Using Topic Modeling Identifying Trends in Technologies and Programming Languages

1Prof.Shital B. Jadhav, 2Akshaya Shelke, 3Mayuri Shinde, 4Megha Peddi
1Assistant Professor, 2,3,4Students
Computer Engineering
Bharati Vidyapeeth College of Engineering for Women, Pune-43.

Abstract: Technology question and answer websites are a great source of technical knowledge. Users of these websites raise various types of technical questions and answer them. These questions cover a wide range of domains in Computer Science like Networks, Data Mining, Multimedia, Multithreading, Web Development, Mobile App Development, etc. Analyzing the actual 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 project, textual data from famous question and answer website called StackOverflow is analyzed using Latent Dirichlet Allocation (LDA) topic modeling algorithm. The results show that these techniques help discover 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 computer science or software engineering.

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

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 programmers 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 StackOverflow can be analyzed to understand the trending topics. Analyzing the actual 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. 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.

Related work:

Latent Dirichlet Allocation is a “generative probabilistic model” of a set of compositions 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 Dirichlet Allocation can be used to generate topics to understand a document’s general theme, and it 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 answers which is nothing but the knowledge. These provide the expertise help to end users in the form of technical help. This kind of sites contains over time, turns into repositories for software engineering knowledge. Such knowledge repositories can be invaluable for gaining insight into the use of specific technologies and the trends of developer discussions.

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.
4. Collaborative Topic Modeling 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, 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.

**Motivation:**

- In mixture of unigrams model, there is only one topic for all the words in documents.
- Probabilistic latent semantic indexing (pLSI) overcomes the limitation and words are grouped to different topics based on number of parameters.
- But this algorithm has limitations like number of parameters grows linearly with number of parameters and there is no natural way to use it to assign probability to a previously unseen document.

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.

**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, data pre-processing, topic modeling, 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 NLP on the text corpus. In the proposed system, the question is searched by the user. Answer list will be provided to the user. 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 going to apply on QA corpus to find trending topic. After getting current trending topic study material will be provided to the user.

![Fig. The Proposed System](image)

**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.

**References:**


