answer Count, Question Fraction and Answer Fraction.

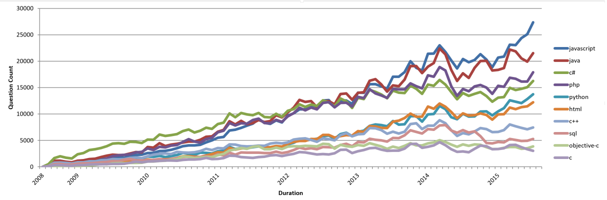


Figure 4: Question count per top ten programming language from 2008 to present on monthly basis

Here we performed basic analysis on the question count of popular programming languages from 2008 to present. Per programming language, question count per month was calculated for the entire period. We noticed following trends:-

- C# was most popular language from 2008 until 2012(in terms of questions asked), after 2012 it lost its popularity.
- Javascript gained its popularity from 2013, currently Javascript tops in terms of number of questions asked.

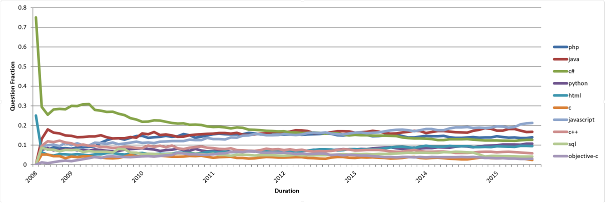


Figure 5: Question fraction per top ten programming language from 2008 to present on monthly basis

Here we performed basic analysis on the question fraction of popular programming languages from 2008 to present. Per programming language, Question Fraction value equals its fraction of questions from the question pool of top ten programming languages on a monthly basis. Per top programming language we calculated this value. We noticed following trends:-

- Initially C# had lots of questions compared to other languages.
- Question fraction value remains between 0 to 0.2 for most of the languages.

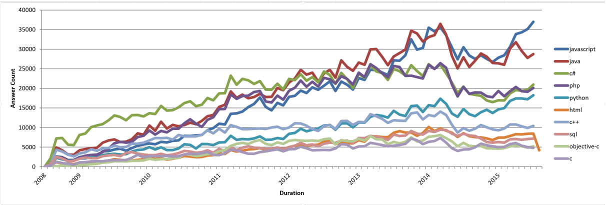


Figure 6: Answer count per top ten programming language from 2008 to present on monthly basis

Here we performed basic analysis on the answer count of popular programming languages from 2008 to present. Per programming language, answer count per month was calculated for the entire period. This is equal to the number of posts which were answers in that month for that programming language. We noticed following trends:-

- C# was the most popular language from 2008 until 2012(in terms of answers provided), after 2012 it lost its popularity.
- Javascript gained its popularity from 2013, currently Javascript tops in terms of number of answers provided.

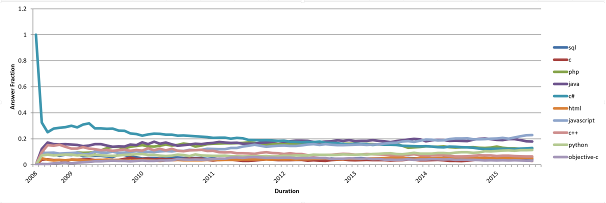


Figure 7: Answer fraction per top ten programming language from 2008 to present on monthly basis

Here we performed basic analysis on the answer fraction of popular programming languages from 2008 to present. Per programming language, Answer Fraction value equals its fraction of answers from the answer pool of top ten programming languages on a monthly basis. Per top programming language we calculated this value. We noticed following trends:-

- Initially C# had lots of answers compared to other languages.
- Answer fraction value remains between 0 to 0.2 for most of the languages.

