Similarities in Challenges faced by Developers: Investigations on Stack Overflow and GitHub

by

Nidhi Pandya
202011069

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

June, 2022

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Nidhi

Nidhi Pandya

Certificate

This is to certify that the thesis work entitled **Similarities in Challenges faced by Developers: Investigations on Stack Overflow and GitHub** has been carried out by **Nidhi Pandya (202011069)** for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

Dr. Saurabh Tiwari

Acknowledgments

I would like to express the deepest gratitude to my thesis guide, Prof. Saurabh Tiwari for providing all the unwavering support, motivation, and help through out my research work. The continuous guidance and trust by him was invaluable during my thesis. His vast knowledge and wealth of experience helped me in the every step of my academic research. He was always there to help and guide me towards to the right direction whenever I got stuck in the work. I am grateful that you accepted me as your student and have continued to show support throughout the thesis work.

I would also like to offer my heartfelt thanks to my parents, and friends for being there for me with continuous support and assistance during every stage of my thesis.

Finally, I would like to thank God for guiding me through all my challenges and giving me strength to complete my thesis work.

Contents

Abstract

A large amount of rich data available in today's world encounters a lot of opportunities to analyze the data and identify some useful patterns from them. However, dealing with such data requires automated frameworks and knowledge of programming languages. Java and Python are the most commonly and widely used programming languages among the developers which it is evident from the queries and issues posted on Stack Overflow and GitHub. Despite the popularity of both Java and Python, the challenges in transitioning from one technology to another technology are hard for individuals and industries. In this thesis work, we aim to investigate similarities in the challenges faced by the developers while dealing with both the programming languages. To achieve this goal, we formulated three research questions (RQs) for understanding the topics and issues asked and faced by developers. To achieve the results, we have used the topic modeling technique Latent Dirichlet Allocation (LDA). We have also identified the temporal trend of asking new questions on Stack Overflow for Java and Python programming languages (PLs). Our results revealed the changing trend, in the year 2015 onwards, from Java to Python and inclined towards Python from the number of the new posts in a Stack Overflow. We performed analysis on 18,892 Stack Overflow questions related to Java and Python PLs and 42,674 issues from 22 different GitHub repositories, 11 for each PL. Our results indicates that questions asked on the Stack Overflow are co-related to issues posted by developers on GitHub during real-time development for a respective PL.

List of Tables

List of Figures

CHAPTER 1
Introduction

1.1 Objective and Problem Description

Fourth-Generation programming languages are been used by developers for over a decade. According to Tahmooresi et al. [20] and Stack Overflow annual survey[1] engaging more than 90,000 developers in 2019 highlighted that Python edged out Java in the overall ranking at the end of 2018, like when Java edged out C# in 2017, and PHP in 2016. In the coming time, Python being one of the trending programming languages intrigues us to investigate the relationship between challenges faced by developers in Java and Python languages.

Java and Python being popular in today's time between developers, can be seen on the developers community website Stack Overflow with the number of questions with tag <Java> are 1836182 and for <Python> are 1918781 till the 31st March 2022[2]. These two tags <Java> and <Python> are in the list of most popular tags on the Stack overflow website which proves these two programming languages are popular between the developers.

Java is one of the most used programming languages (PLs) in the industry and Python is one of the PLs that are becoming popular these days due to its packages and other useful libraries for machine learning, Big Data, Analytics and many others. Inspired by the work in [19], we targeted Stack Overflow and applied the topic modeling to its posts since it is one of the most active question and answer (Q&A) websites used by developers where they can upvote and downvote on already posted questions; post queries, errors and challenging questions. These doubts and queries are discussed and answered by the developer having knowledge of that particular area of topic on the website. We also used the issues in repositories on GitHub in addition to the Stack Overflow data. GitHub is a platform where developers work collaboratively by sharing code, creating issues,

[1] https://insights.stackoverflow.com/survey/2019
[2] https://stackoverflow.com/tags

tracking work progress and synchronizing the overall development work.

In this thesis work, we intend to investigate similarities in the challenges faced by the developers in Python and Java Programming Languages. Our study tries to map the topics asked by developers on Q&A website Stack Overflow to the real-time issues faced in the development by developers on GitHub. Specifically, we intend to conduct a comparative study to identify the similarities in challenges faced by the developers:

- Who wants to transition from Java to Python or vice versa.
- To understand the challenges that are similar in each of the programming languages.
- To understand how questions asked on the Q&A website are related to real-time issues faced in the development by developers on GitHub.

1.2 Motivation and Preliminary Analysis

Though Java and Python are the most popular programming languages(PLs) among the developers, it is important to identify whether there is some commonality between the questions asked for the two PLs. To understand that we mined the questions asked by the developers on Stack Overflow, and analysed the temporal trend. Specifically, we aim to understand ***"What is the temporal trend of asking new questions on Stack Overflow for Java and Python?"***.

Table 1.1: Stack Overflow Temporal Trend Data

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Java	98753	144257	189962	216438	214653	196374	174068	145028	126026	120751	103740
Python	42046	64028	96018	115899	136694	157957	191501	205097	223823	285355	290658

We have collected the data for the past ten years from 2011 to 2021 to understand the trend of asking new questions for both programming language (PL). We used SQL query to mine the count of new posts for a particular PL on the Stack Exchange Data Explorer website. We kept the condition of Tags LIKE <Java> and <Python> for respective PL with years changing from 2011 to 2021. Table 1.1 shows the data we mined for count of new number of questions for both Java and Python.

Figure 1.1 shows that from 2011 to 2015, the number of new posts for both the PL are increasing, while the number of posts for Java is significantly more than the Python. From 2015 onwards the number of new posts in Java starts decreasing, and on the other hand, the number of new posts for Python is in increasing order as before. From the graph, we can see that Java was more popular in 2014 and 2015, but the number of new questions for Java had started decreasing since late 2016. While new questions for Python are only in the increasing order till 2021, implying the popularity shift to Python PL.

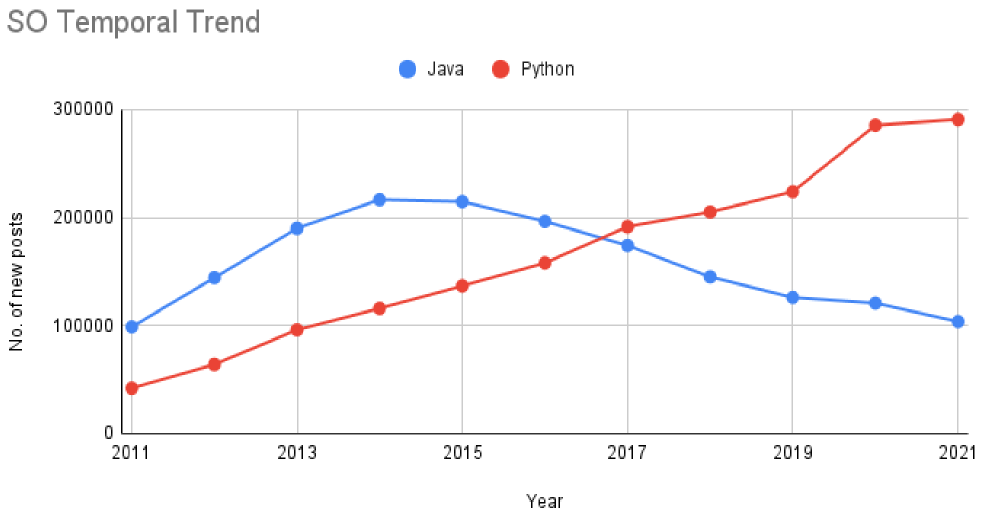
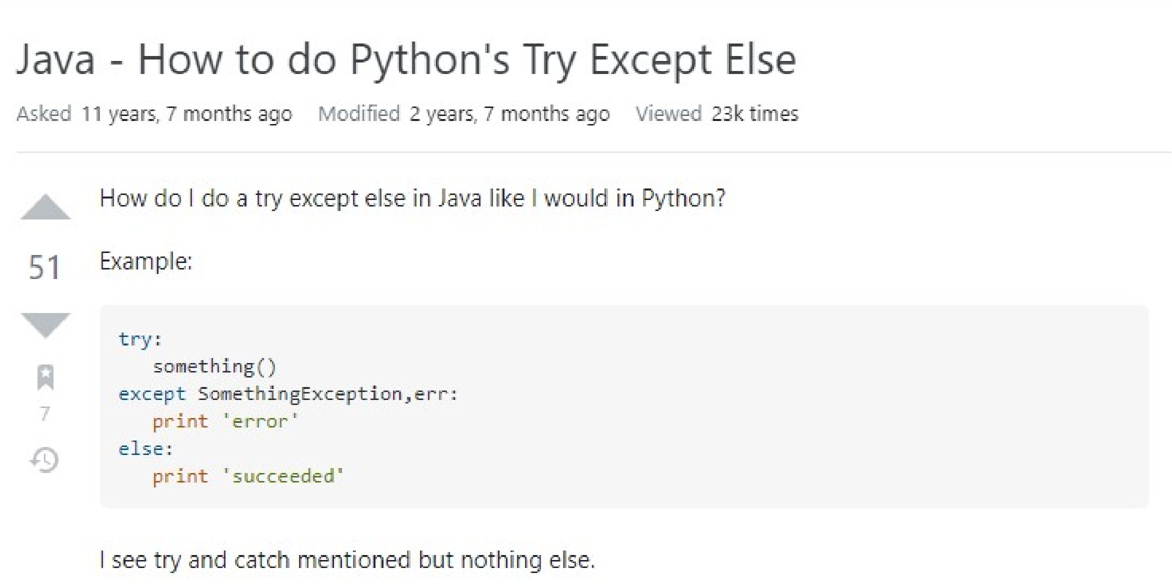


Figure 1.1: Stack Overflow (SO) Temporal Trend

From these results of temporal trend, we found that with time even if the developer is working with Java, after one point the developer may have to learn Python due to its usefulness with libraries and packages supporting machine learning and artificial intelligence which are not entirely present in Java. Java still has its significant importance in the industry and among the developers. But in today's world looking at the increasing popularity of Python we can say that at some point in their careers, developers will need to learn and maybe shift from Java to Python.

