IDENTIFCATION OF TRENDS IN TECHNOLOGIES AND PROGRAMMING LANGUAGES USING TOPIC MODELING

[1]Neha Shinde, [2]Prachiti Thorat, [3]Nidhi Shinde, [4]Shweta Malge, [5]Asst. Prof. S.A.Hadke

1,2,3,4UG Students, Information Technology Department, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India,

5 Professor, Department of Information Technology, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India.

Abstract-
Websites providing technical questions and answers are a great source of knowledge. Technical questions are asked and also answered by the users of questions and answer website. These questions cover a wide range of domains in Computer Science like Networks, Data Mining, Multimedia, Multithreading, Web Development, Mobile App Development, etc. The results show that these techniques help discovers dominant topics in developer discussions.

Keywords-* *Topic modeling, Latent Dirichlet Allocation(LDA), Machine Learning, Natural Language Processing.

INTRODUCTION

The field of computer science field has a large number of technologies. Every day we are introducing the new technologies and the technology is changing in rapid pace. So, in order to keep pace with ever-changing technology, developers sharing their knowledge areas and seeking help from other developers on areas where they have less knowledge. Stack Overflow is one of the question and answer website which provides such a platform. Developers can discuss a wide range of technical topics among themselves and share knowledge. Understanding these topics could allow programming language and tool developers to understand usage trends, commercial vendors to assess the adoption rate of their products, and question and answer sites to perceive the usage patterns of their information content. Textual data of websites such as Stack Overflow can be analyzed to understand the trending topics. Analyzing the textual content of these websites can help computer science and software engineering community better understand the needs of developers and learn about the current trends in technology. In this paper, textual data from famous question and answer website called Stack Overflow is analyzed using Latent Dirichlet Allocation (LDA) topic modeling algorithm. These topics are analyzed to find a number of observations such as popular technology/language, impact of a technology, technology trends over time, the relationship of a technology/language with other technologies and comparison of technologies addressing area of computer science or software engineering.

LITERATURE SURVEY

[1] Y. Chen, R. Dios, A. Mili, L. Wu, and K. Wang. An empirical study of programming language trends. IEEE Software, 22:72--78, May 2005.

Predicting software engineering trends is a strategically important asset for both developers and managers, but it's also difficult, due to the wide range of factors involved and thecomplexity of their interactions. This paper reveals some interesting trends and a method forstudying other important software engineering trends.

[2] J. Dickey, J. Jiang, and J. Kadane. Bayesian methods for censored categorical data. Journal of the American Statistical Association, 82: 773-781, 1987.

Bayesian methods are given for finite-category sampling when some of the observations suffer missing category distinctions.Dickey's (1983) generalization of the Dirichlet family of prior distributions is found to be closed under such censored sampling. The posterior moments and predictive probabilities are proportional to ratios of BC Carlson's multiple hypergeometric functions.

[3]T. Fritz and G.C.Murphy. Using information fragments to answer the questions developers ask. In Proc. Of the 32nd Intl. Conf. on Software Engineering- Volume 1, ICSE'10, pages 175—184, New York, NY, USA, 2010. ACM.

Each day, a software developer needs to answer a variety of questions that require the integration of different kinds of project information. We introduce an information fragment model that automates the composition of different kinds of information and that allows developers to easily choose how to display the composed information.

[4]ChengXiang Zhai**.** Statistical Language Models for Information Retrieval A Critical Review.

Statistical language models have recently been successfully applied to many information retrieval problems. A great deal of recent work has shown that statistical language models not only lead to superior empirical performance, but also facilitate parameter tuning and open up possibilities for modeling nontraditional retrieval problems. The purpose of this survey is to systematically and critically review the existing work in applying statistical language models to information retrieval, summarize their contributions, and point out outstanding challenges.

[5] Margaret-Anne Storey. How do programmers ask and answer questions on the web.

In this paper, we analyze data from Stack Overflow to categorize the kinds of questions that are asked, and to explore which questions are answered well and which ones remain unanswered. Our preliminary findings indicate that Q&A websites are particularly effective at code reviews and conceputal questions.

Design, Description and Implementation

A. Architecture

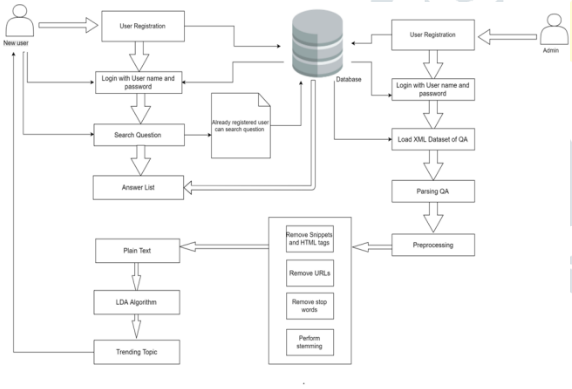

Fig (1): System Architecture

Latent Dirichlet Allocation is a “generative probabilistic model” of a set of composites made up of different parts. In the findings of topic, composites are the documents and its parts are words or phrases. Latent Dirichlet Allocation (LDA) is a generative probabilistic model i.e. topic bag of words model that automatically finds topics in text corpus. This model regards each document as a combination of various topics, and that each word in the document belongs to one of the document’s topics. Latent Dirichlet Allocation is useful when you have a set of documents, and you want to discover patterns within, but without knowing about the documents themselves. Latent Dirichelet Allocation is a Bayesian model in which each item of the collection is modeled as a finite mixure over an underlying set of topics.

