**DOCUMENT QUESTION ANSWERING SYSTEM USING DEEP LEARNING TECHNIQUES**

by

K.Eswar Gowtham 19BEC1135

A.V.N.M Hema Teja 19BEC1025

A project report submitted to

**Dr SUCHETHA M**

**SCHOOL OF ELECTRONICS AND COMMUNICATION**

**ENGINEERING**

in partial fulfilment of the requirements for the course of

**ECE1901 – TECHNICAL ANSWERS FOR REAL WORLD**

**PROBLEMS**

in

**B.Tech. ELECTRONICS AND COMMUNICATION**

**ENGINEERING**

****

**VIT UNIVERSITY, CHENNAI**

**Vandalur – Kelambakkam Road Chennai – 600127**

**APRIL 2022**

**BONAFIDE CERTIFICATE**

Certified that this project report entitled “**DOCUMENT QUESTION ANSWERING SYSTEM USING DEEP LEARNING TECHNIQUES”** is a bonafide work of **K. ESWAR GOWTHAM (19BEC1135)** and **A.V.N.M HEMATEJA (19BEC1025)** who carried out the project work under my supervision and guidance.

**Dr SUCHETHA M**

Associate Professor

School of Electronics Engineering (SENSE),

VIT University, Chennai

Chennai – 600127.

**ACKNOWLEDGEMENT**

We wish to express our sincere thanks and a deep sense of gratitude to our project guide, **Dr Suchetha M,** Associate Professor, School of Electronics Engineering, for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to **Dr Sivasubramanian. A**. V.T, Dean of the School of Electronics Engineering, VIT Chennai, for extending the facilities of the School towards our project and for her unstinting support.

We express our thanks to our Programme Chair **Dr Vetrivelan. P** for their support throughout the course of this project. We also take this opportunity to thank all the faculty of the School for their support and the wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

**ABSTRACT**

Due to covid, education has totally changed to online. So both students and the professors are highly using digitized documents like pdf and text files. Now we need an algorithm that does a semantic search over these documents. So that we can ask a question and get the appropriate answer related to the content and the context of the document.

In this project, we will be doing something similar to that which is a Document Question Answering system. Using this kind of system we can upload any document or enter the text that we need to comprehend and then we can ask any question related to the document as input. Then the appropriate answer is displayed as output. To finish this task we are Deep Learning algorithms such as transformer models.

| **S.NO** | **TABLE OF CONTENTS** | **PAGE NO.** |
| --- | --- | --- |
|  | INTRODUCTION  1.1 OBJECTIVES AND GOALS  1.2 BENEFITS | 6 |
|  | IMPLEMENTATION    2.1 BLOCK DIAGRAM | 6-8 |
|  | RESULTS | 8-12 |
|  | FUTURE WORK | 12-13 |
|  | REFERENCES | 13-14 |

**1. INTRODUCTION**

Question answering (QA) is a well-researched problem in NLP. In spite of being one of the oldest research areas, QA has applications in a wide variety of tasks, such as information retrieval and entity extraction. Recently, QA has also been used to develop dialog systems and chatbots designed to simulate human conversation. Traditionally, most of the research in this domain used a pipeline of conventional linguistically-based NLP techniques, such as parsing, part-of-speech tagging, and coreference resolution. Many of the state-of-the-art QA systems – for example, IBM Watson use these methods. However, with recent developments in deep learning, neural network models have shown promise for QA systems.

**1.1 OBJECTIVES AND GOALS**

To create a successful and completely working website for processing the given document and use deep learning algorithms and to find the corresponding answer for the required question asked from the document.

**1.2 BENEFITS**

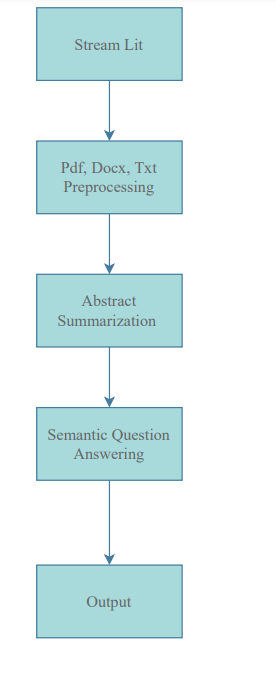
Question Answering provides the perfect solution to get valid and accurate answers to question user asked.

Saves time required to read the whole document.

Helpful for every individual working with digitized documents in their daily life.

**2. IMPLEMENTATION**

**2.1 BLOCK DIAGRAM**

****

**Streamlit:**

Streamlit is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as sci-kit learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc. Using Streamlit software we developed a website in which we incorporated our document question answering system.