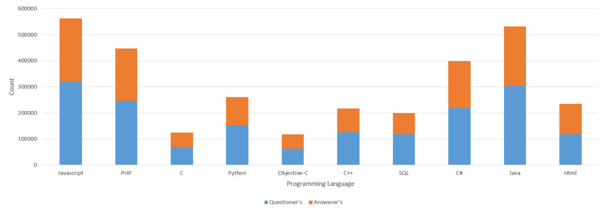


Figure 8: Questioners/Answerers Distribution

Here we performed basic analysis to find the Questioners(Users who ask questions) and Answerers(Users who provide answers to questions). Figure 8 clearly shows that both questioners and answerers are almost equally active on StackOverflow for top ten programming languages. Per top ten programming languages we found:-

- Number Questioners = Number of unique questioners who raised questions related to that particular language.
- Number Answerers = Number of unique answerers who answered questions related to that particular Language.

Topic Analysis

To help find the topics that people frequently ask about, we analyzed post tags that are related to each one of the top ten programming languages. We filtered the tags that marked other programming languages as they are irrelevant to our topic analysis. Table 1 shows the most popular topics for the top ten programming languages. We find that tags are good indicators of the topics that people discuss on StackOverflow. The top ten topics include data structures, platforms, library names or some specific concepts in certain programming languages. This data could greatly benefit programming language designers. They can work on these topics to improve the experience of language developers. They can also try to provide better documentation for these topics.

This could be very useful for improving future versions of programming languages. It could also be useful in developing good documentation and IDEs.

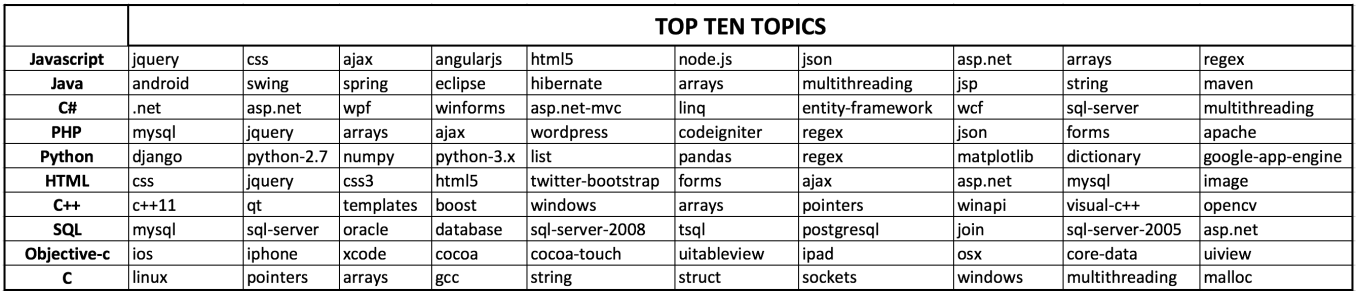
Predicting Time to Answer

One advantage of large scale collaborative sites like Wikipedia and Stack Overflow is the speed at which new information accumulates. Some questions on Stack Overflow seem to get answered immediately while others sit around for a while. As you are writing your question, it would be nice to have an idea of how long it might take to get an answer. How well can we predict the time until a question gets answered? This will likely depend on the subject matter of the question since some programming communities are more active than others.

Approach

We decided to use machine learning techniques and train a classifier to answer this question. We proceeded with Weka for our study. Weka is a collection of machine learning algorithms for data mining tasks developed at the University of Waikato. Since we wanted to predict the time to answer soon after a question is submitted we considered only those attributes which were available when a question is formulated. We selected the following attributes:

- Tag - We restricted the tags only to the top 10 programming languages. We wanted to verify if answering time was dependent on the programming language the question was associated to.
- Creation month - We wanted to check if questions submitted during holiday season had different answering times as compared to other times of the year.
- Body Length - The length of the description field was calculated.
- Number of Tags - Total number of tags associated. We wanted to verify if more tags are used does it reduce the answering time. In general, more tags used more will be the users who get notified about the topic of interest.