From the results of the preliminary analysis of our research, we can see that Java is popular but the number of new questions asked by developers on Stack Overflow is decreasing day by day. On the other hand for Python new questions are increasing over the period of decade. This result significantly states that in the coming time Python will be one of the most popular programming languages. We want to show the similarities between challenges faced by developers and the types of questions developers ask in both programming languages. For getting results in depth we are taking 20,744 issues from Java GitHub repositories and 21,930 from the Python GitHub repositories respectively. We compare these real-time issues with the posts of one of the most famous Q&A websites used by developers Stack Overflow with 8,753 Java PL posts and 10,139 Python PL posts.

The result of topics from these two different datasets will significantly show that there are some similarities between these two programming languages.


Java - How to do Python's Try Except Else

Asked 11 years, 7 months ago Modified 2 years, 7 months ago Viewed 23k times

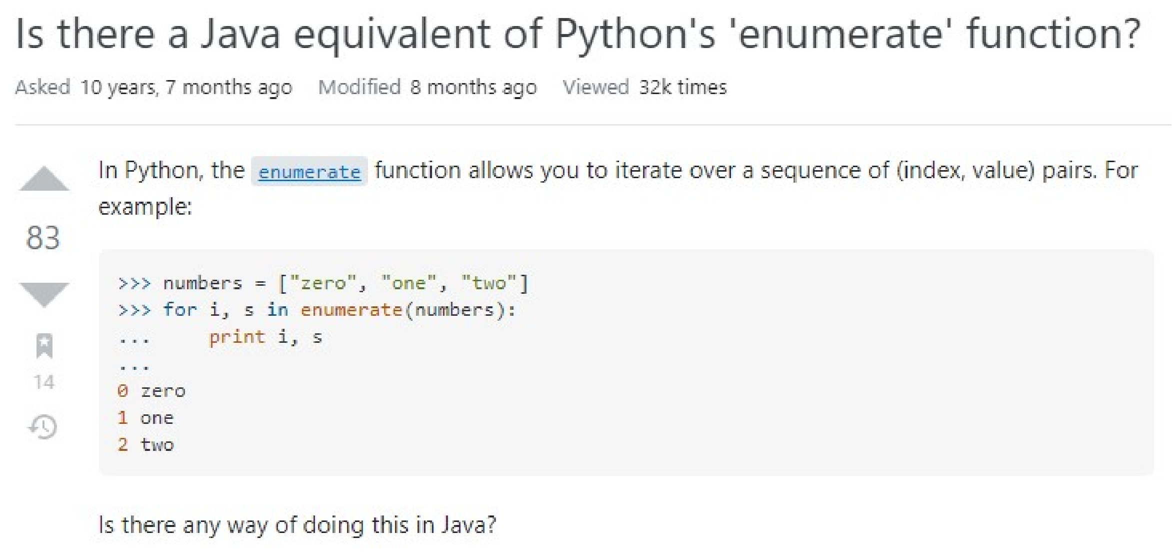
How do I do a try except else in Java like I would in Python?

Example:

```
try:
   something()
except SomethingException,err:
   print 'error'
else:
   print 'succeeded'
```

I see try and catch mentioned but nothing else.

Figure 1.2: Questions on seeking similarities with Java and Python-1


Is there a Java equivalent of Python's 'enumerate' function?

Asked 10 years, 7 months ago Modified 8 months ago Viewed 32k times

In Python, the `enumerate` function allows you to iterate over a sequence of (index, value) pairs. For example:

```
>>> numbers = ["zero", "one", "two"]
>>> for i, s in enumerate(numbers):
...     print i, s
...
0 zero
1 one
2 two
```

Is there any way of doing this in Java?

Figure 1.3: Questions on seeking similarities with Java and Python-2

For instance, Figure 1.2 and Figure 1.3 show the two examples highlighting questions asked co-related to both the programming languages (PLs) stating developers do ask the questions related to both the PL. Figure 1.2 shows a question on *Java - How to do Python's Try Except Else*, which discusses the error handling concept of Try Catch in Java and Python. Similarly, Figure 1.3 shows another question on *Is there a Java equivalent of Python's 'enumerate' function?*, this question shows that a developer is trying to find a similar in-built function in Java as the enumerate function present in Python.

We found these two examples in our dataset that we will be using for our thesis work, giving us motivation and proof that Java and Python programming languages are used by developers interchangeably. As developers do learn other programming language while being expert in one language already. This motivates us to form our research question around these two programming languages.

The results and observations from our thesis work shows the similarities in the challenges faced by developers that will help the developer who want to learn and transition from one programming language(PL) to another, mostly from Java to Python PL. The motivation for our research is to make the transition of developers from one PL to another easier and smooth with our results showing similarities between PLs, also how these topics of discussion are related to real time issues faced by developers. The formulated research questions for our thesis work are :

- **RQ1:** What are the issues for Java and Python that are similar on the Stack Overflow?
- **RQ2:** Do the Stack Overflow questions have a relationship with the GitHub issues?
- **RQ3:** Which topics are mostly asked by the developers for both PLs?

1.3 Thesis Contribution

The idea of identifying the similarities in the challenges faced by developer from the Stack Overflow and GitHub comes up with motivation *'to help developers transition from one programming language (PL) to another smoothly'*. In this thesis work, firstly we understand the temporal trend of new questions on Stack Overflow website for Java and Python PL. As discussed in the above section, we mined the data from Stack Overflow data dump to understand the trend of number of new questions asked in past ten years(2011 to 2021) and see how the Python is in trend with the increasing new number of questions over the decade.

With this thesis work, we identified the similarities in challenges faced by the developers for the Java and Python PL by performing investigations on open source platforms. We understood the co-relation between Java and Python PL using the data dump of Stack Overflow, and identified common topics using Latent Dirichlet Allocation (LDA) [9], topic modeling technique. We also mined issues related data from GitHub repositories for both PLs, and compared the topics achieved from Stack Overflow to the GitHub topic-of-issues for each PL. This analysis helped us to understand the relation between questions asked with the real-time issues.

1.4 Organisation of the Thesis

The rest of the thesis is organised as follows:

- Chapter 2 gives a brief introduction to Stack Overflow, GitHub and topic modeling technique - Latent Dirichlet Allocation (LDA).
- Chapter 3 discuss the existing literature related to our research work.
- Chapter 4 discusses the approach we used for our research work and details about data collection and filtering also data extraction.
- Chapter 5 gives an overview of the results we achieved from our work. It discusses the limitations of our work and how we mitigated them.
- Chapter 6 concludes the thesis as well as the work that can be done in future.

CHAPTER 2

Background

2.1 Stack Overflow and GitHub

Stack Overflow:

Stack Overflow[1] is one of most famous question and answer website used by professional and enthusiast developers. It was launched on 15 September 2008. It is a public platform where developers can find and contribute answers to technical challenges on a wide range of topics in computer programming. Stack Overflow being a free online community, where programmers provide useful answers to other users. When other users up-vote the answers, the persons who provided them acquire popularity based on the quality of the responses.

Developers can ask the question, query, any error or issue faced in the development process; this question becomes the post on the website and other developers with the knowledge related to that particular question can answer and help each other. In the Figure 2.1, we see that one of the developer asks the question on the website and other N developers provide answers to that question. From all the answers given by the other developers on the post, website authorities accepts one of those answer as a valid answer for the question asked in the post. Now this question with solution is viewed by many other developers. The person who views or refer to the question post can up-vote or down-vote the question. For our research work, we are mining these question posts and using them as the stack overflow dataset to achieve our thesis results.

