Escort - Natural Language Processing Based University Students Guidance System

Career Guidance QA System for IT students

2022-179

Final Report - Individual

Parathan Thiyagalingam

IT19125176

B.Sc. (Hons) Degree in Information Technology

(Specialization in Software Engineering)

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2022

Escort - Natural Language Processing Based University Students Guidance System

Career Guidance QA System for IT students

2022-179
Final Report - Individual

Parathan Thiyagalingam
IT19125176

Supervisor
Ms.Hansi De Silva

Co-Supervisor
Mr. Dharshana Kasthurirathne

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

September 2022

DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic, or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

T.Parathan

…………………09.09.2022…………   …..…………..…………………………..…

(Date)   (Parathan T)

The above candidate is carrying out research for the undergraduate Dissertation under my supervision

………………………………………   ………..………………………………

(Date)   (Signature of the supervisor)

ABSTRACT

Career guidance for university students is an essential step. While the students are following their undergraduate degree, they should be aware of what they are learning, how they can improve their skills, and what resources they can browse/study to have in-depth knowledge. On the other hand, getting mentorship from industry people who have experience is also an added advantage for students to perform well during their internship as well as during their permanent employment. By considering these factors, a career guidance system is created for students to interact with the system and get suitable mentors to get mentorship along with a question-answering feature where students can query the system with their career guidance doubts and IT-related theoretical questions from their academic modules.

Keywords: Career Guidance, Question Answering, Career building, Graph Database

ACKNOWLEDGEMENT

My deepest gratitude should offer to our supervisor Ms. Archana Kugathasan who is the main person who gave great support to this research. Without her proper guidance and intensity, this research would not have been a success. Also, our co-supervisor Mr. Dharshana Kasthurirathna was the other person who guided us and helped us to achieve our research goals. Then I would like to offer my appreciative thanks to my research team for supporting me and finishing the research successfully.

Finally, I would like to acknowledge all the lecturers and colleagues of the Sri Lanka Institute of Information Technology who have supported us from the beginning of this research. Without any of their single guidance and suggestions, it would have been complicated to achieve our targets.

Table of Contents

DECLARATION iii

ABSTRACT iv

ACKNOWLEDGEMENT v

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATIONS x

1. Introduction 1

1.1. Background Literature 3

Background 3

Literature Review 4

1.2. Research Gap 5

2. Research Problem 6

3. Research Objectives 8

Main Objective 8

Sub Objective 8

4. Methodology 9

1. Retriever 10

2. Reader 11

3. Models for Reader pipeline 12

4. Data Collection 13

5. Document Passage Retriever and Model Build 13

FAISS Document Store 15

4.1. Commercialization aspects of the product 16

Software Specifications, Research Review, or Design Components 16

Budget and Budget Justification (if any) 16

5. Implementation & Testing 17

5.1 Implementation 17

5.2 Testing 18

6. Results & Discussion 25

6.1. Results 25

6.2. Research Findings 26

6.3. Discussion 27

7. Conclusion 29

References 30

Appendices 32

LIST OF TABLES

Table 1	Retriever Pipeline parameters
Table 2	Reader Pipeline parameters
Table 3	Test Case 1
Table 4	Test Case 2
Table 5	Test Case 3
Table 6	Test Case 4
Table 7	Test Case 5

LIST OF FIGURES

Figure 1 System Overview Diagram

Figure 2 Retriever pipelines Comparison

Figure 3 Annotated Documents List

Figure 4 Annotated Document with Question and Answer

Figure 5 DPR training data file

Figure 6 Front-end of career guidance QA page

Figure 7 Unit Testing of Button

Figure 8 Final Output Answer for a Question: “How to become a Software Engineer?”

Figure 9 Test data accuracy of DPR model built

LIST OF ABBREVIATIONS

QA	Question Answering
IT	Information Technology
SE	Software Engineering
DS	Data Science
UI	User Interface
SQL	Structured Query Language
NLP	Natural Language Processing
API	Application Programming Interface
EC2	Elastic Cloud Computing
AWS	Amazon Web Services

1. Introduction

Providing career guidance during undergraduate studies for students is an essential step taken by all universities and higher educational institutes. Mostly, universities have a career guidance unit. The core functions of a career guidance unit are to organize awareness programs among students about their studies and the demands of their studies worldwide, arrange group discussions, mock interviews, communication with industry experts, sharing knowledge and experience regarding future career opportunities. However, students' utilization of the career guidance unit is scarce. Most students are not aware of the existence of a career guidance unit in their institution. Reaching out to career guidance unit representatives is also a significant problem the students face. During the covid pandemic, students could not reach out to the career guidance unit and participate in programs physically. However, some institutions organize workshops and discussions with industry experts through online platforms such as Zoom and MS Teams. Those discussions were unavailable to students who could not participate or were hesitant to join such extensive discussions.

For career-decision making one should possess accuracy of self-knowledge and career information. The quality of the career information is a crucial factor. The career information should include relevant information about opportunities, occupations, characteristics of jobs, and market demand and supply [3]. The students should have the access to get quality career information, understand the available information, relate the information to their needs and situation, and should convert those into actions.

To ease students' access, a question-answering system regarding career guidance would be the best choice for many universities and institutions. QA is a research area that combines research from different fields with a common subject which are Information Retrieval (IR), Information Extraction (IE), and NLP. QA systems are different from search engines by the following 2 factors. (1) In QA query is the question where in search engine it a keyword is, (2) QA response with a specific answer to a question without providing a list of documents as search engines do.

A considerable number of applications exist regarding career guidance where students can query through web-based or mobile-based interaction. Those existing systems are not specific to a particular industry (IT/ SE / DS) and for the country where this application will solve. Through this application, students can query the system regarding their career doubts and get information regarding the latest findings about their academic-related modules.

1.1. Background Literature

Background

The Question-Answering system has been built for many closed and open domain use cases. A chatbot that answers questions from any domain is an open domain, whereas a chatbot that answers in a specific domain (E.g.: healthcare-related things) is known as a closed domain. Building an open domain QA system needs more data and has many training hours, whereas building a domain-specific QA system is much more valuable and timesaving than an open domain QA system. A QA system can be built by using different methods. For example, rule-based, context-based, using the knowledge base, and database approach.

Due to the advancement of technologies and internet evolution lots of data has been increased drastically. Wikipedia, Microsoft, Google, Apple, and Mozilla etc. companies are interested in building a knowledge base of their product queries as well as collaborating in many open-source projects in the field of NLP. The information from these companies can be used to build up a different type of real-world applications and decision-making solutions for humankind in day to-day life. The common mission of a QA system is to provide a relatable or appropriate and reasonable answer/ answers for the asked question. In most of the QA systems the traditional method is providing answers for the queries from a FAQ (Frequently Asked Questions) page of a webpage or a system. This is a kind of supervised learning method. But there is less amount of system that can understand the input query, pre-process to understand the semantic of what is the real question asked by the user and provides the answer. In the recent advancement in the field of state-of-the-art of NLP different research are being conducted, models build and published. The accuracy of the system that can understand a natural language sentence has increased a lot. Therefore, the scope for the NLP application in the QA system is increasing a lot. The significant fact of this kind of system is that due to the scaling of data, the system should be able to index the data, understand the context and retrieve the answer according to the input query. The optimal way of indexing needs a high-powered computing resource. However, currently the cloud computing and virtual machines provides a lot of opportunities to do the heavy weight computations.

Literature Review

A simple closed-domain question answering system was built for the baseball game during the 1960s’. This was the first attempt as a question-answering system [5].

A system for Question Answering called "YodaQA" was built based on unstructured answers' sources. The primary source for the system is DBpedia ontology and the Freebase RDF dump, and it uses multiple searching techniques such as full-text search, structured search, and document search [1].

