Smart Education Chatbot – A Chatbot Based Question Answering System for Retrieving Answers from PDFs Using BERT

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

by

Prakash Mukesh Sewani (Roll No. 50)

Arman Sunil Budhrani (Roll No. 08)

Bhavesh Jagdish Thadhani (Roll No. 60)

Supervisor:

Prof. Sandeep More

Department of Computer Engineering

Watumull Institute of Electronics Engineering and Computer Technology
Ulhasnagar
2021 – 2022

CERTIFICATE

This is to certify that the project entitled **“Smart Education Chatbot – A Chatbot Based Question Answering System for Retrieving Answers from PDFs Using BERT”** is a bonafide work of

Prakash Mukesh Sewani (Roll No. 50)

Arman Sunil Budhrani (Roll No. 08)

Bhavesh Jagdish Thadhani (Roll No. 60)

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Undergraduate** in **Bachelor of Computer Engineering**.

Prof. Sandeep More Supervisor/Guide	Co-Supervisor/Guide
Prof. Dhananjay Raut Head of Department	Dr. Sunita Sharma Principal

Project Report Approval for B. E.

This project report entitled "***Smart Education Chatbot – A Chatbot Based Question Answering System for Retrieving Answers from PDFs Using BERT"*** by

Prakash Mukesh Sewani (Roll No. 50)

Arman Sunil Budhrani (Roll No. 08)

Bhavesh Jagdish Thadhani (Roll No. 60)

is approved for the degree of **Undergraduate** in **Bachelor of Computer Engineering**.

Examiners

1.--

2.--

Date: /   /2022

Place: Ulhasnagar – Thane

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Prakash Mukesh Sewani (Roll No. 50): ------------------

Arman Sunil Budhrani (Roll No. 08): -------------------

Bhavesh Jagdish Thadhani (Roll No. 60): -------------------

Date: / / 2022

INDEX

Contents Page No.

Table of Contents

Chapter 1 - Introduction

Abstract:

Nowadays the use of Chatbots is very popular in a large scale of applications especially in systems that provide an intelligence support to the user. The use of Chatbots has evolved rapidly in numerous fields in recent years, including Marketing, Supporting Systems, Education, Health Care, Cultural Heritage, and Entertainment. In fact, to speed up the assistance, in many cases, these systems are equipped with Chatbots that can interpret the user questions and provide the right answers, in a fast and correct way. This project presents the realization of a prototype of a Chatbot in educational domain: it has been developed a system to provide support to university students on some courses. The initial purpose has focused on the design of the specific architecture, model to manage communication and furnish the right answers to the student. For this aim, it has been realized a system that can detect the questions and thanks to the use of natural language processing techniques and the ontologies of domain, gives the answers to student. In the proposed system, we present a new way of answering queries (questions) asked by users. The proposed system identifies the user context which triggers the particular intent for a response. Since it is responding dynamically, a desired answer will be fetched for the user. The proposed system utilizes a popular Question Answering algorithm by Google known as BERT. It is a context-based question answering algorithm that tokenizes given queries and generates using modern NLP techniques. The proposed system also utilizes the concepts of context identification and web scraping.

Problem Statement:

A student has to visit various websites in order to complete their assignments, projects, experiments etc. and filter out the content that is needed to be written in the respective papers. We aim at centralizing the workload of students to a singular website (our website) to accomplish these tasks. This will help reduce the stress on students.

The number of students per teacher every year is also increasing tremendously. This decreases the doubt solving efficiency of the teacher since the teacher cannot be present for every student at all times. This also decreases the learning capability of the children. With our project, we also aim to reduce the workload of teachers.

Our Chatbot is built to handle various queries of students ranging from straight forward answers to brief answers. We understand that more than one answers can be correct for a given question, our Chatbot hence specifically asks what kind of answer would the user prefer, short straight to the point answer or a brief one.

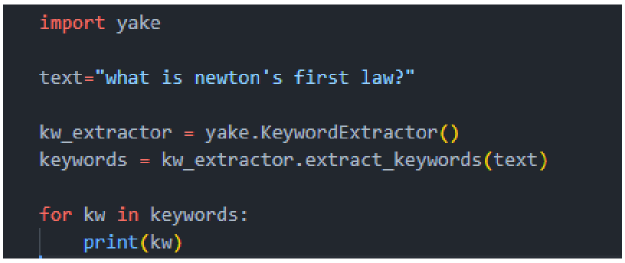
Scope of the Project:

Main scope of this project is to attract students of Standard 11th and 12th CBSE Boards to ask their queries and doubts regarding class sessions. Our website also will reduce the workload of students by fetching web scrapped results from websites as per the user’s requirement.

Chapter 2 - Review of Literature

YAKE!:

YAKE! (Yet Another Keyword Extractor) is a light-weight unsupervised automatic keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text. The system does not need to be trained on a particular set of documents, neither it depends on dictionaries, external-corpus, size of the text, language or domain.



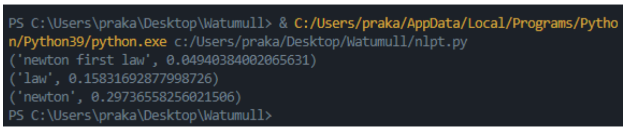
```
import yake

text="what is newton's first law?"

kw_extractor = yake.KeywordExtractor()
keywords = kw_extractor.extract_keywords(text)

for kw in keywords:
    print(kw)
```

Fig.1.1 Demonstration of YAKE!



```
PS C:\Users\praka\Desktop\Watumull> & C:/Users/praka/AppData/Local/Programs/Pytho
n/Python39/python.exe c:/Users/praka/Desktop/Watumull/nlpt.py
('newton first law', 0.04940384002065631)
('law', 0.15831692877998726)
('newton', 0.29736558256021506)
PS C:\Users\praka\Desktop\Watumull>
```

Fig.1.2 Demonstration of YAKE!

As shown in Fig.1.1, code snippet of YAKE! Algorithm is provided with its output in Fig.1.2.

BERT:

BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. In NLP, this process is called attention.

BERT model is designed in such a way that it can read text from both directions, i.e., from left-to-right and right-to-left. Using this bidirectional capability, BERT is pre-trained on two different but related NLP tasks: Masked Language Modelling and Next Sentence Prediction

The objective of Masked Language Model (MLM) training is to hide a word in a sentence and then have the program predict what word has been hidden (masked) based on the hidden word's context. The objective of Next Sentence Prediction training is to have the program predict whether two given sentences have a logical, sequential connection or whether their relationship is simply random.

For our project we are utilizing the BERT Next Sentence Prediction Module to answer user queries.

Next Sentence Prediction:

In the BERT training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. The assumption is that the random sentence will be disconnected from the first sentence.

To help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model:

1. A [CLS] token is inserted at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.
2. A sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2.

3. A positional embedding is added to each token to indicate its position in the sequence. The concept and implementation of positional embedding are presented in the Transformer paper.

To predict if the second sentence is indeed connected to the first, the following steps are performed:

1. The entire input sequence goes through the Transformer model.

2. The output of the [CLS] token is transformed into a 2×1 shaped vector, using a simple classification layer (learned matrices of weights and biases).

