An analysis of developer interest into machine learning

Abhishek Subedi
Universität Passau
subedi01@ads.uni-passau.com

ABSTRACT

Engineers, programmers, developers are using various kinds of online forum to ask about the problems they face during their research and development. Platform like Stack Overflow has proven to be very successful in creating a collaborative environment between users. Researchers are exploring the interest of users to better explore the state-of-use of various topics, tools and platform. In this paper we did a research to understand the interest of users in machine learning topics using the data from Stack Overflow. We used LDA algorithm to model the topics and find the topics of interest from the questions posted by users in Stack Overflow. As a result we came to know that users are more interested in topics like "**Keras**", "**Tensorflow**", "**Neural Network**" etc.

KEYWORDS

stack-overflow, machine learning, LDA, topic of interest

ACM Reference Format:
Abhishek Subedi. 2018. An analysis of developer interest into machine learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Online forums are widely used by various professionals and Stack Overflow is being liked by the software development groups. It constitutes more than 4.5 million questions posted by 1.5 million users and these data can be retrieved through StackOverFlow API or several other APIs. The information topics in the platform ranges from algorithms to programming languages [Allamanis and Sutton 2013]. As the field of software and computer science has expanded tremendously, the technologies used for those sectors are also growing. Developers and engineers face various questions and problems in the process of development, so they seek help from other professionals using platforms like Stack Overflow [Barua et al. 2014b]. Stack Overflow allows users to find answers by querying or posting questions on the platform. The standards of the questions and answers are ranked by the community. Positive votes ensure the questions are good and negative votes ensure the questions are not so good. These votes help to gain the "reputation points" for the questions. Similarly reputations of contributors are measured based on their answers posted for the questions. Researchers examine such forums to find out facts, figures, patterns, interest of the users [An et al. 2008] and analyse users interest and problems. In this research we are focused on finding the topic of interest of users in machine learning.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 RESEARCH OBJECTIVE

In this research we are analysing the developer interest into machine learning. Machine learning in itself is a huge field of artificial intelligence. Developers and engineers are working in various sectors of ML and their problem of interest might vary. Hence we are focused on the research question given below.

2.1 Research questions

- **Which topics related to machine learning are Stack Overflow users interested in?**
  Understanding the interest or the problems of users can help forums like StackOverflow to provide best possible answers. This information also can help the tools and technology vendor companies to improve their products according to the users interest. The information can also benefit publishers of books and research papers to focus the area in which users are more interested in. On the other hand it can be a topic of research that can have a high impact on practice in future.

3 RELATED WORK

We looked at related work in various perspectives: one is obtaining data from Stack Overflow, other that implements

LDA algorithm with the obtained data and third one that directs the popular topics or interest.

StackOverFlow datasets are used in empirical studies on the basis of research criteria. Research done in [Barua et al. 2014a], uses posts data which are made publicly available by Stack Overflow. It constitutes the text content of the posts, type of post, view count, creation date, favorite count and information about the user that created the post. The research used Latent Dirichlet allocation (LDA) to model the popular topics from the datasets. Their main focus of research was to analyze topics and trends in stack overflow and find out what developers are talking about. Similarly, the search [Villanes et al. 2017], explores "what are software engineers asking about android testing on stack overflow?". They mined eighteen thousand questions concerned with "android" [Villanes et al. 2017] between a certain time frame. This research also used Latent Dirichlet Allocation (LDA) algorithm to cluster or categorize the acquired data. They presented the main problems and challenges, engineers and developers are suffering from.

4 RESEARCH DESIGN

From the research question, we were very clear and focus on our research topic. To give a structure to our research we defined a guideline and worked accordingly. The Figure 1 and following descriptions give us an overview of our research design.

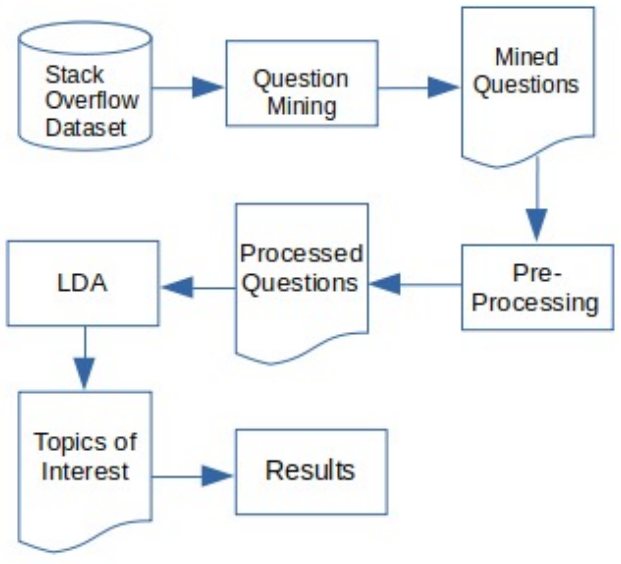

Figure 1

4.1 Research question

What topics related to machine learning are stackoverflow users interested in?