SQL databases are prevalent for their usage. It has been one of the most used databases in the world. As these are widely used, using SQL as a database and querying answers for a question will provide more possibility of getting the answers. However, natural language needs to be converted to SQL queries to query the DB. In this approach [3], SQL and NoSQL databases can be used to query.

Answerbag is a question-answering website where users can ask their questions and get answers. The questions can include fact-checking, entertainment, or any other open-domain questions. Professional researchers and site members can also answer the questions on the website. This site can be considered an expert community system as well.

AskHermes is a medical related question answering system that can handle complex natural language queries and output the suitable answers as a single word or a paragraph in an extractive summary [10]. The structured domain-specific ontology is used to handle complex queries. The several dataset corpuses are collected and have been categorized into 12 main categories using SVM (Support Vector Machine). Document retrieval and passage retrieval methods are used by the system to extract suitable answer as output. After that, summarization technique is used to summarize the retrieved answer and returns to the user.

1.2. Research Gap

The different research that was conducted up to now are based on combining NLP techniques such as entity recognition, pos tagging, creating knowledge bases, and querying them with machine-understandable queries by constructing queries using NLP mapping methods. There is a lack of a system for education-based QA. Especially in the career-guidance system for university students.

Research related to career guidance using an expert system approach [6] was about analyzing the student's personal skills and helping them to choose the right vocational course for their college studies. But this doesn't help the students to get to know about their industry, culture, and latest trends followed in that field. Therefore, a system should be there to assist students to make the following things and prepare students themselves for the job.

2. Research Problem

- **Career guidance system in university and education institutions**:
  The career guidance in the universities and education institutions should be aware of what students are expected about their careers in the future as well as their internship opportunities. According to that, the career guidance unit needs to consider the needs of students and act according to them.

- **The students' perspective on career guidance:**
  80% of the students in educational institutes are not aware of career-guidance units and career-related information provided by the universities until the first half of their study life [8]. If the students could be able to get introduced and can able to find their own career-related information it could be a great chance for many students.

- **Latest technologies advancement among students**:
  Students have a great exposure to the latest technologies and almost every student are using them too. The impact of technologies in the current generation is higher. Therefore, the time students spend on digital devices is higher than before. And the career opportunities available through the internet are immense. Due to the covid pandemic, work-from-home options are becoming trendy. But it is always a doubt for many that students are grabbing those opportunities and performing well or just spending their time on digital devices as a waste of time.

- **Need for an expert system to be available for students anywhere at any time:**
  In the university system, career guidance is mostly a one-way communication where instructions and notices are published through the learning management system. Even though workshops and guidance are organized physically or online, most students are not aware of it and are hesitant to

participate in the programs. Therefore, an online platform that would solve the student's doubts about career guidance will be beneficial.

3. Research Objectives

Main Objective

- Assess the latest NLP techniques used in the building of a QA system
- Comparing each of those techniques and select a practically applicable method as well as an efficient implementation method
- Building a career-guidance QA system using NLP techniques for the IT field

Sub Objective

- Collect career guidance-related corpus (documents, web articles)
- Build up an interactive web UI for users to interact and query the system
- Dockerizing the app for ease of deployment

4. Methodology

The Figure illustrates the overall system architecture diagram of the career-guidance QA system.

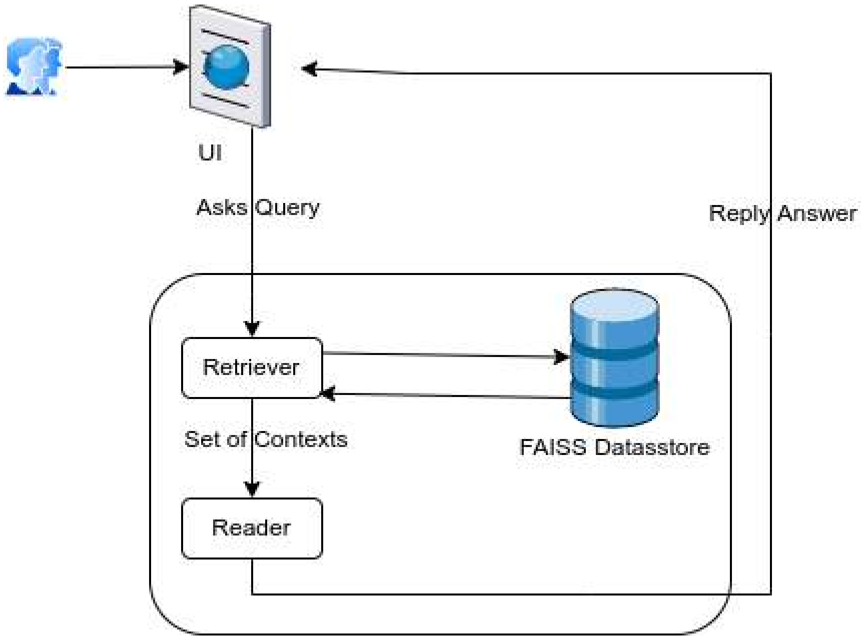


Figure 1: System Overview Diagram

The system consists of a front-end created by React and with the backend written in python. In the backend, the haystack library is used to build the question-answering component. This library consists of the necessary components needed to build the whole system itself. The haystack library contains the following pipeline nodes for the processing of the input text. These nodes can perform the steps like preprocessing, retrieving, or summarization of text along with routing to different pipeline nodes as well. Nodes can be considered as the building blocks that are able to switch from one to another node where one node’s output will be fed as input to the other node. In this research, the following pipelines are used from this library.

1. Retriever

The retriever pipeline node performs Document Retrieval by sweeping through a document store and returns a set of candidate documents that are relevant to the query. If the reader pipeline is also used along with the retriever node, then this will easily sift out irrelevant documents and saves the loading time to return the answer.

The retriever node is tightly coupled with a document store. In this research FAISS is the document store we used. The FAISS store will be specified when initializing the retriever. The retriever will receive the query as an input and checks the documents contained in the FAISS document store and pass the best-fitted documents to the next node.

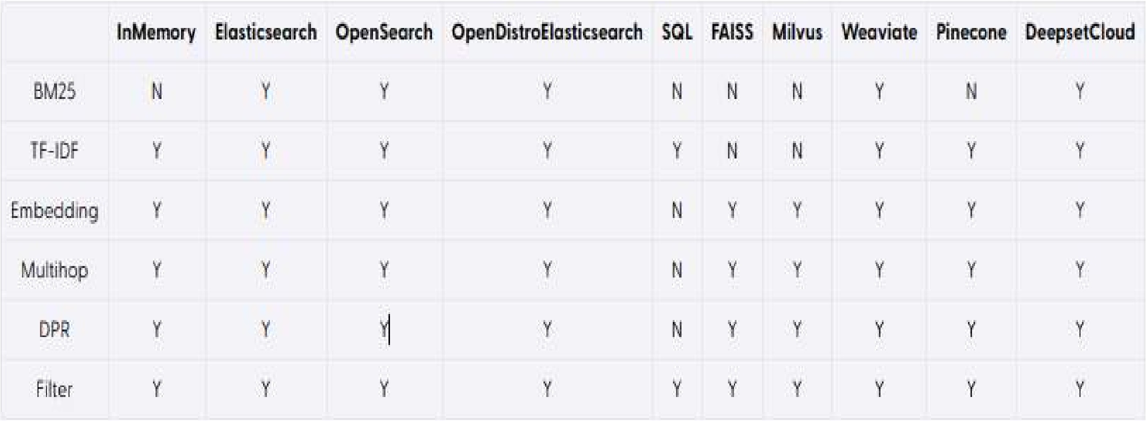
The following Table illustrates the parameters and output the retriever node will accept and output.