3. Calculating the probability of IsNextSequence with softmax.

Review Existing System:

The state of the research on chatbot applications in the education sector was described with according to the reviewed articles. A review of the studies selected for this survey found that chatbots were used in a variety of ways, including teaching and learning, administration, evaluation, advice and research and development, for educational purposes. The introduction of learning pedagogy as the chatbot system in education has personalized online learning and made student learning materials accessible anywhere and anytime. According to their studies, chatbots are good technological innovations that can improve student engagement, cognitive acquisition, and performance.

Today's chatbots typically use two types of models to generate responses, a retrieval-based model and a generation-based model. An on-demand model is a common choice of chatbot because it is easier to develop and maintain, e.g. For example, this model is essentially a matching process between the input question and an output answer. We can use either cosine similarity or a trained CNN, LSTM model to calculate the probability of a match between a question and an answer. All responses from bots are maintained and stored in databases. When an input question is received, this model calculates the probability of a match for the input question and for all responses in the database. The higher the probability of a match, the more likely the correct answer is. Google's email autoresponder suggestion system is based on, the same type of on-demand model. This model is ideal for a quality control system that needs precise and exact answers, because the answers are stored in databases. The quality of responses can also be guaranteed. However, this Model suffers from poor response coverage because natural language is complex. You cannot answer a question or enter a phrase that is not included in the predefined answers. Creating a database containing all kinds of topics, responses and answers is time consuming, difficult to maintain, and difficult to cover all possible responses in natural language. Unlike a retrieval-based model, a generation-based model does not store any predefined responses in a database. Generates a response through the model itself so that it can be used in open or wide coverage applications. The most commonly used generation-based machine learning model is sequence-by-sequence models. Sequence-by-sequence models, known as seq2seq, consist of an encoder and a decoder based on a neural network model. In the training state, the encoder receives word embedding vectors from the question mark string as its input, and decoder receives the word embedding vectors from the response string as its output. By coupling the encoder vector output to the input of the decoder, we can train the model to generate responses through its trained decoder. The retrieval-based model focuses on improving the learning efficiency of E-Learning and the other everyday tasks of a personal assistant.

One of the approaches to this project is [1] paper, in which Farhan M. et al. using a web bot in an e-learning platform, to address the lack of real-time responses for the students. In fact, when a student asks a question on e-learning platform the teacher could answer at a later stage. If there are more students and more questions, this delay increases. Web bot is a web-based Chatbot that predicts future events based on keywords entered on the Internet. In this work Pandora is used, a bot that stores the questions and answers it on XML style language i.e., Artificial Intelligence Markup Language (AIML). This bot is trained with a series of questions and answers: when it cannot provide a response to a question, a human user is responsible for responding. In the last recent years some interesting research works can be found. This was an interesting approach to an Educational Chatbot as it utilised the essentials of Machine as well as Human together.

In [2] paper, Satu S., Chatbot is called Tutorbot because it is functionality backing of didactics done in eLearning environments. It contains some features as natural language management, presentation of contents, and interaction with search engine. Besides, e-learning platforms work is linked to indispensable services to web service. This project utilises various Educational Website APIs, search engines and backend data to provide an appropriate answer to the students.

In [3] paper, Muhammad Rana, while developing EagleBot, categorized user queries into three parts as Unstructured QA, Structured QA, FAQs using Google's Dialogflow. This enabled him to easily fetch answers for the user. For Structured QAs he usually fetched the queries from the data stored in databases. For Unstructured QAs, Muhammad Rana used the information present with the regarding query and web scrapped some results as well and fed it directly to the BERT Algorithm along with user query. This returned the appropriate result. For FAQ QAs, he web-scrapped his University Website for similar topics of discussions and provided user with the links to the forums.

Chapter 3 - Previous Works

In one study, an android application was presented containing education chat-bot designed for visually impaired people (Kumar et al., 2016). The application can be launched with Google Voice Search and is used by asking questions in spoken natural language. The application then converts it into text and run the query against the AIML database or, if no pre-saved answer was found, against the Wikipedia API. Though the application of this chatbot is noble, the retrieval architecture is a simple rule-based one.

Jill Watson, the Georgia Tech teaching assistant chatbot demonstrated the strong viability of chatbots in the educational domain (Goel and Polepeddi, 2016). Jill Watson showed promise as an alternative to teachers in the near future. Georgia Tech Computer Science Professor Dr. Ashok Goel built this chatbot to help students on their assignment related questions in one of his Artificial Intelligence courses. JW1 (Jill Watson version 1) was built using IBM Watson APIs. JW1 had a memory of question answer pairs from previous semesters organized into categories of questions.

Aforementioned chatbot systems work for a very specific domain and don't deal with the whole university domain. As of now, as to the best of our knowledge, there is no such integrated system for answering all types of questions asked in the university domain

Chapter 4 - Limitations

Limitations of Existing Systems:

Repetition: Machines are trained to provide standard answers to user queries. In a situation when a user does not find the answer satisfactory and rephrases the question, the machine still provides the same answer. This is a give-away to the user that they are in fact interacting with a machine and not a live human being. This can prove irritating and prevent users from proceeding any further.

They Can't Make Decisions: Another limitation of chatbots is that they lack decision-making. They don't have the right know-how to differentiate between the good and the bad.

Maintenance and Monitoring: Monitoring ensures that the input and output data from the chatbot are correct and that the system`s operation conforms to the design goals. The accuracy of the information provided by the bot is determined by the input data. You should also consider whether adding new data will interfere with the search experience for existing data. The more data the bot has to process, the longer the search will take. Building a chatbot system is an ongoing process that requires constant monitoring and maintenance, which can be difficult.

Technologies implemented in [1] such as NLP, AIML to provide an appropriate response to the user are heavy. They require too much resources and will not function properly when more than one user is using the bot simultaneously. To overcome this problem, individual instances of the server should be created for each individual user, hence leading to more resources being consumed.

Limitations of Proposed System:

Limitations of PYPDF2:

1. PDF Parsing/Reading Algorithms utilize PDFs that already contain text in them. They cannot read images, pure text must be present in the PDF files
2. As PDF Parsing/Reading Algorithms cannot read images, important diagrams are often skipped which will result in return of misinformation.
3. Another drawback of using such algorithms is that when faced with an equation in the PDF, if proper encoding is not used, the equation will not be represented properly. If encoding is used to solve this error, text errors might occur while dealing with special characters such as " ' ", " ^ "and so on.

Limitations of BERT Algorithm:

BERT algorithm utilizes context for answering user queries. The context provided to the algorithm are faced with certain rules. BERT uses word-piece tokenization. So, when some of the words are not in the vocabulary, it splits the words to its word pieces. For example: if the word playing is not in the vocabulary, it can split down to play, ##ing. This increases the number of tokens in a given sentence after tokenization. Max number of tokens that is algorithm can handle is 512. This is more than enough for question answering but sometimes the context itself is not enough for the answer which leads to improper responses.

