Trends Identification and Analysis of Popular Technologies Using Topic Modelling

[1]Akshita Lakkad, [2]Deep Sanghavi, [3]Vidhi Shah, [4]Prof. Mrs. Stevina Correia
[1,2,3]Student, [4]Assisstant Professor
[1]Department of Information Technology,
[1]Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.

Abstract: Several technical discussion forums have seen a drastic hike in the number and the expertise of the users of the corresponding websites like Stack Overflow, Stack Exchange, GeeksforGeeks, Medium etc. It has been observed that the questions and the queries on these websites cover a wide range of domains. These websites thus, overtime, become knowledge repositories for software engineering community to accurately understand the needs of the developers and understand the current trends in technology. Thus, through this project we aim to discover interesting trends bolstered by precise statistical data by scraping real-time dynamic data from the websites and performing topic modelling for the same. The analysis is done to uncover trends, popularity and impact of technology over time as well as to mine the correlation and dynamics between different technologies.

IndexTerms **– languages, trends, machine learning, topic modelling, latent Dirichlet algorithm, natural language processing.**

I. INTRODUCTION

Majorly, when students take up a particular skill, they are skeptical that these courses might not be of relevance to their desired profile or of minimal importance according to the current trends of the corporate world. This might result in a weak and scattered resume, lessening the chances of bagging the right job let alone their impact on the field of application. Hence, it becomes necessary to design your line of action based on reliable and analyzed data rather than just current fads.

The ever-changing field of technology might often create confusion among emerging developers as to how to cope up with the fast-paced evolution of the upcoming technical trends. Thus, to keep up with this, online forums have been developed to debug codes, solve errors in a collective way. Portals like Stack Overflow and Stack Exchange allow the users to post their queries online and share knowledge. GitHub allows users to collaboratively create and host their projects. Geeks for Geeks is an online portal referred for programming related doubts.

Understanding these widely discussed topics scraped from the afore mentioned portals could allow programming language and tool developers to gain an insight on the actual usage trends and perceive the usage patterns of the data. It is also helpful for the vendors to and product managers to perform analysis of the market. Tool developers can consider this analysis to make an informed decision about the technologies they want to support. Once the textual data is extracted and the topics are recovered, further analysis on the data will help uncover useful evaluations like Popularity, Impact and Trends.

The structure of the rest of the paper is as follows: Section II talks about the prior approach and the existing systems. Section III is about proposed methodology and implementation plan. Section IV presents results, conclusions and the future scope.

II. BACKGROUND AND EXISTING WORK

Large volumes of data can be analysed with simplicity using topic modelling. This also helps in establishing semantic relationships and uncover contextual clues between words with different meanings and multiple usages of the words with similar meanings in a different context.

Topic modelling works on the principle that each document is a collection of topics and each topic is thus a collection of words. It works on the idea that the semantics of a document are driven by the unnoticed hidden latent variables like ‘topics’. The various topic modelling algorithms include Latent Semantic Analysis (LSA), probabilistic Latent Semantic Analysis (LAS), Latent Dirichlet Allocation (LDA). LSA aims at unravelling meaning of words in the document. It is based on the term-frequency inverse document frequency (tf-idf) principle where the word the appears more frequently is given less weightage than the word that appears infrequently. pLSA is based on document probability being a fixed point on dataset while LDA serves as a training data and gives an accurate generalization for new topics.

A review of existing work on trending technologies and most-used programming languages was conducted. A comparison of the above-mentioned algorithms is demonstrated in one of the reviews [5][9]. Analysis of factors that influence the developers to choose a language was done. The study shows how factors like libraries, existing code, domain and availability affect the adoption of a programming language.

Another existing study focuses on analyzing the details of one of the most widely used discussion forum, StackOverflow [6]. It is a huge source of information for both, academics and industrial users. Its analysis can provide useful insights. This paper’s objective is to explore the objectives of the users and find out the behavior of these users in terms of technologies. In this paper, they discuss a few observations regarding StackOverflow and analyze the details of this platform with a special emphasis on human factors [5][6]. They mined StackOverflow repositories to look for patterns and presented the results based on these

observations. The results presented can be generalized for many such other platforms. Cases where the number of answers of people from the same locality were more were observed and discussed in the paper. A survey was conducted to understand the reasons of decline in users' activity [6].

Another paper presents a novel topic modelling-based methodology to track emerging events in microblogs [8]. Their model has an in-fabricated update mechanism dependent on time cuts and actualizes a dynamic vocabulary. LDA is a generative model that learns topics from a collection of documents. The inputs to LDA is a pack of-words which are an accurate representation of the document. A set of descriptive topics are yielded as output to the model [8]. Thus, a document is a distribution of topics and the topics are in turn a distribution of corresponding words. LDA forms the record in a solitary clump to gain proficiency with the subject assignments.

