Scientific knowledge base creation from heterogeneous data

Final Year Project

2018-2022

A project submitted in partial fulfillment of the degree of

BS in Computer Science



Department of Computer Science

Faculty of Computing & Artificial Intelligence (FCAI)

Air University, Islamabad

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type (Nature of project) | | | [✓] Development [ ] Research | | |
| Area of specialization | | | [] WebApp [ ] Mobile App  [✓] AI based [ ] Embedded System | | |
| FYP ID | | | CS-SP-21-29 | | |
| **Project Group Members** | | | | | |
| Sr.# | Reg. # | Student Name | | Email ID | \*Signature |
| (i) | 181058 | Hamza Ali | | 181058@students.au.edu.pk |  |
| (ii) | 181122 | M. Huzaifa | | 181122@students.au.edu.pk |  |

# Plagiarism Certificate

This is to certify that, I Hamza Ali S/D of Saif Ali, group leader of FYP under registration no 181058 at Computer Sciences Department, Air University. I declare that my FYP report is checked by my supervisor.

Date:\_\_\_\_\_\_\_\_\_\_\_\_\_ Name of Group Leader: Hamza Ali Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Supervisor: Shoaib Malik

Designation:

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

HoD: Dr. Mehdi Hassan

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Scientific knowledge base creation from heterogeneous data

**Change Record**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author(s)** | **Version** | **Date** | **Notes** | **Supervisor’s Signature** |
| Hamza Ali  Muhammad Huzaifa | 1.0 |  | <Original Draft> |  |
| Hamza Ali  Muhammad Huzaifa | 2.0 |  | <Changes Based on Feedback from Supervisor> |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**APPROVAL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project Supervisor** | | |  |  |
| Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |  | | |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |
| **Project Manager** | | | |  |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |  |

|  |  |
| --- | --- |
| **Chair Department** | |
|  |  |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

# Dedication

I dedicate my dissertation to my family, friends, and my teachers. A special thanks to my parents who acknowledged my each decision, and helped me in achieving what I am today.

Also, I want to dedicate my work to my friends, that gave unique ideas throughout the university. I appreciate what they all have done for me when troubled. Thanks to Huzaifa, I’ve managed to polish my programming and documentation skill.

I dedicate this work to my teachers that were always available when I had any trouble with my studies, and they responded positively.

# Acknowledgments

I am thankful for my teachers that gave me an opportunity to work on this project. I would be expressing my gratitude to Sir Shoaib for helping us out with this project as a supervisor. I must be thankful to my parents for what I am now, because of him. Without their support, I couldn’t complete the project on my own.

Table of Contents

[1. Introduction 1](#__RefHeading___Toc4741_2815074626)

[1.1 Background 1](#__RefHeading___Toc2420_2653754408)

[1.2 Motivations and Challenges 1](#__RefHeading___Toc2180_2132206801)

[1.3 Goals and Objectives 2](#__RefHeading___Toc2418_2653754408)

[1.4 Project Plan 2](#__RefHeading___Toc2416_2653754408)

[1.5 Work Breakdown Structure 2](#__RefHeading___Toc2414_2653754408)

[1.6 Roles & Responsibility Matrix 3](#__RefHeading___Toc2412_2653754408)

[1.7 Gantt chart 4](#__RefHeading___Toc2410_2653754408)

[1.8 Report Outline 4](#__RefHeading___Toc2422_2653754408)

[2. Literature Review 5](#__RefHeading___Toc2634_2653754408)

[2.1. Introduction 5](#__RefHeading___Toc2136_2132206801)

[2.2. Why this project 5](#__RefHeading___Toc2160_2132206801)

[2.3. Working 5](#__RefHeading___Toc2158_2132206801)

[3. Software Requirement Specifications 8](#__RefHeading___Toc2632_2653754408)

[3.1. Introduction 8](#__RefHeading___Toc2156_2132206801)

[3.2. Overall Description 9](#__RefHeading___Toc2154_2132206801)

[3.3. External Interface Requirements 10](#__RefHeading___Toc2152_2132206801)