Position in a Pipeline	At the beginning of a querying Pipeline
Input	Query
Output	Documents
Classes	BM25Retriever ElasticsearchRetriever **DensePassageRetriever** TableTextRetriever EmbeddingRetriever TfidfRetriever ElasticsearchFilterOnlyRetriever

Table 1: Retriever Pipeline parameters

BM25Retriever, ElasticsearchRetriever, DensePassageRetriever, TableTextRetriever, EmbeddingRetriever, TfidfRetriever, ElasticsearchFilterOnlyRetriever are the different type of retrievers that can be used to implement the retriever model. For this research, DensePassageRetriever was used.

The Table: illustrated the document store capability.



	InMemory	Elasticsearch	OpenSearch	OpenDistroElasticsearch	SQL	FAISS	Milvus	Weaviate	Pinecone	DeepsetCloud
BM25	N	Y	Y	Y	N	N	N	Y	N	Y
TF-IDF	Y	Y	Y	Y	Y	N	N	Y	Y	Y
Embedding	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
Multihop	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
DPR	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
Filter	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Figure 2: Retriever pipelines Comparison

2. Reader

The reader pipeline node takes a question and a set of documents as input to the pipeline and returns the answer by selecting a suitable phrase within the documents. The reader can also be known as an open-domain QA system in Machine learning terminology.

This means the output documents from the retriever pipeline will be parsed as input to the retriever pipeline.

Position in a pipeline	Generally, after a Retriever
Input	Documents, Question
Output	Answer
Classes	FARMReader TransformersReader TableReader

Table 2: Reader Pipeline parameters

Followings are the advantages of using Reader pipeline in this research:

- Build on top of using Latest transformer-based language models
- Strong semantics
- Syntactic structure
- Uses the State-of-the-art in QA processes like SquAD (Stanford Question Answering Data) and Natural Questions

Reader pipeline in this library provides all the components of end-to-end, open-domain QA system. Followings are some of the components included:

- Model weights loading
- Tokenization
- Span prediction
- Candidate aggregation
- Embedding computation

3. Models for Reader pipeline

Hugging Face platform has different models for reader pipelines. Each model has a different level of strength and weakness.

Following BERT-based models are compatible with the reader pipeline:

- BERT
- RoBERTa
- ALBERT
- MiniLM
- XLM
- DistilBERT
- XLM-RoBERTa
- DeBERTa

4. Data Collection

The data for the research project was gathered from social media posts, medium articles, and different IT experts’ interviews and talks in a text format.

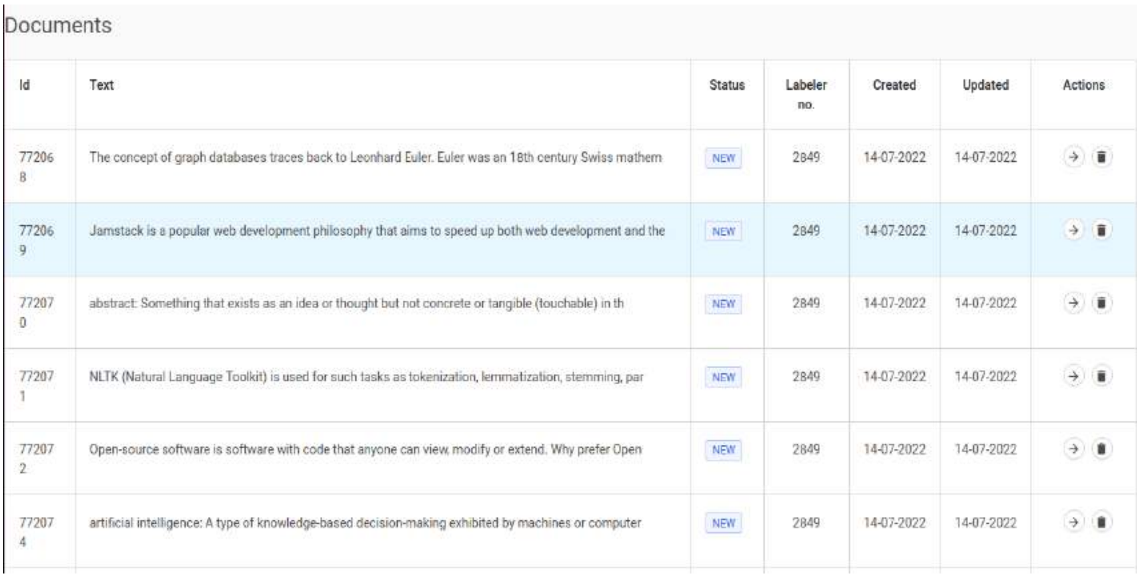
5. Document Passage Retriever and Model Build

a. Annotating the document

The collected data were annotated with questions and with the respective answers. The content of the dataset was preprocessed and removed special characters.

b. Train & Test dataset

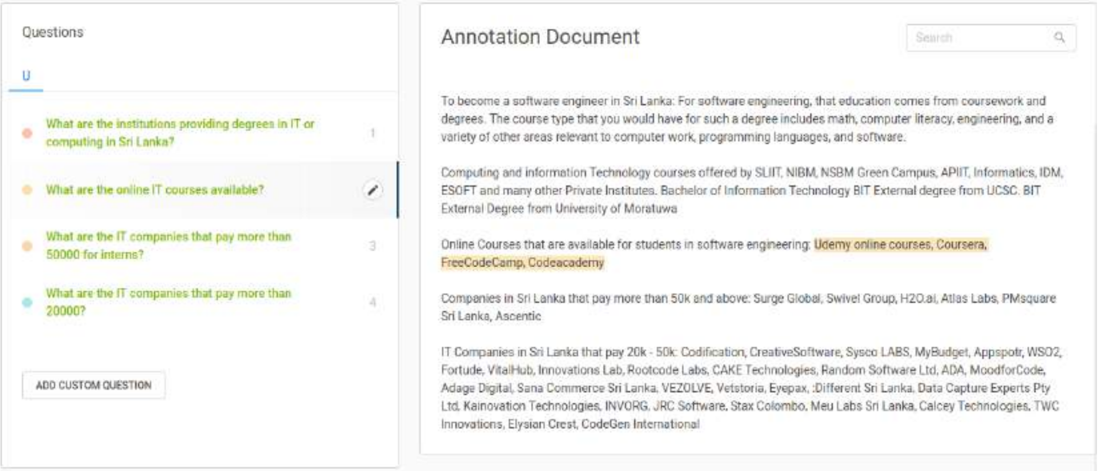
The 2 datasets were prepared for the testing purpose separately. Both test and train datasets were annotated with a question and its respective answer.



Documents

Id	Text	Status	Labeler no.	Created	Updated	Actions
772068	The concept of graph databases traces back to Leonhard Euler. Euler was an 18th century Swiss mathem	NEW	2849	14-07-2022	14-07-2022	→ 🗑
772069	Jamstack is a popular web development philosophy that aims to speed up both web development and the	NEW	2849	14-07-2022	14-07-2022	→ 🗑
772070	abstract: Something that exists as an idea or thought but not concrete or tangible (touchable) in th	NEW	2849	14-07-2022	14-07-2022	→ 🗑
772071	NLTK (Natural Language Toolkit) is used for such tasks as tokenization, lemmatization, stemming, par	NEW	2849	14-07-2022	14-07-2022	→ 🗑
772072	Open-source software is software with code that anyone can view, modify or extend. Why prefer Open	NEW	2849	14-07-2022	14-07-2022	→ 🗑
772074	artificial intelligence: A type of knowledge-based decision-making exhibited by machines or computer	NEW	2849	14-07-2022	14-07-2022	→ 🗑

Figure 3: Annotated Documents List



Questions

U

- What are the institutions providing degrees in IT or computing in Sri Lanka? 1
- What are the online IT courses available?
- What are the IT companies that pay more than 50000 for interns? 3
- What are the IT companies that pay more than 20000? 4

Annotation Document

To become a software engineer in Sri Lanka: For software engineering, that education comes from coursework and degrees. The course type that you would have for such a degree includes math, computer literacy, engineering, and a variety of other areas relevant to computer work, programming languages, and software.