B. Implementation

The experimental platform is built on using JAVA8 and MYSQL. Experiments are performed on computer with 2.4GHz CPU and 2GB RAM.

C. Working of LDA algorithm

LDA is an unsupervised approach used for finding and observing the bunch of words in large clusters of text. There are five steps involved in the implementation :
1)Data Extraction
2)Pre-Processing
3)Topic Modeling
4)Post Processing
5)Inferring result and creating visualization of trends

D. System Overview

- This application is used to take input from the user in the form of the post (Question). The system will provide an answer according to the question.
- To identify current Trending Topic in technologies and programming language NLP and machine learning is used.
- After identifying Trending Topic the Notes can be provided to the user.

II. Results

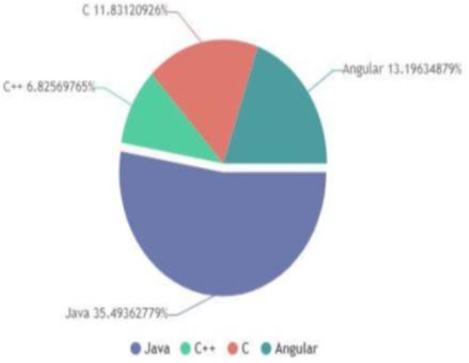

Fig .(2): Result

The above pie chart shows the expected precise output. It shows which topic(technology or programming language) is trending and by what percentage. In the above pie chart, Java is the most trending topic by (35.49%), the second most trending topic is angular (13.19%), the third most trending topic is C(11.83%), and the fourth most trending topic is C++(6.82%).By this graphical representation, user can have an idea on which topic he should be an expert according to the trend shown in the pie chart.

III. Future Scope

Using the LDA algorithm the communication gap can be reduced between an expert of technical field and non-expert while ensuring topic modeling in order to analyze the trends in technologies. So, we proposed a system to minimize the communication gap between them and help to take different domain knowledge.

IV. Conclusion

Text data of Stack Overflow website was analyzed using well known topic modeling algorithm called LDA. The analysis was done on stack Overflow dataset. Dataset contains user and developers discussion post in the form of Question Answers. The topics are meaningfully labeled based on top words allocated by LDA. Result will show top word technology i.e. trending topic. The results of this analysis will help both developers and commercial vendors track latest trends in technology and programming languages.

.

V. REFERENCES

[1] Blei, D. M., Ng, A., Jordan, M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research, pp. 993-1022 Volume 3, 2003.

[2] Meyerovich, L. A., Rabkin,A. S. Empirical analysis of programming language adoption. ACM SIGPLAN Notices – OOPSLA '13, pp. 1-18, Volume 48, Issue 10, Oct 2013.

[3] Barua, A., Thomas, S. W., Hassan, A. E. What are developers talking about? An analysis of topics and trends in StackOverflow. In Empirical Software Engineering, pp.619-654, Vol 19, Issue 3, 2014.

[4] Tyler Doll, LDA Topic Modeling (2018).

[5] Tegawendé F.Bissyaé Popularity Interoperability, and Impact of Programming Languages in 100,000 Open Source Projects.

[6]Stack Exchange - https://archive.org/details/stackexchange

[7]Natural Language toolkit - http://www.nltk.org

[8] Thomas W.Jones, Topic Modeling (2019).

[9]Machine Learning Toolkit - http://mallet.cs.umass.edu/topics.php

[10] Tegawendé F.Bissyaé Popularity Interoperability, and Impact of Programming Languages in 100,000 Open Source Projects.

[11] Chong Wang Collaborative Topic Modeling for Recommending Scientific Articles.

[12] David M. Blei Latent Dirichlet Allocation

[13] Bissyand, T., et al: Popularity, Interoperability, and Impact of Programming Languages in 100,000 Open Source Projects: Computer Software & Apps. Conference (COMPSAC), pp. 303-312, 2013.

[14] Wang, C., Blei, D. M. Collaborative topic modeling for recommending scientific articles. In ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, pp. 448-456, August 2011.

[15] David Andrzejewski, Xiaojin Zhu, Mark Craven, Ben Recht.