[3.4. System Features 11](#__RefHeading___Toc2150_2132206801)

[3.5. Domain Requirements 13](#__RefHeading___Toc2148_2132206801)

[4. Use Case Analysis 14](#__RefHeading___Toc2630_2653754408)

[4.1. Use Case Model 14](#__RefHeading___Toc2146_2132206801)

[4.2. Flow Diagram 15](#__RefHeading___Toc2144_2132206801)

[4.3. Sequence diagram 16](#__RefHeading___Toc2142_2132206801)

[4.4. Use Cases Description 16](#__RefHeading___Toc2140_2132206801)

[5. Proposed Approach 18](#__RefHeading___Toc2628_2653754408)

[5.1. Data Set Exploration 18](#__RefHeading___Toc2138_2132206801)

[6. Business Plan 19](#__RefHeading___Toc2626_2653754408)

[6.1. Business Description 19](#__RefHeading___Toc2426_2653754408)

[6.2. Market Analysis & Strategy 19](#__RefHeading___Toc2424_2653754408)

[6.3. Products/Services Description 19](#__RefHeading___Toc2162_2132206801)

[7. Conclusion 20](#__RefHeading___Toc2624_2653754408)

List of Figures

Table of Figures

[Figure 1: Gantt Chart 4](#Figure!0|sequence)

[Figure 2: Use Case Model 14](#Figure!1|sequence)

[Figure 3: Flow Diagram for Knowledge Base 15](#Figure!2|sequence)

[Figure 4: Sequence Diagram 16](#Figure!3|sequence)

[Figure 6: Sequence Diagram 16](#Figure!6|sequence)

[Figure 5: Sequence Diagram 16](#Figure!5|sequence)

[Figure 6: Proposed Approach 18](#Figure!8|sequence)

List of Tables

Index of Tables

[Table 1: Roles and Responsibility 3](#Table!2|sequence)

[Table 2: Use Case: Answer a Query 17](#Table!3|sequence)

[Table 3: Table 5: Use Case: Posting a Question 17](#Table!4|sequence)

Abstract

Solving a user’s query has been tough since the use of internet, as internet is the source of information and spams. For getting solution, one relies on the google but sometimes it gives false information via spamming tags. For that, one needs to have a search engine that would yields solution from authentic sites and scholarly articles. For that, we would be using web crawling, NLP and numerous data processing python tools for giving authentic solutions for students, teachers and any other tech seekers. Our web-app would simply input search query along with certain categories like Linux, Linux → server (sub category). After the input field, user’s queries would be entertained. In some cases, if the tools fails to gather relevant data, that queries can either be posted to our forums, which would be held till certain time, and after then (unsolved) it will be posted to stack overflow. This process will give out authentic solutions to user’s queries.

List of Abbreviations and Symbols

|  |  |
| --- | --- |
| **KB** | Knowledge Base |
| **NLP** | Natural Language Processing |
| **UI** | User Interface |

Glossary

|  |  |
| --- | --- |
| **Knowledge Base** | A self-serve online library of information about particular product, service, department, or topic. |
| **\*** | will be adding more till completion of this report |

1. Introduction

Google has become the best query solver for its reputable solving algorithm, but many sites owners uses fake tags for ranking up, due to which user can’t ensure the solution to be generic. Therefore they rely on reputed authentic sites like stack overflow, Quora and many more, but the problem is that many tech seeker don’t know where to start from. For solving their problem, we are creating a web app that would aid them in solving their queries. Our app would rely on major authentic sites and academic datasets for crawling answers, and would be shown if found, else it’ll be send to our forums / stack overflow. For this project, we are using front-end for the web design, python and Mongo DB for back-end.

* 1. Background

Techseeker (seniors) uses only reputed sites for their solution, be it the new technology or some help with the research, but for our juniors, they need certain guidance to distinguish between the reputed sites and unauthentic sites. For that, we are developing a knowledge base for scientific documents, which would include differnet tech terminologies, from AI, BigData, and Linux to certain small queries. A knowledge base is a self-serve customer service library that includes information about a product, service, or some specific topic. A knowledge base allows you to create self-service customer support content around recurring topics, issues, and themes. A knowledge base doesn’t simply have to be a reservoir of help center articles. It can also include functions like a frequently-asked questions (FAQ) section, a user forum, articles, white papers, how-to articles, video tutorials, case studies, and dictionaries or glossaries — essentially, anything that helps customers understand and use your products or services.

