# Answer features extraction from StackOverflow - An Analysis of StackExchange Questions and Answers

**Jardel Batista Gonçalves**[1], **Rafael Teodósio Pereira**[1], **Simone Regina Ceolin**[1], **Renato Preigschadt de Azevedo** [1]

[1]GRIPI - CTISM – Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brazil

{jardel.goncalves,rafatp,sceolin,r}@redes.ufsm.br

***Abstract.*** *Question & Answering (QA) sites (e.g., StakeExchange (SE) and StackOverflow (SO)) provide a platform where users can ask questions on specialized topics and get feedback provided by users who can have knowledge on the subject. With more than 2.9 billion answers submitted in 2017 is a challenge to get the best answer. In this paper we present preliminary findings based on an analysis of data from a Q&A site, StackOverflow StackExchange, concentrating on discovering characteristics present in the answers that have the Python tag. Identify characteristics of the preferred answers can give us a way to classify which answer has a better probability to be chosen as an accepted answer. We used a quantitative approach to analyze the data and discover if some characteristics are true. We analyzed 5 features and their representativeness, such as scope awareness, explained codes, presence of code, length and time after the question been made. Our findings shows what features are relevant in the answers from the extracted patterns. The exploratory study used in this paper could improve the understanding about characteristics of preferred answers.*

## 1. Introduction

Question & Answering sites (e.g., StakeExchange (SE) and StackOverflow (SO)) provide a platform where users can ask questions on specialized topics and get are feedback provided by users who can have knowledge on the topic. The community sizes in these sites often are very large, in the order of thousands or even millions of users, so is difficult to know characteristics that will make a question or an answer more acceptable. There are prior studies to examine the features that make a good answer in the StackExchange Sites [Mehdi Nasehi et al. 2012], [Bhanu and Chandra ], [Fu and Fan ], [Honsel et al. 2015]. These articles focused on classifying characteristics in the source code fragments presented, not on various features that can be identified in the answers, such as *the use of time, answer length*, among others.

We did an analysis of the Python community on the Stack Overflow site because it has questions and answers that contain several contents like commands, configuration files, scripts, as well as source code. We processed data related to Python programming language also for pragmatic reasons, given the size of StackOverflow dataset.

The structure of the paper is as follows: Related Work is presented in Section 2; in Section 3 some relevant aspects on StackOverflow data are presented; in Section 4 the proposed experiments to extract characteristics is described; The experiments and the discussion for the obtained results are then presented in Section 6 and in Section 7 conclusions and future works are discussed.

## 2. Related Work

There are several studies that explored aspects of Stack Exchange questions and answers. Honsel et al. [Honsel et al. 2015] made a list of myths gathered through a group of developers. They defined nine myths in which all the developers in the group believe. The study tried to prove if the myths are real or not. Four myths about StackOverflow are true, such as new users violate rules more often, positively voted questions are more likely to get an answer. In the study presented by Nasehi et al. [Mehdi Nasehi et al. 2012] the authors made a qualitative analysis on a set of 163 questions of Java subject. The data was gathered by a crawler instead of using the data available through the StackOverflow dump. It was created a list of common good and low-quality attributes, based on the score of the answers.

The work proposed by [Gruetze et al. 2016] tries to identify a relationship between the tags, and topic shifts to the probability of the question be answered. Mamykna et al. [Mamykina et al. 2011] conduct interviews with the creators of StackOverflow and some users (beginners and advanced) to understand what make the SO different from others CQAS. A gender representation and participation in the SO is presented in the paper [Vasilescu et al. 2012], where the authors try to detect the gender of the users. After the detection of the gender, there were discussions about the participation of woman in the StackOverflow.

## 3. Stack Overflow

Stack Overflow is one of the 166 Stack Exchange Community[1]. StackExchange is an Online Social Question and Answering site which allows users to post questions and answers to questions already asked. Stack Overflow contains questions and answers to programming languages. Python programming language is one among several languages discussed on the SO site. We decided to analyze the StackOverflow among another Community Question Answer site (CQAS) because of the public availability of the data, as well as being regularly updated.

StackOverflow works as a regular CQAS with users posting questions, answers, commenting and voting positively or negatively in posts and comments. Users can only modify their posts, having the author of the question the responsibility for choosing an answer as a correct one. Users can register in the SO to be able to vote and maintain a reputation, gaining rights and badges based on the reputation.