4.2 Data Strategies

As we have set a research question, data from stackoverflow that can justify the research question are mined for pre-processing and analysis. First we gather the data, analyze it and present the result supporting the research question as shown in Figure 2.

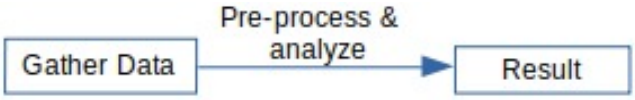

Figure 2

Since empirical research makes use of data, we follow a Purposive data sampling strategy, where we gather data according to predefined and relevant criteria to the research question [Mack et al. 2005]. Our predefined criteria is to get only the data tagged with keyword "machine-learning".

4.3 Data mining

To mine the dataset from stack overflow we used StackAPI. We were focused on the questions that were posted in the Stack Overflow by users and only the questions which are tagged with 'machine-learning' keywords. The StackAPI fetches the most recent 500 questions from Stack Overflow. We considered only the question posts because it is the starting point of users interest and are also able to represent the interest of users. So at the end we had 500 different questions which are the most recent post tagged with 'machine-learning'.'

4.4 Pre-processing

Since the extracted data are not well structured and clean to be used for further processing, we need to pre-process the data. The data must be pre-processed such that it is ready to use for input to the algorithms and visualisation. After the data mining is done, we pre-process the data in following steps.

- First the data are loaded in the pandas data structure with files in json format.
- Since the json data file is huge, only the questions are extracted and set on pandas data structure.
- All the irrelevant data like stop words, numerics, symbols are removed from the text file using regex in python.
- Finally we have clean questions exported in .csv file.

The following Figure 3 give and overview of pre-processing.

2	tensorflow: how to save/restore a model?
2	tensorflow how to save restore a model

Figure 3

4.5 LDA Implementation

We have applied Latent Dirichlet Allocation (LDA), which is a statistical tool for modelling topics which are of main interest. LDA is one of the most used and successful algorithms applied in various sectors for document modelling. LDA depicts text documents as distribution of probability with respect to the topics that are identified . This helps to automatically find out the topics of interest from the questions and analyse the interest of users. We visualise the topics discovered and do qualitative analysis of the knowledge gained. As an example, if the algorithm models "neural network" as one of the topic then, we can say that "neural network" is a topic of interest to the users. As LDA uses document probability distribution with respect to all the topic individually, we can get an idea about the involvement of topic in a document. That mean its shows the topics that the documents are talking about. The figure below gives and graphical model for LDA [Tong and Zhang 2016].

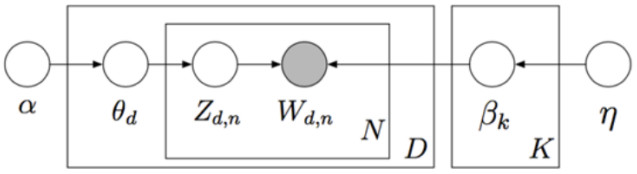

Figure 4

The following notation gives more understanding of LDA.

- α : proportion parameter
- η : topic parameter
- β_K : distribution over vocabulary
- θ_d : topic proportion for d-th document
- $\theta_{d,k}$: topic proportion for topic k in d document
- Z_d : topic assignment for d document
- $Z_{d,n}$: topic assignment for n word in d document
- W_d : noticed word for document d
- $W_{d,n}$: n word in document d, ie. an element from vocabulary which is fixed

With the given notations, the approach of LDA directs to the given joint distribution of the observed and hidden variables.

- $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = a * b * c$
  - $a = \pi_{i=1}^{K} p(\beta_i)$
  - $b = \pi_{d=1}^{D} p(\theta_i)$

Table 1: Some most frequent words

Frequent words
Keras
Tensorflow
Neural Network

  - $c = \pi_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})$

The distribution depicts the amount of dependencies in number. Here $Z_{d,n}$: topic assignment rely on topic distribution for each document θ_d. Also word $w_{d,n}$ rely on assignment of topic $Z_{d,n}$ and all the topics $\beta_{1:K}$ [7].