* 1. Motivations and Challenges

Every newbie in the tech world needs help to take certain direction for the career, they need to know what will yield useful data from the rest. They would be needing a knowledge base that can extract the required information from such sites and provide them the soltuion in easy UI for everyone. We would be extracting data from wikipedia using Wikipedia API and devlepod the relation among the text where nodes represent entities and edges represent relationship among them. Grammatical patterns were extracted from the given text using NLP library SpaCy, and neuralcoref. After developed pattern we developed knowledge graph.

* 1. Goals and Objectives

By taking input and categories from the user, our goals will be:

* Extracting the required information for specific topic from available datasets,
* Extraction of entity pairs from grammatical patterns from text using NLP library SpaCy,
* Developing relationships among text,
* Knowledge graph formation
  1. Project Plan

Our project is divided into 3 components, each being dependent on the others.

* FYP – 1
* FYP – 2
* FYP – 3

At each phase, we have set certain objectives or challenges to overcome in our process. Our FYP-1 was to gather required research for the project, while FYP-2 and FYP-3 revolves around the documentation and the practical part, where we programmed to obtain results from input queries via Wikipedia scraping. Also, our project was solely research and development based, so we didn’t invested a penny for our project, as majority of the tools required are open source. For the Gantt chart, visit [section](#_toc545) [1.7](#_toc13).

* 1. Work Breakdown Structure

**PHASE 1:** Project Idea finalized + documentation

**PHASE 2:** Knowledge graph creaion from wikipedia + other files

**PHASE 3:** Frontend development and linking

**PHASE 4:** TESTING + deployment

**PHASE 5:** Deployment

* 1. Roles & Responsibility Matrix

Table 1: Roles and Responsibility

|  |  |  |
| --- | --- | --- |
| Responsibility | (R1) Hamza Ali | (R2) M Huzaifa |
| Research | ✔ | ✔ |
| Finding Code | ✔ |  |
| Understanding | ✔ | ✔ |
| Implementation | ✔ |  |
| Documentation |  | ✔ |
| Visual representation |  | ✔ |

* 1. Gantt chart
  2. Report Outline

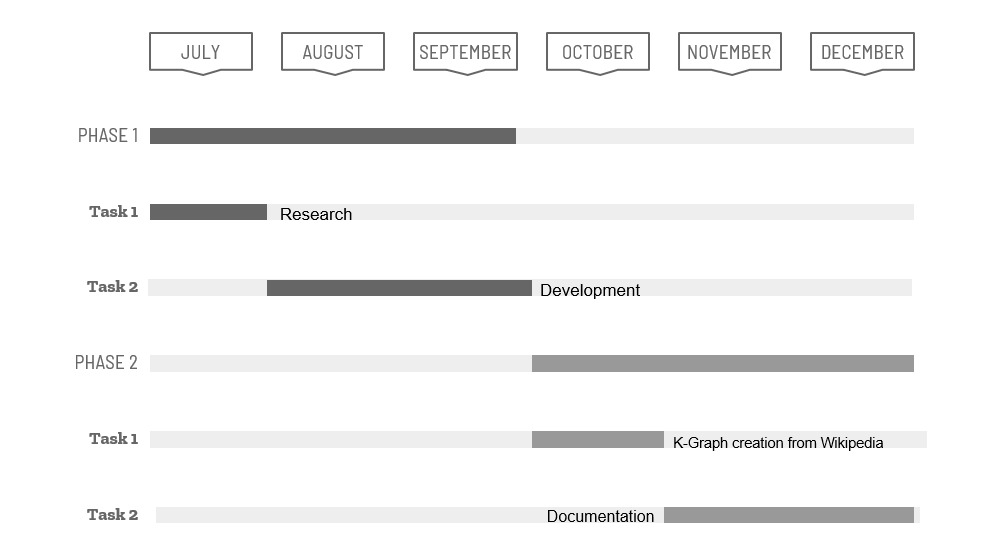


Figure 1: Gantt Chart

Our thesis report will be including each in-depth details of following topics.