GitHub:

GitHub[2] is a code hosting platform for version control and collaboration. It allows you and others to collaborate on projects from anywhere. The primary advantage of GitHub is its version control system, which enables seamless collaboration

[1] https://stackoverflow.com/

[2] https://github.com/

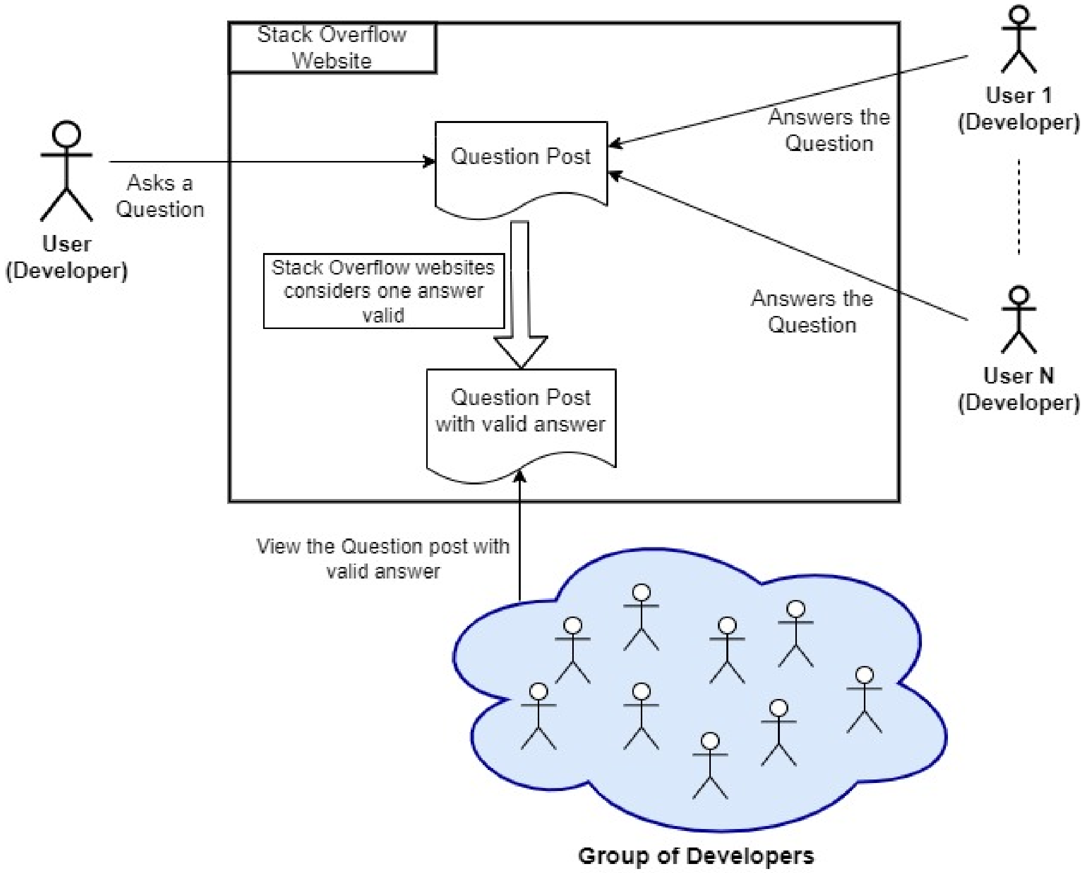


Figure 2.1: Stack Overflow Question Post Overview

without jeopardising the original project's integrity. The projects and repositories on GitHub are examples of open-source software.

Figure 2.2 shows a simple steps about how the issues are handled on GitHub. From the figure we can see that a person with responsibility like project leader will be creating the issues in the project repository, these issues will be assigned to one or more other members of the team. When the issue is created its state is OPEN stating the work is ongoing for the particular issue, once the work stated in issue is completed the issue's state changes to CLOSED.

In the GitHub repository, contributor with write access can create and assign issues to other contributors. Issues allow you to keep track of your progress on GitHub, where project development takes place. GitHub Issues is most commonly used for reporting bugs and requesting features. It may hold discussions, assist with the processing of support requests, or even provide comments on documentation. We mined the repositories and then issues from the respective repository which will be used as GitHub dataset for our thesis work.

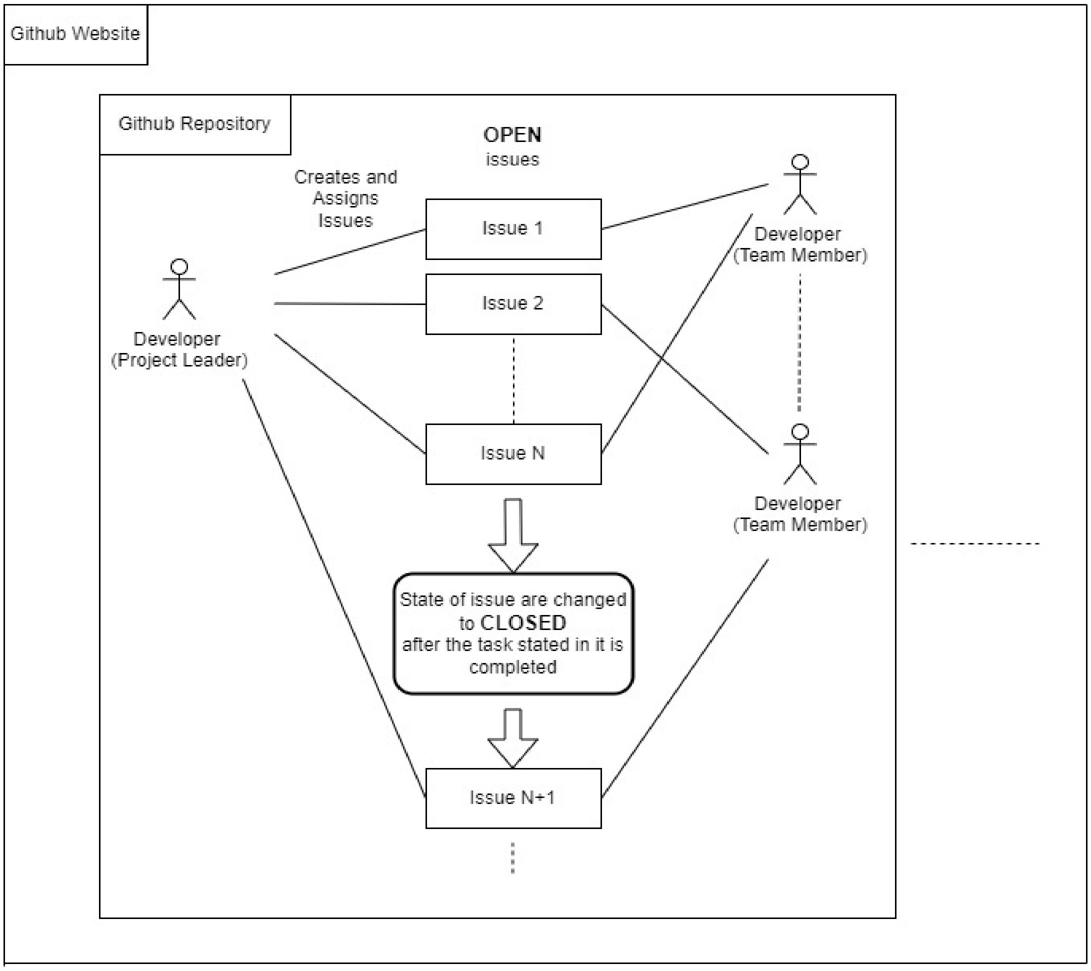


Figure 2.2: GitHub Website Issues Overview

2.2 Topic Modeling - LDA

LDA [9] is a popular topic modeling technique to extract topics from the data. topic modeling is a type of statistical modelling for discovering the abstract topics that occur in a collection of documents. The term “latent” in LDA means the topics that we want to extract from the data are hidden topics and it is yet to be discovered. Many of the previous literature works have used this topic modeling technique to achieve results for their research work.

Figure 2.3 shows the working of LDA topic modeling. Collection of documents are the post questions on Stack Overflow and issues on GitHub. With the internal working of LDA, we mention K to identify number of topics which are formed from the cluster of words. Each of word in the cluster has a frequency attached to it and results into distribution of topics. The LDA makes two key assumptions: i) Documents are a mixture of topics, and ii) Topics are a mixture of words.

In the Figure 2.4, the yellow box refers to all the documents in the corpus (rep-

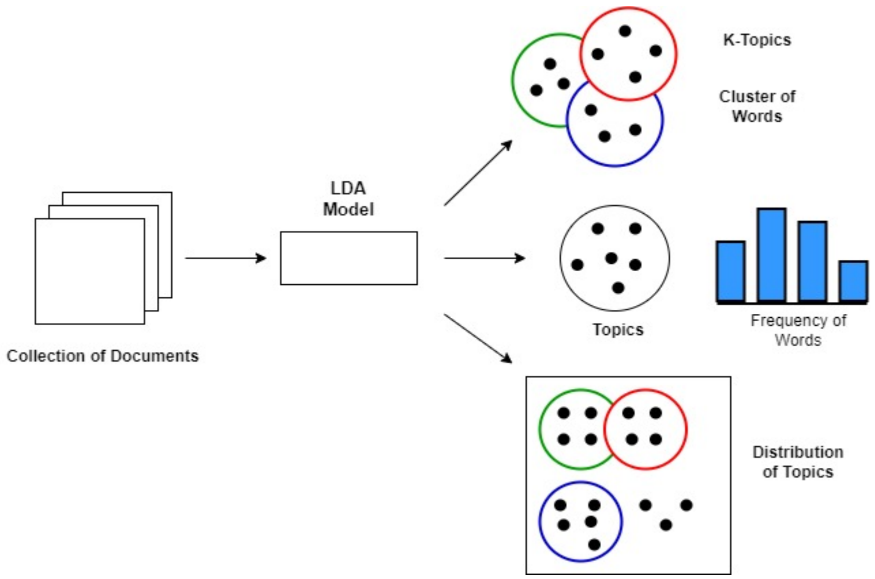


Figure 2.3: Latent Dirichlet Allocation(LDA)

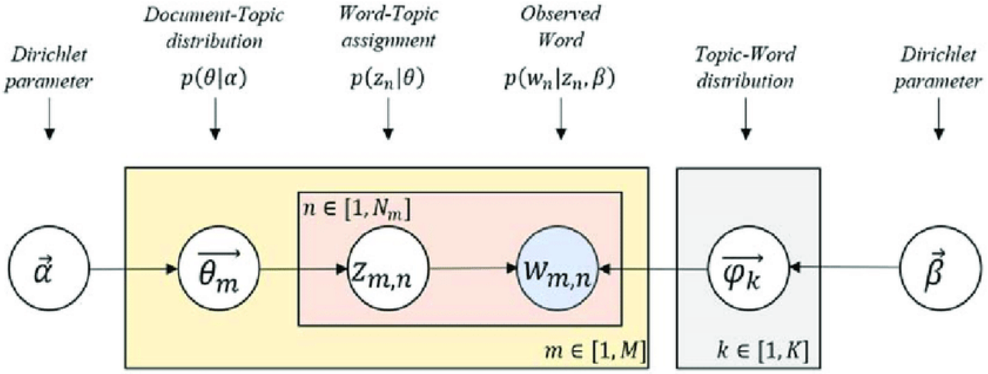


Figure 2.4: Vector Space of LDA [13]

resented by M). Next, the peach color box inside it is the number of words in a document, given by N. Inside this peach box, there can be many words. One of those words is w, which is in the blue color circle [16].

