Introducing a System that Identifies Trends in Technologies and Programming Languages using Topic Model Algorithm

Megha Peddi Akshaya Shelke Mayuri Shinde

Guide Name: Prof. Shital.B.Jadhav

Abstract :

There are various sources to get technical knowledge. Discussion is the best way for achieving it i.e technology related question and answer are a great source of technical knowledge where users of these websites put on various types of technical question and answer them. The questions raised can cover a wide range of domains like Networks, Machine Learning etc. as well as various technologies. Analyzing the actual textual content of these website can help the software community better understand the needs and learn about the current trends in technology. In this project, We consider the data from famous Question and Answer section website called Stack Overflow and is analyzed using one of the Topic Modeling Algorithm (Latent Dirichlet Allocation). The results show that these techniques help discovers dominant topics in developer discussions. These topics are analyzed to find a number of interesting observations such as popular technology/language, impact of a technology, technology trends over time, a relationship of a technology/language with other technologies and comparison of technologies addressing an area of software engineering.

I. Introduction

Computer science field has a large number of technologies. Every day we are introducing the new technologies and they are changing in rapid pace. So, in order to keep pace with ever-changing technology, developers share their knowledge areas and seek help from other fellow developers on areas where they have less knowledge. Question and answer websites like Stack Overflow 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 topic. Here, we analyze data from website and generate a trending topic that would help students to know about new technologies. The output would be in form of dashboard which would make it easier to visualize and understand the trends.

Keywords
Topic modeling, Latent Dirichlet Allocation (LDA), Machine Learning, Natural Language processing.

.

II. RELATED WORK

Latent Dirichlet Allocation is a “generative probabilistic model” of a set of composites made up of different parts. In finding 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 Dirichlet Allocation can be used to generate topics to understand a document’s general theme, and is often used in recommendation systems, document classification, data exploration, and document summarization. Additionally, Latent Dirichlet Allocation is useful in training predictive, linear regression models with the topics and occurrences.

1. What are developers talking about? An analysis of topics and trends in Stack Overflow

Description: Stack overflow site provides programming question and answers which is nothing but the knowledge. These provide the expertise to help end users technically . As developers use multiple platforms for development and hence through this discussion it can pinpoint the major areas of interest for developers.

2. Empirical Analysis of Programming Language Adoption

Description: Some programming languages become extensively trendy while others fail to grow beyond their role or fade away altogether. Understanding this process is a initial step towards enabling language designers and advocates influencing its outcome and overall language use. This paper uses study methodology to recognize the factors that lead to language adoption.

3. Popularity Interoperability, and Impact of Programming Languages in 100,000 Open Source Projects

Description: Programming languages have been proposed even before the era of the modern computer. As years have gone, computer resources have increased and application domains have expanded, leading to the proliferation of hundreds of programming languages, each attempting to improve over others or to address new programming paradigms and to become popular. What is the impact of the programming language used can also be analyzed?

4. Collaborative Topic Modelling for Recommending Scientific Articles

Description: Newly formed online communities of researchers sharing citations provide a new way to solve this problem. In this paper, an algorithm is developed to recommend scientific articles to users of an online community.

5. Latent Dirichlet Allocation

Description: We describe Latent Dirichlet Allocation, is a generative probabilistic model i.e. topic bag of words model for collections of distinct data such as text corpus. Latent Dirichlet Allocation is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics.

III. MOTIVATION

Motivation was driven by the problems faced by us, as we being students we face problems for choosing technology for our career. We thought of why not introducing a system that would take down the burden of students while choosing technologies. The question answer websites provides reviews/ comments in the form of text. We only see the tags for technologies but we don't have any system which will find current topic which is in trends now days. So from only text we cannot find the current trending topic. For that purpose the system is proposed which will identify the current trending topic.

IV. MATHEMATICAL MODEL

S= {s,e,I,X, F, O}

Where, S = Proposed system.

s = Start of the system

1. Register/Login into the system

I = Input of system (Search Text).

e = End of the System

X = Functions of the system.

O = Output of the system (Answered of trending topic).

X = {x1, x2,x3}

- x1= Pre-processing of Data.
- x2= Create Document
- x3= Trending Topic.

A=Success of System = 1. Trending Topic = According to the input by user.

F= Failure of System

1. Login Failure
2. Database connection

- First, user provides Input as a post or a question to the system.
- System extracts the words related to technology from the post entered ,processes it and then finds trending topic
- Let X be the document created by the machine

- Input from user will be compared with the data available in Database/Document, Result will be the trending topic as per the data available.

V. SYSTEM ARCHITECTURE

In the proposed system, we are going to find the current trending topic based on Text. There are five main steps involved in the implementation: data extraction(taking the stack overflow dump data available on stackexchange website), data pre-processing(Remove the snippets, stopwords, url, stemming), topic modeling(Gibbs sampling), post-processing, inferring results and creating visualization of trends. Topic modeling using LDA algorithm to analyze the trends in technologies and languages. From text corpus we need to extract topics, it will be done performing Natural Language Processing on the text corpus after this we pre-process the data by removing stopwords, url and unwanted content. After pre-processing a document is created and when topic modelling is applied on the processed document then it forms another text document which is topic wise distribution (Topic and Topic membership). Then user searches a question which will be the test data to our system and related answers/answer list will be provided to the question searched by system. As the machine is already trained previously by the stack overflow dataset. On the basis of question and answer trending topic will be detected. The LDA algorithm is applied on QA corpus to find trending topic. After getting current trending topic it is visualized using graphs & dashboards and accordingly study material will be provided to the user.