* Literature Review
* SRS (Software Requirement Specifications) document
* Use Case Analysis
* Our Approach (proposed)
* Business Plan

1. Literature Review
   1. Introduction

Getting authentic information has been quite difficult now-a-days due to large number os scams/ fake meta tags. The huge internet is useful if one knows how to use it for themselves. For a developer/ tech seeker, one would always consider relevant and up to date solutions for their queries, therefore they depends on getting that without wasting time in browsing through each links. Similarly, for newbies, they would get many links for their queries, each with relevant or fake meta tags, but they can’t distinguish between the both. In both cases, they want a platform that can provide authentic and up to date solutions to their problems. They just need to provide the queries and relevant categories to our server, and let the server handles the rest, which would crawls the web and academic datasets for the solutions. Provided the case that if the solution could not be found, then the queries (with unique tokens) would be posted to our forums where our experts will answer it for them.

* 1. Why this project

There are many sites that provide same functionality, but newbies don’t know about them. Some are (of close resemblance) open-KB that don’t have enough datasets. Stack-overflow, Quora, and many more. Many tech seekers needs quick solution within a click away, that’s what we are trying to achieve in our project. We would be using help of large academic datasets with recent articles, and the internet, if solution is not provided in the datasets. In worst case, it’ll be sent to our forum for our experts to solve.

* 1. Working

For the time being, we are utilizing the help of Wikipedia that is scraping each links of topic to find relevant data, and printing it out, along with different entities that it found with their respective knowledge graph. Our project is divided into 3 parts

* + 1. Web Scraping

Our main goal was to scrape the data in the internet. We have used provided datasets for this task. For that, we have to submit the topic name and let it do its job. For this, we have used the given libraries.

***Pandas***

Pandas is used for data analysis and its manipulation. In our code, we have used it for tabular data of each list that has to be printed. Our list consists of all pages with relevant text, categories and topic. This list will be added to source that can be printed later in the code.

***Concurrent.futures***

This is the best part of our code that runs our task in parallel making process execution faster. From our code, we have assigned 5 threads for our task

* + 1. Getting entity pairs

For this task, we have selected the first page of our topic, that is most relevant, As spacy and neuralcoref only works well with spacy 2.1.0, neuralcoref 4.0, Python 3.7.0, so we have to install it as per requirements, in order for it to work. Also the downside of spacy 2.1.0, we have to add filter spans function that only works with spacy above 2.1.4. After all requirements, we would create pairs for first page, let's see how it works. To begin with, we would use regex to replace next line and numbers, as we don't need it for the nlp.

Then we would use neuralcoref to replace each pronoun with the main person (resolving co-reference clusters), after that, we would create unwanted tokens that would include particle, determiners, stop words. Then, we would resolve each entity if it's not in unwanted token or stop words. We may confirm each entity along with its type after the refinement part. After refinement, the program would tokenize each entity, and at the last, we would use pandas Data frame for printing the data accordingly.

* + 1. Creating knowledge graph

By using NLP library SpaCy, we extracted entity pairs from large amount of text. The spaCy library comes particularly handy when dealing with big amounts of text. It may be used to create data extraction and natural language understanding systems, as well as to pre-process text for deep learning.

The dependencies can be mapped in a directed graph representation:

* Words are the nodes.
* The grammatical relationships are the edges.

The get\_entity\_pairs function defines entity pairs as entities/noun chunks with subject — object dependencies connected by a root verb “**from\_pandas\_edgelist (df, source='source', target='target', edge\_attr=None, create\_using=None, edge\_key=None)**. This function returns a graph from Pandas DataFrame containing an edge list. The Pandas DataFrame contain at least two columns of node names and zero or more columns of edge attributes. **layout = nx.spring\_layout (k\_graph, k=0.15, iterations=20)** This function creates a graph where k controls the distance between the nodes and iterations is the number of times simulated annealing is run. If we don’t use this function then the nodes will be overlapping each other which would be challenging to understand. The next function nx.draw\_networkx draws the graph with Matplotlib with options for node positions, labeling, titles, and many other drawing features means we can set the property of each element according to our choice. The network\_edge\_label creates a connection between vertices and then we show it up.