According to LDA, every word is associated with a latent (or hidden) topic, which here is stated by Z. Now, this assignment of Z to a topic word in these documents gives a topic word distribution present in the corpus that is represented by theta. From the figure 2.4, the LDA model has two parameters that control the distributions: (i) Alpha controls per-document topic distribution, and (ii) Beta controls per topic word distribution. While,
M: is the total documents in the corpus,
N: is the number of words in the document,
w: is the Word in a document,
z: is the latent topic assigned to a word,
theta : is the topic distribution,

LDA models parameters: Alpha and Beta. We can say that Latent Dirichlet Allocation (LDA) does two tasks: it finds the topics from the corpus, and at the same time, assigns these topics to the document present within the same corpus.

In our thesis work, Natural Language Processing technique latent Dirichlet allocation (LDA) is used to achieve topics from the datasets. The first step for applying LDA topic model is data pre-processing. In data pre-processing we will first perform tokenization that is to split the sentences into words, convert words to lowercase and remove the punctuation's and stop words from the datasets using the genism and NTLK [15] library of Python in the code. Next step is to make a bag or cluster of words from the dataset.

The number of topics (K) is a user-specified parameter in the code that provides control over the granularity of the discovered topics. Larger values of K will produce finer-grained, more detailed topics while smaller values of K will produce coarser-grained, more general topics. From this cluster of words, top words are identified and on the basis of them the topic label is given to justify the broad vision of that cluster of words. The topic labels were given manually by one of our researchers of this thesis work.

CHAPTER 3
Literature Review

3.1 Stack Overflow and GitHub Studies

Several studies have highlighted the importance of data on Stack Overflow and GitHub websites. Scoccia et al. [19] reported the challenges faced by the web developers and mapped real-time GitHub issues with the questions asked on the Stack Overflow, discussing the common domains used by developers for developing web apps. The researchers try to find out what topics related questions developers ask for desktop web apps on the Stack Overflow Q&A website, and also how prevalent these topics are to the issue's topic of GitHub. The results of their work show that there is a significant relationship between Stack Overflow dataset topics and GitHub topics.

Tahmooresi et al. [20] focused on identifying how Python is an ever-growing programming language based on the questions asked in the Stack Overflow, while this study does not talk about the trend followed by the developers working with Python. The authors identify a list of 100 topics from the Stack Overflow dataset; also show the temporal trend analysis for Python on the Stack Overflow Q&A website.

In our study, we are also trying to find a relationship between Stack Overflow and GitHub dataset topics for Java and Python programming language (PL). These two programming languages have significant importance in the industry at the moment, so correlating the challenges faced by developers on Stack Overflow with real-time issues on GitHub will help the developers to learn a new PL and transition from one PL to another, knowing about the challenges they would face in the learning process of a new language.

Bajaj et al. [5] reported the results of the questions asked by web developers on Stack Overflow Q&A website related to HTML, CSS and JavaScript programming languages, with identifying the common topic of discussions. With this research, researchers were trying to understand the common challenges and issues faced by web developers. In their work, they highlighted hot topics, temporal trends, category of topics between three languages, main technical challenges and the relationship between mobile web development to web-related topics.

Vasilescu et al. [21] shows the association between software development and crowd-sourced knowledge. This work shows the association between Stack Overflow activity to real-time GitHub development. Researchers focused on a person's activity as a GitHub committer and Stack Overflow question's asker and answerer. One of the main pointers the results of the study shows is that active GitHub committers ask fewer questions and provide more answers than others and vice versa. This research also highlights other few interesting pointers while correlating GitHub committer to Stack Overflow asker and answerer.

The work of Barua et al. [8] highlights trending topics among developers on the Stack Overflow Q&A website. The authors have identified other interesting analyses on (1) the topics of discussion among developers, (2) how the question of one topic answer/triggers an answer to another question, (3) how developers interest changes with time and also how an interest in programming language changes over time. Their work gives insight into programming languages, hot topics and interest in discussion for developers. Kochhar et al. [12] investigate the effect of programming language on development and software quality. They studied a large number of GitHub projects and showed how a programming language has a significant impact on software quality. The above two works of Barua et al. [8], and Kochhar et al. [12] showed how impactful the dataset of the Stack Overflow Q&A website and GitHub is website are, and can be used for further research and analysis.

Xiong et al. [25] tries to mine the behaviour of a developer between two platforms Stack Overflow and GitHub. There were many interesting results in which one of which was that for most of the developers, the topics of their contents in GitHub are similar to their questions and answers discussed on Stack Overflow. Oliveira et al. [17] did an analysis on the Stack Overflow and GitHub about how the software development was during the COVID-19 pandemic. The main questions related to the coronavirus on Stack Overflow were classified as how-to, concerning web scraping(data mining) and data visualization/processing, using programming languages Python, JavaScript, and R. While, in GitHub the repositories are in the majority of Machine Learning projects, using JavaScript, Python, and Java programming languages.

From the above-discussed research works, we got the inspiration to work on these two resourceful datasets of the Stack Overflow and GitHub website. We want to map the real-time issues faced by developers on GitHub to one of the most used Q&A website Stack Overflow questions asked by the developers community. Also, the result of our study will help developers who want to transition from one programming language to another.

3.2 Topic Modeling studies on Stack Overflow

There are many studies in the support of applying the topic modeling algorithms to the Stack Overflow data and accomplishing insightful results. Barau et al. [8] was the first to use topic modeling for investing the general topics on Stack Overflow discussed by the developers' community, the details of this work are discussed above. Other studies as discussed by Bajaj et al. [5] investigated the common challenges faced by web developers and Tahmooresi et al. [20] work related to Python programming language, used LDA topic modeling for achieving the results. Venkatesh et al. [22] investigated the challenges faced by web developers working in Web APIs and concluded results using LDA topic modeling technique.

For mobile development, the challenges faced by developers in mobile applications development were discussed by Linares-Vasequez et al. [14] used LDA topic modeling to achieve results. The researcher applied a topic modeling approach to detect hot topics from questions on mobile development in this work. According to the findings, the majority of the questions of this work were about general questions and compatibility difficulties. While Rosen et al. [18] analyzed 13,232,821 posts to examine what mobile developers ask about on Stack Overflow and explored the unique problems that exist on various mobile platforms using LDA. According to the result of this work, developers are inquiring about app distribution, mobile APIs, data management, sensors and context, mobile tools and user interface development.

Work by Yang et al. [27] discusses Software security, the researchers conducted a large scale study on Stack Overflow security-related data. They obtained topics related to security using Latent Dirichlet Allocation (LDA) tuned using Genetic Algorithm (GA), based on the study's findings they derive significant conclusions for scholars, educators, and practitioners. Alshangiti et al. [3] and Bangash et al. [7] talks about Machine Learning. Alshangiti et al. [3] goal is they want to learn about the challenges that developers have while developing machine learning applications and give suggestions for how to make the process easier. Many interesting results using LDA were concluded by the authors of this work that the majority of issues occur during the data preparation and model deployment phases. Furthermore, the implementation of ML is significantly more complex for developers than the conceptual part. Bangash et al. [7] work concludes that some machine learning topics are significantly more discussed than others, and others need more attention, this topic was achieved using LDA topic modeling on Stack Overflow data.

Many research works investigated the challenges and needs of developers in multiple fields: Bagherzadeh et al. [4] is about big data, Haque et al. [10] researched virtualization and Wan et al. [23] on the blockchain, Bandeira et al. [6] work talked about micro-services, Ahmed et al. [2] about concurrency. Other studies explored the usage of bio-metric APIs by Jin et al. [11], Adbellatif et al. [1] work was about challenges in chat-bot development.

To the best of our knowledge, no study has explored the challenges faced by developers in Java and Python on Stack Overflow and compared it with the real-time platform GitHub. We believe that our study will be helpful to practitioners and developers to understand the difficulties of Java and Python programming languages (PLs) and help to transition from one PL to another.

CHAPTER 4
Methodology

4.1 Goal and Research Questions

The goal of this research work is to identify the similarities and challenges faced by the developers between two different platforms. The study may also help developers who want to transition from Java to Python programming language. We intend to map the queries or challenging questions asked by developers on Stack Overflow to the issues developers face in real-time on GitHub in the development process using Java or Python as the core programming language. Here, we are focusing on the data from the two well-known websites among developers: Stack Overflow and GitHub. We have formulated three research questions (RQs) to investigate the similarities in challenges faced by developers. The formulated RQs are:

- **RQ1:** What are the issues for Java and Python that are similar on the Stack Overflow?
- **RQ2:** Do the Stack Overflow questions have a relationship with the GitHub issues?
- **RQ3:** Which topics are mostly asked by the developers for both PLs?