**Document Preprocessing:**

In this step, we will take pdf or txt formatted documents as input from the website then we convert this pdf and docx file into an easily editable document format which is .txt using a technique called OCR (Optical Character Recognition)

We need OCR because it enables us to convert that frozen text into machine-readable data so that it is searchable. This means OCR enables you to find the specific information in your documents that can be copied and pasted for other uses.

**Abstract Summarization:**

It is the process of creating a condensed form of text document which maintains significant information and general meaning of the source text. Automatic text summarization becomes an important way of finding relevant information precisely in large text in a short time with little effort using sequence to transformer models.

Automatic summarization reduces reading time. When researching documents, summaries make the selection process easier. Automatic summarization improves the effectiveness of indexing. Automatic summarization algorithms are less biased than human summarizers.

**Semantic Question Answering:**

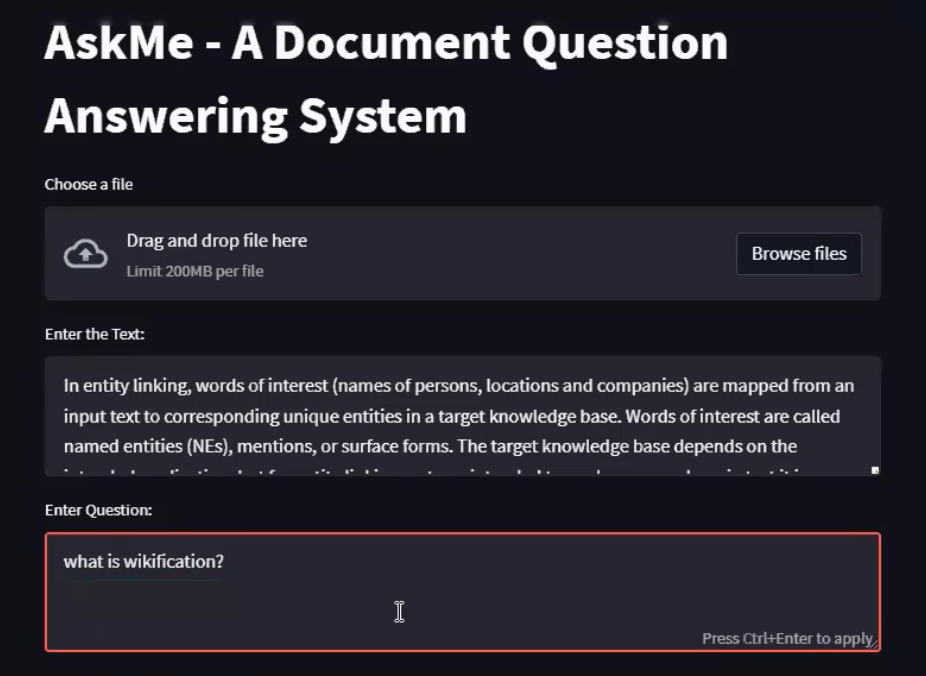
This is the most important task for the question-answering process. Here both the document and question are converted into vector format then question-related data will be checking the closest related data from the document with help of the long former neural network method and identify the appropriate answer according to the question.

Then the identified data will be in vector format which will be again converted into user-friendly text format and this will send as the output into the website from the backend.

**3. Results**

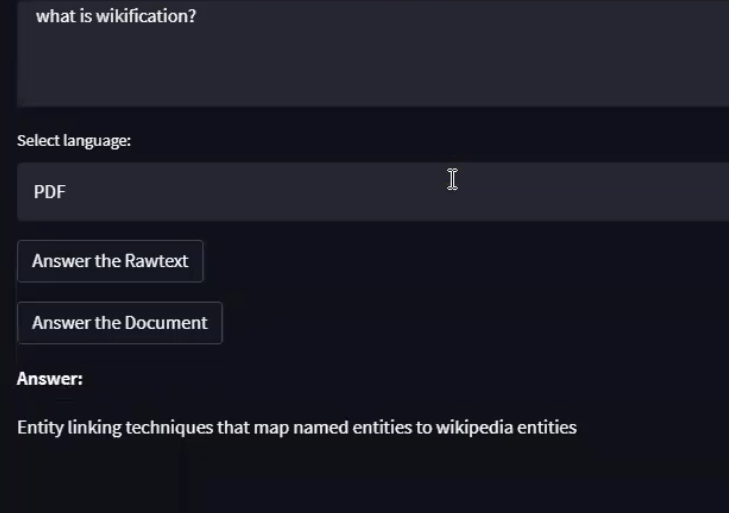
In our project, basically we can two types of input which are giving the data in the form of raw text or we can upload a pdf or txt format files from our local device.