1. Software Requirement Specifications
   1. Introduction
      1. Purpose

Purpose is to develop a knowledge base that can entertain the query given to it on the basis of provided datasets. It will work by developing relationship and forming entity pairs among them. If the query can’t be resolved from provided documents then experts will answer that particular query.

* + 1. Document Conventions

The main chapter headings are labeled as Chapter no. and chapter name with font size 20 **BOLD** and font style Time New Romans. The headings inside the chapter are **BOLD** with font size 16 and labeled as 1.1, 1.2, 1.3, and the subheadings among them are labeled as 1.1.1, 1.1.2 **BOLD** and so on.

* + 1. Intended Audience and Reading Suggestions

This report is intended for developers, project managers, users, testers, and writers. Chapter defines basic overview of the project by reading chapter 1 users can study about basic functionality etc. Chapter 3 describes purpose, audience type and what will be the scope of the project. Chapter 4 describes functionality. Chapter 5 defines flow of project, diagrams to visualize the project so that developers, testers can understand the flow of project. Chapter 6 defines business model.

* + 1. Product Scope

Knowledge base will be part of platform which will respond to user query on the basis of dependencies and relations. It can be any forms like FAQs section, suggestion on search etc. The benefit of knowledge base will be that user will get answer to his solution on the spot without waiting for response from the expert. If query is unable to get resolved by knowledge base itself then expert will answer.

* 1. Overall Description
     1. Product Perspective

Our goal was to scrape the data in the internet and provided datasets. For that, we have to submit the topic name and its categories, and let it do its job. For this, we have used the given libraries.

***Pandas*** is used to analyze and manipulate data.

***Concurrent.futures*** to run tasks in parallel

* + 1. User Classes and Characteristics

**Getting entity pairs**

With the help of libraries: Spacy, neuralcoref we developed relations among data and finally got entity pairs.

**Knowledge graph creation**

After extraction of entity pairs the function nx.draw\_networkx draws the graph with Matplotlib with options for node positions, labeling, titles, and many other drawing features. The network\_edge\_label creates a connection between vertices and then we show it up.

* + 1. Operating Environment

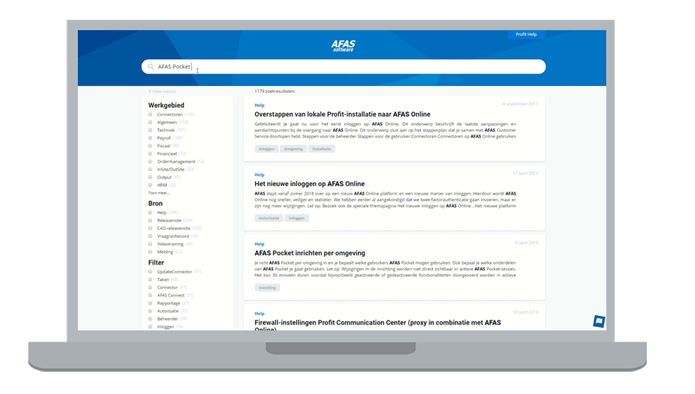
It can be integrated to any software or website, but we are currently developing it for a social media platform. It does not require any special system requirements.

* + 1. Design and Implementation Constraints

Only owners can update/remove datasets on the basis of which query will be resolved. The users can only access the service. If they didn’t get the answer then they can post their question which will be answered by expert.

* 1. External Interface Requirements
     1. User Interfaces

User interface will include a search bar where user will put his query and there will be several other options like selecting a category and further sub category to get more accurate answer. After search answers will be shown to the user. There will be another option to post a question for which you need an expert answer.



* + 1. Hardware Interfaces

Knowledge base does not require any special hardware to run. It utilizes a smaller amount of memory to provide you results. Initially, we are running it on our systems where our system is operating as a server to answer the user queries.