One principle distinction between their subject demonstrating and the online LDA is the exchange of parameters from a formerly learnt model to a refreshed model. At each model update, the word circulation of topic changes, anyway a balanced correspondence between words is kept. Topic would thus always advance as new documents are handled. On each update, newly emerged topics are counted [8]. The model is first tested on a synthetic dataset to study the strength and then moved on to raw Twitter feeds to detect recent topics [8].

After a thorough research, we found out that LDA algorithm fits our requirement. Hence in this paper we use LDA algorithm to analyze StackOverflow, GitHub and other technical forums' data to thus unravel hidden insights and trends about the various tools and programming languages. The methodology used by us is described in the below sections.

III. METHODOLOGY OF PROPOSED SYSTEM

Our aim at building the system was to guide the users to make a well-informed decision before taking up a technology and completely rely on the outputs to make further decisions for the progress of a particular technology. The project extracts data dynamically from web portals like StackOverflow, GeeksForGeeks, StackExchange, etc. to monitor the discussions ongoing for a technology. The students can use system to determine the trend of various technologies over a period of time and then make a statistically well-formed decision. It can also be used by Subject Matter Experts to scan through the response received by a particular technology and then provide their expertise wherever needed. Figure 1 represents the various modules of the system.

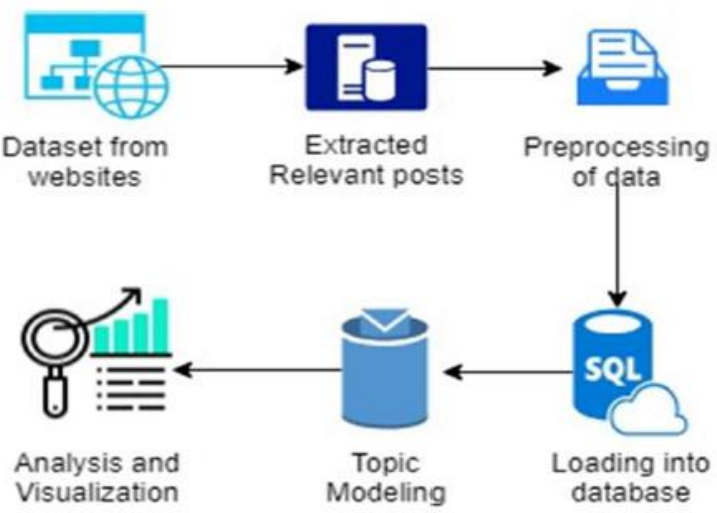

Figure 1: Data flow through various modules in the system

3.1 Web Scraping and Extraction of Data

Data is scraped from a myriad of datasets and various technical discussion forums, StackOverflow being the major source. A data dump is extracted by running a script through the website. This data dump file from StackOverflow consists of all the user questions and answers is extracted based on the tags, the occurrence and frequency of the tags and timestamp. The data is then pre-processed to shape it into a required format.

Also, data from other discussion forums like Geeks for Geeks, GitHub are also extracted in the similar way. Python libraries like BeautifulSoup are used for scraping purposes. The data is scraped according to relevance based on popularity of tags, the count of the tags, the upvotes, the number of times a language has been mentioned, maximum number of projects using a language and other important parameters.

3.2 Data Preprocessing

The data thus dynamically collected by running script every hour, is exported to the Amazon Web Services (AWS) for storage. The data pre-processing stage thus plays a vital role in bringing the data to a uniform format before other data manipulation algorithms are performed. The cleaning and pre-processing includes the following steps:
a. All the words are transformed into lowercase alphabets.

b. Punctuation marks are removed.
c. Words that have fewer than three characters are removed.
d. All the stopwords are removed.

Stopwords are the words in the document of least significance and can be ignored by search queries. For example, words like 'the', 'an', 'and' etc. can be removed before analysis.

Tokenization: It is essentially taking a sentence, text or a set of text and breaking it up into individual words. These individual words are usually called as tokens. The tokens thus generated are used as inputs for various other types of analytical tasks.

Lemmatization and Stemming: Stemming and Lemmatization are widely used in tagging systems, indexing, SEOs, Web search results, and information retrieval. For example, searching for fish on Google will also result in fishes, fishing as fish is the stem of both
words [10].
In stemming, the inflection in words is reduced to their root forms. A group of words is mapped to a stem where the stem may or may not be a valid word.
Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language.