Chapter 5 - System Design

Proposed System:

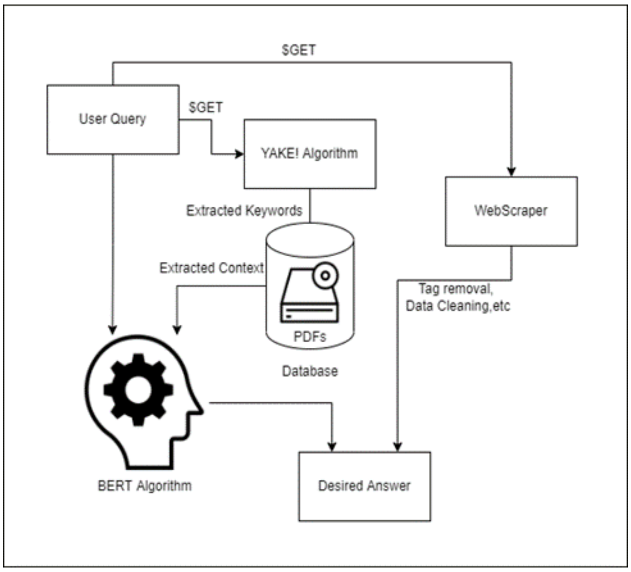


Fig.2 Proposed System Architecture

Proposed System consists of 4 major modules, namely, YAKE! Algorithm, Web scrapper, PDF Context Extraction, BERT Algorithm. User query is first passed on to the YAKE! Algorithm which extracts the necessary keywords for the PDF Context Extraction Module. Which in turn passes the User Query and acquired Context to BERT Algorithm. If the Answer does not match the user expectations, the user query is passed on to the web-scrapping algorithm which fetches the desired output.

System Architecture:

Technologies Used:

1. Python
2. HTML/CSS
3. JavaScript
4. Flask
5. Transformers
6. BERT
7. YAKE!
8. Web Scrapping
9. Wikipedia

Current System:

CPU: AMD Ryzen 5 3450H.

GPU: Nvidia GTX 1650Ti.

RAM: 8Gb.

Storage: 256 Gb.

Minimum System Requirements:

CPU: AMD 3rd Gen Duo Core or Intel Equivalent and Above.

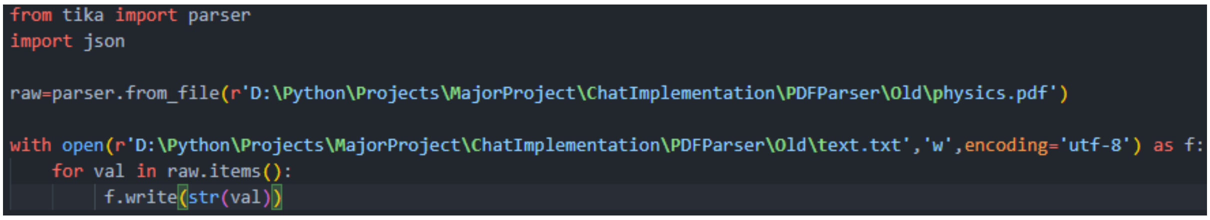
GPU: Not Required.

RAM: Minimum 3 Gb free.

Internet Connection: Not Mandatory.

PDF Reader Module:

We are using PyPDF2 to parse through PDF documents containing Text. This is achieved through Tika Library that contains the PyPDF2 parser.



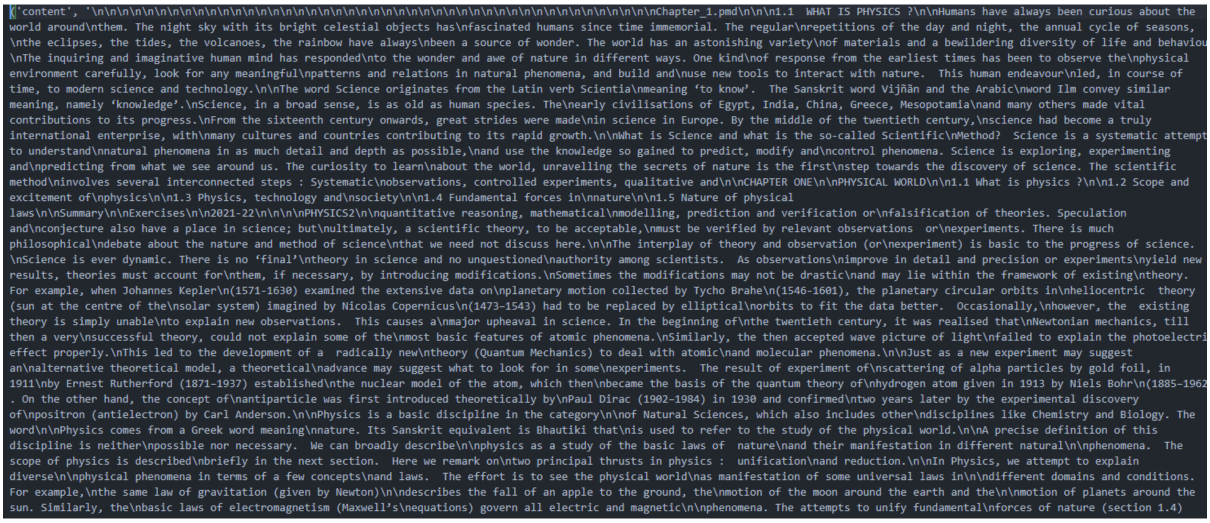
```
from tika import parser
import json

raw=parser.from_file(r'D:\Python\Projects\MajorProject\ChatImplementation\PDFParser\Old\physics.pdf')

with open(r'D:\Python\Projects\MajorProject\ChatImplementation\PDFParser\Old\text.txt','w',encoding='utf-8') as f:
    for val in raw.items():
        f.write(str(val))
```

Fig.3.1 PyPDF2 Demonstration

This module reads the contents of the PDF and stores it in a text file named text.txt. This file is then later used for further parsing and for context identification.