* + 1. Communications Interfaces

When the user question is answered, the user will be notified. If his query is not solved, then he’ll be notified via email when solved by the experts.

* 1. System Features

We are using academic datasets for this project

* + 1. System Feature 1

Academic datasets

#### Description

Datasets contain all the data from where query is to be resolved. The query will be entertained on the basis of relations among the data stored in datasets.

#### Stimulus/Response Sequences

The user asks for some solution. It will be answered after searching from datasets. If the answer to the question is found it will be shown to the user.

#### Functional Requirements

The functional requirements associated with this feature will be:

* Must be authentic
  + 1. System Feature 2

Answer to query

#### Description

#### The user will post a question

#### Stimulus/Response Sequences

#### When user will post a question it would be forwarded to expert. After answering the question the user will be notified.

#### Functional Requirements

The functional requirements associated with this feature are:

* Question type
* Must select category and subcategory (e.g: Windows => Windows 11)
  + 1. Security Requirements

Datasets should not contain personal details of any individual. It shouldn’t reveal any personal detail of the organization to its users because it can may cause private information disclose to the public.

* + 1. Usability Requirements

A knowledge base must have following usability factors:

* Effective
* Error tolerant
* Efficient
* Must be maintainable
  + 1. Reliability Requirements

We can say the Knowledge base is reliable if:

* It must respond to user query
* The answer to the query must be accurate
  + 1. Maintainability/Supportability Requirements

Maintainability defines the time required for a solution or its component to be fixed, changed to increase performance, speedup or other qualities, or adapted to a changing environment. It will be maintainable if it is adaptable to above mentioned features.

* + 1. Portability Requirements

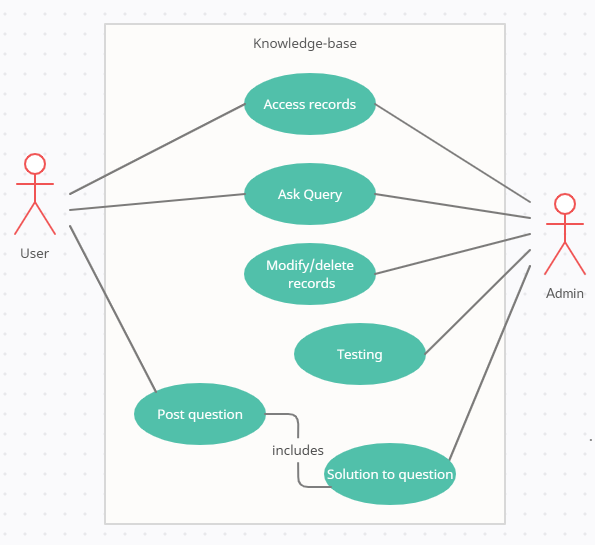
Portability requirements include:

* Operating systems and their versions,
* Network specifics,
* Browsers and their versions, and
* Devices and other hardware requirements.
  + 1. Efficiency Requirements