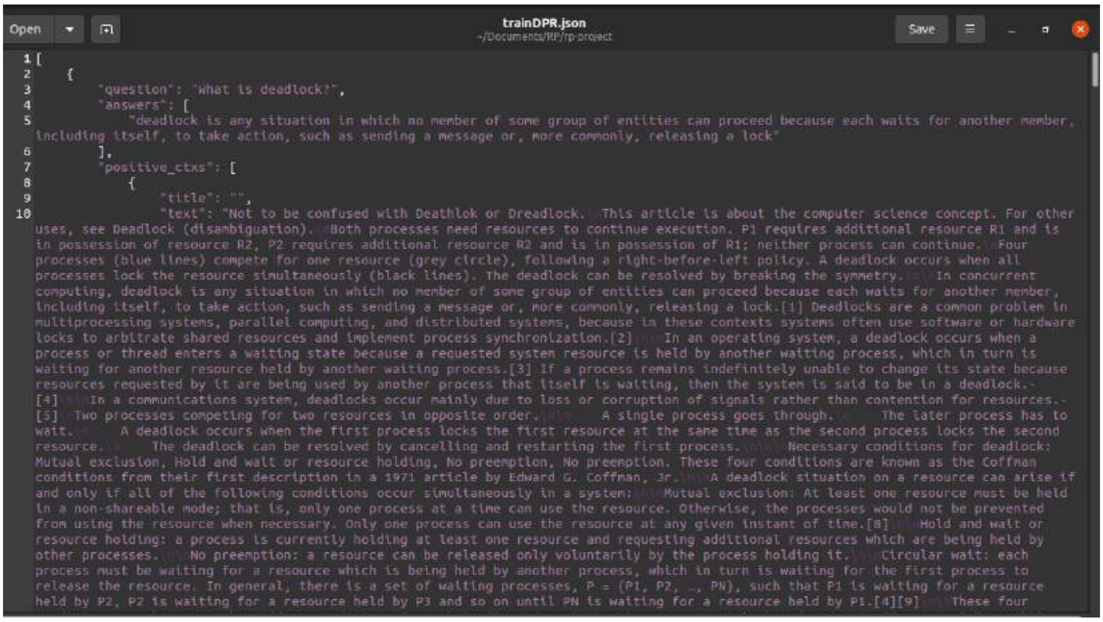
Computing and information Technology courses offered by SLIIT, NIBM, NSBM Green Campus, APIIT, Informatics, IDM, ESOFT and many other Private Institutes. Bachelor of Information Technology BIT External degree from UCSC. BIT External Degree from University of Moratuwa

Online Courses that are available for students in software engineering: Udemy online courses, Coursera, FreeCodeCamp, Codeacademy

Companies in Sri Lanka that pay more than 50k and above: Surge Global, Swivel Group, H2O.ai, Atlas Labs, PMsquare Sri Lanka, Ascentic

IT Companies in Sri Lanka that pay 20k - 50k: Codification, CreativeSoftware, Sysco LABS, MyBudget, Appspotr, WSO2, Fortude, VitalHub, Innovations Lab, Rootcode Labs, CAKE Technologies, Random Software Ltd, ADA, MoodforCode, Adage Digital, Sana Commerce Sri Lanka, VEZOLVE, Vetstoria, Eyepax, :Different Sri Lanka, Data Capture Experts Pty Ltd, Kainovation Technologies, INVORG, JRC Software, Stax Colombo, Meu Labs Sri Lanka, Calcey Technologies, TWC Innovations, Elysian Crest, CodeGen International

Figure 4: Annotated Document with Question and Answer



```
[
    {
        "question": "What is deadlock?",
        "answers": [
            "deadlock is any situation in which no member of some group of entities can proceed because each waits for another member,
including itself, to take action, such as sending a message or, more commonly, releasing a lock"
        ],
        "positive_ctxs": [
            {
                "title": "",
                "text": "Not to be confused with Deathlok or Dreadlock. This article is about the computer science concept. For other
uses, see Deadlock (disambiguation). Both processes need resources to continue execution. P1 requires additional resource R1 and is
in possession of resource R2, P2 requires additional resource R2 and is in possession of R1; neither process can continue. Four
processes (blue lines) compete for one resource (grey circle), following a right-before-left policy. A deadlock occurs when all
processes lock the resource simultaneously (black lines). The deadlock can be resolved by breaking the symmetry. In concurrent
computing, deadlock is any situation in which no member of some group of entities can proceed because each waits for another member,
including itself, to take action, such as sending a message or, more commonly, releasing a lock.[1] Deadlocks are a common problem in
multiprocessing systems, parallel computing, and distributed systems, because in these contexts systems often use software or hardware
locks to arbitrate shared resources and implement process synchronization.[2] In an operating system, a deadlock occurs when a
process or thread enters a waiting state because a requested system resource is held by another waiting process, which in turn is
waiting for another resource held by another waiting process.[3] If a process remains indefinitely unable to change its state because
resources requested by it are being used by another process that itself is waiting, then the system is said to be in a deadlock.-
[4] In a communications system, deadlocks occur mainly due to loss or corruption of signals rather than contention for resources.-
[5] Two processes competing for two resources in opposite order. A single process goes through. The later process has to
wait. A deadlock occurs when the first process locks the first resource at the same time as the second process locks the second
resource. The deadlock can be resolved by cancelling and restarting the first process. Necessary conditions for deadlock:
Mutual exclusion, Hold and wait or resource holding, No preemption, No preemption. These four conditions are known as the Coffman
conditions from their first description in a 1971 article by Edward G. Coffman, Jr. A deadlock situation on a resource can arise if
and only if all of the following conditions occur simultaneously in a system: Mutual exclusion: At least one resource must be held
in a non-shareable mode; that is, only one process at a time can use the resource. Otherwise, the processes would not be prevented
from using the resource when necessary. Only one process can use the resource at any given instant of time.[8] Hold and wait or
resource holding: a process is currently holding at least one resource and requesting additional resources which are being held by
other processes. No preemption: a resource can be released only voluntarily by the process holding it. Circular wait: each
process must be waiting for a resource which is being held by another process, which in turn is waiting for the first process to
release the resource. In general, there is a set of waiting processes, P = {P1, P2, …, PN}, such that P1 is waiting for a resource
held by P2, P2 is waiting for a resource held by P3 and so on until PN is waiting for a resource held by P1.[4][9] These four
```

Figure 5: DPR training data file

FAISS Document Store

Faiss is a library for efficient similarity search and clustering of dense vectors [9]. Based on the DPR model built, the data are indexed as vectors and stored. Whenever the query comes, the query will be converted to a vector and then similar vectors are retrieved from the stored document store.

4.1. Commercialization aspects of the product

This system has scope in every other field such as business, entrepreneurship, and other fields where state university and non-state university students could get the advantage of it. The system could be given to educational institutes as an education subscription, where institutions must pay based on monthly users/queries made to the system. This system has quite a great opportunity in the vision of commercializing for general purposes.

Software Specifications, Research Review, or Design Components

The current backend service for the QA system is hosted on AWS cloud. The current hardware specification is Ubuntu 20.04 LTS operating system with T2 instance memory optimized EC2 instances. The current memory of the EC2 instance is 32 GB with 8 core CPU (central processing unit).

Python 3 can be used for the backend deployment of models.

Haystack AI is a python library for building the QA system.