RQ1 focuses on the similarities between the topics of Java and Python posts on the Stack Overflow website. We are considering the Stack Overflow dataset and after applying the LDA topic modeling from the acquired results we will compare the topic list of both the programming languages respectively.

Mapping real-time issues on GitHub with the doubts and queries discussed on Stack Overflow is one of our motivates for this research work. In RQ2 we are considering all the datasets of GitHub and Stack Overflow for both PLs and compare their resulting topic list. The comparison will show the similarities between the real-time topic-of-issues on GitHub and queries and questions discussed by developers on the Q&A Stack Overflow platform for a respective PL.

From the results of RQ1 and RQ2, we will be able to identify the top topics for both languages. In RQ3, we manually analyze all the topics we got from both the datasets - GitHub and Stack Overflow and reach the results for RQ3. These topics will be identified as mostly asked or top topics if they are present in all three or at least two datasets for respective PL. We also identified some topics irrespective of the programming language; as they were noticed in two or more datasets but not in one particular PL suggesting that it can be a common topic for both PLs.

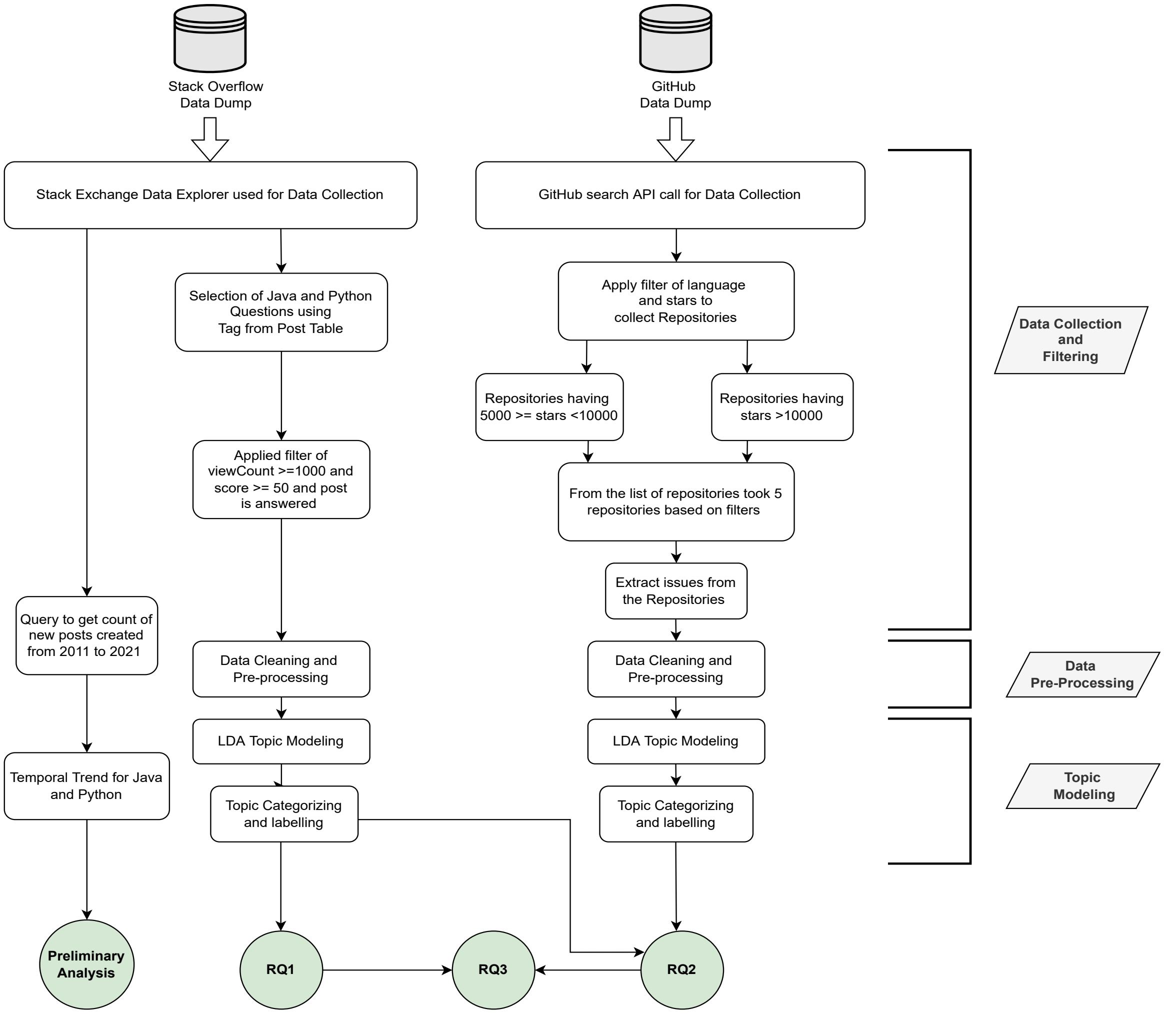

Stack Overflow Data Dump
GitHub Data Dump
Stack Exchange Data Explorer used for Data Collection
GitHub search API call for Data Collection
Selection of Java and Python Questions using Tag from Post Table
Apply filter of language and stars to collect Repositories
Data Collection and Filtering
Repositories having 5000 >= stars <10000
Repositories having stars >10000
Applied filter of viewCount >=1000 and score >= 50 and post is answered
From the list of repositories took 5 repositories based on filters
Extract issues from the Repositories
Query to get count of new posts created from 2011 to 2021
Data Cleaning and Pre-processing
Data Cleaning and Pre-processing
Data Pre-Processing
LDA Topic Modeling
LDA Topic Modeling
Temporal Trend for Java and Python
Topic Categorizing and labelling
Topic Categorizing and labelling
Topic Modeling
Preliminary Analysis
RQ1
RQ3
RQ2

Figure 4.1: Research Process

Figure 4.1 shows the overview of the study design with the following steps of data collection and filtering, data pre-processing and topic modeling to achieve results.

4.2 Data Collection and Filtering

4.2.1 GitHub Data Collection and Filtering

GitHub Search API[1] is used to collect the data of issues posted in repositories using the Java and Python programming languages(PLs) on GitHub. We are trying to collect issues from the repositories based on the stars stated for a particular repository, a repository's star feature indicates how many users have marked the repository as a favourite on GitHub. We have collected the GitHub repositories of Java and Python programming languages with conditions : i) stars of repository>10000 and creation year from 2016 to 2020 ii) stars of repository 5000 to 10000 and creation year from 2016 to 2020.

Dataset of repositories for GitHub contains

- High Popularity repositories (stars of repository$>$10000)
- Middle level Popularity repositories (stars of repository 5000 to 10000)

Both types of repositories together will give a broad idea and wider understanding of the issues faced by the developers in real-time development.

Java: The total number of repositories for Java programming language with stars > 10000 was 159. Next, we applied a filter to the repository's creation date, which had to be between 2016 and 2020, to ensure that the repository had at least two years' worth of issues data, which reduced the number of repositories to 71. Then, we apply the filters on number of open issues ranging from 200 to 1200 to cover a wide range of repositories. We have found several repositories in different languages other than English, hence we discarded them from our selection batch. Therefore, we have selected a total of 5 repositories : uCrop, material-components-android, lottie-react-native, graal, ghidra with *10,809* issues in our dataset for Java PL with stars$>$10000. Table 4.1 shows the detail about the total number of issues in these Java repositories.

Similarly, for repositories with stars 5000 to 10000, 175 repositories with stars >5000 and creation dates ranging from 2016 to 2020 were found. After using the stars > 5000 and stars <= 10000 filters, we narrowed down to 98 repositories. Then, we apply the filters on several open issues ranging from 200 to 1200 to cover a

[1] https://docs.GitHub.com/en/rest

Table 4.1: GitHub Java Repository and Issue Details

Repository (Stars > 10000) Total : 10,809		**Repository (5000 < Stars <= 10000) Total : 9,935**	
Repository Name	**Total Issues**	**Repository Name**	**Total Issues**
uCrop	775	camerakit-android	630
material-components-android	2330	react-native-svg	1682
lottie-react-native	802	capacitor	4422
graal	3682	AndroidPdfViewer	1054
ghidra	3220	epoxy BottomBar	1266 881

wide range of repositories. We have found several repositories in different languages other than English, hence we discarded them from our selection batch. Therefore, we have selected a total of 6 repositories : camerakit-android, react-native-svg, capacitor, AndroidPdfViewer, epoxy, BottomBar which gives in total *9,935* issues in our dataset for Java PL with stars>5000 and stars<=10000 shown in Table 4.1. So, the total number of issues for Java programming language for the GitHub dataset is **20,744**.

Python: We found a total of 215 repositories for Python programming language with stars>10000. After that, we applied a filter to the repository's creation date between 2016 and 2020 to ensure that the repository had at least two years' worth of data, resulting in 112 repositories. Then, we apply the filters on several open issues ranging from 500 to 1000 to cover a wide range of repositories. We selected 5 repositories : darkflowpytorch_geometric, Zappa, fairseq, bert, face_recognition with *11,129* issues in our dataset for the Python language with stars>10000. Table 4.2 shows the repository details and the total number of issues in these Python repositories.