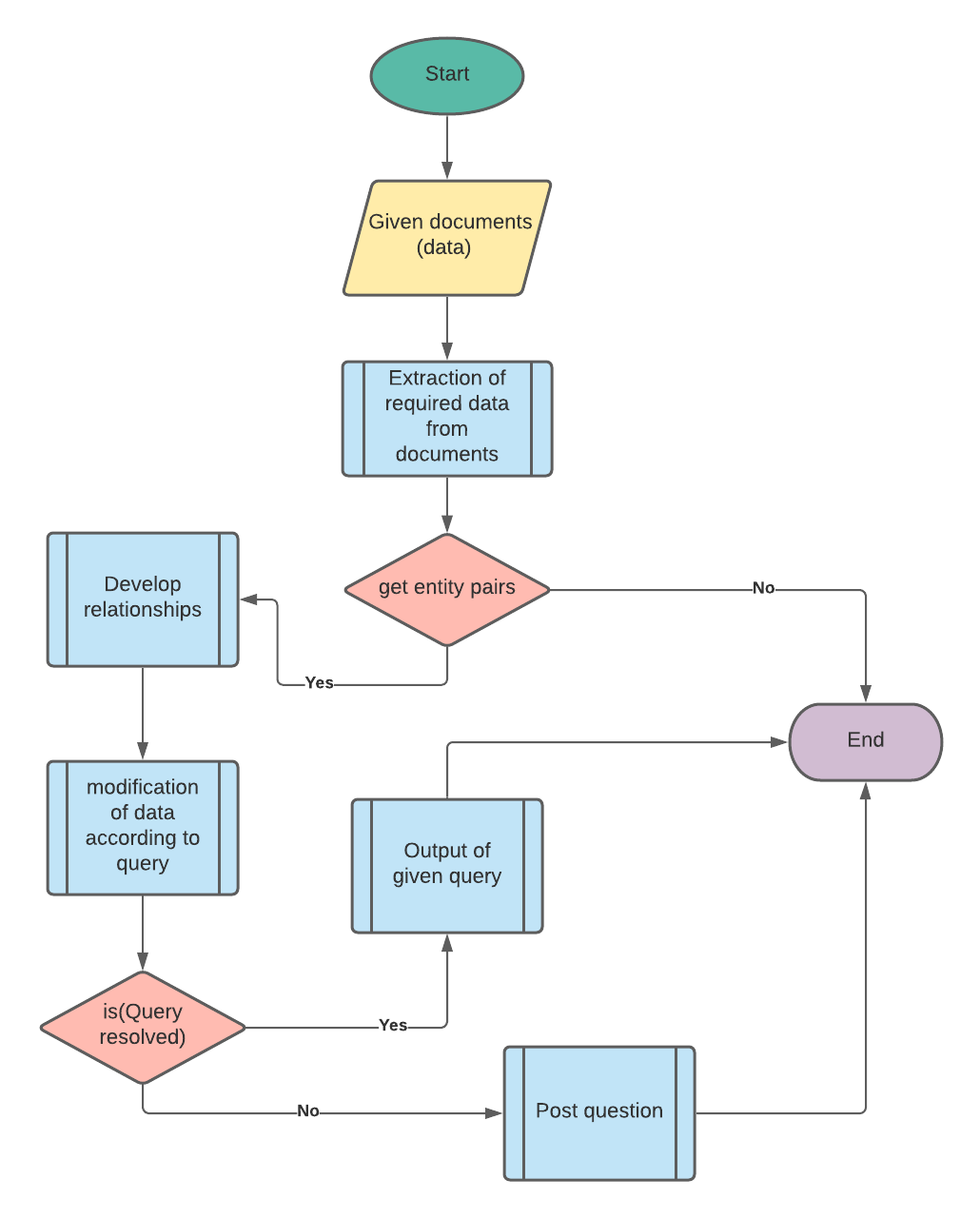
We can say the Knowledge base is reliable and efficient if:

* It must respond to user query
* The answer to the query must be accurate

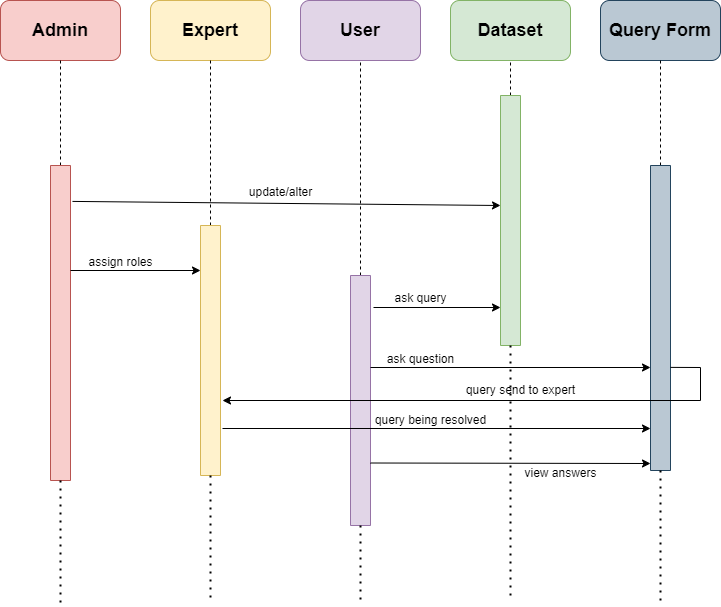
1. Use Case Analysis
   1. Use Case Model