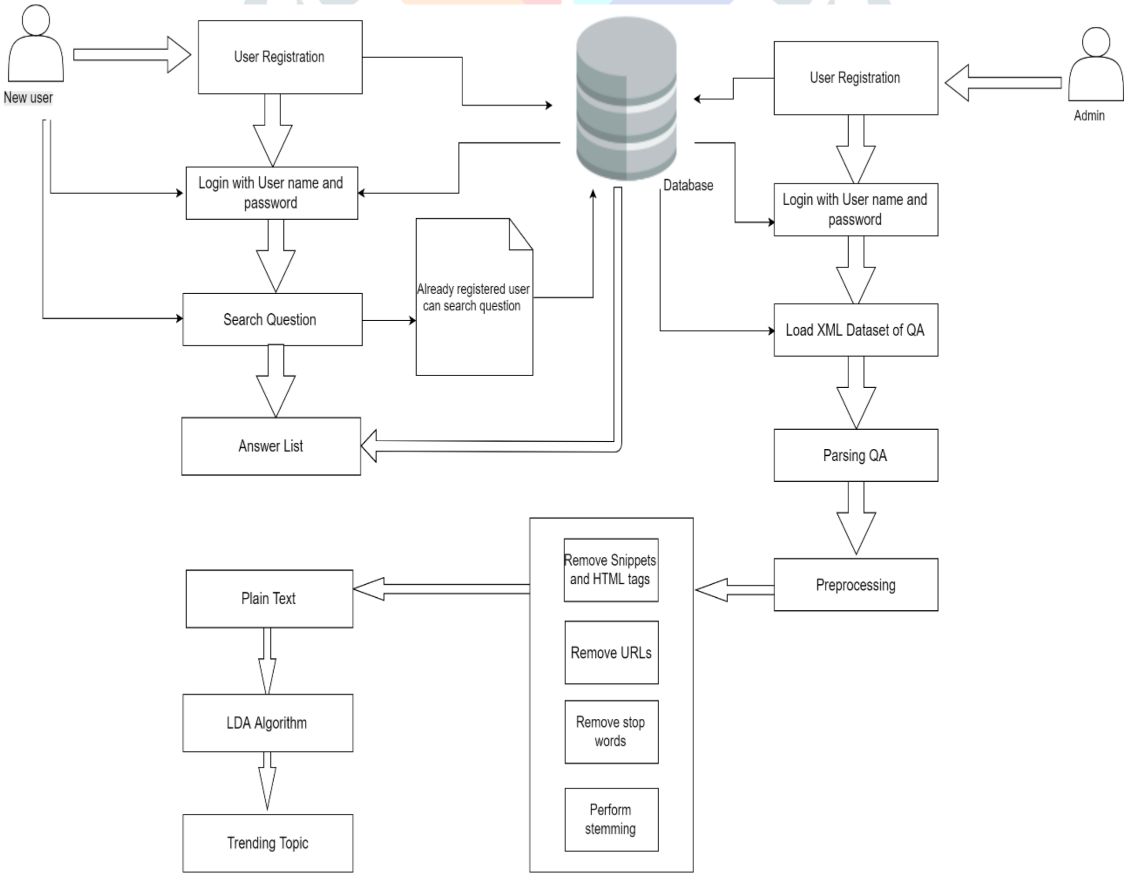

Fig. The Proposed System

VI. RESULTS:

As a probabilistic module, LDA can be readily embedded in a more complex mode. LDA model would help to discover the topics such as Problem/Solution, Analyze/Improve Approach and QA and Links. LDA can be applied to a smaller time interval of posts which may help reveal topics that emerge during specific days. So that, latest discussion about questions and answers can be considered. The frequency of topic gives indication of the rise or fall of interest in a particular topic. Topic Relationship helps to make grouping together closely coupled topics. Our system would help the users to search the questions and get related answer according to the question. If answer is not found in the dataset then that question will be displayed on home page of the system. Users can post our answers to the unanswered questions. The answers of the unanswered questions will be sent to that particular person via mail. Also, system will graphically show the trending topics with the help of Pie-Chart. And also whatever links related to the topic will be provided.

VII. 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.

VIII. ACKNOWLEDGEMENT

It gives us great pleasure in presenting the preliminary project report on **‘Introducing a system that identifies trends in technologies and programming language using a topic modeling algorithm’.**

We would like to take this opportunity to thank my internal guide for giving me all the help and guidance we needed we are really grateful to them for their kind support. Their valuable suggestions were very helpful.

we are also grateful to Head of Computer Engineering Department, for his indispensable support and suggestions.

IX. 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] 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.

[5] 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.

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

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

[8] Porter, M. F. An algorithm for suffix stripping. In Readings in information retrieval. Morgan Kaufmann, pp. 313-316, 1999.

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

[10]https://medium.com/@tomar.ankur287/topic-modelingusing-lda-and-gibbs-sampling-explained-49d49b3d1045 for Gibbs sampling