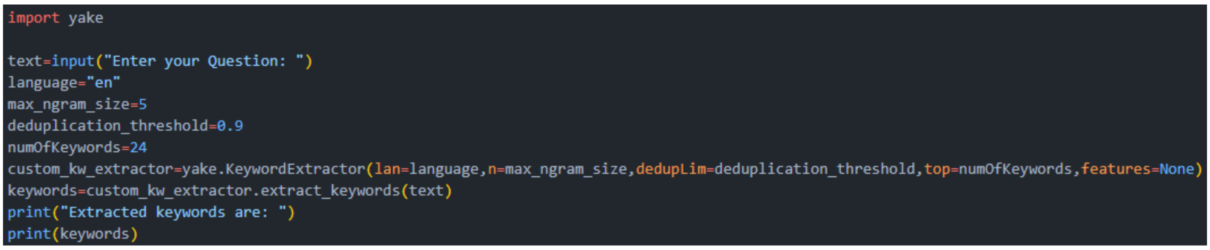
```
('content', '\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nChapter_1.pmd\n\n\n1.1  WHAT IS PHYSICS ?\n\nHumans have always been curious about the
world around\nthem. The night sky with its bright celestial objects has\nfascinated humans since time immemorial. The regular\nrepetitions of the day and night, the annual cycle of seasons,
\nthe eclipses, the tides, the volcanoes, the rainbow have always\nbeen a source of wonder. The world has an astonishing variety\nof materials and a bewildering diversity of life and behaviou
\nThe inquiring and imaginative human mind has responded\nto the wonder and awe of nature in different ways. One kind\nof response from the earliest times has been to observe the\nphysical
environment carefully, look for any meaningful\npatterns and relations in natural phenomena, and build and\nuse new tools to interact with nature.  This human endeavour\nled, in course of
time, to modern science and technology.\n\nThe word Science originates from the Latin verb Scientia\nmeaning ‘to know’.  The Sanskrit word Vijñān and the Arabic\nword Ilm convey similar
meaning, namely ‘knowledge’.\nScience, in a broad sense, is as old as human species. The\nearly civilisations of Egypt, India, China, Greece, Mesopotamia\nand many others made vital
contributions to its progress.\nFrom the sixteenth century onwards, great strides were made\nin science in Europe. By the middle of the twentieth century,\nscience had become a truly
international enterprise, with\nmany cultures and countries contributing to its rapid growth.\n\nWhat is Science and what is the so-called Scientific\nMethod?  Science is a systematic attempt
to understand\nnatural phenomena in as much detail and depth as possible,\nand use the knowledge so gained to predict, modify and\ncontrol phenomena. Science is exploring, experimenting
and\npredicting from what we see around us. The curiosity to learn\nabout the world, unravelling the secrets of nature is the first\nstep towards the discovery of science. The scientific
method\ninvolves several interconnected steps : Systematic\nobservations, controlled experiments, qualitative and\n\nCHAPTER ONE\n\nPHYSICAL WORLD\n\n1.1 What is physics ?\n\n1.2 Scope and
excitement of\nphysics\n\n1.3 Physics, technology and\nsociety\n\n1.4 Fundamental forces in\nnature\n\n1.5 Nature of physical
laws\n\nSummary\n\nExercises\n\n2021-22\n\n\n\nPHYSICS2\n\nquantitative reasoning, mathematical\nmodelling, prediction and verification or\nfalsification of theories. Speculation
and\nconjecture also have a place in science; but\nultimately, a scientific theory, to be acceptable,\nmust be verified by relevant observations  or\nexperiments. There is much
philosophical\ndebate about the nature and method of science\nthat we need not discuss here.\n\nThe interplay of theory and observation (or\nexperiment) is basic to the progress of science.
\nScience is ever dynamic. There is no ‘final’\ntheory in science and no unquestioned\nauthority among scientists.  As observations\nimprove in detail and precision or experiments\nyield new
results, theories must account for\nthem, if necessary, by introducing modifications.\nSometimes the modifications may not be drastic\nand may lie within the framework of existing\ntheory.
For example, when Johannes Kepler\n(1571-1630) examined the extensive data on\nplanetary motion collected by Tycho Brahe\n(1546-1601), the planetary circular orbits in\nheliocentric  theory
(sun at the centre of the\nsolar system) imagined by Nicolas Copernicus\n(1473–1543) had to be replaced by elliptical\norbits to fit the data better.  Occasionally,\nhowever, the  existing
theory is simply unable\nto explain new observations.  This causes a\nmajor upheaval in science. In the beginning of\nthe twentieth century, it was realised that\nNewtonian mechanics, till
then a very\nsuccessful theory, could not explain some of the\nmost basic features of atomic phenomena.\nSimilarly, the then accepted wave picture of light\nfailed to explain the photoelectri
effect properly.\nThis led to the development of a  radically new\ntheory (Quantum Mechanics) to deal with atomic\nand molecular phenomena.\n\nJust as a new experiment may suggest
an\nalternative theoretical model, a theoretical\nadvance may suggest what to look for in some\nexperiments.  The result of experiment of\nscattering of alpha particles by gold foil, in
1911\nby Ernest Rutherford (1871–1937) established\nthe nuclear model of the atom, which then\nbecame the basis of the quantum theory of\nhydrogen atom given in 1913 by Niels Bohr\n(1885–1962
. On the other hand, the concept of\nantiparticle was first introduced theoretically by\nPaul Dirac (1902–1984) in 1930 and confirmed\ntwo years later by the experimental discovery
of\npositron (antielectron) by Carl Anderson.\n\nPhysics is a basic discipline in the category\n\nof Natural Sciences, which also includes other\ndisciplines like Chemistry and Biology. The
word\n\nPhysics comes from a Greek word meaning\nnature. Its Sanskrit equivalent is Bhautiki that\nis used to refer to the study of the physical world.\n\nA precise definition of this
discipline is neither\npossible nor necessary.  We can broadly describe\n\nphysics as a study of the basic laws of  nature\nand their manifestation in different natural\n\nphenomena.  The
scope of physics is described\nbriefly in the next section.  Here we remark on\ntwo principal thrusts in physics :  unification\nand reduction.\n\nIn Physics, we attempt to explain
diverse\n\nphysical phenomena in terms of a few concepts\nand laws.  The effort is to see the physical world\nas manifestation of some universal laws in\n\ndifferent domains and conditions.
For example,\nthe same law of gravitation (given by Newton)\n\ndescribes the fall of an apple to the ground, the\nmotion of the moon around the earth and the\n\nmotion of planets around the
sun. Similarly, the\nbasic laws of electromagnetism (Maxwell’s\nequations) govern all electric and magnetic\n\nphenomena. The attempts to unify fundamental\nforces of nature (section 1.4)
```

Fig.3.2 PyPDF2 Demonstration

As you can see the data parsed from the PDF is stored in a single array position. This makes it impossible to find the necessary context. Hence a Tokenizer is necessary which is implemented as below.

Keyword Extraction Module:

We have used the YAKE! Model to extract keywords from a given sentence. This is necessary since Context Identification is made much easier by just passing the necessary keywords as opposed to passing the whole Query to the parsed PDF file above.

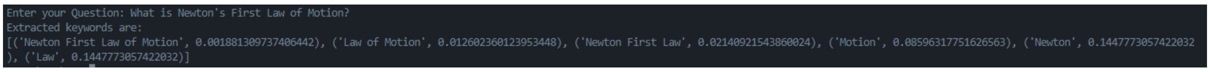


```
import yake

text=input("Enter your Question: ")
language="en"
max_ngram_size=5
deduplication_threshold=0.9
numOfKeywords=24
custom_kw_extractor=yake.KeywordExtractor(lan=language,n=max_ngram_size,dedupLim=deduplication_threshold,top=numOfKeywords,features=None)
keywords=custom_kw_extractor.extract_keywords(text)
print("Extracted keywords are: ")
print(keywords)
```

Fig.4.1 YAKE! Algorithm

Fig.4.1 demonstrates the code snippet for the YAKE! Algorithm. Various parameters can be changed depending upon the type of output required by the user. max_ngram_size denotes the number of words the extracted keyword should contain. If you keep increasing this value, the probability of the group of words extracted being a keyword keep on decreasing.