3.3 Topic Modelling Framework

Topic Modelling, as the name suggests, it is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus [1]. While the traditional rule-based text mining algorithms use dictionary-based searching techniques for detecting keywords, topic modelling follows an unsupervised learning approach. It is useful for detecting relevant topics from large clusters of documents. Out of the many algorithms available, LDA is used for this study. It is advantageous in comparison with other algorithms because of its better adaptability and easier generalization to new documents. LDA is an efficient representation of topic models. It is a probabilistic model where documents are represented as a collection of topics and each topic is characterized by a distribution of words. Probabilistic rules are used to analyze the structure of the topic. Thus, understanding the semantics and the underlying implications of the words.

The following figure show the basic working model of the LDA algorithm. Here, a topic distribution of document is chosen randomly from a Dirichlet distribution. From this mixture of topic, a random topic is selected. Now, from another Dirichlet distribution a word distribution of the same topic is selected. Finally, a word is selected from the word distribution. This is the basic working of an LDA model.

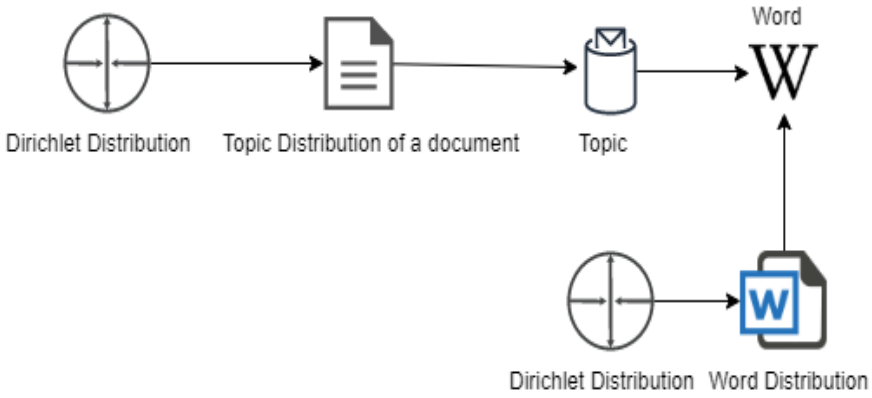

Figure 2: Basic LDA Model

In this analysis, we have used the LDA implementation provided by the Gensim package in python. Once the number of topics is provided to the algorithm, it obtains a good composition of topic-keyword pair. Word dictionary and corpus are the two main inputs. A corpus consists of an accurate mapping of the word to its frequency of occurrence. After application of the package libraries, the final model consists of a defined number of topics which in turn is a collection of multiple keywords. These keywords have weights assigned to it.

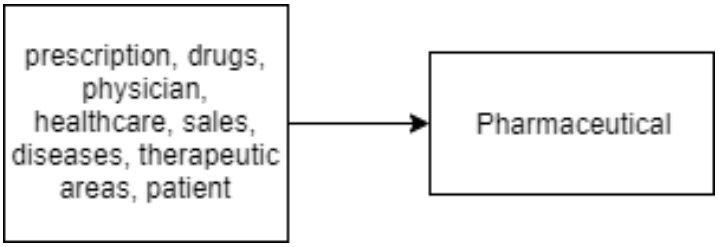

Figure 3: Inferencing a topic from keywords

Thus, based on the above implementation methodology, the most popular tags occurring frequently from the discussion forums are analyzed to determine the trends, the interaction pattern between the topics and thus helps the vendors to evaluate the rate at which their product is being adopted and received by the customers.

3.4 Visualization and Analysis

After successfully applying the model, the obtained data is represented in the form of graphs, multi-line graphs, network diagrams to gain a better insight on the same. Various python libraries like seaborn, matplotlib and pyplot are used for the same.

A comparative analysis of the trends of languages over a period is performed to gain better insights on the rise or fall of tools and languages. The figure below illustrates a line graph that shows a steep incline of python language in terms of popularity and use cases.