Table 4.2: GitHub Python Repository and Issue Details

Repository (Stars > 10000) Total : 11,129		**Repository (5000 < Stars <= 10000) Total : 10,801**	
Repository Name	**Total Issues**	**Repository Name**	**Total Issues**
pytorch_geometric	2507	darkflow	1204
Zappa	2202	apex	1257
fairseq	3826	sentence-transformers	1362
bert	1249	keras-yolo3	766
face_recognition	1345	PySimpleGUI	5048
		tflearn	1164

Similarly, for repositories with stars 5000 to 10000 for Python PL, we found 347 repositories with stars>5000 and creation dates from year 2016 to 2020. Next, we found 223 repositories by applying the filter of stars>5000 and stars<=10000. After that we apply the same filter we applied for repositories with stars>10000 which is open issues must be from 500 to 1000. We selected a total of 6 repositories : dark-flow, apex, sentence-transformers, keras-yolo3, PySimpleGUI, tflearn having *10,801* issues of Python language with stars>5000 and stars<=10000. For the details of repositories and the total number of issues in these Python repositories refer Table 4.2

Finally, we found the total issues for the Python programming language for the GitHub dataset are **21,930**. The total number of issues of the GitHub dataset will be **42,674** combined with both the programming languages respectively.

4.2.2 Stack Overflow Data Collection and Filtering

Stack Overflow data was collected from the data dump available on the website Stack Exchange Data Explorer[2]. All the details related to the post of questions asked by developers are available in the *Posts* table of the online data dump. With the use of SQL query we are able to retrieve the required data for the research work.

There are 1821216 posts regarding the Java programming language(PL), and 1869754 posts regarding the Python PL which we got by applying the filter on the Tags field of Posts table as '<Java>' or '<Python>' till 18th January 2022. Table 4.3 show the details about the filters applied to achieve desired Stack Overflow dataset for both PLs.

Table 4.3: Stack Overflow Data Collection and Filtering

Filtering Conditions	**Java**	**Python**
All Posts(Till 18th January 2022)	1821216	1869754
AcceptedAnswerId is Not NULL filter	914811	973744
AcceptedAnswerId is Not NULL and Score>=50 and ViewCount>=1000	8753	10,139

The first filter we applied on the column *AcceptedAnswerId* of the Posts table. This field contains the user ID that provides the correct answer to the question

[2] https://data.stackexchange.com/stackoverflow/query/new

mentioned in the post. The answer then is verified by the Stack Overflow engine before accepting the answer as correct. We apply a filter that the AcceptedAnswerId field is not NULL, which means that we will only consider posts which have already been answered and verified by the Stack Overflow website. We got 914811 posts for Java PL and 973744 posts for Python PL after applying the above-explained filter.

Second filter applied is on fields – *Score* and *ViewCount*. A Score is calculated by the total number of upvotes on the post, minus the total number of downvotes on the same post. The ViewCount denotes the total number of people who have viewed that post. We applied a filter of Score>=50 and ViewCount>=1000 on the above data and found *8753* Java PL posts and *10,139* Python PL posts for analysis. In total **18,892** post questions are considered as dataset of Stack Overflow.

4.3 Data Extraction

Here, we describe the steps undertaken to extrapolate results from the three datasets to answer the three research questions.

Data Pre-processing : In data pre-processing, we first perform tokenization split the sentences into words, convert words to lowercase and remove the punctuation and stop words from the datasets using the genism and NLTK [15]. We performed these steps on all the data of our datasets (i.e., Stack Overflow posts and GitHub issues) to achieve the respective titles (topic labels). The title (topic label) for sure has to be representative of full data content and have lesser noise that will skew the results for our analysis [26][1]. The removing of the *stop words*, i.e., the process of removing the commonly used words in the English language, such as ’is’, ’the’, ’in’, ’of’ and others which do not significantly affect the semantics of sentence and it will potentially reduce the noise is a significant step. We considered the NLTK stop words [15] list for the data pre-processing.

Topics identification: To identify the topics from our datasets, we have used the topic modeling technique *Latent Dirichlet Allocation* (LDA) [9], widely used in software engineering studies to identify topics from the cluster of the words. The main idea of using the LDA algorithm is to identify the "topic" from the cluster of words that occurred frequently in the data. The number of topics (K) is a user-specified parameter in the code that provides control over the granularity of the discovered topics. Larger values of K will produce finer-grained, more detailed topics while smaller values of K will produce coarser-grained, more general topics. For our research purpose, we kept K=15 for all six datasets to achieve accurate results after experimenting with values of K as K=10 to K=20.

Naming topics: Right after the identification of results from LDA, we name the topic labels from the cluster of words. The naming of the topics was done by one of the author of thesis, who has a good knowledge of Java and Python programming language. From this cluster of words, top words are identified and the topic label is given to justify the broad vision of that cluster of words. The topic labels were given manually by the two authors through discussion and collaboration. Initially, the primary author understood the words in the cluster and accordingly decide on the topic label for that cluster of words. Thereafter, the second author reviewed the labels and finalize the topics for all the datasets in consultation.

CHAPTER 5
Analysis Results

5.1 RQ1: What are the issues for Java and Python that are similar on the Stack Overflow?

Using the dataset of Stack Overflow for both the PLs, we applied LDA topic modeling keeping K = 15 for our research. In the results, when we were labeling the topic clusters, we identified similar topics repeating. In the end we concluded the 10 topics after combining the similar clusters for both PL. Table 5.1 shows the list of Java Topics and Table 5.2 demonstrates list of topics of Python PL in which the topics in **Bold** are common with the Stack Overflow dataset for the other PL and topics in *Italic* are common with the GitHub dataset for the same PL.

Table 5.1: List of Topics Java Stack Overflow

Topic Label	**Top Keywords**
Basic Syntax	string, convert, variable, return, method, array, static
Web Protocols/Framework	JUnit, interface, implement, http, request, post, queries
Android Programming	android, build, handle, application, work, gradle, kotlin
Error Handling	final, block, handle, throw, error, thread, case
File Handling	file, object, path, resource, add, folder, write
Data Interchange Formats	json, default, jackson, initiate, regex, size, define
Date and Time	date, properties, time, util, hashmap, serial, libraries
Open-source Integrated Development Environment (IDE)	spring, maven, eclipse, boot, change, version, option
Code Processing	Java, stream, code, project, package, perform, jdbc
Inbuilt Functions	differ, type, enum, char, mean, sort, double

Table 5.2: List of Topics Python Stack Overflow

Topic Label	**Top Keywords**
Code Processing	create, script, pass, data,check, loop, directories
OOPS Concept	object, variable, number, class, attribute, instance, contain
Open-source Integrated Development Environment (IDE)	plot, dictionaries, datetime, notebook, jupyter, timezone, iPython
Python Packages	panda, file, column, dataframe, numpi, format, package
Basic Syntax	list, element, append, insert, array, convert, write
Web Protocols/Framework	django, line, dict, get, tuple, filter, assign
Error Handling	Python, error, import, request,local, open, libraries
Installation	value, install, method, version, mock, vitrualenvironment, call
Editor	module, name, command, window, pycharm, access
Graph plotting Library	string, function, matplotlib, type, specific argspar, axis, size

Topic Overview: Here, we discuss and describe the common topics that we identified for both programming languages (PLs). The topics that are common in both the PLs are *Basic Syntax, Code Processing, Web Protocols/Framework, Error Handling, open-source Integrated Development Environment (IDE)*. For better understanding, we discuss few examples for each of the common topics in the lists from the Stack Overflow dataset shown in Table 5.1 and 5.2.

Basic Syntax: questions in this topic discuss the fundamental basic syntax for both the PLs Java and Python. For Java, an example of this kind of question is *"How can I format a String number to have commas and round?"*, in which the developer asks for help to understand how can he format a string number which will have commas and round in the resulting value. *"How to remove the last character from a string?"*, here developer asks questions about how he can remove the last character from a given string. These two examples from the Java set show question-related to *string* top keyword. *"Why does Java's Arrays.sort() method use two different sorting algorithms for different types?"*, in this question developer details why Arrays.sort() method of the array has different algorithms for sorting for different types. The above question is related to *array* top keyword for Topic Basic Syntax as we can view in Table 5.1 for Java PL.

In Python, the example of this kind of question for Basic Syntax topic is *Dif-*

ference between del, remove, and pop on lists in which the developer asks about the difference between different types of method to delete elements in the list, this question is for the *list* top keyword. *"How to transform negative elements to zero without a loop?"*, the developer wants help to know how to convert negative elements to zero without using a loop. The above question was from the dataset related to the element top keyword as in Table 5.2.

The questions we observed in the dataset related to the Basic Syntax topic are relatively simple to be answered for the developers with a little bit of experience in the respective language Java or Python. The keywords related to the topic maybe non-identical for both the languages but it can be significantly observed that they are concluding to Basic Syntax of the particular programming language respectively.

Code Processing: this topic comprises the questions regarding code processing. An example regarding the same for Java PL is *"What code analysis tools do you use for your Java projects?"*, here developer asks about code analysis tools for Java projects. Another example is *"How to start a transaction in JDBC?"*, in this question developer, is asking about the transaction starting in JDBC. The first question was from the *code* and the second question from *JDBC* keyword.