### 3.1. Access to Data

SE provides an anonymized data dump through the Archive.org Site [2]. All user contributions are made over the Creative Commons cc-by-sa 3.0 license[3] allowing the data to be made available for any purpose, even commercially. The data is available in two forms: a direct download and a torrent file. There are more than 300 files to download, with

[1] www.stackexchange.com

[2] https://archive.org/details/stackexchange

[3] http://creativecommons.org/licenses/by-sa/3.0/

at least one compressed 7z file[4] available for each subject available in StackExchange. Some subjects have more than one file, given the size of the dataset. The size of all the datasets is 51 GB. Only the StackOverflow dataset has 37 GB of compressed data.

### 3.2. Data Format

Each SE file has at least 8 XML files: Votes, Tags, Users, PostLinks, Posts, PostHistory, Comments, and Badges. The Users file contains the information about the users, like Display Name, Creation Data, and other information. The Badges file includes a relationship between badges and users. The tags used in the SE are inside the Tags file. The contents of the questions and answers are into Posts file. The Comments file contains comments produced by users of SE about the questions and answers that are inside the Posts file. The Posts file has the following information: Id, PostTypeId, ParentID, AcceptedAnswerId, CreationDate, Score, ViewCount, Body, OwnerUserId, LastEditorUserId, LastEditorDisplayName, LastEditDate, LastActivityDate, CommunityOwnedDate, Title, Tags, AnswerCount, CommentCount, FavoriteCount. The field PostTypeId define if the post is a question or an answer.

### 3.3. Statistics

There are approximately 8.7 million registered users in the StackOverflow CQAS. All these users posted nearly a 15.7 million questions, 24.3 million answers, and 65.3 million comments in these questions and answers. As our focus in this study is Python Q&A, the following statistics only concerns the Python subject. The SO contains almost 1 million question and 1.5 million answers, but only 44.8% of the questions are marked as answered by the authors. This low percentage of not-accepted answers is due, maybe, to the need of a manual action from the user that needs to select the preferred response and mark it; to corroborate this interpretation notice that the percentage of the unanswered question was only 13.8%. The average answers per question ratio is 1.6, with one question being answered 191 times. The user selected an answer as a correct one in less than two hours in 72% of the answered questions.

## 4. Experiment Setup

In this section, we present and explain the experiment setup made to analyze the StackOverflow data.

### 4.1. Sample Selection

All data available in the *StackOverflow data dump* as of March 2018 was explored, but only questions with accepted answers were analyzed. Only the data from the file Posts.xml are addressed as we are focused on extract features from questions and answers. We processed only data that has a relation with the Python programming language for pragmatic reasons.

## 5. Data Processing

We downloaded the StackExchange programming data from StackOverflow. Firstly we extract the Questions that have answers from the Posts file. Next, we select only questions

[4] http://www.7-zip.org/7z.html

and answers that has a Python tag associated with the pair Question → Answer. After the extraction of all Question → Answer pairs, we inserted the data into a MySQL [Bell 2012] relational database. All these steps are explained in more detail in Subsection 5.1.

### 5.1. Data Preparation

To process only questions that have answers and have the tag *python*, a Python script was developed to be able to handle the size of the XML file (almost 100 Gigabytes). Firstly the script processes the XML file Posts with the tag PostTypeId equal to 1, that is, Questions, and creates an array with the ids. The entry is also duplicated in the result file, for later processing if the entry also has a *Python* tag. If the entry is an answer, the tag PostTypeId is equal to 2 but is only processed if their ID is in the previously created array. This is necessary to select only answers picked by the User that made the question. This entry is also duplicated in the result XML file. After the execution of that script only pairs Question → Answer with *Python* tag are available in the output file.

After the preprocessing phase, the data was loaded into a SQL database, to allow a simple way to select specific data. We then selected all Questions that contain a valid answer.

### 5.2. Evaluation Parameter Definition

To be able to evaluate the StackOverflow a few features were selected based on a review of the literature presented in Section 2. To analyze the collected data, the data extracted and stored in the database, it was necessary to define the following parameters:

- Score: the difference between positive and negative votes for an answer.
- Accepted Answer: an answer that is identified by the user that posted the question.
- Long Answer: an answer that has at least 1000 non-code characters.
- Explained code: a code block that is presented in segments with textual explanations in between, instead of a simple code block.
- Scope Awareness: an answers that refers to portions of the original question text.

## 6. Experiment Results

To analyze the data from the StackOverflow, we have conducted a statistical analysis of several features presented in the data. To be able to extract features we looked into the related works that have done qualitative analysis.

We decided to rank the answers by quality, looking into a relation between score and answer age, answer size, explained code and scope awareness. The score was the measure used to compare the quality of answers because this value is community curated. The higher the score, the more helpful is the answer [Mehdi Nasehi et al. 2012].

We extracted 545.746 answered questions related to Python programming language from SO dataset. Table 1 presents a few statistics concerning the score of the answers. The percentile was used to extract score under distinct part of the population to compare the behavior of analyzed features.