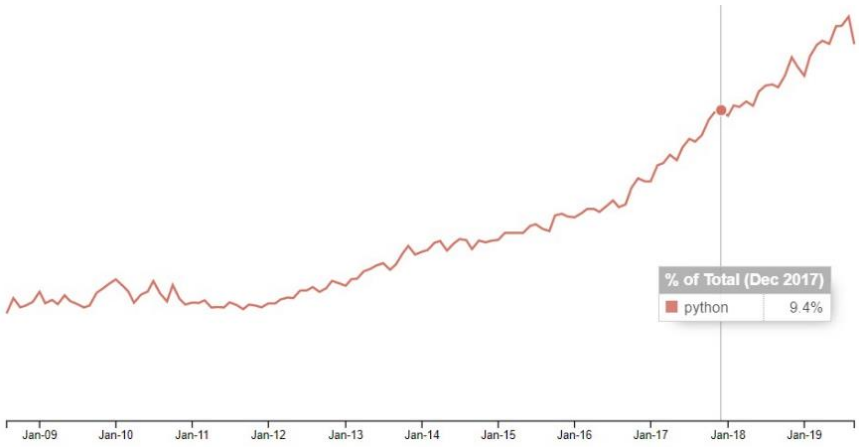

Figure 4: Line Graph of Python

IV. Results and Observations

According to the graphs plotted, a myriad of insights can be drawn regarding the decline or growth of a language over time. As observed, python has a steep growth in the recent times. Whereas, C programming language has observed constant slope over time. C# on the other hand has undergone a decline in the recent times. Due to introduction of new libraries and easier use R is also preferred as a visualization tool over matlab.

The figure below shows the multi-line graph comapring various languages and their trends over time.

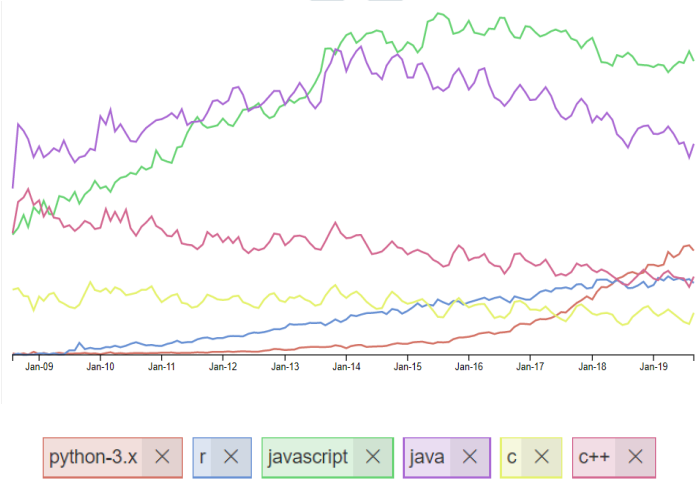

Figure 5: Multi-Line Graph

The data visualization tools were placed in contrast to observe the following results as figure 6. It shows how python is now emerging to be an efficient tool with the libraries like seaborn and matplotlib to design creative and innovative plots and graphs.

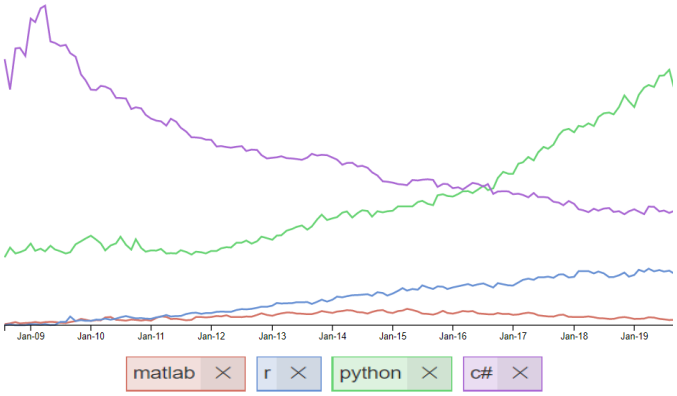

Figure 6: Comparison of visualization tools

V. Conclusion and Future Scope

The result of the analysis obtained thus can be used by various companies to perform a thorough market analysis before launching their product. It can help the tool developers to be in tandem with the enormous changes happening in the industry. Thus, we can conclude that modelling data and processing it can reveal huge chunks of useful information.

The scope of this project can be further expanded to analyze data for a specific time apart from a longer stretch of time. This can help us understand a sudden burst in the trend. The project can further be expanded to include more technical discussion forums across the internet.

References

[1] SHI, J., FAN, M. and LI, W. (2010). Topic Analysis Based on LDA Model. Acta Automatica Sinica, 35(12), pp.1586-1592
[2] Stack Exchange - https://archive.org/details/stackexchange
[3] Natural Language toolkit - http://www.nltk.org
[4] 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.
[5] V. Johri and S. Bansal, "Identifying Trends in Technologies and Programming Languages Using Topic Modeling," *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, 2018
[6] Understanding and evaluating the behaviour of technical users, Springer Open, 2017.
[7] Identifying Trends in technologies and programming languages using Topic Modelling,
IEEE International Conference, 2018.
[8] Online Trend Analysis with Topic Models, National Institute of Informatics, Japan.
[9] Rajasundari, T. & Palaniappan, Subathra & Kumar, Parambalath. (2017). Performance analysis of topic modeling algorithms for news articles. Journal of Advanced Research in Dynamical and Control Systems.
[10] Balakrishnan, V. and Ethel, L. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances.