ProbExpert - A novel learning aid platform for reducing time spent on learning and finding answers.

2021-155

Prob Expert

Team

Supervisor
Ms. Dinuka Wijendra

Co-Supervisor
Ms. Anjalie Gamage

Kamal Thennakoon
IT18004564

Dasun Ekanayake
IT1801360

Ashen Ranasinghe
IT18011012

Thanura Marapana
IT18078992

Introduction

- Focused on creating a better learning aid platform to the learners and contributors.
- One to one learning
- give newbies a solid idea of how they should shape their path
- Providing a platform to users regardless of their knowledge degree to interconnect with each other.

What is one to one learning in ProbExpert?

- Find a mentor for your requirement.
- Schedule a video call using ProbExpert coins.
- At the end, mentor will receive ProbExpert coins from the mentee.

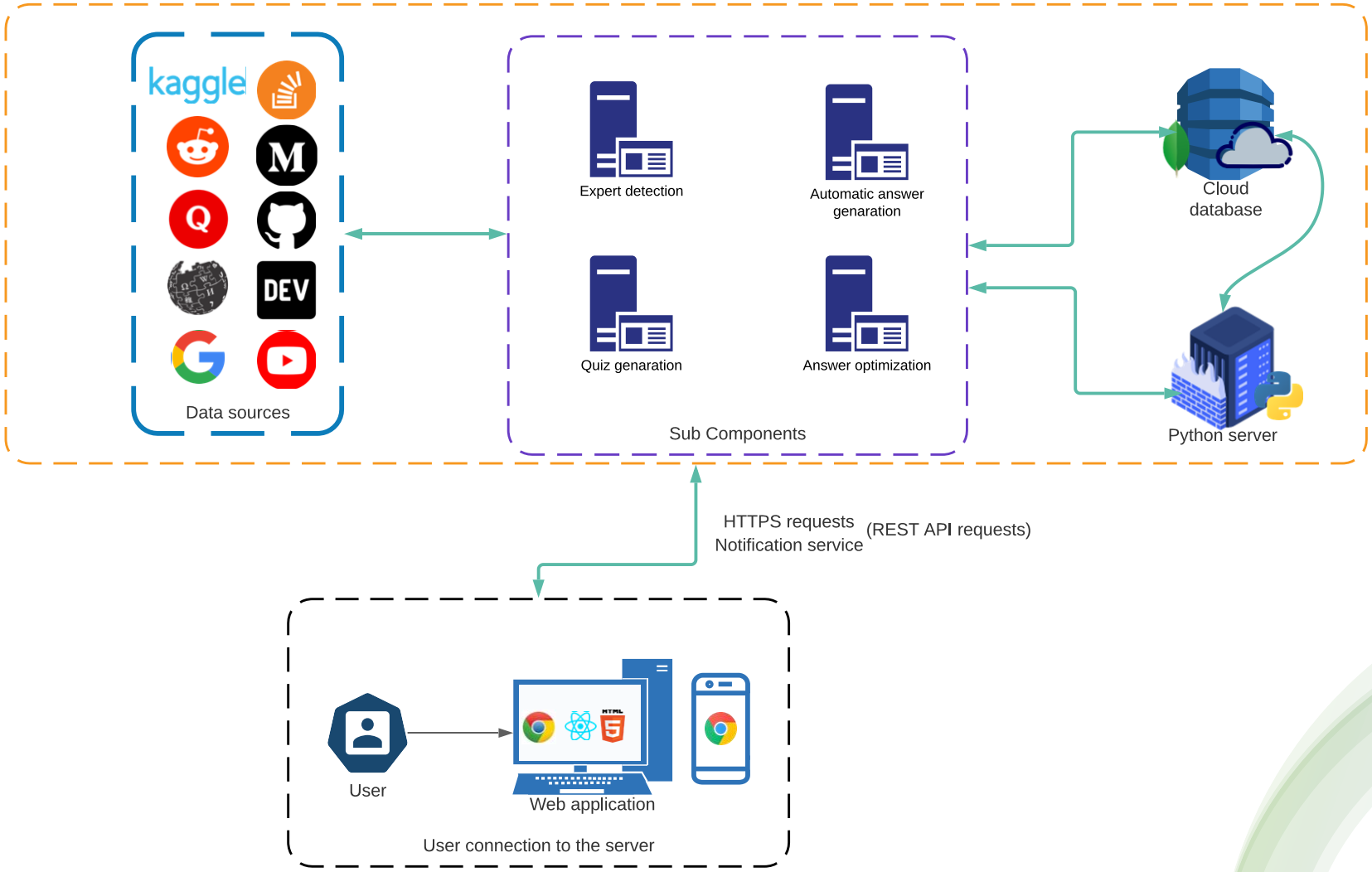
System overview

- Portfolio generation
- Automatic answer generation from public information
- Structured-type quizzes for knowledge checking
- Optimized answers
- Each user can earn ProbExpert coins through their contribution to the platform.

Objectives

- Detecting users' proficiency level along with an auto-generated portfolio.
- Form an answer automatically using public data from multiple sources for user’s questions.
- Optimal answer generation through up-voted answers.
- Formulating structured type questions for knowledge checking, using existing answered questions of the platform

Overall System Diagram



System budget

Component	Amount (USD)	Amount (LKR)
Yearly cost		
Digital Ocean hosting	240	48000
Cloud database (Mongodb)	110	22000
Push notification	60	11000
Domain name	11	2100
Total	421	83100

Commercialization

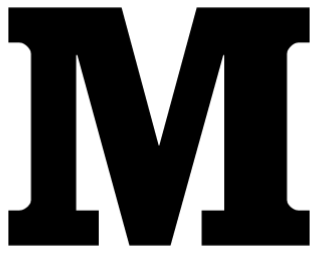
- If user ran out of points user will be able to purchase points for connecting with experts
- Organizations will be able to publish sponsored advertisements via the platform.
- IT firms can advertise their job openings.

IT18004564| Thennakoon T.M.K.H.B

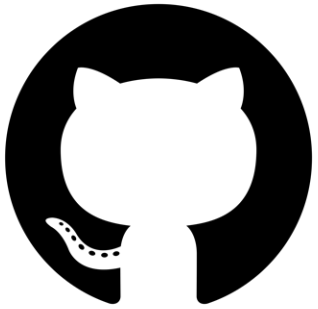
B.Sc. (Hons) Degree in Information Technology Specialized in Software Engineering

Introduction

Background/Research Gap



follower count



follower count

No specific evaluation method

Reputation based

- Most of developer's social platforms don't have a method to evaluate their users.
- Platforms like Stack Overflow uses reputation-based system.

Background/ Research Gap

Prevailing methods for detect experts in general social platforms.

- Page Rank
- HITS
- Mutual Reinforcement approach
- Hybrid approach of Wen-Chen
- Topic sensitive probabilistic model

Research Gap

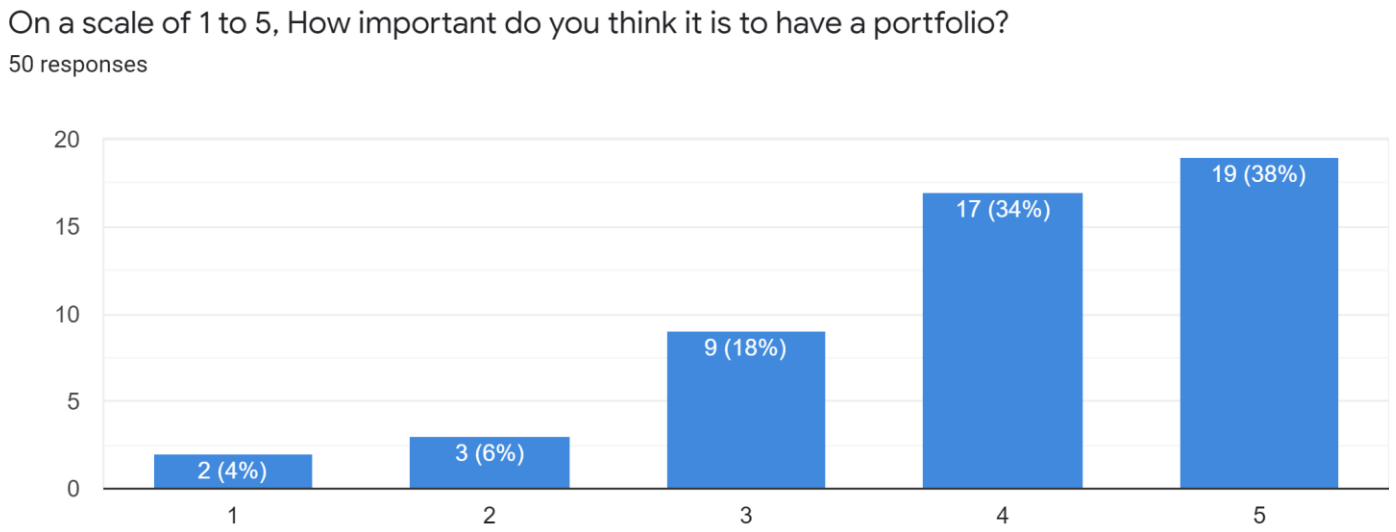
Platform	User-level	Publications	Open-Source contributions	Detailed Bio	Blogs
Stack Overflow	Based on the platform Karma	None	None	Medium	None
LinkedIn	Based on endorsements	manually	None	Good	None
Medium	None	None	None	Low	Platform Dependent
GitHub	None	None	Available	Low	None
ProbExpert	Based on all the platforms activities and behaviors	Available	Available	Good	Platform Independent

Research Problem

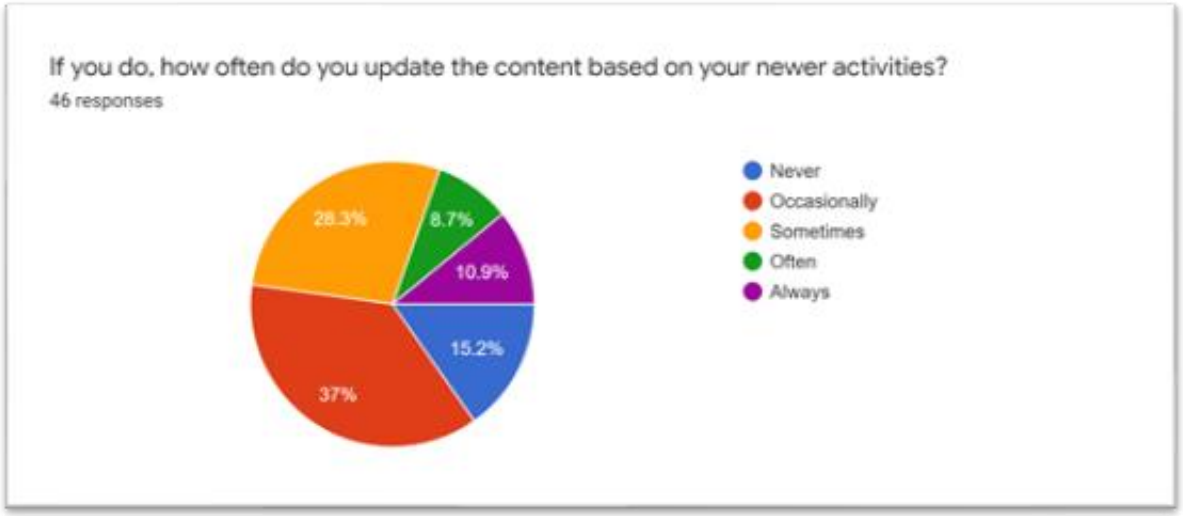
- No way to find an expert
- No platform to showcase proficiency and the contributions to the dev communities.
- Recruiters still don’t have a method to evaluate their candidates in a single place.
- No methods to find a skilled unemployed developers in a short time.
- Newbies don’t have a proper way to follow an expert of their field as a role model, to shape their career paths.

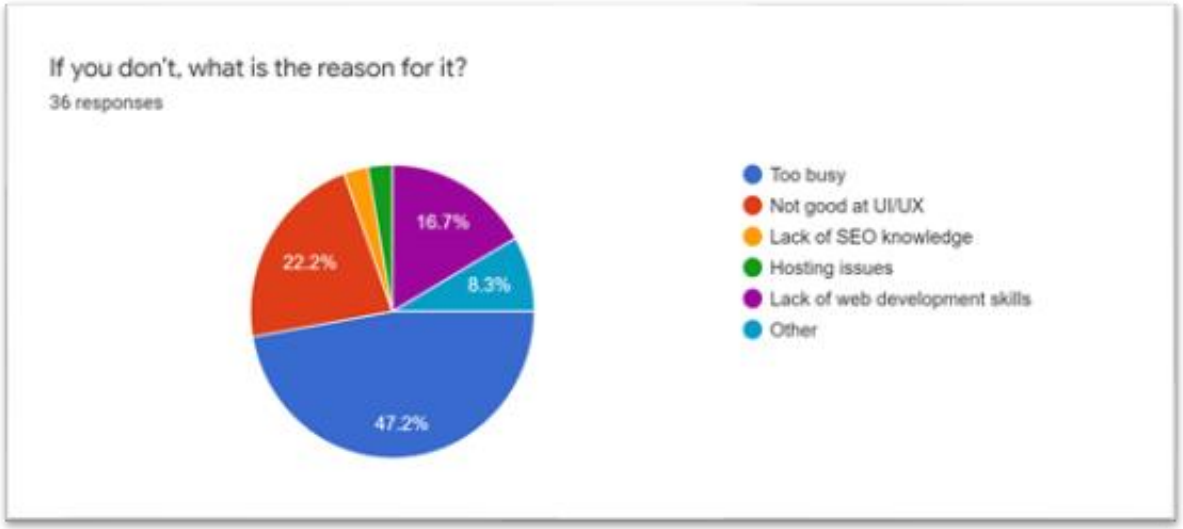
Research Problem

How Important is it to have a Portfolio?



Research Problem





Reason behind not being able to maintain an updated portfolio

- Busy lifestyle
- Need of continuous maintenance.

Specific and Sub Objectives

- **Detecting users' proficiency level along with an auto-generated portfolio**
  - Web Crawler to scrape user data
  - Two Statistical Models to cluster users.
  - Generating a Portfolio based on gathered data.

Research Methodology

Methodology

Web Crawler

Since above platforms do not provide public APIs to collect data, We proposed an advanced web crawler to scrape public user data.

- Python
- Scrapy
- Splash

Methodology

Statistical Models

- Three Unsupervised classification models to cluster users based on their behaviors.

  1. Q&A platform data
  2. Social Coding platform data
  3. Academics/ Professional data

- Algorithms: Random Forest, J48, SMO

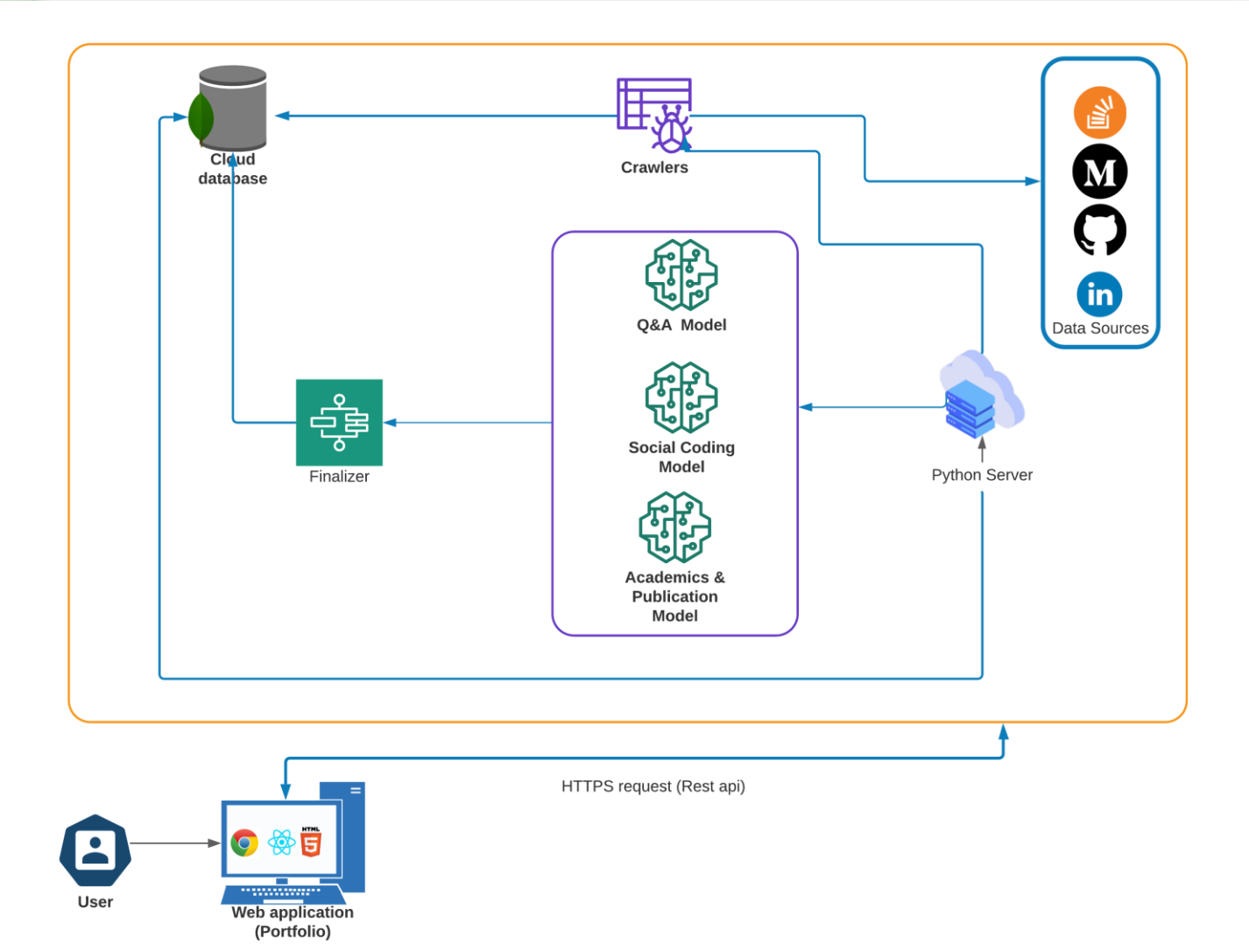
Methodology

Portfolio

- Generates the user profile by filtering out unnecessary details from the gathered data set.
- Filtered details will be ordered precisely on a timeline using the timestamps.
- Portfolio will be develop using next.js to maximize the search engine results.

Technologies: React.js, Next.js

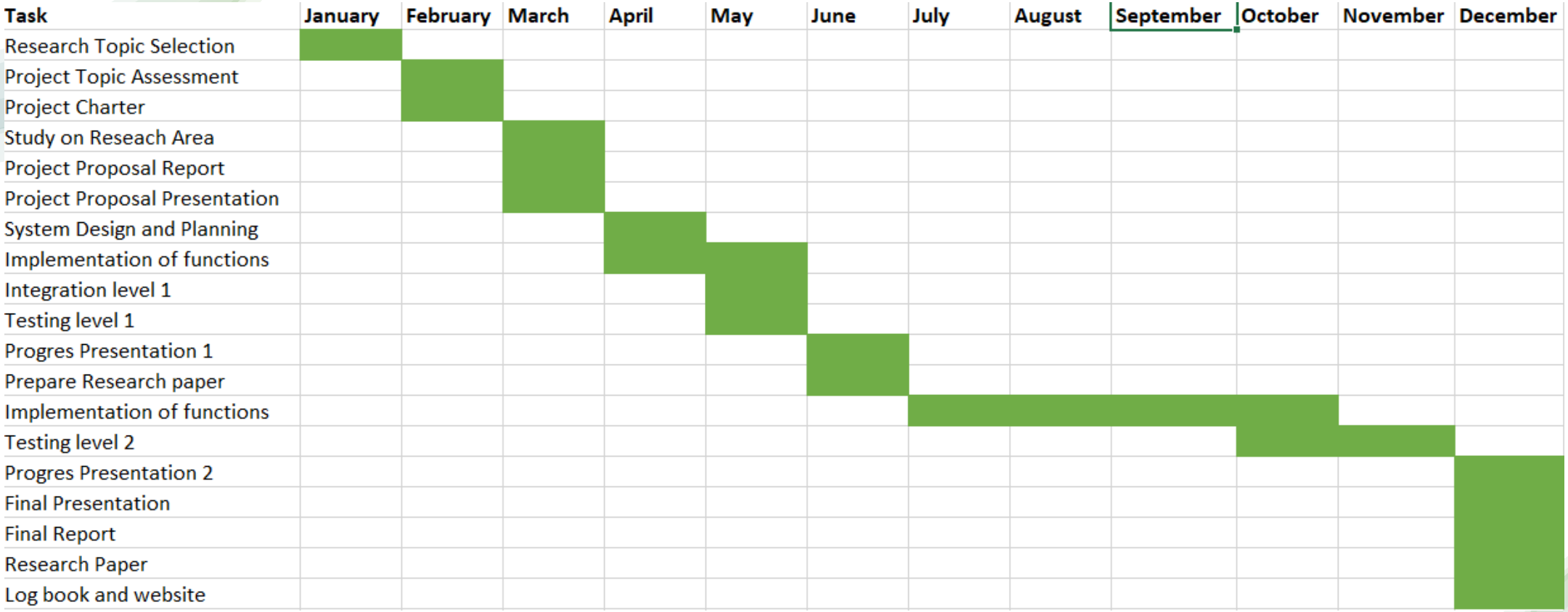
System Diagram



Commercialization

- Leading organizations can advertise their job openings.
- Premium accounts for recruiters to find unemployed skill full candidates.

WBS – Gantt Chart



References

[1] C. Hauff and G. Gousios, "Matching GitHub developer profiles to job advertisements," *IEEE Int. Work. Conf. Min. Softw. Repos.*, vol. 2015-Augus, pp. 362–366, 2015, doi: 10.1109/MSR.2015.41.

[2] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise Networks in Online Communities: Structure and Algorithms."

[3] A. Dargahi Nobari, M. Neshati, and S. Sotudeh Gharebagh, "Quality-aware skill translation models for expert finding on StackOverflow," *Inf. Syst.*, vol. 87, Jan. 2020, doi: 10.1016/j.is.2019.07.003.

IT18013610
Ekanayake P.M.D.P

B.Sc. (Hons) Degree in Information Technology Specialized in Software Engineering

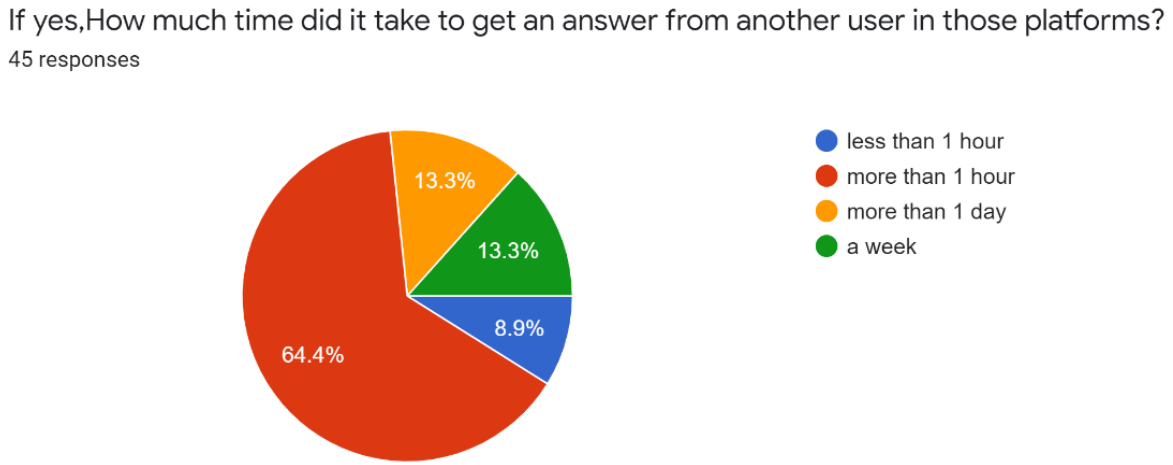
Introduction

Background

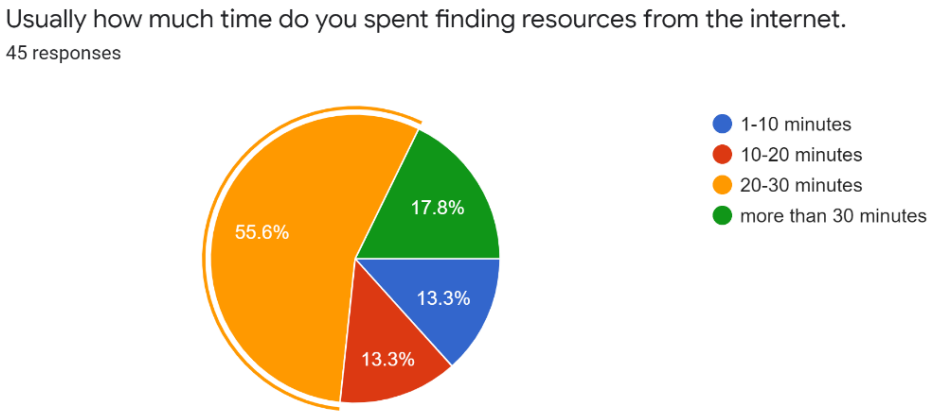
- Learning programming should be easy and possible with readily accessible information and resources.
- However, due to the vast number of available resources, it is particularly hard to find the most appropriate resources or information.
- Best way to find relevant resources is ask from an expert of the field.
- People cannot find expert all the time.
- Question and answering platforms have been used by people to find answers to their questions.
- Or users may search their question directly on the internet to find resources like tech blogs.

Background/Research Gap

- When using question and answering platforms users must wait for other users' answers.
- Most users have to wait at least 1 hour to get an answer from another user.
- Sometimes the first answer will not contain the information question asker has been looking for.



Background/Research Gap



- When searching questions in internet, users have to visit multiple websites that search engines provide.
- 80% percent of users visit at least 3 websites to find a complete answer.
- Majority of users spent more than 20 to 30 minutes to find an answer or resource from the public websites.

Research Gap

Platform	Features				
	Find similar question	Provide automatic answer	Code Support	Answers from multiple sources	One to one communication
Quora	Yes	No	No	No	No
Stackoverflow	Yes	No	Yes	No	No
Stack exchange	Yes	No	Yes	No	No
ProbExpert	Yes	Yes	Yes	Yes	Yes

Research Questions

- How to reduce waiting time spent on getting an answer from QandA platforms.
- How to create an answer from multiple resources automatically
- How to find YouTube video and GitHub repositories relevant to the question.
- Reduce the time spent on multiple visits to different web sites for gather information.
- How to provide access to users with low network bandwidth.

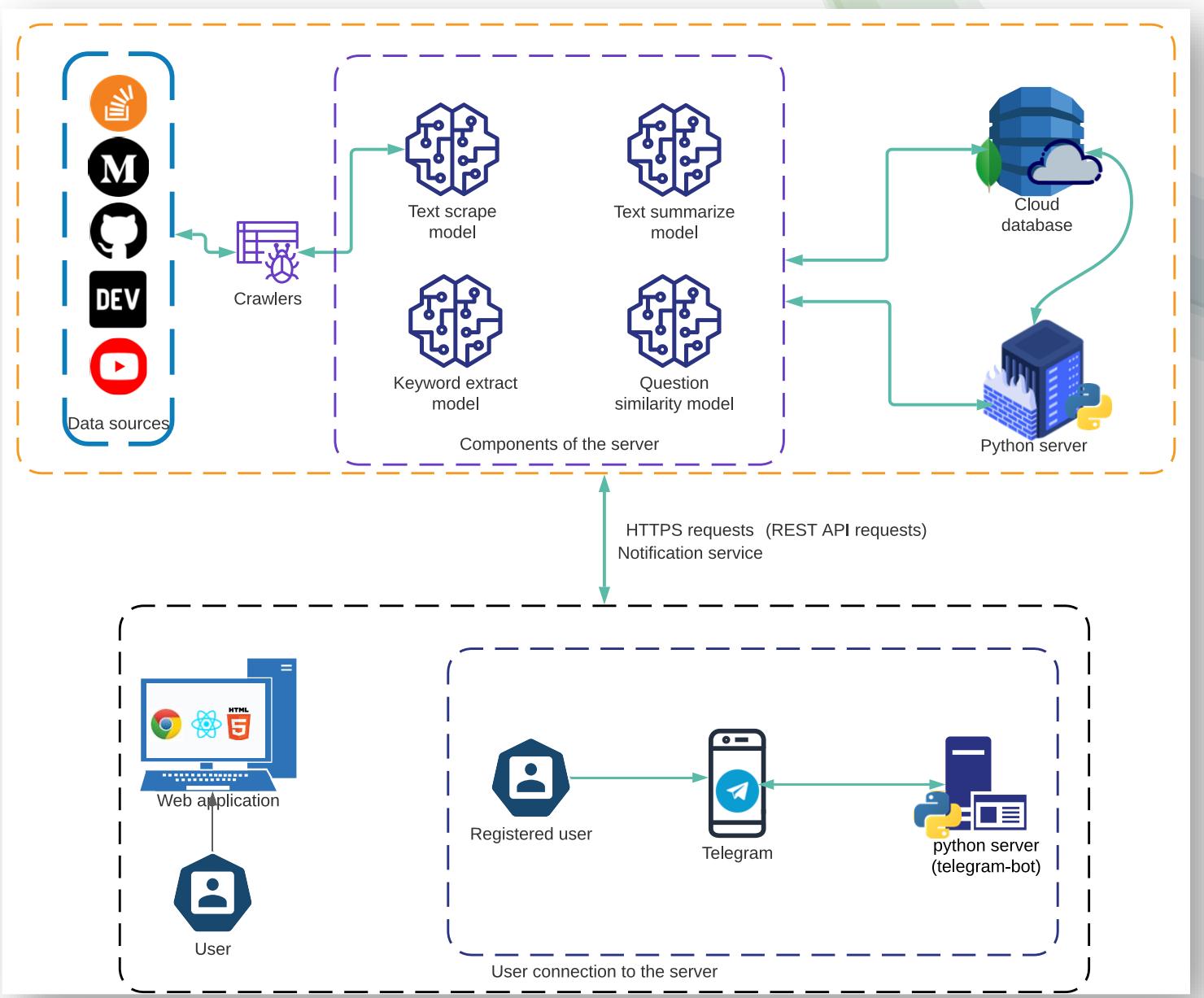
Specific and Sub Objectives

- Form an answer for user's question automatically using information gathered from public websites.
  - Keyword extraction.
  - Find similar questions in database
  - Scraping information.
  - Summarizing gathered information.
  - Find relevant YouTube video.
  - Find relevant GitHub repository.

Research Methodology

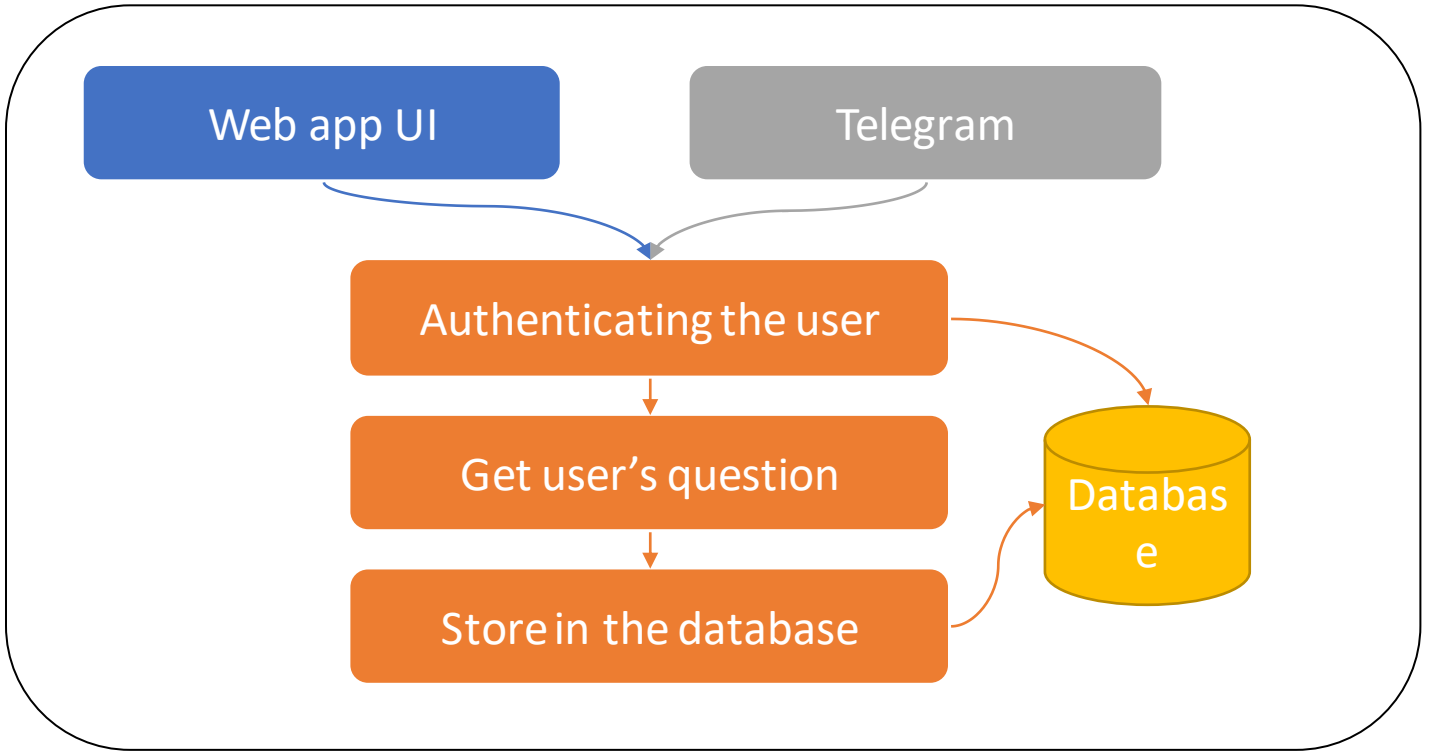
IT18013610 | Ekanayake P.M.D.P | 2021-155

System diagram



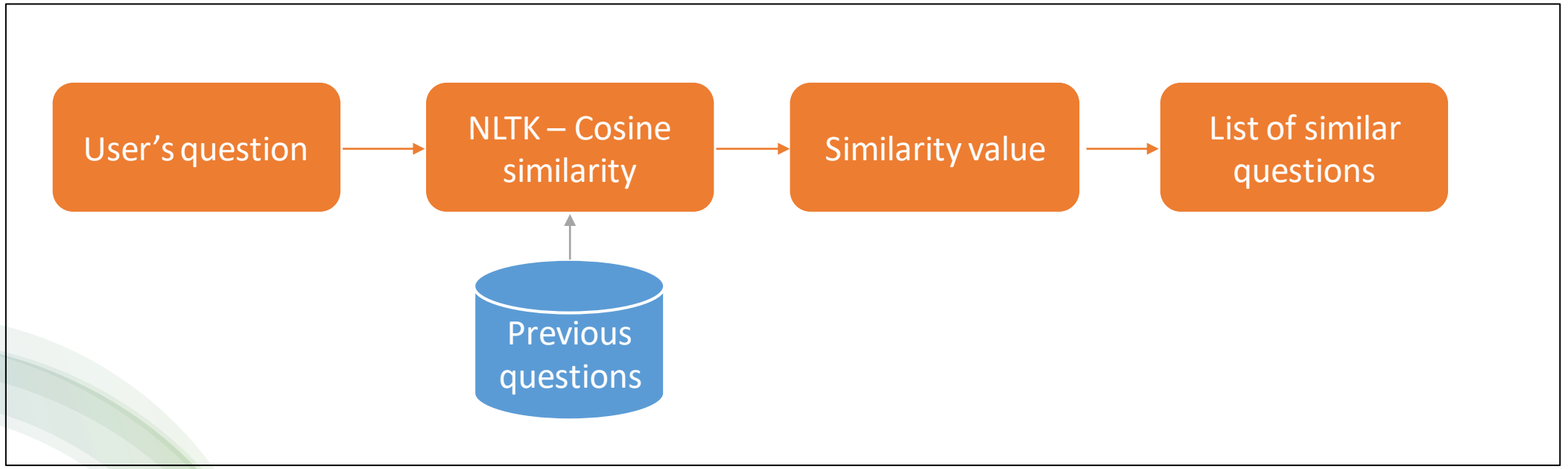
Technologies, techniques, algorithms

- Authenticating and getting user's question.
  - React js, Python, Telegram api, JWT, mongoDB



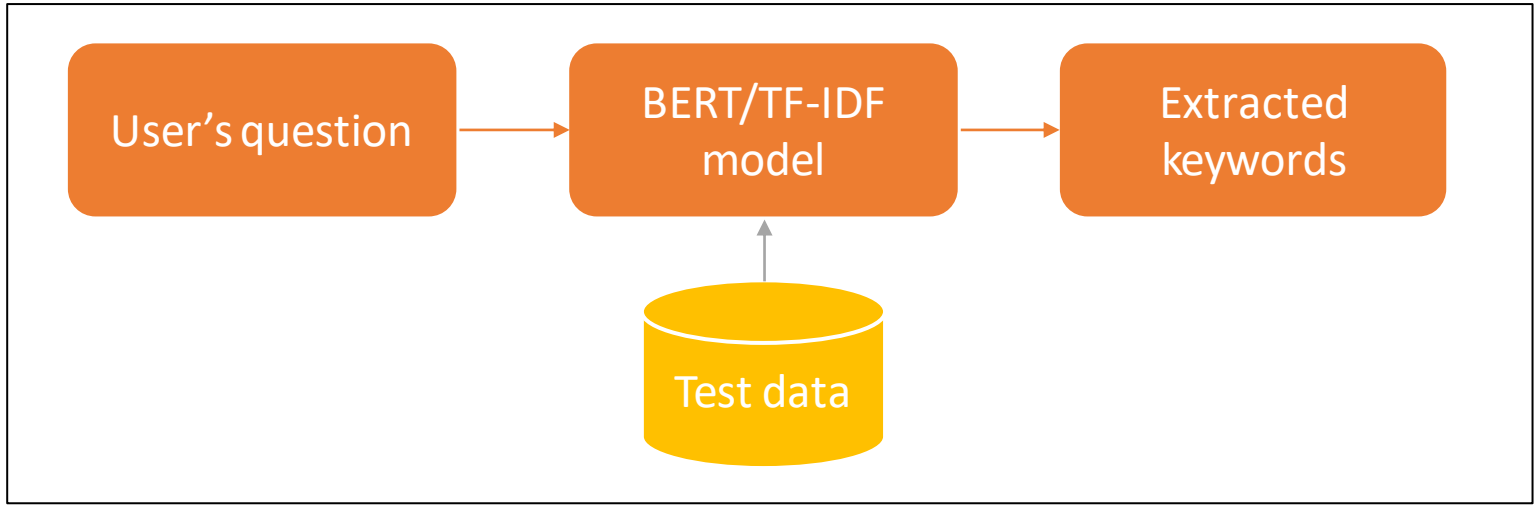
Technologies, techniques, algorithms

- Find similar questions in database using NLTK and cosine similarity (unsupervised algorithms – n-gram).



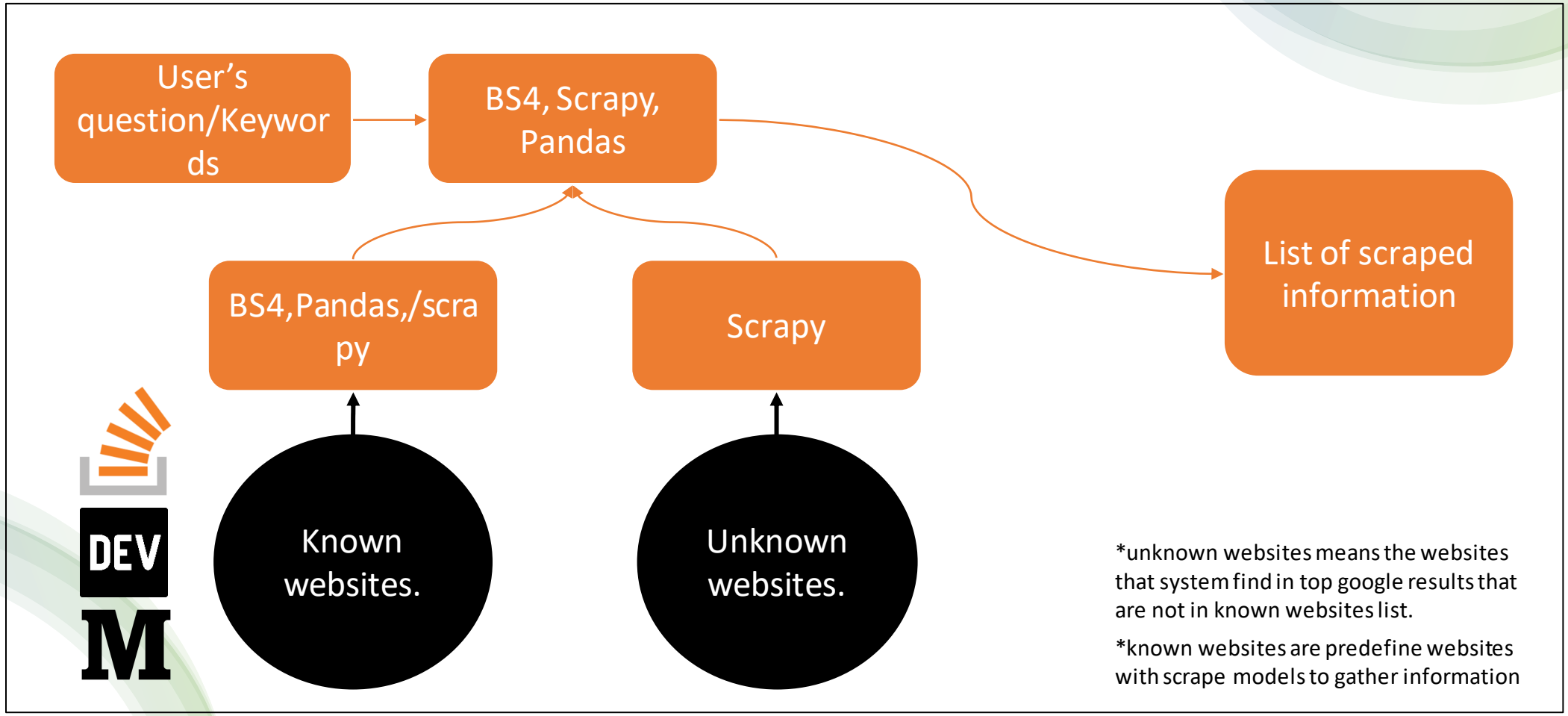
Technologies, techniques, algorithms

Use BERT, TF-IDF for keyword extraction.



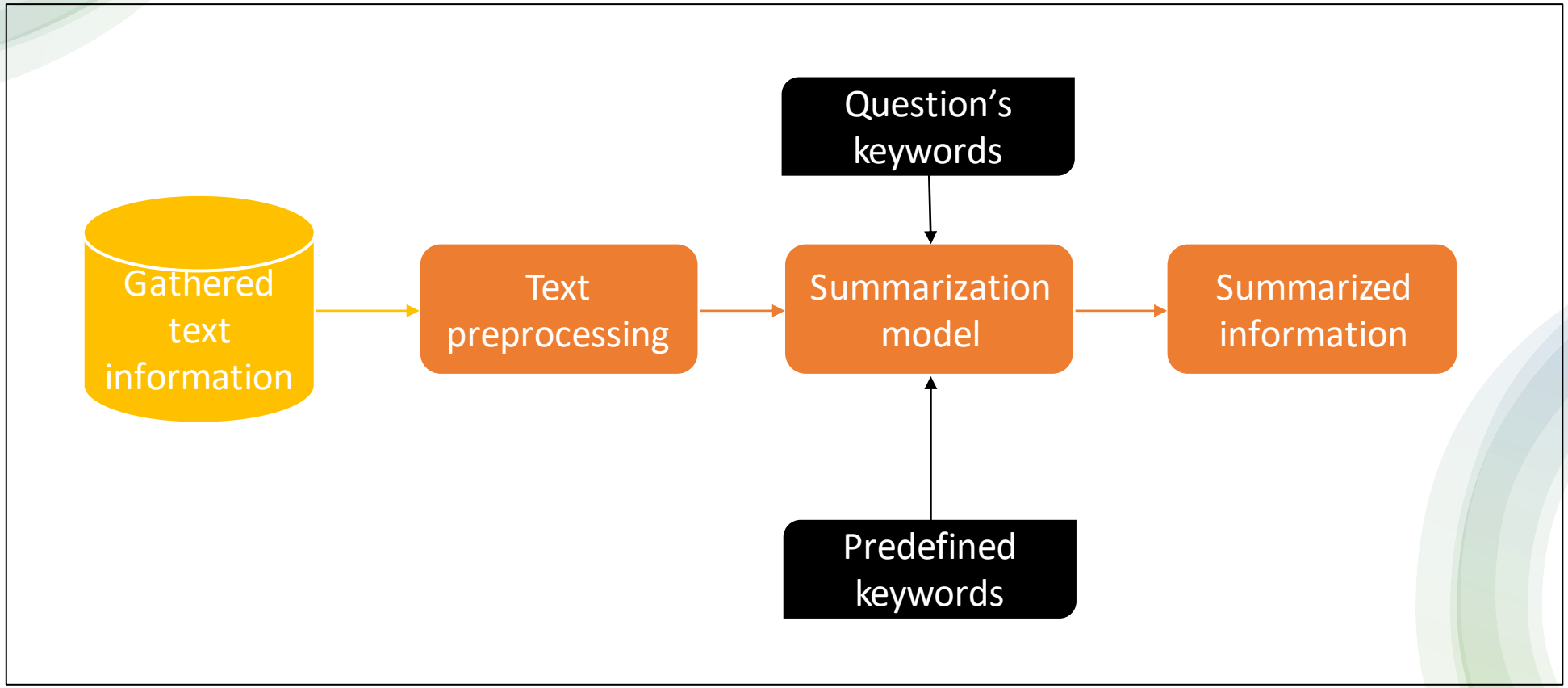
Technologies, techniques, algorithms

- Beautiful soup4, scrapy, pandas for scrape information.

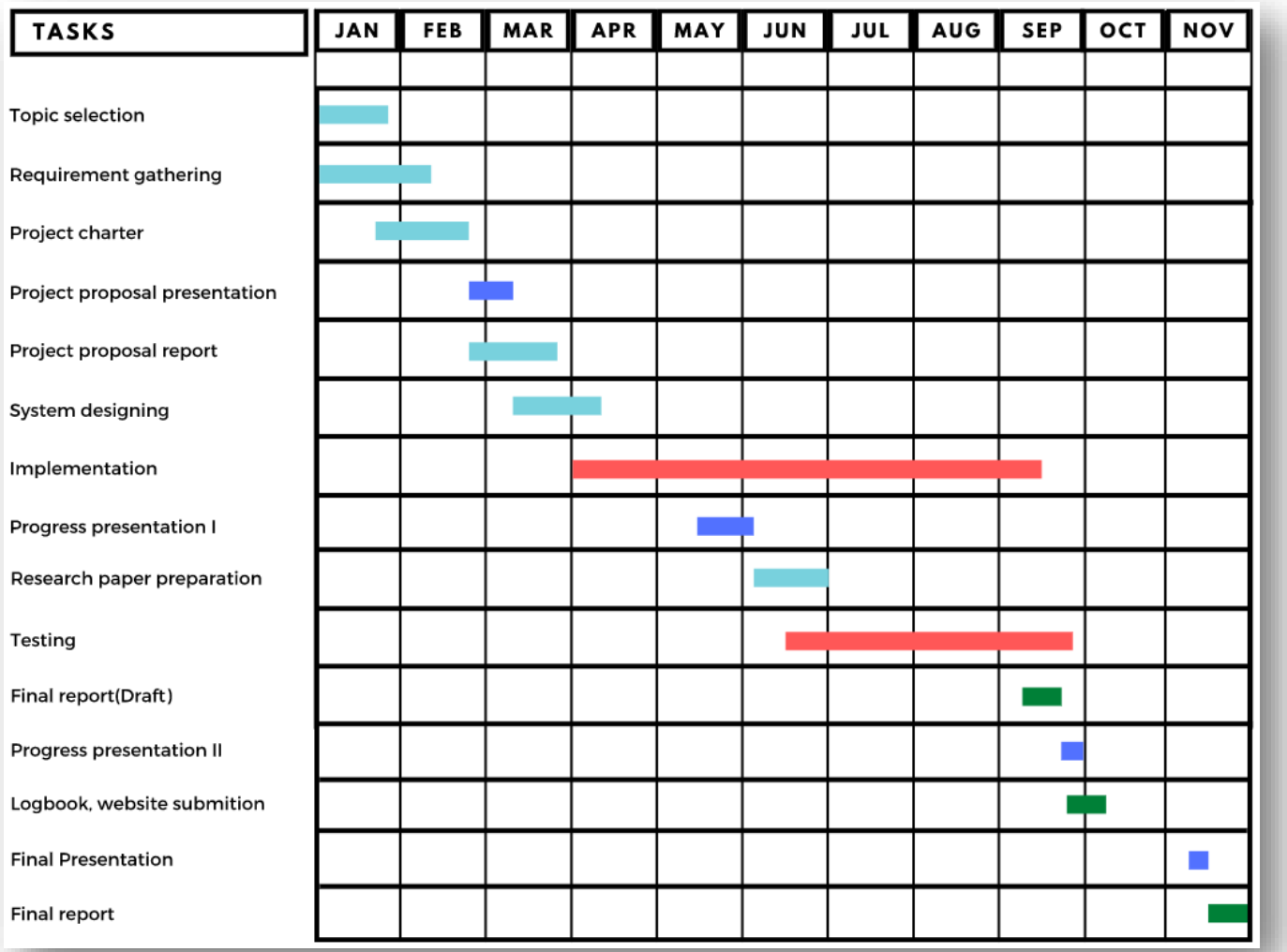


Technologies, techniques, algorithms

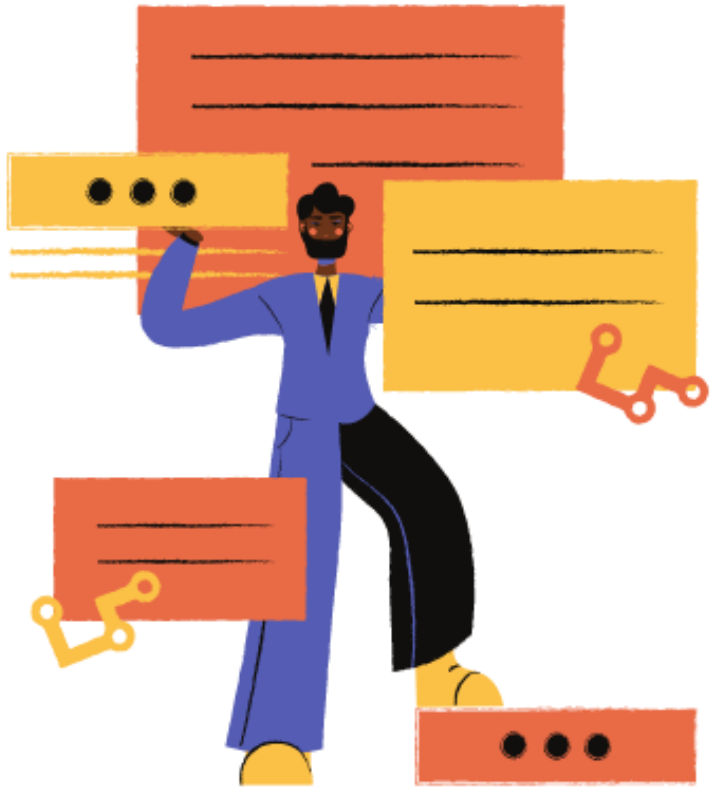
- NLTK, SpaCy, TF-IDF, Query heap algorithm for text summarization.



Work breakdown structure – Gantt chart



Supportive information



Commercialization

- Users will get 3 free automatic answers for a month.
- Monthly or early subscription service will be available for users who need to get more than 3 automatic answers in one month.

References

[1] W. Song, M. Feng, N. Gu and L. Wenyin, "Question Similarity Calculation for FAQ Answering," Third International Conference on Semantics, Knowledge and Grid (SKG 2007), Xi'an, China, 2007, pp. 298-301, doi: 10.1109/SKG.2007.32.

[2] I.Ali, R. Y. Chang and C. Hsu, "CRED: Credibility-Enabled Social Network Based Q&A System for Assessing Answers Correctness," 2020 IEEE Wireless Communications and Networking Conference (WCNC), Seoul, Korea (South), 2020, pp. 1-6, doi: 10.1109/WCNC45663.2020.9120750.

[3] H. Li, Z. Xing, X. Peng and W. Zhao, "What help do developers seek, when and how?," 2013 20th Working Conference on Reverse Engineering (WCRE), Koblenz, Germany, 2013, pp. 142-151, doi: 10.1109/WCRE.2013.6671289.

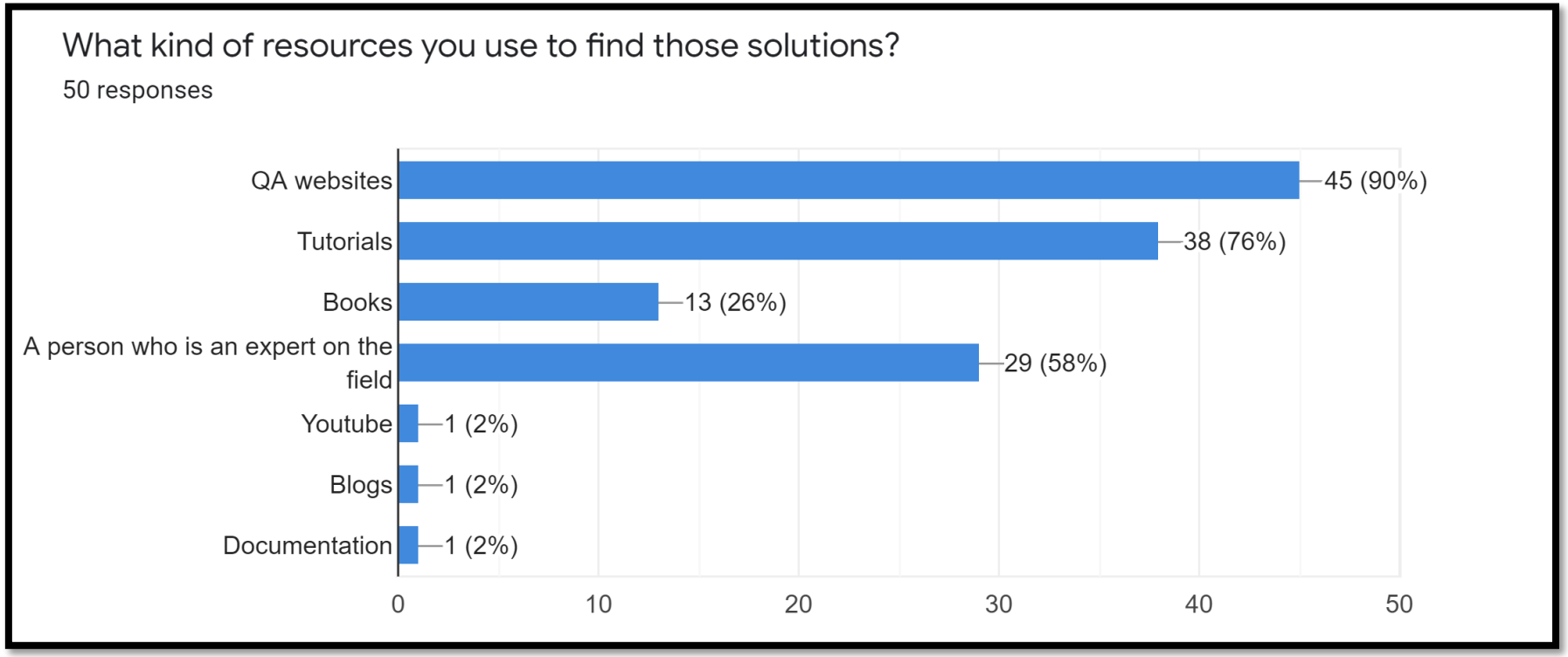
IT18011012 | Ranasinghe R.M.A.K

B.Sc. (Hons) Degree in Information Technology Specialized in Software Engineering

Introduction

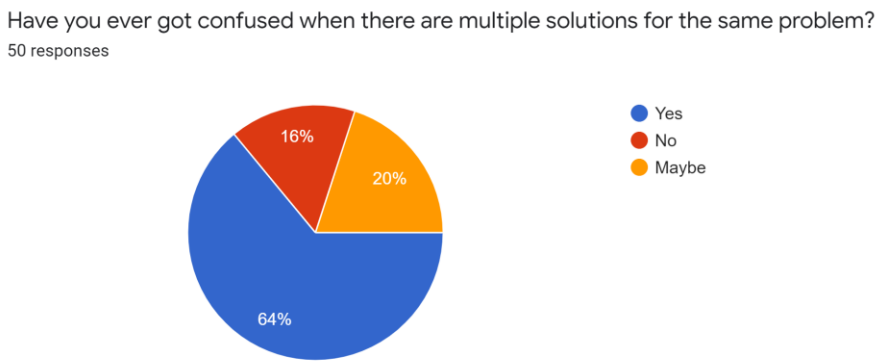
Background

- People mostly use Q&A platforms to get solutions for their problems rather than other resources.



Background

- Most of the time the given answers maybe not working.
- 80% of users must go back and forth at least 4 times to find a solution.
- 60% of users get confused because there are more than 1 answer are available for the same problem.



Research Gap

- When creating an optimized answer from a set of given answers, there are few things to be considered.
  - Keyword extraction
  - Summarization
  - Answer quality
  - Merge answers into a single answer

- Many researches are mainly concerned with summarization, keyword extraction and answer quality.
- There are no research available on how to merge answers into a single optimized answer

Research Gap

Platform	Features		
	Answer Up-vote System	Provide optimized answers	Answer comparison
Quora	Yes	No	No
Stackoverflow	Yes	No	No
ProbExpert	Yes	Yes	Yes

Research Question

Reduce wasting time on finding a solution that works for a problem

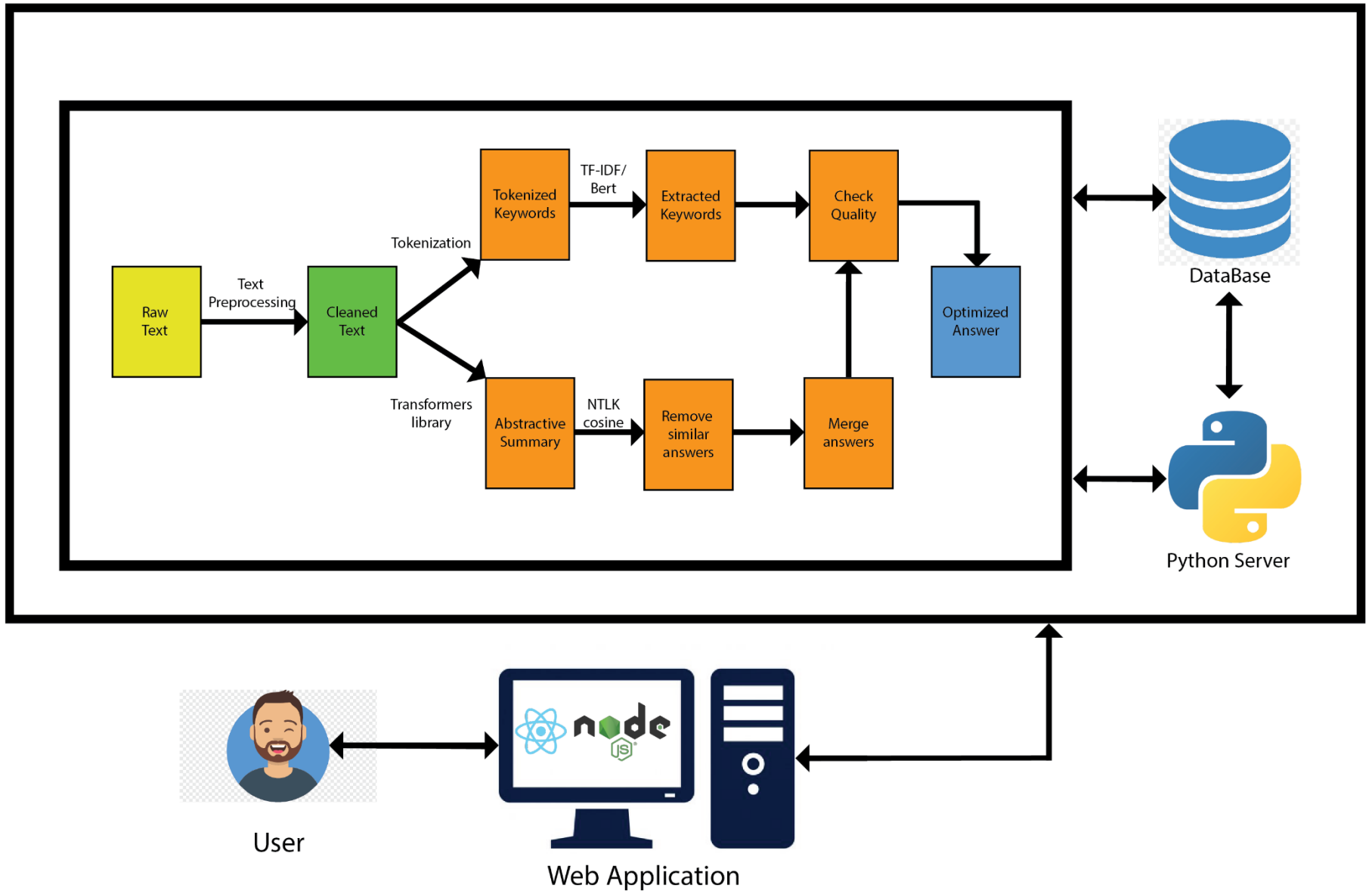
How to combine all the alternative solutions into a one optimized answer without harming the semantic information

Specific and Sub objectives

- Generating an optimized answer using the top voted answers
  - Text preprocessing
  - Keyword extraction
  - Summarize answers individually
  - Check and remove similar answers
  - Merge answers together
  - Check quality of the merged answer

Research Methodology

System Diagram



Technologies, Techniques and Algorithms

Python, React JS, Node JS

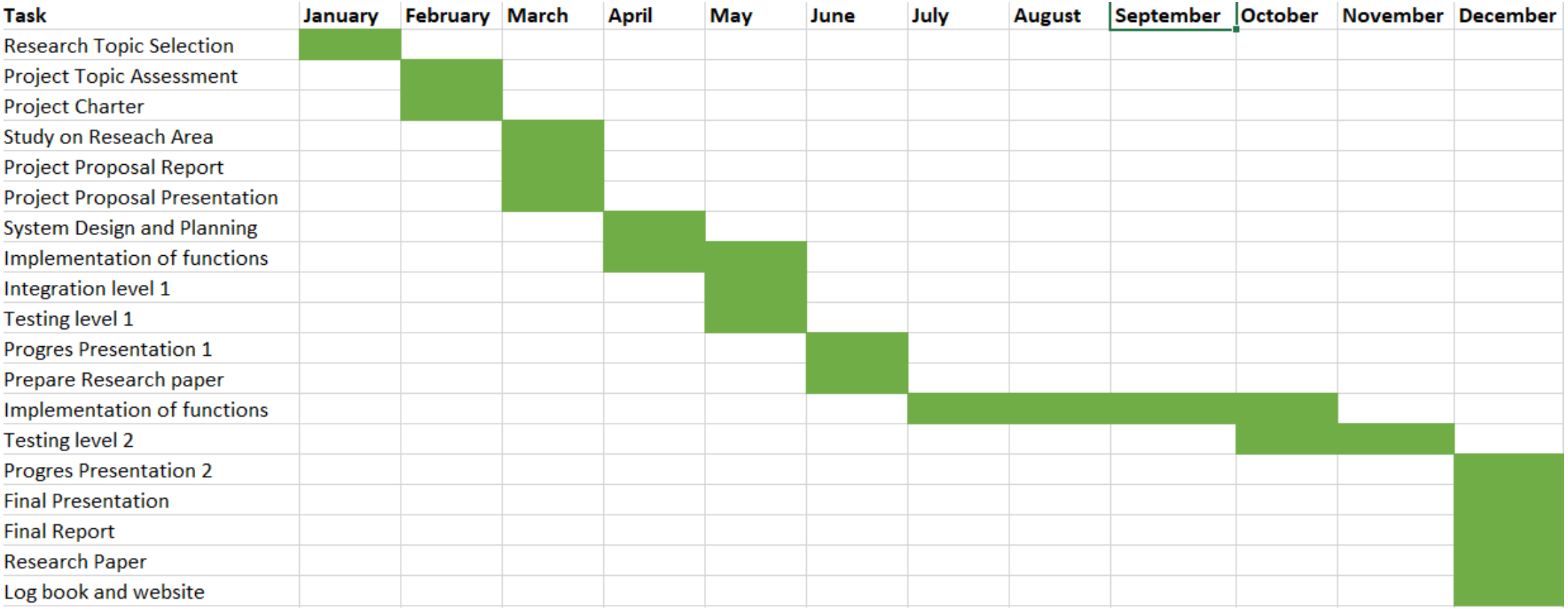
Database - Mongo DB

Machine Language

- NLP
  - NLTK
  - TF-IDF
  - Hugging face's transformers library
  - BERT

Cloud service - Digital ocean

WBS


Task
January
February
March
April
May
June
July
August
September
October
November
December
Research Topic Selection
Project Topic Assessment
Project Charter
Study on Reseach Area
Project Proposal Report
Project Proposal Presentation
System Design and Planning
Implementation of functions
Integration level 1
Testing level 1
Progres Presentation 1
Prepare Research paper
Implementation of functions
Testing level 2
Progres Presentation 2
Final Presentation
Final Report
Research Paper
Log book and website

Supportive information

Commercialization

- To view the optimize answer users have to unlock using ProbExpert coins.

References

[1] O. Tas and F. Kiyani, “A survey automatic text summarization,” *Pressacademia*, vol. 5, no. 1, pp. 205–213, Jun. 2017, doi: 10.17261/pressacademia.2017.591.

[2] “The automatic creation of literature abstracts | IBM Journal of Research and Development,” *IBM Journal of Research and Development*, 2021.

[3] E. Yulianti, R.-C. Chen, F. Scholer, W. B. Croft, and M. Sanderson, “Document Summarization for Answering Non-Factoid Queries,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 15–28, Jan. 2018, doi: 10.1109/tkde.2017.2754373.

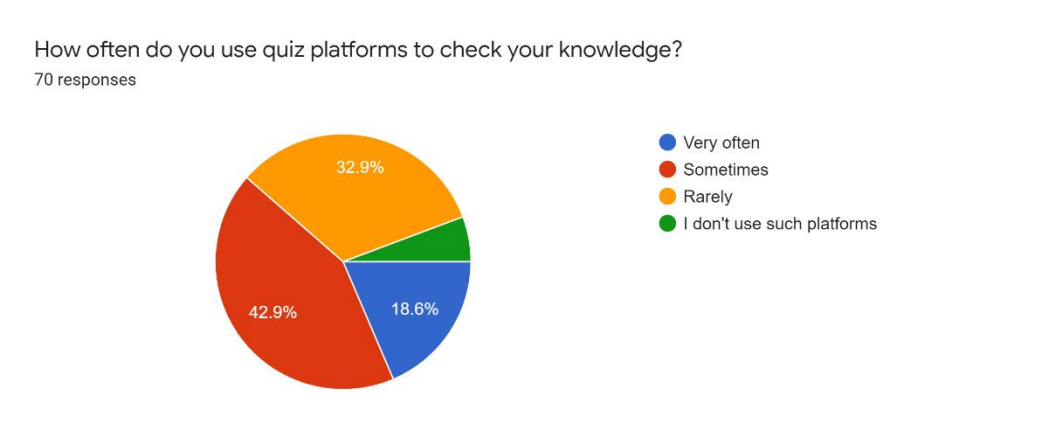
IT18078992 | Marapana S.K.C.W.K.M.R.T.S.B.

B.Sc. (Hons) Degree in Information Technology Specialized in Software Engineering

Introduction

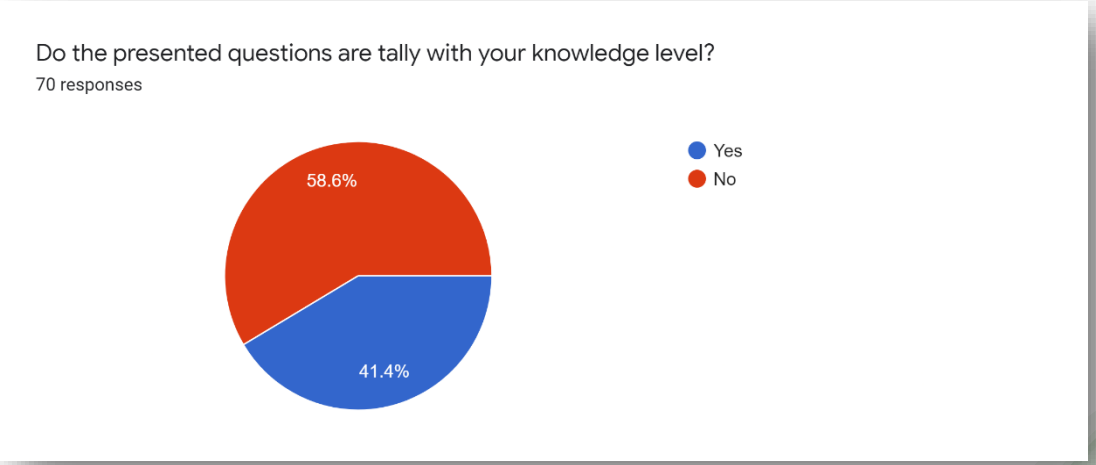
Background/Research Gap

- Most of the programming related learners use quiz platforms for knowledge checking, most of the times.



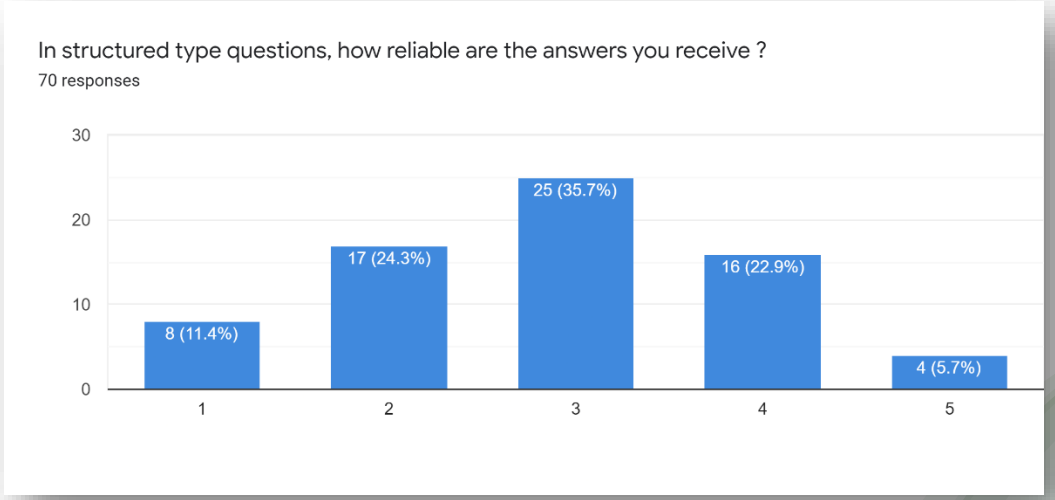
Background/Research Gap continued..

- But most of the learners feels that the questions they get are not adaptive to their knowledge level



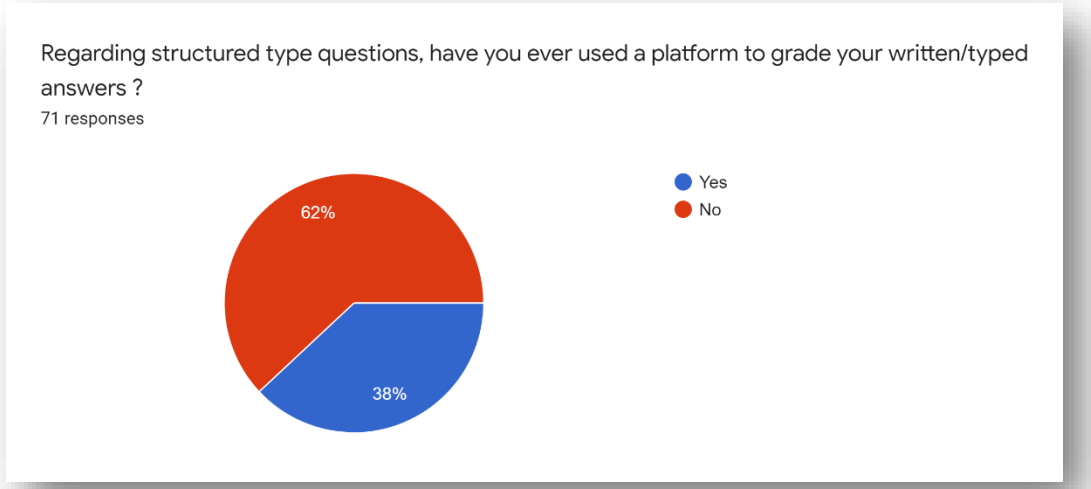
Background/Research Gap continued..

- Most users do not think the received answers for the structured-typed questions are reliable.



Background/Research Gap continued..

- Most of the users have not used a platform to grade their structured type answers.



Background/Research Gap continued..

When recruiting many organizations pays a strong eye to individuals' theoretical knowledge

Many platforms just focus on polishing coding skills only and not offers the facility to check knowledge on theoretical questions.

Most popular learning aid platforms like StackOverflow, Quora, and HackerRank do not provide many features to knowledge checking.

Background/Research Gap continued..

Platform	Features			
	Provide structured quizzes for knowledge checking	Provide good quality answers for users	Answer comparison	Scoring method
StackOverflow	No	No	No	No
Quora	no	No	No	No
HackerRank	Coding questions only	Yes (Coding only)	No	Yes (With test cases, coding only)
ProbExpert	Yes	Yes	Yes	Yes

Research Question

- How to check theoretical knowledge using structured type questions in more accurate way.
- Adaptive questions on each learner's knowledge level

Specific and Sub objectives

- Formulating structured type questions for knowledge checking, using existing answered questions of the platform based on user level
- Formulate questions from platform's existing user questions
- Formulate answers from optimal answers and top 4 voted answers and transform user's answer for similarity checking
- Similarity checking using cosine similarity and score assignment

Research Methodology

Methodology

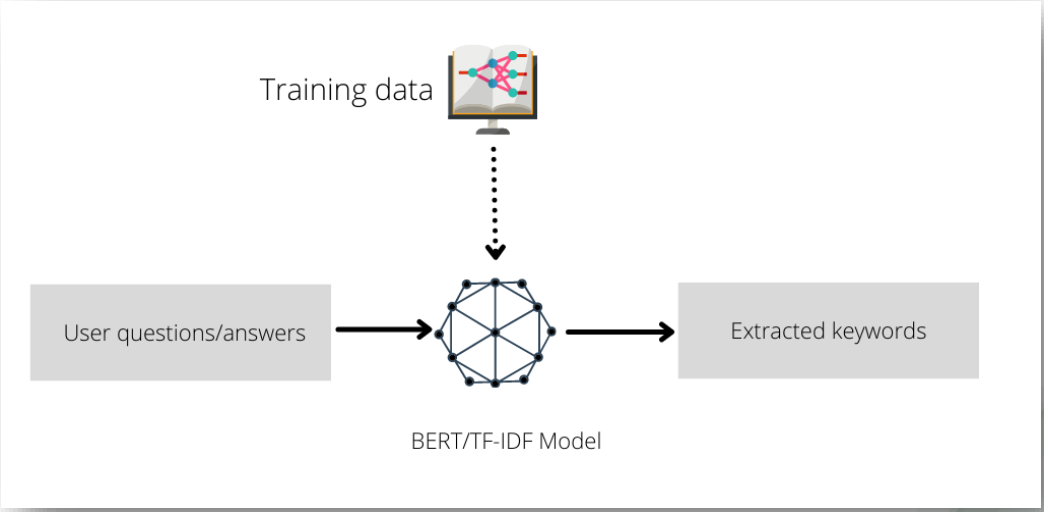
- The use of word embedding techniques TF-IDF, Word2Vec helps to handle synonyms or words with similar meanings, therefore enables to the formulation of a common answer.

  - Perform text pre-processing techniques, uniformity of cases
  - removal of stop words (using NLTK library), punctuations, and non-ASCII characters
  - Lemmatization using lemmatization algorithm
  - Feature extraction - using Word2Vec (Unsupervised algorithms- Skip-gram) for word embedding operations we can use Gensim library

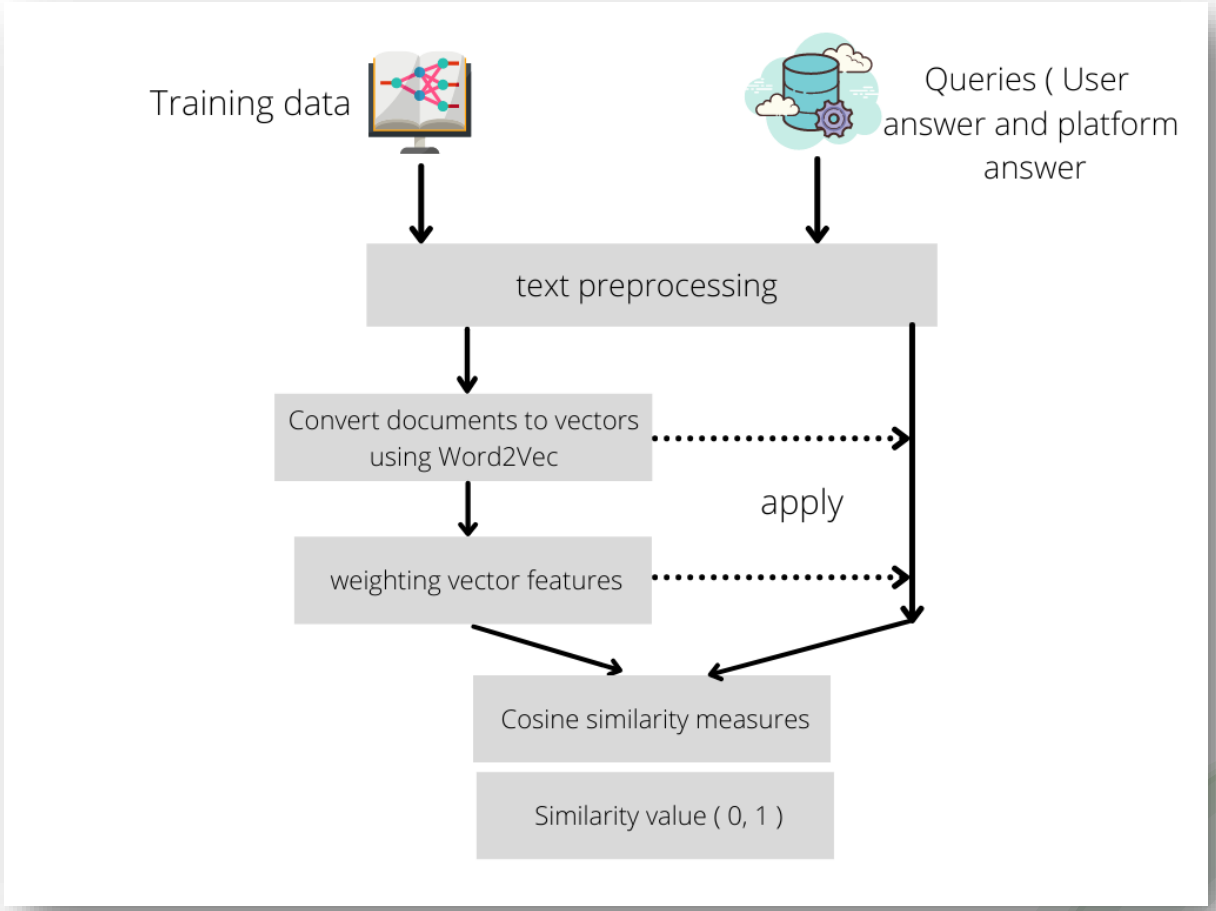
- WebCrawler for questions scraping
  - If user's desired questions are not available. Relevant set of questions will be scraped and displayed.

Methodology continued..

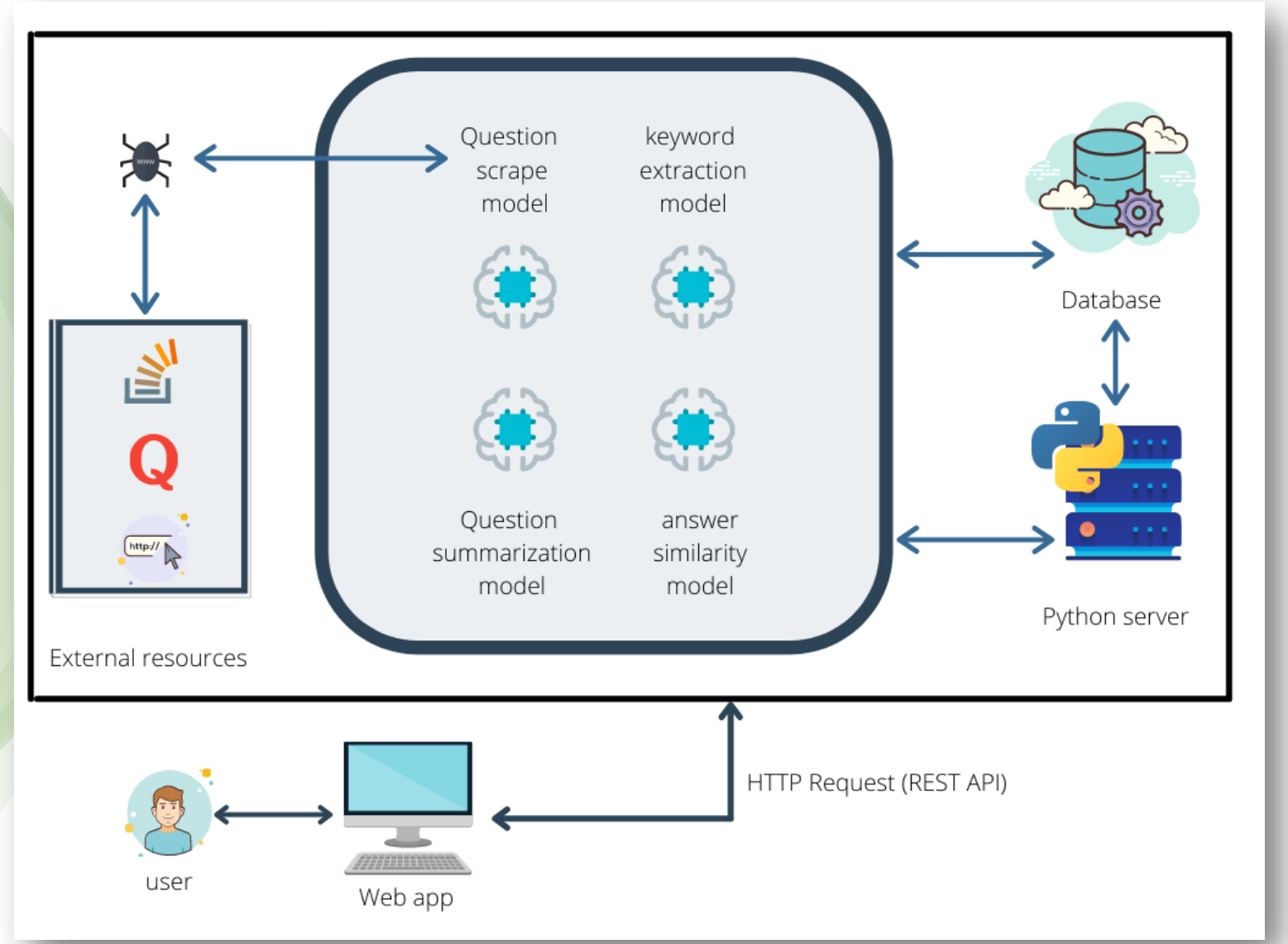
- Keyword extraction
  - We need to extract important keywords which helps to formulate quality answers and questions. This task can be achieved with BERT by making a keyword extraction model



- Similarity checking
  - Cosine similarity can be used in this scenario to determine the similarity level between the user answer and the platform's answer. Two vectors will be checked and based on the cosine similarity score; a score can be given for the answer.



System Diagram



Technologies, Techniques and Algorithms

Keyword extraction – BERT/TF-IDF

Word embedding – Word2Vec, Skip-gram algorithm, NLTK Library, Lemmatization algorithm, Gensim library

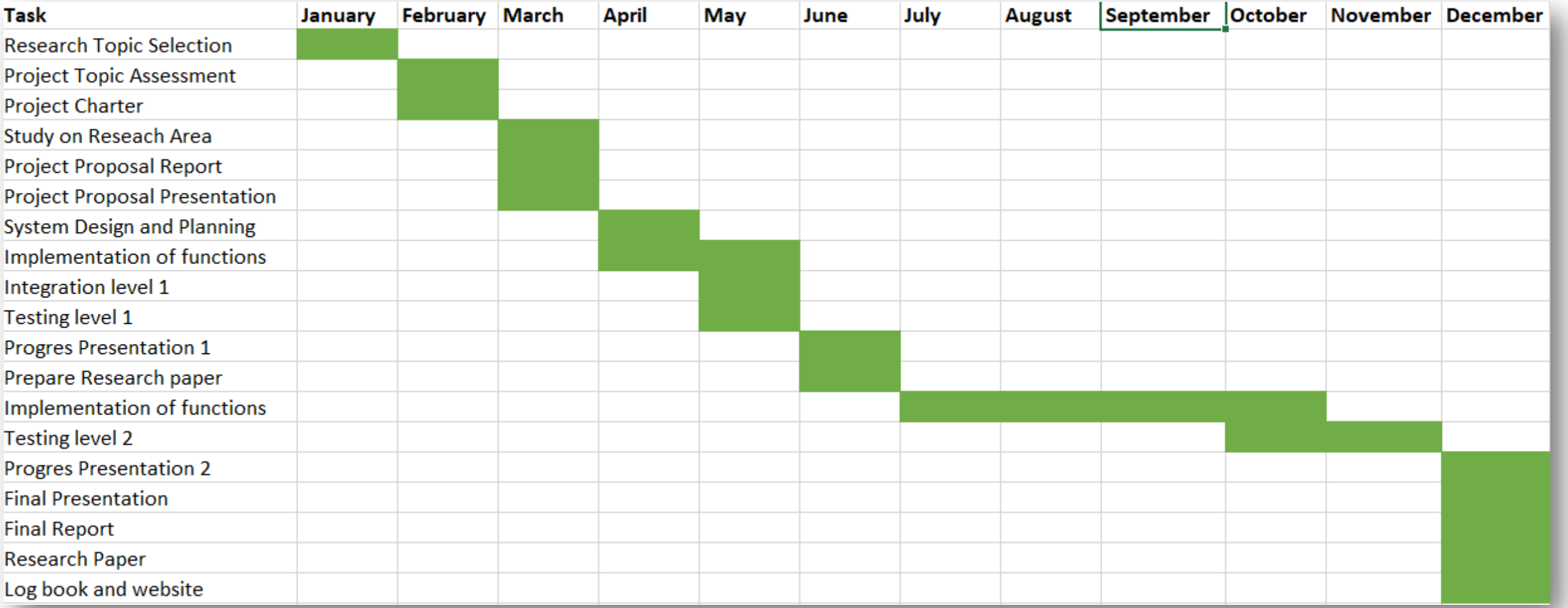
Similarity checking – Cosine similarity

Web crawling – Beautiful soup4, scrappy, pandas

Digital ocean cloud and AI service

Python, Reactjs, Nodejs, MongoDB

WBS



Task	January	February	March	April	May	June	July	August	September	October	November	December
Research Topic Selection												
Project Topic Assessment												
Project Charter												
Study on Reseach Area												
Project Proposal Report												
Project Proposal Presentation												
System Design and Planning												
Implementation of functions												
Integration level 1												
Testing level 1												
Progres Presentation 1												
Prepare Research paper												
Implementation of functions												
Testing level 2												
Progres Presentation 2												
Final Presentation												
Final Report												
Research Paper												
Log book and website												

Supportive information

Commercialization

- Subscription based accounts for users to access unlimited questions and answers.
- Organizations packages for knowledge checking of their candidates/employees

References

[1]Devlin, J.; Chang, M.-W.; Lee, K. & Toutanova, K. (2018), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding' , cite arxiv:1810.04805Comment: 13 pages .

[2] X. Jin, S. Zhang and J. Liu, "Word Semantic Similarity Calculation Based on Word2vec," 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), Hangzhou, China, 2018, pp. 12-16, doi: 10.1109/ICCAIS.2018.8570612.

[3] K. Jayakodi, M. Bandara and D. Meedeniya, "An automatic classifier for exam questions with WordNet and Cosine similarity," 2016 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2016, pp. 12-17, doi: 10.1109/MERCon.2016.7480108.