5 IMPLEMENTATION AND RESULTS

5.1 Demographics

As we analysed the questions only from the Stack Overflow, we obtained some demographics that depicts and overview of highly interested topics. The Figure 5 is a word cloud which illustrates the words that are most common words occurring in the questions.

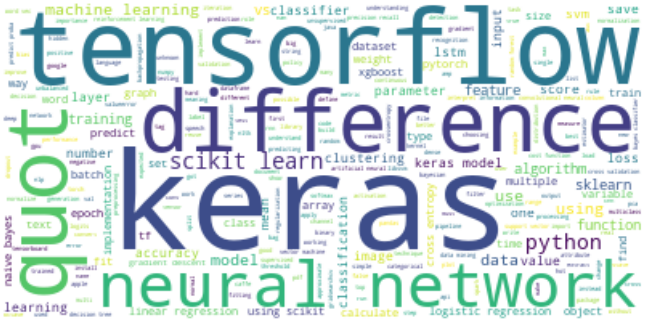

Figure 5

Though the word cloud does not model a topic from the questions, it certainly does provide us an idea about the most widely used keywords to communicate and talk about machine learning. The Figure 6 give another overview of how frequently the word are occurring. The figure-5 and figure-6 also partially explains the interest of users because users will be talking about the topic that they are more interested in. But here our aim is to find out the topic of interest using topic modelling techniques called Latent Dirichlet Allocation (LDA).

From the demographics above we can say that some of the words given in the Table 1 are more frequent words.

5.2 LDA Result

Table 2, shows the topics that are modelled by LDA using questions as dataset. Below is the output that has been

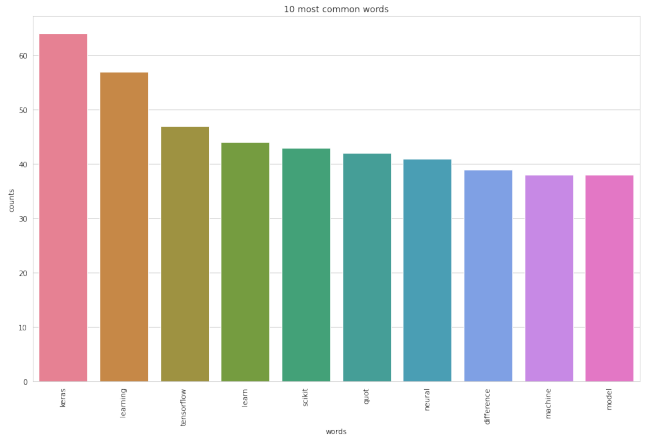

Figure 6

Table 2: Topics Modelled by LDA

Topic No.	Topics
1	Scikit, learn, Tensorflow, Pytorch, Entropy, Differences
2	Linear Regression, Keras, Features, machine learning, Python
3	Naive Bayes, Classifier, Classification, SVM, Word
4	Neural Network, Data, Model, Keras, Scikit
5	Graph, Neural Network, Tensorflow, Keras, machine learning

achieved implementing our research approach on our dataset. Our main target here is to answer our research questions. The following is our research question and its answer.

- **Which topics related to machine learning are Stack Overflow users interested in?**
  - By the application of LDA, 50 topics were discovered by our research approach. Some topic like "**linear**" & "**regression**", "**neural**" & "**network**", "**naive**" & "**bayes**" were modelled separately. These keywords are more of same concept and are not complete with each other in machine learning. Hence these kinds of topics are combined and presented in our result. To have a good and better result, 20 topics that are mostly frequent among all topics are shown in Table 2. In some of the clusters, the topics of interest were also frequent. Topics like "**Keras**", "**Tensorflow**", "**Neural Network**" are the most repeatable topics in more clusters, clustered by LDA. In a general sense of today technological development of machine learning, these topics are the current popular topics engineers and developers are looking forward too.

5.3 Evaluation of Validity

The validity of an experimental research is to ensure the research is on the right direction and no false prediction and result comes from the output of the research . Certain criteria and guideline have to be applied to make sure that threats to validity are minimised by the research approach that have been adopted. Now we evaluate the threats to validity for our research with respect to various validity types [Runeson and Höst 2009].

5.3.1 Internal validity. This validity is associated with causal relation analysis. As the experimenter is doing research on a topic, either one factor is influencing an investigated factor or there is always a threat that the inquired factor is also affected by other factor. The researcher must be aware to what extent it is influencing the inquired factor, so as to minimise internal validity.