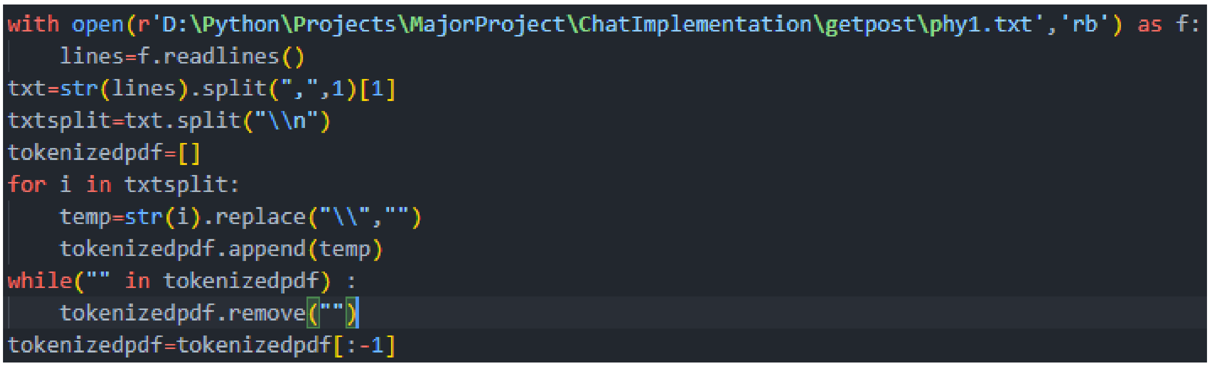
```
Enter your Question: What is Newton's First Law of Motion?
Extracted keywords are:
[('Newton First Law of Motion', 0.001881309737406442), ('Law of Motion', 0.012602360123953448), ('Newton First Law', 0.02140921543860024), ('Motion', 0.08596317751626563), ('Newton', 0.1447773057422032
), ('Law', 0.1447773057422032)]
```

Fig.4.2 YAKE! Algorithm

As you can see in fig.4.2, if the number of words in a keyword decrease, the probability of the extracted set of words being a keyword increase.

Context Identification Module:

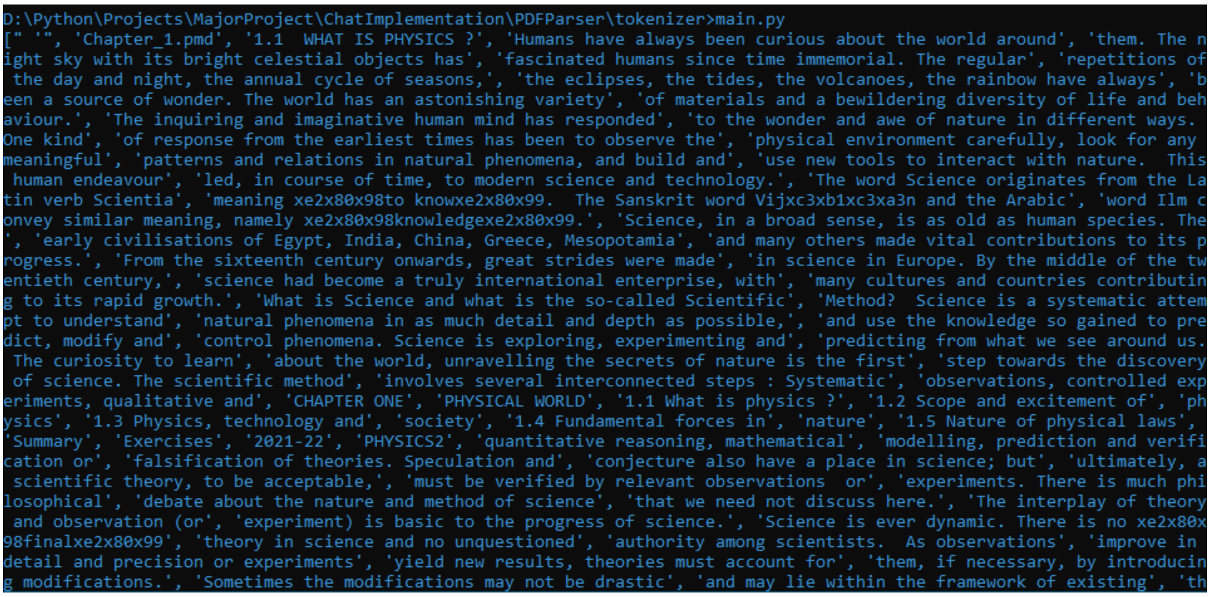
This module utilizes the parsed PDF Contents and keywords to fetch the required context from the PDF for the BERT Algorithm.



```
with open(r'D:\Python\Projects\MajorProject\ChatImplementation\getpost\phy1.txt','rb') as f:
    lines=f.readlines()
txt=str(lines).split(",",1)[1]
txtsplit=txt.split("\\n")
tokenizedpdf=[]
for i in txtsplit:
    temp=str(i).replace("\\","")
    tokenizedpdf.append(temp)
while("" in tokenizedpdf) :
    tokenizedpdf.remove("")
tokenizedpdf=tokenizedpdf[:-1]
```

Fig.5.1 PDF Tokenizer

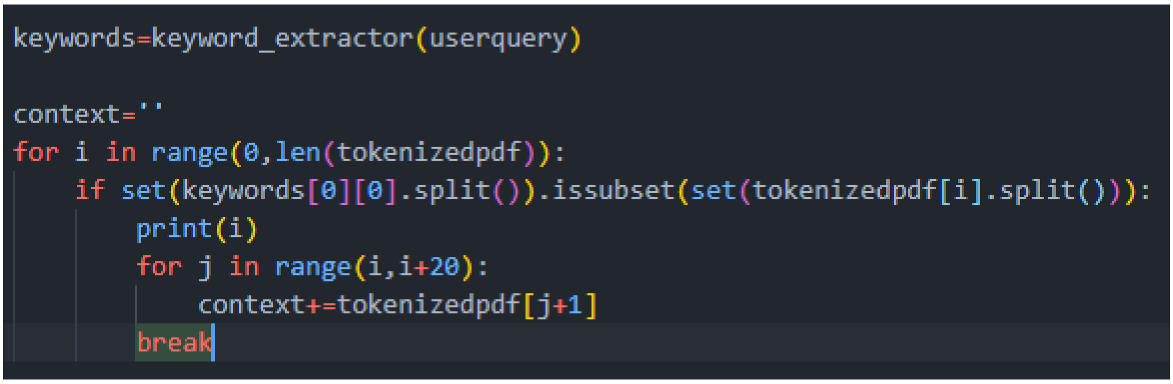
This is a text tokenizer. It converts all the text present in the parsed PDF, i.e. at a single array position, into multiple sentences present at different positions of the new tokenizedpdf array.



```
D:\Python\Projects\MajorProject\ChatImplementation\PDFParser\tokenizer>main.py
[" '", 'Chapter_1.pmd', '1.1  WHAT IS PHYSICS ?', 'Humans have always been curious about the world around', 'them. The n
ight sky with its bright celestial objects has', 'fascinated humans since time immemorial. The regular', 'repetitions of
 the day and night, the annual cycle of seasons,', 'the eclipses, the tides, the volcanoes, the rainbow have always', 'b
een a source of wonder. The world has an astonishing variety', 'of materials and a bewildering diversity of life and beh
aviour.', 'The inquiring and imaginative human mind has responded', 'to the wonder and awe of nature in different ways.
One kind', 'of response from the earliest times has been to observe the', 'physical environment carefully, look for any
meaningful', 'patterns and relations in natural phenomena, and build and', 'use new tools to interact with nature.  This
 human endeavour', 'led, in course of time, to modern science and technology.', 'The word Science originates from the La
tin verb Scientia', 'meaning xe2x80x98to knowxe2x80x99.  The Sanskrit word Vijxc3xb1xc3xa3n and the Arabic', 'word Ilm c
onvey similar meaning, namely xe2x80x98knowledgexe2x80x99.', 'Science, in a broad sense, is as old as human species. The
', 'early civilisations of Egypt, India, China, Greece, Mesopotamia', 'and many others made vital contributions to its p
rogress.', 'From the sixteenth century onwards, great strides were made', 'in science in Europe. By the middle of the tw
entieth century,', 'science had become a truly international enterprise, with', 'many cultures and countries contributin
g to its rapid growth.', 'What is Science and what is the so-called Scientific', 'Method?  Science is a systematic attem
pt to understand', 'natural phenomena in as much detail and depth as possible,', 'and use the knowledge so gained to pre
dict, modify and', 'control phenomena. Science is exploring, experimenting and', 'predicting from what we see around us.
 The curiosity to learn', 'about the world, unravelling the secrets of nature is the first', 'step towards the discovery
 of science. The scientific method', 'involves several interconnected steps : Systematic', 'observations, controlled exp
eriments, qualitative and', 'CHAPTER ONE', 'PHYSICAL WORLD', '1.1 What is physics ?', '1.2 Scope and excitement of', 'ph
ysics', '1.3 Physics, technology and', 'society', '1.4 Fundamental forces in', 'nature', '1.5 Nature of physical laws',
'Summary', 'Exercises', '2021-22', 'PHYSICS2', 'quantitative reasoning, mathematical', 'modelling, prediction and verifi
cation or', 'falsification of theories. Speculation and', 'conjecture also have a place in science; but', 'ultimately, a
 scientific theory, to be acceptable,', 'must be verified by relevant observations  or', 'experiments. There is much phi
losophical', 'debate about the nature and method of science', 'that we need not discuss here.', 'The interplay of theory
 and observation (or', 'experiment) is basic to the progress of science.', 'Science is ever dynamic. There is no xe2x80x
98finalxe2x80x99', 'theory in science and no unquestioned', 'authority among scientists.  As observations', 'improve in
detail and precision or experiments', 'yield new results, theories must account for', 'them, if necessary, by introducin
g modifications.', 'Sometimes the modifications may not be drastic', 'and may lie within the framework of existing', 'th
```