For Python PL, an example of some questions are *"How to specify Python version used to create Virtual Environment?"*, in which developers ask about the Python version to create a virtual environment and *"Is there a difference between 'continue' and 'pass' in a for loop in Python?"*, developer discusses continue and pass in for loop. The above discussed questions were from the topic *create* and *loop*. From the Table 5.1 and Table 5.2, we see that Code processing have different kind of top keywords discussing Java, stream, code, project, perform, JDBC for Java PL and create, pass, run, execute, data, check, and loop for Python PL respectively. We did not go into much detail for this topic, the keywords are different for both of the programming languages symbolizing the code processing related keywords and so it is something researchers in future can work on by going into detail on this topic.

Web Protocols/Framework: this topic aggregates those questions which developers ask about the Web Protocols and Frameworks used in Web development. In Java, we have a question as *"How do you create an asynchronous HTTP request in Java?"*, where a developer needs a help to create an asynchronous HTTP request using Java PL. Other questions are as such *How to send HTTP request in Java?* and

"How do I get a list of all HttpSession objects in a web application?" about HTTP. However, we did not find a significant amount of web frameworks related data in Java dataset but the data we found were mainly related to HTTP and request related. As we can see above examples were also related to the *http* and *request* keyword.

Now in Python, web-related frameworks like Django is prominent in the dataset, while filtering and GET request are significantly discussed by developers. An example of this kind of question is *"Can you change a field label in the Django Admin application?", "Change a Django form field to a hidden field"*, developers are discussing web development related Django problems. *"How can I get href links from HTML using Python?"* and *"get request data in Django form"*, here developers discuss get requests to acquire data.

The top words and in details examples for Java and Python are quite different if we go in-depth of the questions asked by developers, which symbolizes web development differences in both the programming languages. We can conclude from the questions that Django is one of mostly high-level web frameworks used by Python developers and in Java developers we see that their questions are mostly related to HTTP.

Error Handling:these are the questions explaining the type of issues faced by developers in error handling concepts. Java-related examples for this type of question are "*Why use finally instead of code after catch*", "*Exception thrown inside catch block - will it be caught again?*", "*Does a finally block always run?*" and many others related to try-catch and finally concept having *final* and *block* as the keywords. "*Why does String.valueOf(null) throw a NullPointerException?*" and *"How can I solve "Java.lang.NoClassDefFoundError"?* shows example of questions developers asked regarding some error they faced while development in Java language of *throw* and *error* keywords.

Python examples for error handling are as "*pip install MySQL-Python fails with EnvironmentError: mysql_config not found*", "*ImportError: No module named 'encodings'*" explaining the questions regarding install and import error faced by Python developers having *import* and *error* keywords. Other than that we searched examples for exception keyword even though it is not in the top keywords for error handling in Python we got some insightful examples such as "*How can I write a 'try'/'except' block that catches all exceptions?*" and "*How to catch and print the full exception traceback without halting/exiting the program?*", which shows that the try-catch and finally concept of error handling is a concept used by both Java and Python developers.

Open-source Integrated Development Environment (IDE): this topics collects questions related to open source IDE used by developers working in Java and Python PL. The examples of Java questions related to this topics are "*What is the difference between @Inject and @Autowired in Spring Framework? Which one to use under what condition?*", "*How can I disable the Maven Javadoc plugin from the command line?*" and "*What is the single best free Eclipse plugin for a Java developer*". These examples are mainly discussing about *Spring, Maven and Eclipse* integrated development environments (IDE), which states that Java developers use these three IDEs commonly for the development purpose.

Python question examples for open-source IDE are "*How to change working directory in Jupyter Notebook?*", "*Inserting image into IPython notebook markdown*" and "*How to display full output in Jupyter, not only last result?*", these examples manifest that Jupiter and IPython are the notebooks most used by the Python developers. "*matplotlib: overlay plots with different scales?*" and "*How to make Django's DateTimeField optional?*" examples of these questions are related to *plot* and *datetime* keywords, however their were some more example showing that Python developers work using matplotlib library and django open-source web framework as discussed in above two questions.

5.2 RQ2: Do the Stack Overflow questions have a relationship with the GitHub issues?

To answer this question, we identified topics from the GitHub dataset for both programming languages. We kept K=15 while achieving the results using LDA topic modeling, but once the cluster were formed from both the GitHub datasets during the manual assessment we combined a pair of automatically generated topics that were found to be semantically identical. Resulting into 11 topics for Java PL and 12 topics for Python PL respectively. The resulting 11 topics for Java PL are displayed in Table 5.3 and 12 topics for Python PL in Table 5.4 where the topics in *Italic* are common with the STACK OVERFLOW dataset for the respective PL.

Topics overview: The topics we achieved from GitHub dataset differ in the concept from the topics achieved from Stack Overflow dataset. Many new topics were observed from GitHub dataset topic list showing difference between real-time development issues topic and questions on Q&A website.

Table 5.3: List of Topics Java GitHub

Topic Label	**Top Keywords**
Image Processing	page, release, method, camera, image, remove, load
Implementation	version, implement, zoom, function, search, mode, core
Android Programming	android, upgrade, bind, data, load, handle, gradle
Error Handling	crash, execute, change, crash, state, problem, avoid
Package	update, depend, package, differ, invalid, export, header
Deployment	issue, render, beta, module, leak, native, properties
Code processing	test, add, libraries, feature, help, source, code
Web Functions/Protocol	install, typo, graalvm, project, progress, http, perform
Documentation	update, differ, readme, package, document, doc, typo
Support	generate, support, plugin, processor, compile, feature, check
Basic Syntax	class, Java, create, object, initiate, set, instance

Table 5.4: List of Topics Python GitHub

Topic Label	**Top Keywords**
Model Training	latest, train, model, generate, interact, architecture, dataset, machine
Code Processing	issue, code, source, check, process, window, path, project
Installation	iinstall, problem, directories, attribute, valid, function, import, pysimplegui
Machine Learning Training	train, dataset, custom, machine, test, output, dictionaries, predict, roberta, checkpoint
Data preprocessing	preprocess, data, batch, implement, iteration, dict, match, value
Error Handling	error, module, transform, fix, support, runtiimeerror, crash, problem
Documentation	document, show, link, custom, update, readme, input, contain
AI - Artificial Intelligence	test, dictionaries, predict, tflearn, tensorflow, model, text, list, index, time, tensor, network
File Handling	file, save, type, checkpoint, stop, browse, sentence
Graph plotting Library	data, add, graph, true, disable, test, warn, popup
OOPS Concept	object, return, parameter, attribute allow, print, pass
Updation	update, element, change, fix, issue, version, typo, default

We report that **5 out of 11** identified topics of **Java PL** (i.e., *Android Programming, Error Handling, Code processing, Web Protocol/ Functions, Basic Syntax/ OOPS*) are similar to the topics achieved by Stack Overflow dataset for Java PL as represented in Table 5.3. From Table 5.4. we can see for **Python PL**, we outline that **5 out 12** recorded topics(i.e., *Code Processing, Installation, Error Handling, Graph plotting Library, OOPS Concept*) are also present in Stack Overflow dataset topic list of Python PL. These common topics identified between the Stack Overflow and GitHub topics provides an evidence that the problem domains that are stated by these topics are challenges faced by the developers while working real-time in Java and Python development frequently.

Stack Overflow and GitHub Comparison: From the topics achieved by the GitHub dataset for Java PL, we notice a few different topics which are not present in the list of topics from the Stack Overflow dataset. *Image Processing Documentation*, *Deployment* and *Package* are the new topics which are observed in the Java GitHub dataset as seen in Table 5.3. Similarly, for Python PL from Table 5.4 for the GitHub dataset new observed topics are related to Machine Learning such as *Model Training, Data Pre-processing, Machine Learning and AI-Artificial Intelligence* which are not present in the topics we achieved from Stack Overflow dataset. This is an interesting observation that developers are working on Machine Learning and AI-Artificial Intelligence topics in real-time development, still, questions related to the same topics are not significantly present in the Stack Overflow dataset suggesting many developers are still in the learning process of these two topics.

The above discussed list of topics demonstrates the wide range of topics covered by the Question and Answer (Q&A) website questions and real-time issues. Some of the topics are similar which we stated above while many topics differ and other than that we noticed that GitHub dataset deals with more diversified aspects and cover different concepts than Stack Overflow.

5.3 RQ3: Which topics are mostly asked by the developers for both PLs?

From the results we achieved in the RQ1 and RQ2, we were able to conclude some topics that were mostly asked topics by the developers.

Common Topics between Java and Python : The top topics which are present in all the dataset topic lists(i.e. Table 5.1, Table 5.2, Table 5.3, Table 5.4) are stated below for the Java and Python programming languages(PLs):

- Code Processing: Code processing consists of simple coding related questions, these are kind of general questions on problems developers face in development process, making this mostly asked topic as it was also present in all the datasets.

- Error Handling: This topic consists of the question regarding the different types of errors developers faced while coding in Java language. Other than development errors, developers also have many questions regarding try-catch and finally the concept of Exception related to Error Handling in the Stack Overflow dataset.

Java: Most of top topics for java are discussed above in common topics, the top topic only for the Java programming language(PL) is listed below :

- Android Programming: The android related questions were present in all the datasets and Java PL is used in android development. From which we can conclude that Android Programming is one of the important topics in which Java developers face issues while development and is one of the most asked topics by Java developers.

Python: The top topics for the Python PL are listed below :

- Machine Learning: Python being one of the languages in which developers code for machine learning concepts, makes this topic significant. The questions regarding machine learning were present only in the GitHub issues, which are real-time queries/issues developers face while developing. This is one of the topics in which Python plays an important role and developers are interested to learn code for the same, which makes this topic one of the most asked topics by Python developers.