Budget and Budget Justification

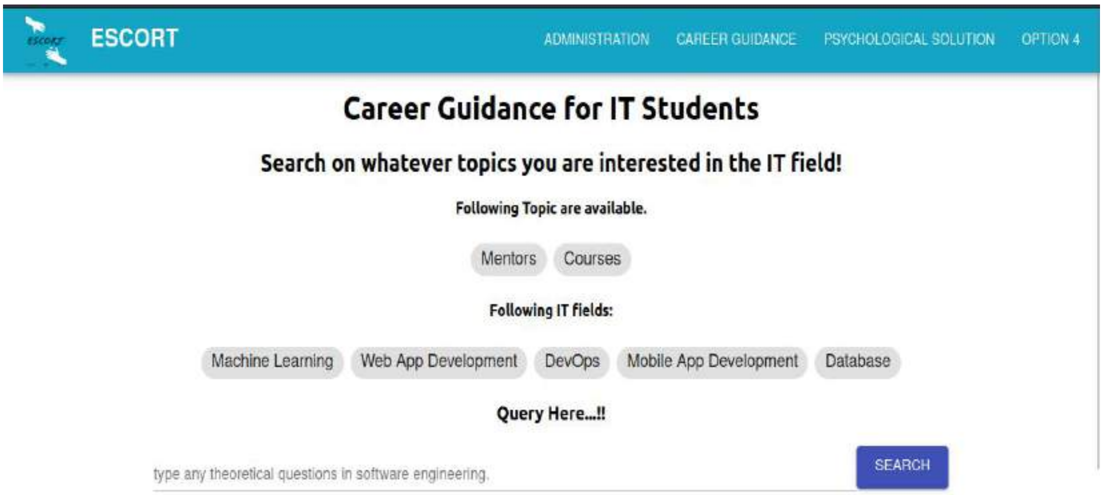
The backend service memory needs higher memory usage. Therefore, the development server is hosted on an AWS memory-optimized server.

5. Implementation & Testing

5.1 Implementation

The procedures to build up a deep learning model to answer the question based on the query was all done in this stage. The retriever model was built from the collected data. For that, each document's content was annotated with questions and their respective answers. After that, the Dense Passage Retriever model was built for this specific domain of career guidance QA. The python language was used in the whole model training. Using the built model, the dataset was fed into the datastore as vector embeddings. FAISS was used as a data store. Whenever a question comes to the system the question sentence will be vectorized and then compared with the vectors in the FAISS datastore. Similar contexts from the datastore will be retrieved and the answers will be sent back to the user.

A. Front-end user interaction for query question



Career Guidance for IT Students

Search on whatever topics you are interested in the IT field!

Following Topic are available.

Following IT fields:

Query Here...!!

Figure 6: Front-end of career guidance QA page

B. Dockerizing the backend

For ease of deployment, a docker image was built. Docker is a platform as a service (PaaS) that uses virtualization technology to provide software as containers. This docker image eases the method of deployment of the backend service in a short period of time. The only thing we need to do is pull the latest code from the version control to the server, build the docker image and start the docker container service.

When building the docker image, the DPR model will be rebuilt, and the latest dataset will be fed to the FAISS data store as vector embeddings.

5.2 Testing

1. Unit Testing

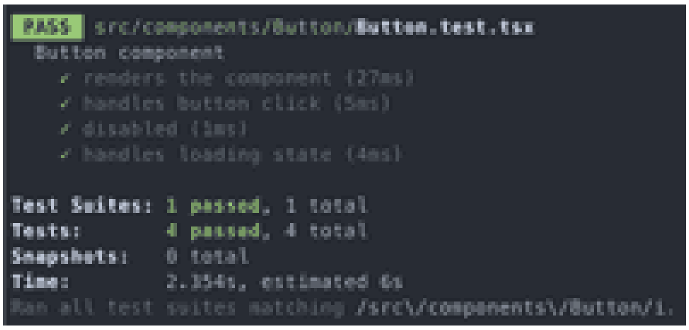
Unit testing is a best practice to reduce the bug by 40-80%. Following are the advantages of unit testing:

- Improves application architecture and code maintainability
- Leads to better API composability by focusing on developer experience before implementation
- Provides safety when implementing a new feature or refactoring existing code

Unit tests work especially for pure functions. Functions that are given the same input will provide the same output.

In the front end, the React components are separated and tested one by one.

- Wrapping components, passing props to the components
- Isolating application logic and business rules
- Isolate side effects using container components



```
 PASS  src/components/Button/Button.test.tsx
  Button component
    ✓ renders the component (27ms)
    ✓ handles button click (5ms)
    ✓ disabled (1ms)
    ✓ handles loading state (4ms)

Test Suites: 1 passed, 1 total
Tests:       4 passed, 4 total
Snapshots:   0 total
Time:        2.354s, estimated 6s
Ran all test suites matching /src\/components\/Button/i.
```

Figure 7: Unit Testing of Button

2. **Module Testing**

   Module testing is another type of testing that allows developers to test modules one by one. Most of the time due to time constraints the whole system is tested as one module at a time. If that worked well, then no problem. But, if one module failed or stopped responding to a request it is a problem to be considered. Module testing was done before combining each system together with one single frontend and backend code base.

3. **Integration Testing**

   There are 2 types of Integration testing

   I. The Big Bang Approach: In this testing, all units and modules are bounded together and tested as a whole. Since the whole ESCORT system is a slightly bigger system after implementing each system the source code was tested by making it altogether.

   II. The Incremental Approach: Each component was tested one by one and checked whether they return the correct response as expected.

4. System Testing

As the name suggests, after integrating the whole 4 systems together the testing was conducted. This ensures that each system has its own expected output for the input given

5. Test Cases

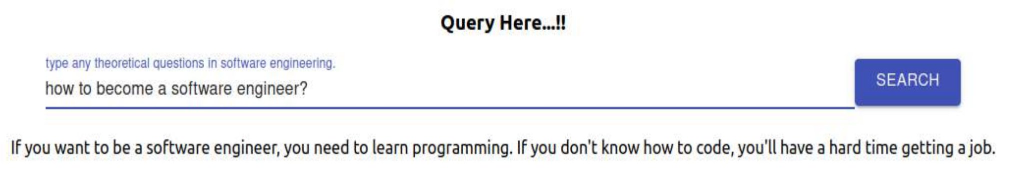
Test case	Test case 001
Description	User asks a question on “How to become a software engineer?”
Summary	When user asks a question on career-guidance related question on “How to become a software engineer?” The system replies with the correct answer
Pre-condition	-
Post-condition	User got the correct response
Test procedure	1. Go to Escort website 2. Choose Career Guidance from the top navigation bar 3. Type the question 4. Click ‘Search’ Button
Test input	“How to become a software engineer?”
Expected result	Any answer that is relevant to becoming a software engineer
Actual result	 Query Here...!! type any theoretical questions in software engineering. how to become a software engineer? SEARCH If you want to be a software engineer, you need to learn programming. If you don't know how to code, you'll have a hard time getting a job.
Test result	Pass

Table 3: Test Case 1

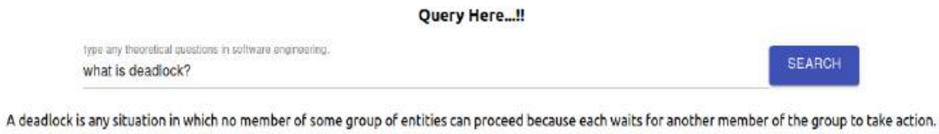
Test case	Test case 002
Description	User asks a question on “What is deadock?”
Summary	When user asks a question on any IT field related module question like “What is deadock?” The system replies with the correct answer
Pre-condition	-
Post-condition	User got the correct response
Test procedure	1. Go to Escort website 2. Choose Career Guidance from the top navigation bar 3. Type the question 4. Click ‘Search’ Button
Test input	“What is deadock?”
Expected result	Any answer that is relevant to deadlock situation
Actual result	 Query Here...!! type any theoretical questions in software engineering. SEARCH what is deadlock? A deadlock is any situation in which no member of some group of entities can proceed because each waits for another member of the group to take action.
Test result	Pass