Fig.5.2 PDF Tokenizer

The PDF is now tokenized and is now ready to be used for context Identification.

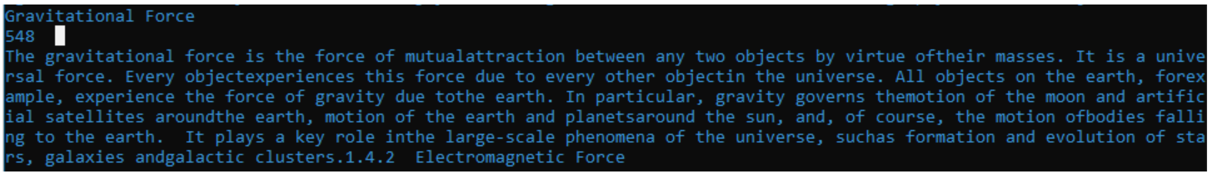


```python
keywords=keyword_extractor(userquery)

context=''
for i in range(0,len(tokenizedpdf)):
    if set(keywords[0][0].split()).issubset(set(tokenizedpdf[i].split())):
        print(i)
        for j in range(i,i+20):
            context+=tokenizedpdf[j+1]
        break
```

Fig. 5.3 Context Extraction Algorithm

Context Identification algorithm utilizes the Keywords that are acquired from the Keyword Extraction Module and the Tokenized PDF array which is acquired above. This then prints the necessary content for the said Keyword.



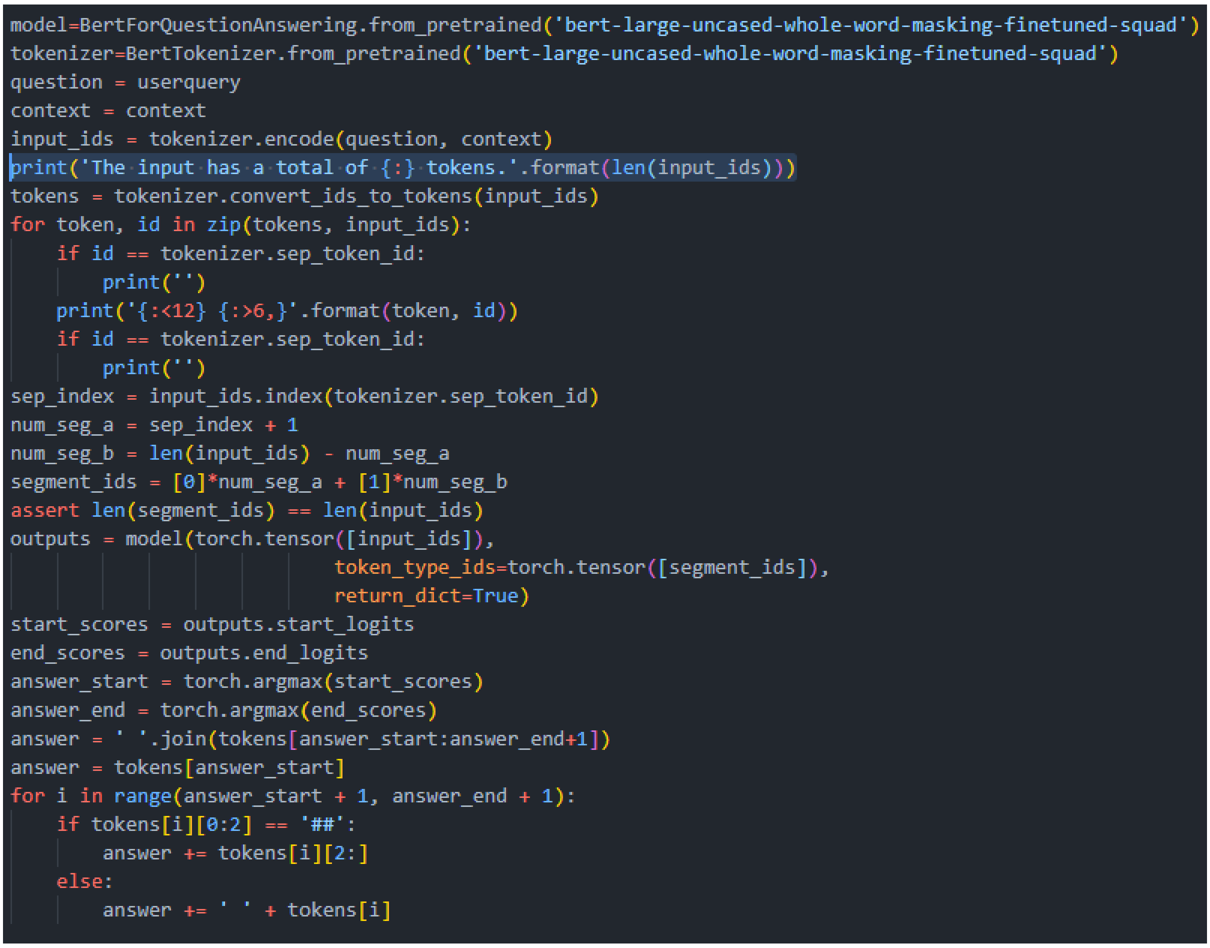
```
Gravitational Force
548
The gravitational force is the force of mutualattraction between any two objects by virtue oftheir masses. It is a unive
rsal force. Every objectexperiences this force due to every other objectin the universe. All objects on the earth, forex
ample, experience the force of gravity due tothe earth. In particular, gravity governs themotion of the moon and artific
ial satellites aroundthe earth, motion of the earth and planetsaround the sun, and, of course, the motion ofbodies falli
ng to the earth.  It plays a key role inthe large-scale phenomena of the universe, suchas formation and evolution of sta
rs, galaxies andgalactic clusters.1.4.2  Electromagnetic Force
```