**Input:**

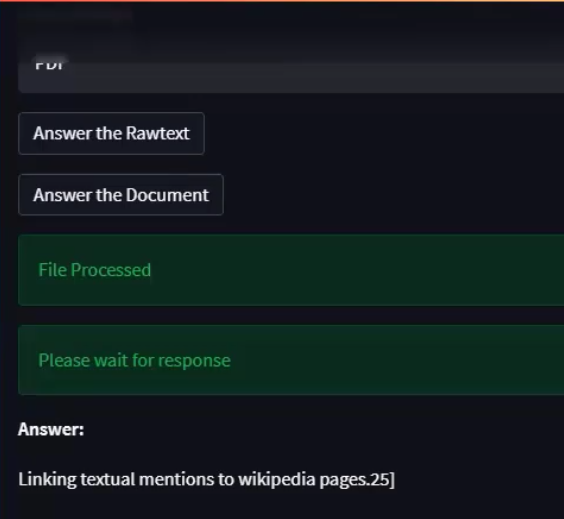
****

From the figure, we can see some raw text about Wikipedia entity linking is given as an input, then a question was asked about wikification related to wikification and the output given was correct by our Question Answering system as you can see below.

**Output:**

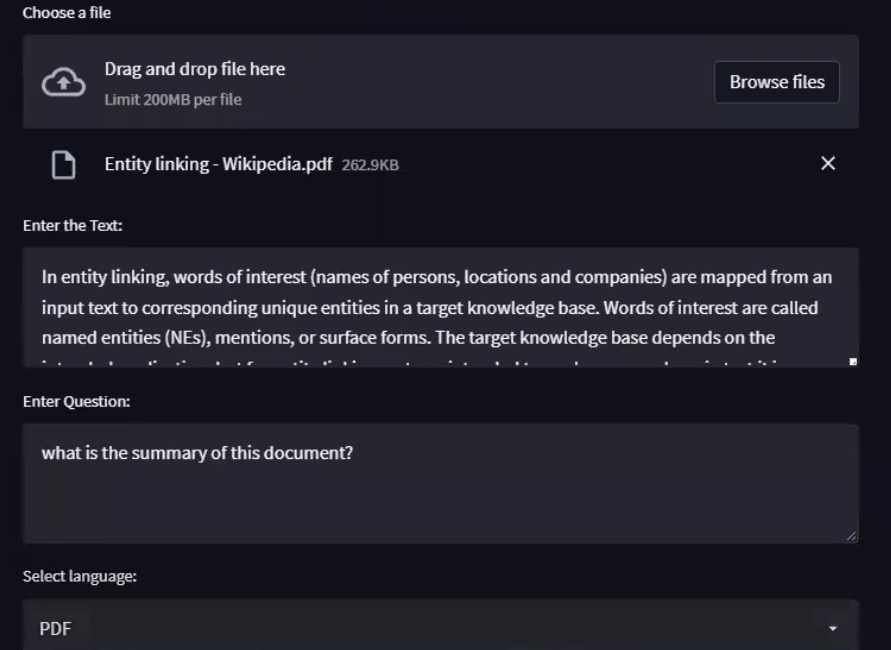


Now, we are the local pdf file as input related to the same topic we have before. Here we have to choose the required file path after clicking on the Browse files button then enter the question, finally select the pdf format in the dropdown, and the output given is an appropriate answer.



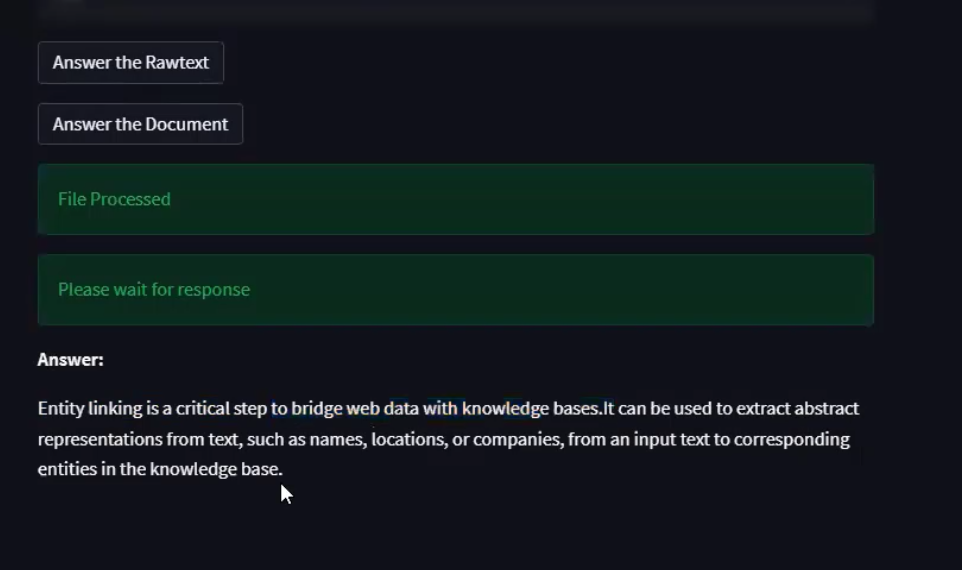
Our next step is answering the general questions like “what is the summary of the document?”, ”what can you infer?” etc. As we can see from the below screenshot the question asked is general for the given input document.