Table 2 presents the measures used to compare the SO data between several features. The analyzed features sometimes present a direct relation to the score as in mean

|  | Mean | Median | min | max | .25 percentile | .50 percentile | .75 percentile |
|---|---|---|---|---|---|---|---|
| Score | 2.0 | 4.78 | -38 | 11384 | 1 | 2 | 3 |

**Table 1. Score values of StackOverflow extracted data**

| Answer Score | <= 1 | > 1 && <=2 | > 2 && <= 3 | > 3 |
|---|---|---|---|---|
| Answer Count | 261312 | 97824 | 57342 | 129268 |
| Mean Answers Count | 1.54 | 1.69 | 1.84 | 2.55 |
| Explained Code | 63.32% | 70.65% | 72.17% | 72.55% |
| Scope Awareness | 2.50% | 2.18% | 2.12% | 1.90% |
| Long Answer | 28.73% | 31.18% | 31.62% | 32.68% |
| Median seconds after question | 2496 | 1274 | 940 | 783 |
| Mean Answers Comments | 1.69 | 1.93 | 2.02 | 2.45 |

**Table 2. Values of features from answers extracted from StackOverflow**

answers count, explained code, long answer and mean answer comments. In the features Scope awareness and median seconds after question, the relation is inversely proportional.

The best answers were quickly answered, have more comments, more explained code blocks, and, slightly more content than the answers with smaller scores. Over the 545.746 answered questions almost 12.2% do not have programming code. Figure 1 shows a histogram graph with the time elapsed after the questions, evidencing that the majority of accepted answers were answered in less than 26 minutes (50.41%) and 16.13% in less than 5 minutes.

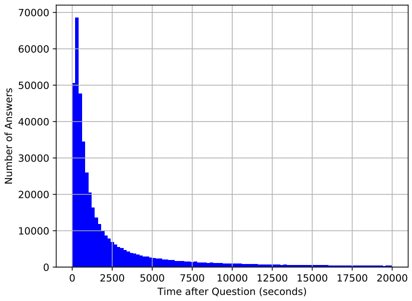

**Figure 1. Histogram of the elapsed time after the question until the user selected a correctly answer**

## 7. Conclusion and Future Work

In conclusion, our study has provided valuable insights into the features that contribute to better answers on platforms like Stack Overflow. By conducting a quantitative analysis inspired by various research papers, we have advanced our understanding of what constitutes a high-quality answer within the context of Question and Answering (Q&A) systems.

The findings from our experiments have allowed us to identify recurring patterns and emphasize the importance of specific features in crafting effective answers. This knowledge serves as a foundation for future research and improvements in Q&A systems.

Looking ahead, our future work will encompass a broader spectrum of subjects, with a particular focus on computer networks. We intend to extract relevant data from Stack Overflow to build and refine a specialized Question and Answering system tailored to the field of computer networks. This endeavor will not only deepen our understanding of domain-specific Q&A dynamics but also provide a practical application of our research findings.

Additionally, we plan to expand our analytical methods by conducting more in-depth statistical analyses. This will enable us to gain further insights into the nuances of effective answers and refine our feature extraction process, ultimately enhancing the performance of Q&A systems in various domains. Through these ongoing efforts, we aim to contribute to the continuous improvement of knowledge sharing and problem-solving on platforms like Stack Overflow. We will also conduct a more extensive analysis to compare and evaluate whether the findings remain consistent when changing the programming language being analyzed.

## References

Bell, C. (2012). *Expert MySQL*. Apress, Berkely, CA, USA, 2nd edition.

Bhanu, M. and Chandra, J. Exploiting Response Patterns for Identifying Topical Experts in StackOverflow.

Fu, H. and Fan, Y. Music Information Seeking via Social Q&A: An Analysis of Questions in Music StackExchange Community.

Gruetze, T., Krestel, R., and Naumann, F. (2016). Topic shifts in stackoverflow: Ask it like socrates. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S., editors, *Natural Language Processing and Information Systems*, pages 213–221, Cham. Springer International Publishing.

Honsel, V., Herbold, S., and Grabowski, J. (2015). Intuition vs. truth: Evaluation of common myths about StackOverflow posts. In *IEEE International Working Conference on Mining Software Repositories*.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. (2011). Design lessons from the fastest q&#38;a site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2857–2866, New York, NY, USA. ACM.

Mehdi Nasehi, S., Sillito, J., Maurer, F., and Burns, C. (2012). What Makes a Good Code Example? A Study of Programming Q&A in StackOverflow.

Vasilescu, B., Capiluppi, A., and Serebrenik, A. (2012). Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*, pages 332–338.