Fig.5.4 Context Extraction Algorithm

548 denotes the position at which the keyword “Gravitational Force” was found in the tokenizedpdf array. This context is then passed on to the BERT Algorithm.

BERT Algorithm:

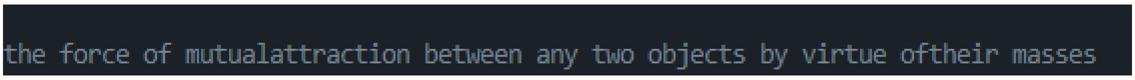
BERT Model utilizes 24 Encoders that utilize its own tokenizer to match keywords of the context and user query to its own vocabulary. This module finds the relationship between the user query and context and then returns it as an answer.



```
model=BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
tokenizer=BertTokenizer.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
question = userquery
context = context
input_ids = tokenizer.encode(question, context)
print('The input has a total of {:} tokens.'.format(len(input_ids)))
tokens = tokenizer.convert_ids_to_tokens(input_ids)
for token, id in zip(tokens, input_ids):
    if id == tokenizer.sep_token_id:
        print('')
    print('{:<12} {:>6,}'.format(token, id))
    if id == tokenizer.sep_token_id:
        print('')
sep_index = input_ids.index(tokenizer.sep_token_id)
num_seg_a = sep_index + 1
num_seg_b = len(input_ids) - num_seg_a
segment_ids = [0]*num_seg_a + [1]*num_seg_b
assert len(segment_ids) == len(input_ids)
outputs = model(torch.tensor([input_ids]),
                            token_type_ids=torch.tensor([segment_ids]),
                            return_dict=True)
start_scores = outputs.start_logits
end_scores = outputs.end_logits
answer_start = torch.argmax(start_scores)
answer_end = torch.argmax(end_scores)
answer = ' '.join(tokens[answer_start:answer_end+1])
answer = tokens[answer_start]
for i in range(answer_start + 1, answer_end + 1):
    if tokens[i][0:2] == '##':
        answer += tokens[i][2:]
    else:
        answer += ' ' + tokens[i]
```

Fig. 6.1 BERT Algorithm

In fig.6.1 it is visible that the user query and context is passed on to BERT's Tokenizer then broken down further to find the exact relationship between the query and context.



```
the force of mutualattraction between any two objects by virtue oftheir masses
```

Fig. 6.2 BERT Algorithm

The answer returned is now up to the user to decide whether it is up to its expectations or not.

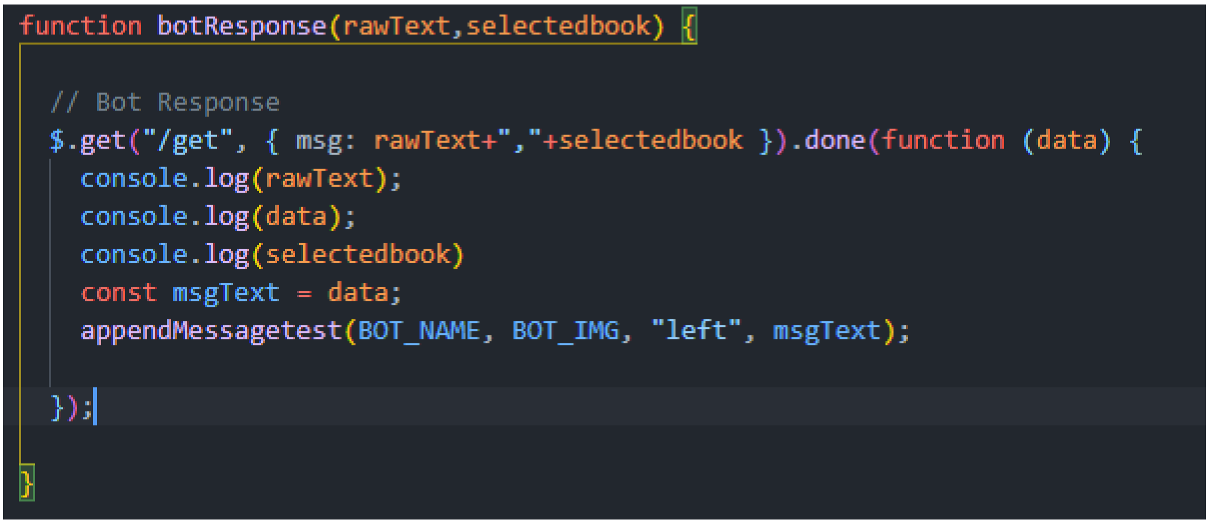
Web Scrapper Module:

This is a simple module as an alternative to BERT Algorithm. If the BERT Algorithm is not able to generate an appropriate answer for the User, the user query is then passed onto the web scrapped which scrapes the top website that is recommended by google once the user query is passed onto the Web Scrapper Module. This also utilizes the Wikipedia API.

Chapter 6 - Implemented System

app.py Module:

app.py Module consists of every module stated in the System Design. This module also consists the User Interface and the necessary get-post methods to fetch user queries from the User Interface.



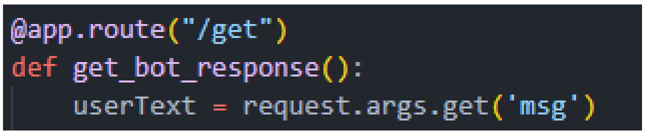
```
function botResponse(rawText,selectedbook) {

  // Bot Response
  $.get("/get", { msg: rawText+","+selectedbook }).done(function (data) {
    console.log(rawText);
    console.log(data);
    console.log(selectedbook)
    const msgText = data;
    appendMessagetest(BOT_NAME, BOT_IMG, "left", msgText);

  });

}
```