Table 4: Test Case 2

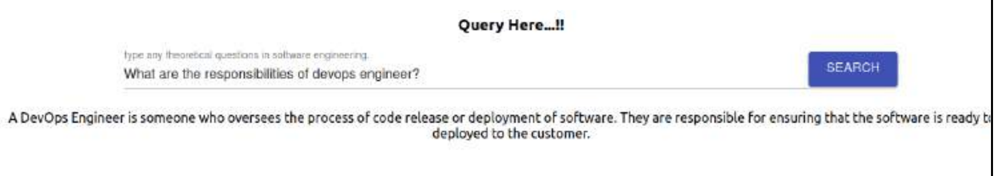
Test case	Test case 003
Description	User asks a question on “What are the responsibilities of devops engineer?”
Summary	When user asks a question on career-guidance related question on “What are the responsibilities of devops engineer?” The system replies with the correct answer
Pre-condition	-
Post-condition	User got the correct response
Test procedure	1. Go to Escort website 2. Choose Career Guidance from the top navigation bar 3. Type the question 4. Click ‘Search’ Button
Test input	“What are the responsibilities of devops engineer?”
Expected result	Any answer that is relevant to responsibilities of a devops engineer
Actual result	 **Query Here...!!** type any theoretical questions in software engineering. What are the responsibilities of devops engineer? A DevOps Engineer is someone who oversees the process of code release or deployment of software. They are responsible for ensuring that the software is ready t deployed to the customer.
Test result	Pass

Table 5: Test Case 3

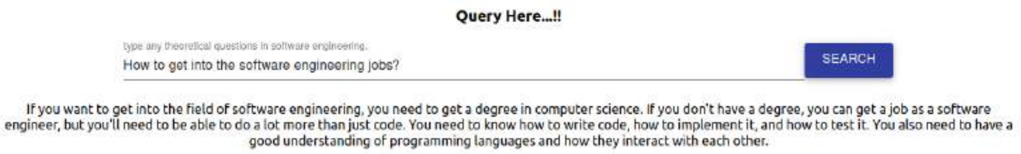
Test case	Test case 004
Description	User asks a question on “How to get software engineering jobs?”
Summary	When user asks a question on career-guidance related question on “How to get software engineering jobs?” The system replies with the correct answer
Pre-condition	-
Post-condition	User got the correct response
Test procedure	1. Go to Escort website 2. Choose Career Guidance from the top navigation bar 3. Type the question 4. Click ‘Search’ Button
Test input	“How to get software engineering jobs?”
Expected result	Any answer that is relevant to becoming a software engineer
Actual result	 Query Here...!! type any theoretical questions in software engineering. How to get into the software engineering jobs? If you want to get into the field of software engineering, you need to get a degree in computer science. If you don't have a degree, you can get a job as a software engineer, but you'll need to be able to do a lot more than just code. You need to know how to write code, how to implement it, and how to test it. You also need to have a good understanding of programming languages and how they interact with each other.
Test result	Pass

Table 6: Test Case 4

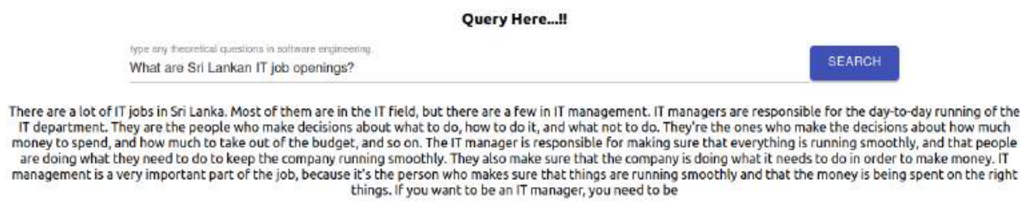
Test case	Test case 005
Description	User asks a question on “What are the Sri Lankan IT jobs openingns?”
Summary	When user asks a question on career-guidance related question on “How to become a software engineer?” The system replies with the correct answer
Pre-condition	-
Post-condition	User got the correct response
Test procedure	1. Go to Escort website 2. Choose Career Guidance from the top navigation bar 3. Type the question 4. Click ‘Search’ Button
Test input	“What are the Sri Lankan IT jobs openingns?”
Expected result	Any answer that is relevant to becoming a software engineer in Sri Lanka
Actual result	 Query Here...!! type any theoretical questions in software engineering SEARCH What are Sri Lankan IT job openings? There are a lot of IT jobs in Sri Lanka. Most of them are in the IT field, but there are a few in IT management. IT managers are responsible for the day-to-day running of the IT department. They are the people who make decisions about what to do, how to do it, and what not to do. They're the ones who make the decisions about how much money to spend, and how much to take out of the budget, and so on. The IT manager is responsible for making sure that everything is running smoothly, and that people are doing what they need to do to keep the company running smoothly. They also make sure that the company is doing what it needs to do in order to make money. IT management is a very important part of the job, because it's the person who makes sure that things are running smoothly and that the money is being spent on the right things. If you want to be an IT manager, you need to be
Test result	Pass

Table 7: Test Case 5

When noticing the Test case 005, it gives only 1 sentence as a relevant to the question asked other are related to IT manager job responsibilities. These needs more dataset so that, the model can be able to pick more reliant information from the documents by the reader model.

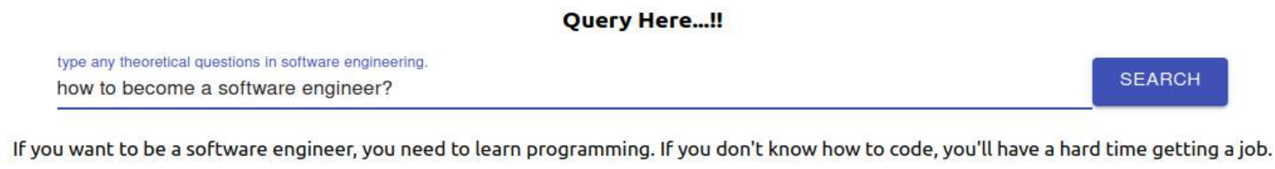
6. Results & Discussion

6.1. Results

Career guidance for students during their undergraduate studies is an inevitable step that both university and students must take care of. Unfortunately, neither the education institutions nor students take those steps carefully. Career guidance should need to be made public for each one. The students should be able to get access to the system and be able to clear their doubts. A manual system like mentors or representatives chatting over the internet will not be going to work anymore. Each person has got busy and has filled within a time frame. Therefore, considering these things in mind, a QA system for a career guidance system is very important.

But having a career guidance QA system for all the faculties, fields, and departments at the same time will be a tedious task. Therefore, as the initial step of the career guidance QA system for IT, students are the better way to implement it.

The implemented system was able to provide suitable answers to the questions asked. When a user enters the question and clicks the Search button from the front end, the query will be passed to the backend for processing and send back the answer to the questions



Query Here...!!

type any theoretical questions in software engineering.