	TOP TEN TOPICS									
Javascript	jquery	css	ajax	angularjs	html5	node.js	json	asp.net	arrays	regex
Java	android	swing	spring	eclipse	hibernate	arrays	multithreading	jsp	string	maven
C#	.net	asp.net	wpf	winforms	asp.net-mvc	linq	entity-framework	wcf	sql-server	multithreading
PHP	mysql	jquery	arrays	ajax	wordpress	codeigniter	regex	json	forms	apache
Python	django	python-2.7	numpy	python-3.x	list	pandas	regex	matplotlib	dictionary	google-app-engine
HTML	css	jquery	css3	html5	twitter-bootstrap	forms	ajax	asp.net	mysql	image
C++	c++11	qt	templates	boost	windows	arrays	pointers	winapi	visual-c++	opencv
SQL	mysql	sql-server	oracle	database	sql-server-2008	tsql	postgresql	join	sql-server-2005	asp.net
Objective-c	ios	iphone	xcode	cocoa	cocoa-touch	uitableview	ipad	osx	core-data	uiview
C	linux	pointers	arrays	gcc	string	struct	sockets	windows	multithreading	malloc

Table 1: The most popular topics for the top ten programming languages

Data Pre-processing A question can have multiple answers. Since we were interested only in the time taken for first answer, we mapped all the questions to their first answers and discarded all the other answers. Since the dataset was very huge this was a challenging step. We then calculated delta answer for all the questions. Delta answer is the difference between creation time for first answer and creation time of parent question. For some of the questions we found that delta answer was either zero or negative. Since answers should follow questions we discarded these records.

We introduced a new nominal class attribute called Time-Answer for questions. We decided three values for this nominal class:

- less6 - All questions which received first answer in less than 6 minutes.
- bet6and20 - All questions which received first answer between 6 minutes and 20 minutes.
- 20andmore - All questions that received first answer within 20 minutes or more.

Each of the questions were tagged with a Time-Answer label based on the above logic.

Why did we categorize questions specifically under these ranges? After few hit and trials we selected the above ranges so that all the questions were evenly distributed across all three ranges. We noticed a large percentage of the questions received their first answer within the first 6 minutes. This suggested that StackOverflow community is highly active.

Total Questions after data pre processing consisted of 4490947 records. Since the data set was very large for Weka GUI, we used a subset of the data having 449000 records. We used a skip mechanism to create this subset. In our skip method we selected every 10th record from the questions set and rejected others. The reason we chose a skip method is because when we considered the first 500000 chronological records we noticed that the records were biased towards C# language. This trend was noticed because most of the early questions were restricted primarily to C# language as shown earlier as part of temporal trend analysis. So to get a better coverage of the overall data we chose the skip method.

Weka accepts data in a special .arff format. This added several more steps to our data pre-processing step. Since these steps were redundant and took lot of time we developed a python script which generated Weka .arff source file on the fly. It greatly reduced the time spent for preprocessing data for Weka.

Evaluation

Two classifiers were used for predicting time until answer. The accuracy while predicting nominal class Time-Answer was used for rating the effectiveness of our method. We ran a 10 fold cross-validation over our entire subset data. We ran the experiment in two iterations:

- all attributes - Our training data consisting of all the attributes.
- body length and number of tags attributes only - Our training data consisting of only these attributes. All other attributes for questions were discarded.

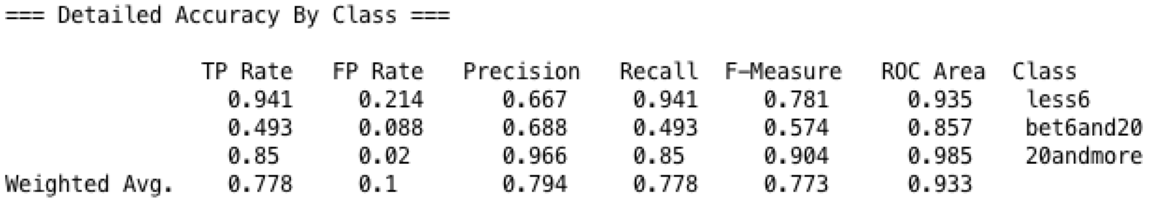
Naive Bayes Classifier We used the Naive Bayes classifier with all the attributes in the first experiment. We got a weighted average F-Measure of 0.379. In the second iteration after normalization and with only body length and number of tags as attributes we got a weighted average F-measure of 0.773. Table 2 highlights our results. There was a significant jump in effectiveness of our classifier by dropping some of the attributes.