Figure 2: Use Case Model

* 1. Flow Diagram

Figure 3: Flow Diagram for Knowledge Base

* 1. Sequence diagram

Figure 5: Sequence Diagram

* 1. Use Cases Description
     1. Answer a query

Table 4.4.1: Answer a query

|  |  |
| --- | --- |
| **Title** | Answer a query |
| **Requirement** | To answer a user query |
| **Rational** | Get answer |
| **Restriction or Risk** | Checking for answer |
| **Dependency** | Pc, datasets, internet |
| **Priority** | Safety, timing |
| **Output** | Answer to the query |

Table 2: Use Case: Answer a Query

* + 1. Post a question

Table 4.4.2: Post a question

|  |  |
| --- | --- |
| **Title** | Post a question |
| **Requirement** | To get answer a question from expert |
| **Rational** | Get answer |
| **Restriction or Risk** | May be time taking |
| **Dependency** | Knowledge of expert |
| **Priority** | timing |
| **Output** | Answer to the given question |

Table 3: Table 5: Use Case: Posting a Question

1. Proposed Approach
   1. Data Set Exploration

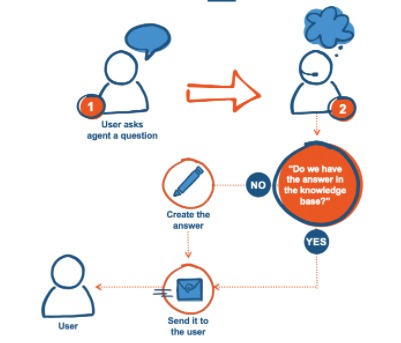


Figure 6: Proposed Approach

In this step, we explored our data-set and found the following limitations:

* Tackling of large datasets
* Web Crawling (request approval)
* Faster execution with higher accuracy

1. Business Plan

This project aims to provide academic knowledge to not only student, teach but all tech seekers along the globe. It would benefit every individual that wants to learn about new technology or any programming language. In short, if one wants to learn about latest tech, or solving errors in their code, they can just switch to our web-app and find their solution right away.

6.1. Business Description

As for the business, we are aiming it to be an **open-source** where each experts in the tech-world can integrate with the code and make it bug free and more efficient.

6.2. Market Analysis & Strategy

There are many sites as described that proficient developers are using for their day-to-day work, but for starters, they need to research a lot to find trust-worthy sites. For that, we need to help them in finding solutions to their queries faster and with easy access. Other developers can contribute to the project to enhance functioning of the project.

6.3. Products/Services Description

We are aiming to provide following services,

* Giving solution to user’s queries
* If found, user is notified with the solution
* If not found, user will be notified via email when his query would be solved via experts.

1. Conclusion

Creation of knowledge base will not only benefit every students, or teacher but every tech geeks around the internet. Searching out for any solution from installation and new technology is troublesome when one doesn’t know what the internet has to offer on the back of each website. There is a dire need of platform that would provide solutions to all academic problems from its datasets and crawling features in short span. For this, we will be creating a simple web-app with easy UI that would only need user’s input and categories for optimizing search for web and datasets. Our datasets would only be the combination of recent tech-related articles, where user may search for different developing languages. For the crawling, we will be utilizing the authentic and trust-worthy sites rated by many developers. Both of these features will be used to enhance the solution quality to each queries. There is a possibility where internet and datasets fail to give out quality solution, in that case, that query would be forwarded to the experts. In short, user will need to input the query and it’ll be entertained with faster and accurate execution.

References

* <https://towardsdatascience.com/auto-generated-knowledge-graphs92ca99a81121>
* <https://www.hubspot.com/knowledge-base>
* <https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/>
* <https://neptune.ai/blog/web-scraping-and-knowledge-graphs-machine-learning>