- Model Training: This topic was also significantly derived from the GitHub datasets, while model training is an important part of Artificial Intelligence and Machine Learning concepts which use Python language for development making this one of the top topics.

- Data Pre-possessing: Data pre-processing is one of the important steps in machine learning, data mining, data science and others. Python developers working on the above-stated concepts have a significant amount of questions while working on the data pre-processing step making this one of the top topics.

- Installation: Installation is a topic which is present in all three datasets, showing the developers who are starting to work with Python language face problems in the installation process of IDE or setting the path to the virtual environment for Python PL making this top topic.

- Graph Plotting Library: Python developers discusses questions related to graph plotting libraries on Stack Overflow and GitHub. *Matplotlib* and *axis* keywords in Stack Overflow list; while *graph*, *data, warn* and *popup* in GitHub list shows that developers use graph to demonstrate their result in Python.

Other than the topics stated as common for both programming language(PL) and for the particular PL, there were some topics which we noticed while working on RQ1 and RQ2 results.

- Documentation: Documentation of any programming language is helpful for developers working together on the same project. From our GitHub dataset for both PL, we can see that developers have the documentation topic from which we conclude it as one the top topics.

- OOPS Concept: In Java, the OOPS concept can be considered in the basic syntax, while the Python developers face problems with the class, object and method concepts.

- File Handling: File handling is the topic we got from Java Stack Overflow list and Python GitHub list. This shows that Python developers in real-time requires file handling while there are many questions by Java developers on Stack Overflow for file handling; concluding this topic is discussed by both PL developers.

- Web Protocols/Frameworks: Web protocols or framework topic was derived from the questions asked by developers while working on web development in their respective language. This topic may not be present in all the datasets while it is present in both the Java PL datasets(i.e. Stack Overflow and GitHub) and in Stack Overflow Python dataset demonstrating web development is an important skill and developers in both the PL Java and Python ask questions about the concerned topic, we added it as one of the top topics.

5.4 Limitations and Threats to Validity

Internal Validity: In our work, we have relied on the Stack Overflow tags generated by the Stack Overflow website to identify posts related to Java and Python PL. With such filtering, there is a possibility that some posts might have been missed during our post selection as they may be mislabeled. Another potential risk resides in the selection of optimal value for the Number of topics (K) value in code for LDA [9] topic modeling. Selecting the value of K for dataset such that it is optimal and leads to identifying the topics, where the value of K is not too narrow or not too general to extract meaningful topics. We mitigated this risk by experimenting with different values from 10 to 20, as those were the optimal value of K for our size dataset. We kept the value of K=15 for the LDA code, which resulted in 10, 11 or 12 topics for different dataset topics.

Construct Validity: intends to map the relation between theory and observation [24]. After applying the LDA topic modeling to our data, the potential threat is the labelling of the automatically generated clusters of words. We mitigate this threat by having named the topics by the primary author who understood the words in the cluster and accordingly decide on the topic label for that cluster of words. While naming the topic, the researcher manually analyzed the words in clusters and reviewed the most relevant documents in addition to the automatically generated keywords in a cluster to name the label. The second author reviewed the labels and finalize the topics for all the datasets in consultation.

External Validity: regarding the generalizability of the obtained results [24]. In our research work, we mainly relied on the data gathered from the Stack Overflow website to obtain the required results to explain issues faced by developers in Java and Python PL. However, this data might not be complete or comprehensive enough for all the difficulties faced by developers in Java and Python PL. To mitigate this risk we also investigated the issues from the real-time used platform GitHub, we collected issues from the repositories used by Java and Python developers where they discuss their problems while project development.

CHAPTER 6
Conclusions and Future Work

In this thesis, we performed analysis on 18,892 Stack Overflow questions related to Java and Python programming languages(PLs) and 42,674 issues from 22 different repositories, 11 for each PL on GitHub. Error Handling and Code Processing are two topics in which developers face challenge on both the platforms(i.e. Stack Overflow and GitHub) for both the PLs. Our results show that Android Programming is one of the top significant topic for Java developers. The topics related to Machine Learning and Artificial Intelligence - AI are prominent in real-time development for Python developers. Results from RQ1 and RQ2 concludes that developers do face similar challenges in both PLs, and real-time development issues with respect to a PL are co-related with questions asked on Q&A website.

In future, one can aim to use a large set of data to get more accurate results for the formulated RQs. One could also work on more than just two programming languages, try to find and help developers when they want to transition from one PL to another. From the achieved results, in future, the researchers can go into the details of common topics – Error Handling and Code Processing to understand how they are different and similar to each other in the context of respective PL.

References

[1] A. Abdellatif, D. Costa, K. Badran, R. Abdalkareem, and E. Shihab. Challenges in chatbot development: A study of stack overflow posts. In *Proceedings of the 17th International Conference on Mining Software Repositories*, pages 174–185, 2020.

[2] S. Ahmed and M. Bagherzadeh. What do concurrency developers ask about? a large-scale study using stack overflow. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*, pages 1–10, 2018.

[3] M. Alshangiti, H. Sapkota, P. K. Murukannaiah, X. Liu, and Q. Yu. Why is developing machine learning applications challenging? a study on stack overflow posts. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11. IEEE, 2019.

[4] M. Bagherzadeh and R. Khatchadourian. Going big: A large-scale study on what big data developers ask. In *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 432–442, 2019.

[5] K. Bajaj, K. Pattabiraman, and A. Mesbah. Mining questions asked by web developers. *11th Mining Software Repositories, MSR 2014 - Proceedings*, 05 2014.

[6] A. Bandeira, C. A. Medeiros, M. Paixao, and P. H. Maia. We need to talk about microservices: an analysis from the discussions on stackoverflow. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 255–259. IEEE, 2019.

[7] A. A. Bangash, H. Sahar, S. Chowdhury, A. W. Wong, A. Hindle, and K. Ali. What do developers know about machine learning: a study of ml discussions on stackoverflow. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 260–264. IEEE, 2019.

[8] A. Barua, S. W. Thomas, and A. Hassan. What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, 19:619–654, 2012.

[9] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. volume 3, pages 601–608, 01 2001.

[10] M. U. Haque, L. H. Iwaya, and M. A. Babar. Challenges in docker development: A large-scale study using stack overflow. In *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11, 2020.

[11] Z. Jin, K. Y. Chee, and X. Xia. What do developers discuss about biometric apis? In *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 348–352. IEEE, 2019.

[12] P. S. Kochhar, D. Wijedasa, and D. Lo. A large scale study of multiple programming languages and code quality. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, volume 1, pages 563–573, 2016.

[13] J. Lee, J.-H. Kang, S. Jun, H. Lim, D. Jang, and S. Park. Ensemble modeling for sustainable technology transfer. *Sustainability*, 10:2278, 07 2018.

[14] M. Linares-Vásquez, B. Dit, and D. Poshyvanyk. An exploratory analysis of mobile development issues using stack overflow. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 93–96. IEEE, 2013.

[15] E. Loper and S. Bird. Nltk: the natural language toolkit. *CoRR*, cs.CL/0205028, 07 2002.

[16] S. Neha. Topic modeling and latent dirichlet allocation (lda) using gensim and sklearn. volume https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/, 2021.

[17] P. A. M. Oliveira, P. A. S. Neto, G. Silva, I. Ibiapina, W. L. Lira, and R. M. Andrade. Software development during covid-19 pandemic: an analysis of stack overflow and github. In *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, pages 5–12. IEEE, 2021.

[18] C. Rosen and E. Shihab. What are mobile developers asking about? a large scale study using stack overflow. *Empirical Software Engineering*, 21(3):1192–1223, 2016.

[19] G. L. Scoccia, P. Migliarini, and M. Autili. Challenges in developing desktop web apps: a study of stack overflow and github. In *2021 IEEE/ACM 18th Mining Software Repositories (MSR)*, pages 271–282, 2021.

[20] H. Tahmooresi, A. Heydarnoori, and A. Aghamohammadi. An analysis of python's topics, trends, and technologies through mining stack overflow discussions, 04 2020.

[21] B. Vasilescu, V. Filkov, and A. Serebrenik. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *2013 International Conference on Social Computing*, pages 188–195, 2013.

[22] P. K. Venkatesh, S. Wang, F. Zhang, Y. Zou, and A. E. Hassan. What do client developers concern when using web apis? an empirical study on developer forums and stack overflow. In *2016 IEEE International Conference on Web Services (ICWS)*, pages 131–138. IEEE, 2016.

[23] Z. Wan, X. Xia, and A. E. Hassan. What is discussed about blockchain? a case study on the use of balanced lda and the reference architecture of a domain to capture online discussions about blockchain platforms across the stack exchange communities. *IEEE Transactions on Software Engineering*, (01):1–1, 2019.

[24] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[25] Y. Xiong, Z. Meng, B. Shen, and W. Yin. Mining developer behavior across github and stackoverflow. In *SEKE*, pages 578–583, 2017.

[26] B. Xu, Z. Xing, X. Xia, and D. Lo. Answerbot: Automated generation of answer summary to developers' technical questions. pages 706–716, 10 2017.

[27] X.-L. Yang, D. Lo, X. Xia, Z.-Y. Wan, and J.-L. Sun. What security questions do developers ask? a large-scale study of stack overflow posts. *Journal of Computer Science and Technology*, 31(5):910–924, 2016.