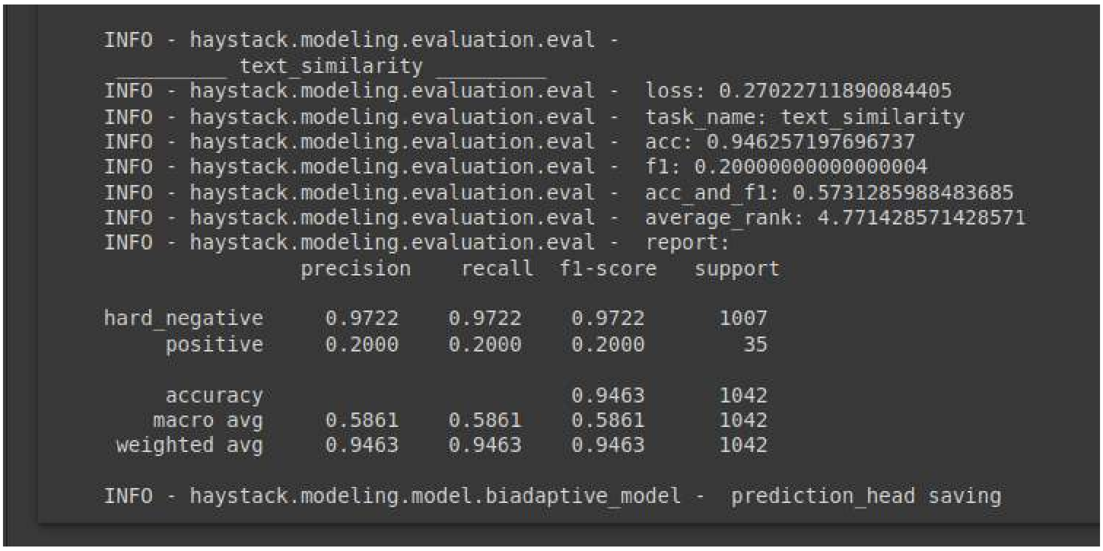
how to become a software engineer?

If you want to be a software engineer, you need to learn programming. If you don't know how to code, you'll have a hard time getting a job.

Figure 8: Final Output Answer for a Question: “How to become a Software Engineer?”

Experts can input the data and can train the retriever model to finetune to get the most relevant answer for the question in future implementation.

The retriever model was tested with the test data and the text similarity accuracy of 0.94% was obtained. Here we used Dense Passage Retriever to build the model.



```
INFO - haystack.modeling.evaluation.eval -
  _________ text_similarity _________
INFO - haystack.modeling.evaluation.eval -  loss: 0.27022711890084405
INFO - haystack.modeling.evaluation.eval -  task_name: text_similarity
INFO - haystack.modeling.evaluation.eval -  acc: 0.946257197696737
INFO - haystack.modeling.evaluation.eval -  f1: 0.20000000000000004
INFO - haystack.modeling.evaluation.eval -  acc_and_f1: 0.5731285988483685
INFO - haystack.modeling.evaluation.eval -  average_rank: 4.771428571428571
INFO - haystack.modeling.evaluation.eval -  report:
                precision    recall  f1-score   support

hard_negative     0.9722    0.9722    0.9722      1007
     positive     0.2000    0.2000    0.2000        35

     accuracy                         0.9463      1042
    macro avg     0.5861    0.5861    0.5861      1042
 weighted avg     0.9463    0.9463    0.9463      1042

INFO - haystack.modeling.model.biadaptive_model -  prediction_head saving
```

Figure 9: Test data accuracy of DPR model built

Accuracy is a type of performance measure that calculates the ration between the correctly predicted answers and the total answers that were tested. If the accuracy is high, then our model performs well. In here, 0.94625% accuracy has achieved, which is a fair enough to publish for the public usage. Even though, this model might get overfitted with the questions as these questions are mostly related to the career-guidance. Adding more data to the system and training the retriever model will increase the accuracy of retrieving answers.

6.2. Research Findings

Considering the need for a career-guidance QA system among university students to prepare for their industrial experience and professional life, we decided to develop a QA system for university students related to career guidance. As a first step, we decided to provide it for IT students. The goal is to provide an informative, useful, and concise platform for university students to build up their professional life by querying their doubts regarding career guidance.

The data required to build up this system were collected from different websites, blogs, YouTube interviews, and podcasts as a text file. NLP techniques and the latest libraries in NLP such as FAISS are used in this research project to provide efficient and reliable answers to students.

Collecting more data, feeding it to the system and training the retriever model will increase the possibility of getting more relevant answers from the system. Currently, the system includes not only on the career-guidance related questions, but also contains few IT course modules notes as well. Where the students can be able to query the questions on their modules. For, now the notes are mostly generalized questions from database, devops and front-end development.

6.3. Discussion

In the individuals, they will feel worried and concerned about their career, and job opportunities that are available to them and their future. From the young age of a child guidance of joining a school is a concern of its parents. For that decision-making, parents need to refer to the analysis report of the schools, and quality of education. Other than that, seeking advice from other people. Therefore, from childhood until death guidance is a significant aspect. Career guidance is important during undergraduate studies and for the working person as well. To choose a better career path one should have an accessible opportunity to get guidance.

As a subsystem of ESCORT, a QA system on career guidance was implemented. According to the results gained it provided satisfiable answers to the queries asked. The QA system on career guidance for Sri Lankan students is a new one. Therefore, it has provided a solution for the students who are currently undergraduates to define their path and how they could improve their skills to perform well in their industry experience.

The implemented system was tested on different devices as well. Since the front end is a browser-based responsive app, the end users were able to use the system without any hustle.

7. Conclusion

As university students, they should be able to have an opportunity to know about their career opportunities and industry-related doubts. The university itself is not able to provide a one-to-one session and clear students' doubts. A QA system related to career guidance will be a good approach for universities and educational institutions. This system itself provides an accuracy rate of 94% and the answers given were satisfiable in terms of the MVP. Through this system, students are able to possess their careers and are able to change their life.

References

[1] Baudiš, Petr, YodaQA: a modular question answering system pipeline. In POSTER 2015-19th International Student Conference on Electrical Engineering, 2015, September

[2] Prof. Pooja Malhotra, Yash Kapadia, Krishna Saboo, and Ankita Sarda, QUESTION-ANSWERING SYSTEM USING NATURAL LANGUAGE PROCESSING WITH NLIDB APPROACH. International Journal of Current Research Vol. 9, Issue, 09, pp.57575-57577, September 2017

[3] Isa Ado Abubakar, Career Guidance, Participation of Students and its Implication for Kano, Nigeria, The Malaysian Online Journal of Educational Science, 2018

[4] Shivani Singh, Nishtha Das, Rachel Michael, Dr. Poonam Tanwar, The Question Answering System Using NLP and AI, International Journal of Scientific & Engineering Research Volume 7, Issue 12, December-2016 ISSN 2229-5518

[5] Bert F. Green , Alice K. Wolf , Carol Chomsky , Kenneth Laughery, Baseball: an automatic question-answerer, In Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEEACM Computer Conference. ACM, New York, NY, USA, IRE-AIEE-ACM ’61 (Western), pages 219– 224

[6] Winston Ojenge, Lawrence Muchemi: Career Guidance Using Expert System Approach, 2008

[7] Ali Mohamed Nabil Allam, Mohamed Hassan Haggag, The Question Answering Systems: A Survey, International Journal of Research and Reviews in Information Sciences (IJRRIS) Vol. 2, No. 3, September 2012