The internal validity could be occurred due to the unwanted values in the dataset, which are minimized by applying certain data cleaning procedure mentioned above in the pre-processing section. The terms that are occurring mostly in the text might influence the result and from the result we can see that there are similarity in between the interested topics and the most frequent topics. Hence we are able to ensure Internal validity.

5.3.2 External Validity. When an empirical experiments are performed the output of a particular research is seen beyond the research and analysed to what extend it holds true outside the context of study. In more general terms, to what degree the output of our research can be generalised. The generalization of research can be seen in many different scenarios: ie. is it generalizable from lab or research group to the mass population and in various research. And whether it generalize from the research study to real life problems. It must not be hypothesised but replication must be applied to make it in track.

We have used only the dataset from Stack Overflow that are tagged with keyword "machine-learning". Stack Overflow being on of the top forum for programmers and developers, still further experiments is required with other popular QA websites. This is one possible threat of our result. So before concluding that it ensures the external validity, we must apply the research in various other popular QA websites. On the other hand the dataset are the most recent one, where topics are modelled only for those data and the data are not used for future predictions.

5.3.3 Construct validity. while doing experiments and research, there must not be distinction between what researcher wants to do and what is being investigated as per the research questions. Here our research was mainly focused on the finding the topic of interest in machine learning from Stack Overflow users. The research methodology and the result of our experiments is also directed to the research question. Hence was able to ensure construct validity.

6 CONCLUSION

Our research methodology was focused on finding the topics related to machine learning in Stack Overflow and explore the topics of interest of the users. We used statistical topic modelling algorithm called Latent Dirichlet Allocation (LDA), it is one of the most widely used statistical tool for modelling topics in various sectors for document modelling.

Developers and engineers encounter various problems in the process of development. Having knowledge and keeping track of every problems becomes quite difficult. Hence platform like Stack Overflow provides a medium where millions of users come up with questions and answers for their problems. On the other hand experimenters are able to use the dataset freely from websites like Stack Overflow to do research on various trends and topics.

This experiment gave us overview of interest of users and developers in Stack Overflow. From the topics modelled, we encountered that topics like "**Keras**", "**Tensorflow**", "**Neural Network**" are interesting for users. This information can be beneficial to QA websites like Stack Overflow to make their service better by enhancing their results. Technology and tool vendors like google who are designing frameworks like "**Tensorflow**" can enhance their tools as per the requirement and need of users experience. Researcher and publishers of books and research materials can be focused on the topics of interest of users. This can also be a topic of research to explore more and benefit the future that can yield high impact on practice.

7 FUTURE WORK

As we performed the research in very short period of time, our area of research question was very certain and limited with small dataset. So we intend to increase this criteria and make the research big enough to explore more queries. Since we only worked on Stack Overflow dataset, this also need to be expanded to various QA platforms with increased size of data, as a result of which external validity can be ensured.

REFERENCES

Miltiadis Allamanis and Charles Sutton. 2013. Why, When, and What: Analyzing Stack Overflow Questions by Topic, Type, and Code. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (San Francisco, CA, USA) *(MSR '13)*. IEEE Press, 53–56.

Zol An, G Ongyi, Georgia Koutrika, Jan Pedersen, and Hector Garcia-Molina. 2008. Questioning Yahoo! Answers. (01 2008).

Anton Barua, Stephen Thomas, and Ahmed E. Hassan. 2014a. What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering* 19 (06 2014). https://doi.org/10.1007/s10664-012-9231-y

Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. 2014b. What Are Developers Talking about? An Analysis of Topics and Trends in Stack Overflow. *Empirical Softw. Engg.* 19, 3 (June 2014), 619–654. https://doi.org/10.1007/s10664-012-9231-y

Natasha Mack, Cynthia Woodsong, Kathleen M. MacQueen, Greg Guest, and Emily Namey. 2005. Qualitative research methods: a data collectors field guide.

Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering* 14 (2009), 131–164. https://doi.org/10.1007/s10664-008-9102-8

Zhou Tong and Haiyi Zhang. 2016. A Text Mining Research Based on LDA Topic Modelling. *Computer Science Information Technology* 6, 201–210. https://doi.org/10.5121/csit.2016.60616

Isabel K. Villanes, Silvia M. Ascate, Josias Gomes, and Arilo Claudio Dias-Neto. 2017. What Are Software Engineers Asking about Android Testing on Stack Overflow?. In *Proceedings of the 31st Brazilian Symposium on Software Engineering* (Fortaleza, CE, Brazil) *(SBES'17)*. Association for Computing Machinery, New York, NY, USA, 104–113. https://doi.org/10.1145/3131151.3131157