**Input:**



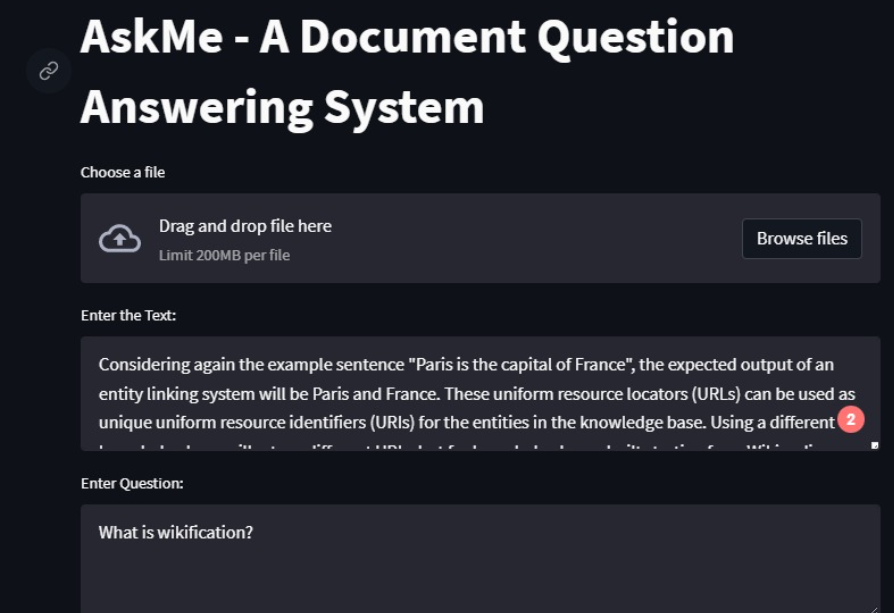
Our deep learning model also gives the appropriate output for these questions as we can from the below figure.

**Output:**

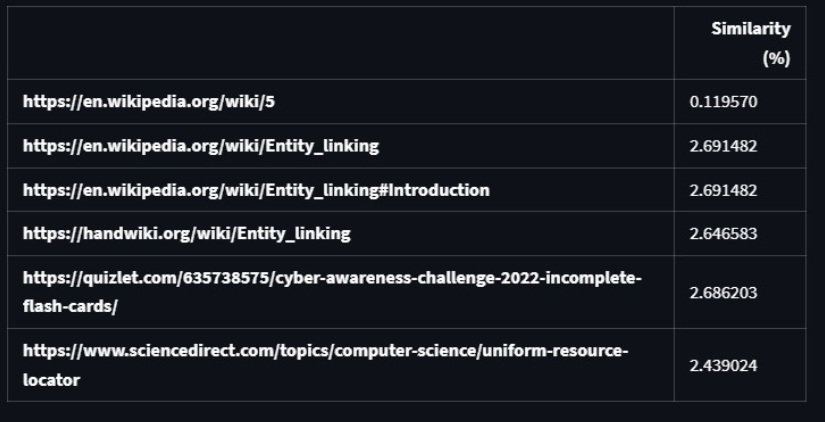


We have also added an additional feature to our website which plagiarism checker which shows the similarity index between the document and the available open web resources.

**Input:**



**Output:**



**4. Future Work:**

In the future, we can extend the document question answering from single to multiple and also to the internet in general. We can use different methods like TF-IDF, ELMO e.t.c for this purpose. But we need to increase the GPU power for faster computation, without which the latency is created.

We can also increase this into continuous plagiarism where the person will be evaluated based on the text he/she is typing dynamically which will reduce the computational power usage and reduces latency.

**5. References**

1. A Review of Question Answering Systems

<https://www.researchgate.net/publication/333627091_A_Review_of_Question_Answering_Systems>

1. Document Collection Visual Question Answering

<https://www.researchgate.net/publication/351221969_Document_Collection_Visual_Question_Answering>

1. A literature review on question answering techniques, paradigms and systems

<https://doi.org/10.1016/j.jksuci.2018.08.005>

1. A Unified Model for Document-Based Question Answering Based on Human-Like Reading Strategy

<https://ojs.aaai.org/index.php/AAAI/article/view/11316/11175>

1. A Question Answering System on Regulatory Documents

<https://ebooks.iospress.nl/pdf/doi/10.3233/978-1-61499-935-5-41>