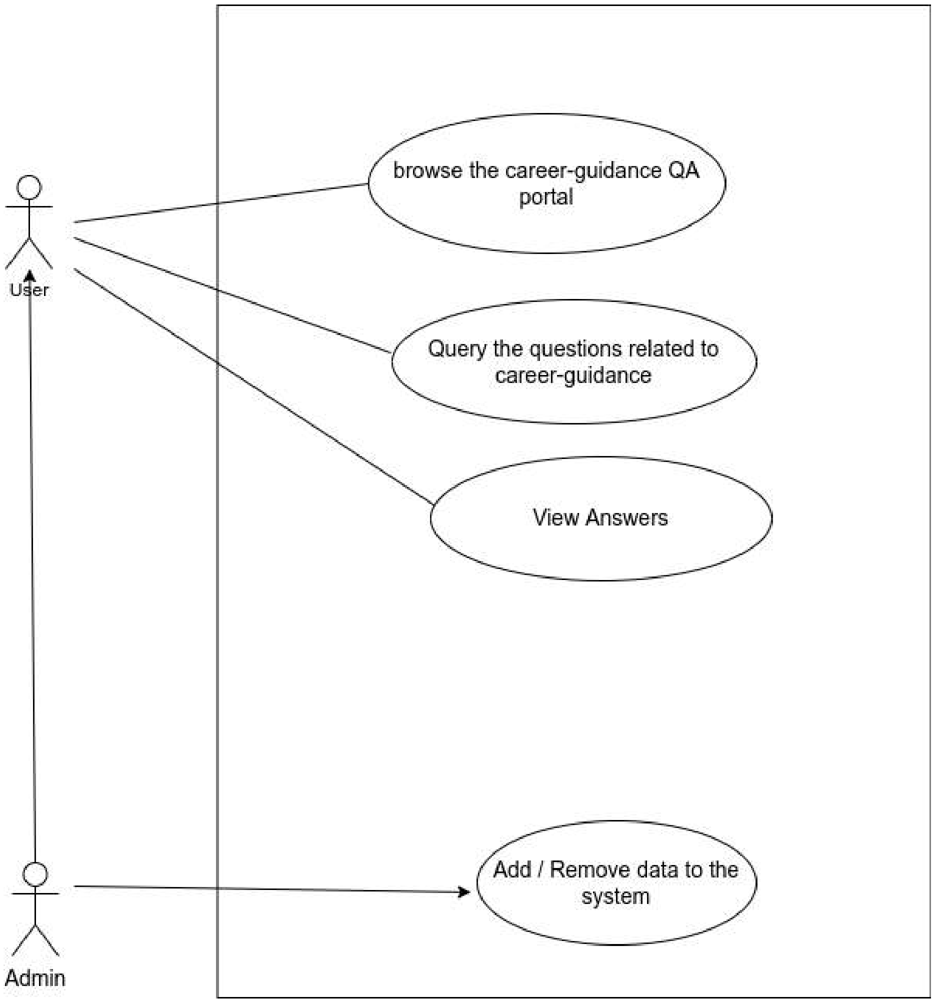
[8] Dr. Radhika Kapur, Career Guidance and Student Counseling, Research Gate

[9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih, Dense Passage Retrieval for Open-Domain Question Answering, 2020

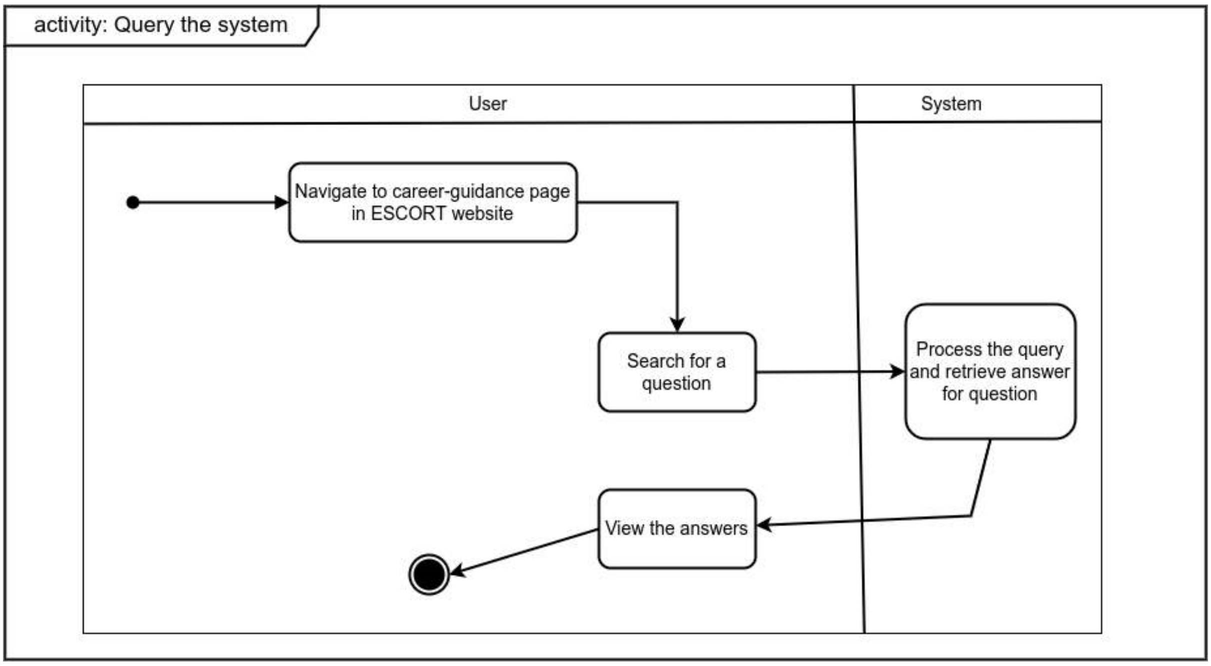
[10] Cao, YongGang, et al. "AskHERMES: An online question answering system for complex clinical questions." Journal of biomedical informatics 44.2 (2011): 277-288.

Appendices

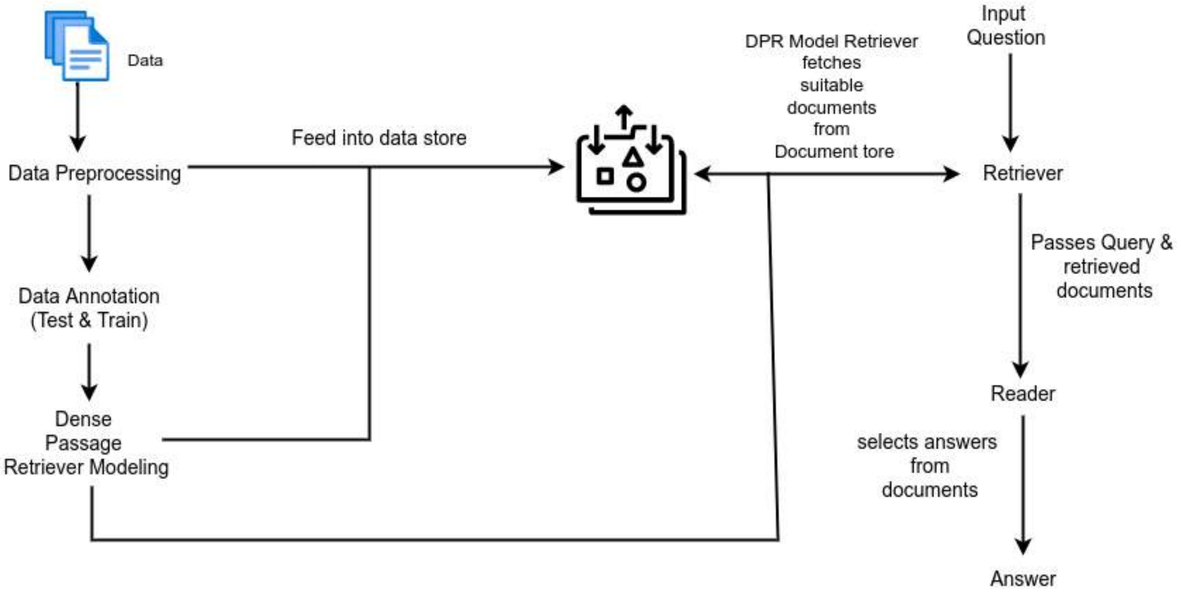
Appendix A: Use Case Diagram


browse the career-guidance QA portal
User
Query the questions related to career-guidance
View Answers
Add / Remove data to the system
Admin

Appendix: B: Activity Diagram


activity: Query the system
User
System
Navigate to career-guidance page in ESCORT website
Search for a question
Process the query and retrieve answer for question
View the answers

Appendix C: System Diagram


Data
Data Preprocessing
Data Annotation (Test & Train)
Dense Passage Retriever Modeling
Feed into data store
DPR Model Retriever fetches suitable documents from Document tore
Input Question
Retriever
Passes Query & retrieved documents
Reader
selects answers from documents
Answer