J48 Classifier We used the J48 tree based classifier with all the attributes in the first experiment. We got a weighted average F-Measure of 0.403. In the second iteration after normalization and with only body length and number of tags as attributes we got a weighted average F-measure of 0.81 as highlighted in Table 3. Thus a similar trend of sudden jump in effectiveness of classifier was observed when some of the attributes were dropped and only body length and number of tags were used as key attributes.

We thus noticed that attributes like tags of programming languages associated, month of creation did help our classifier. Instead by using only body length and tag length we were able to get good results post normalization.

Conclusion

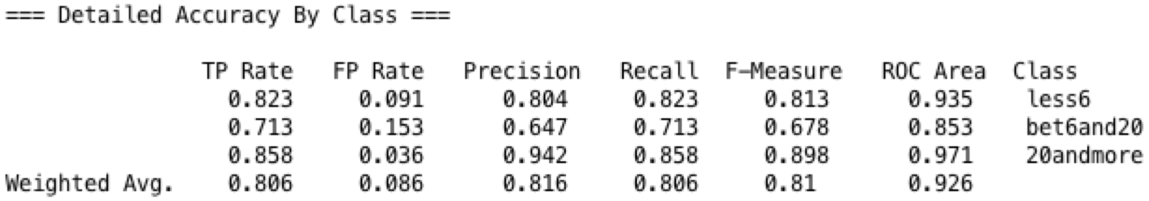
Through our study we were successfully able to find promising trends in scope of top programming languages. From our study we were able to find different activity levels per top programming languages and also the most discussed topics. We also noticed that in many cases the most discussed



```
=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                  0.941     0.214      0.667     0.941     0.781      0.935    less6
                  0.493     0.088      0.688     0.493     0.574      0.857    bet6and20
                  0.85      0.02       0.966     0.85      0.904      0.985    20andmore
Weighted Avg.     0.778     0.1        0.794     0.778     0.773      0.933
```

Table 2: Iteration 2 results using Bayes Classifier



```
=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                  0.823     0.091      0.804     0.823     0.813      0.935    less6
                  0.713     0.153      0.647     0.713     0.678      0.853    bet6and20
                  0.858     0.036      0.942     0.858     0.898      0.971    20andmore
Weighted Avg.     0.806     0.086      0.816     0.806     0.81       0.926
```

Table 3: Iteration 2 results using J48 Classifier

topics were also the most difficult aspect of those programming languages. There is a greater adoption of scripting based languages over the years and this can be proved by the increased popularity of JavaScript. We were successfully able to predict time to answer by training our classifier. We noted that even though we used lesser number of attributes in one experiment our F- Score showed better results. Our study shows that StackOverflow is truly an active community where majority of the questioners get first answers within the first 6 minutes.

Future Work

We wish to add more attributes and study how this data from our machine learning module could be used to develop better interfaces and suggest best text length for description field for faster answering rate. Inclusion of Stanford NER for analysis of data. We also plan to contribute to the Weka community our On the Fly .arff file generator. Also, we wish to showcase our results on an website.

References

*A*llamanis, M. and Sutton, C. 2013. Why, When and What: Analyzing Stack Overflow Questions by Topic, Type and Code. In 10th Working Conference on Mining Software Repositories. Mining Challenge. *IEEE*: 53-56.

*B*hat, V.; Gokhale, A.; Jadhav, R.; Pudipeddi, J.; and Akoglu, L. 2014. Min(e)d your tags: Analysis of question response time in stackoverflow. In Proceedings of ASONAM 2014. *IEEE*: 328-335.

*L*ist of Programming Languages in Alphabetical Order, http://www.scriptol.com/programming/list-programming-languages.php *SCRIPTOL*

*M*achine Learning Group at the University of Waikato, Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/ *University of Waikato*