Fig.7.1 GET Method

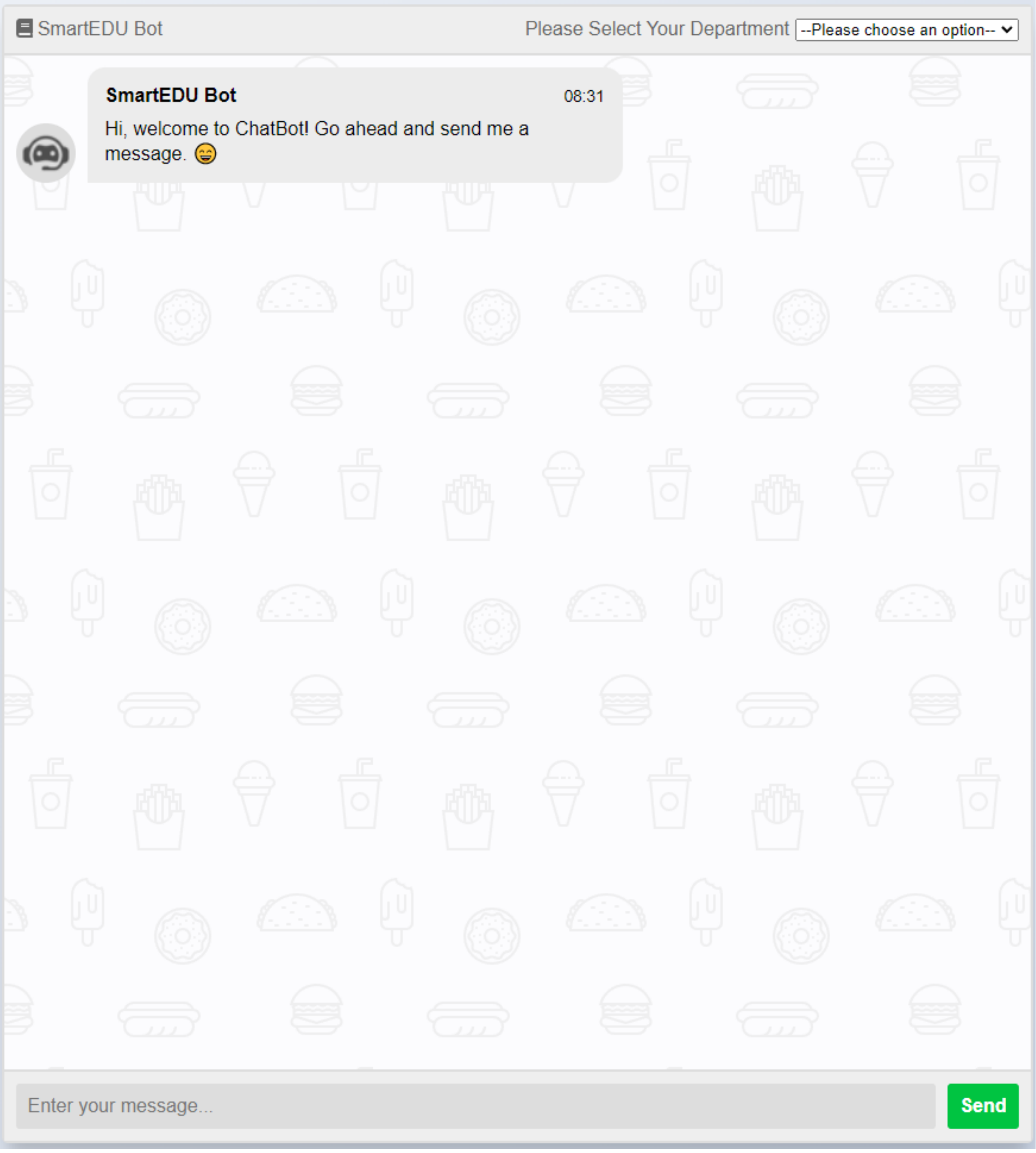


```
@app.route("/get")
def get_bot_response():
    userText = request.args.get('msg')
```

Fig. 7.2 GET Method

This GET Method fetches the User Query which consists of the user query and information about the department in which the user wishes to ask the question.

User Interface:



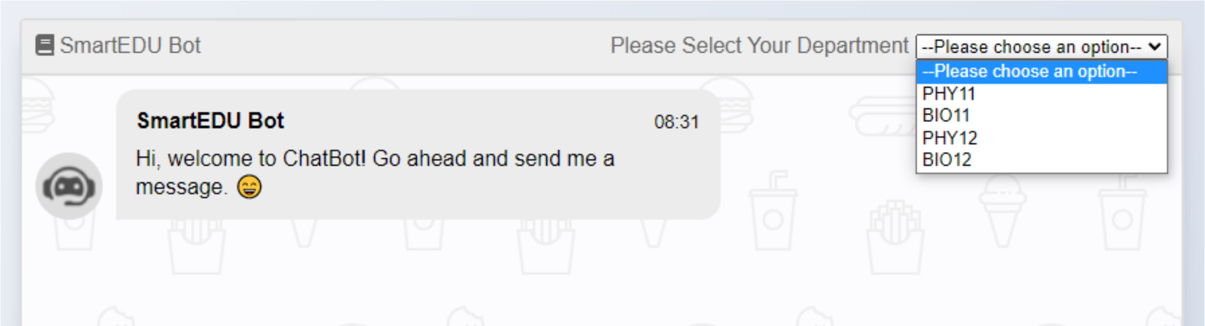
SmartEDU Bot

Please Select Your Department --Please choose an option--

SmartEDU Bot 08:31

Hi, welcome to ChatBot! Go ahead and send me a message. 😄

Fig.8.1 User Interface



SmartEDU Bot

Please Select Your Department --Please choose an option--

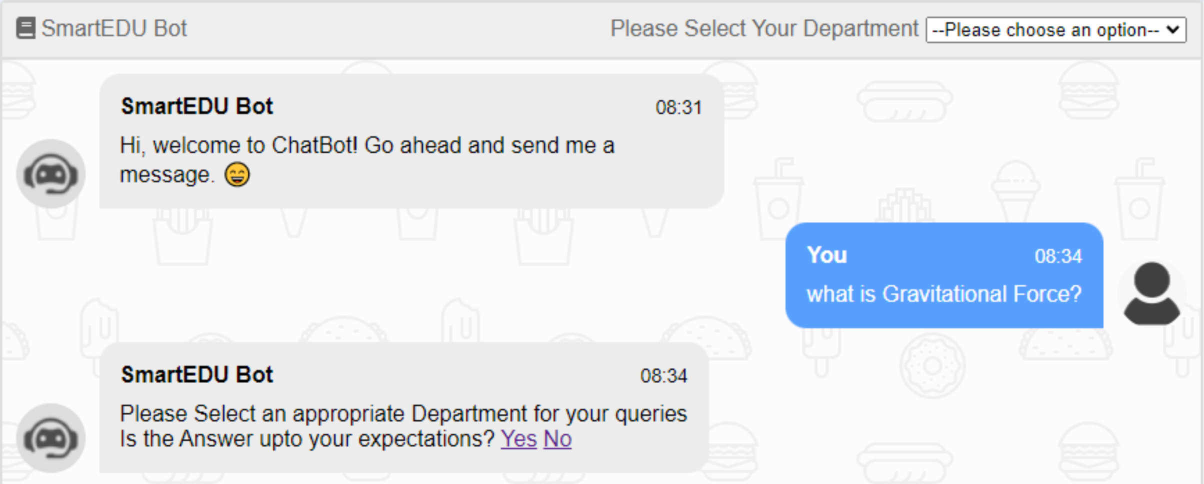
SmartEDU Bot 08:31

Hi, welcome to ChatBot! Go ahead and send me a message. 😄

Fig.8.2 User Interface

Chapter 7 - Results

Outputs:

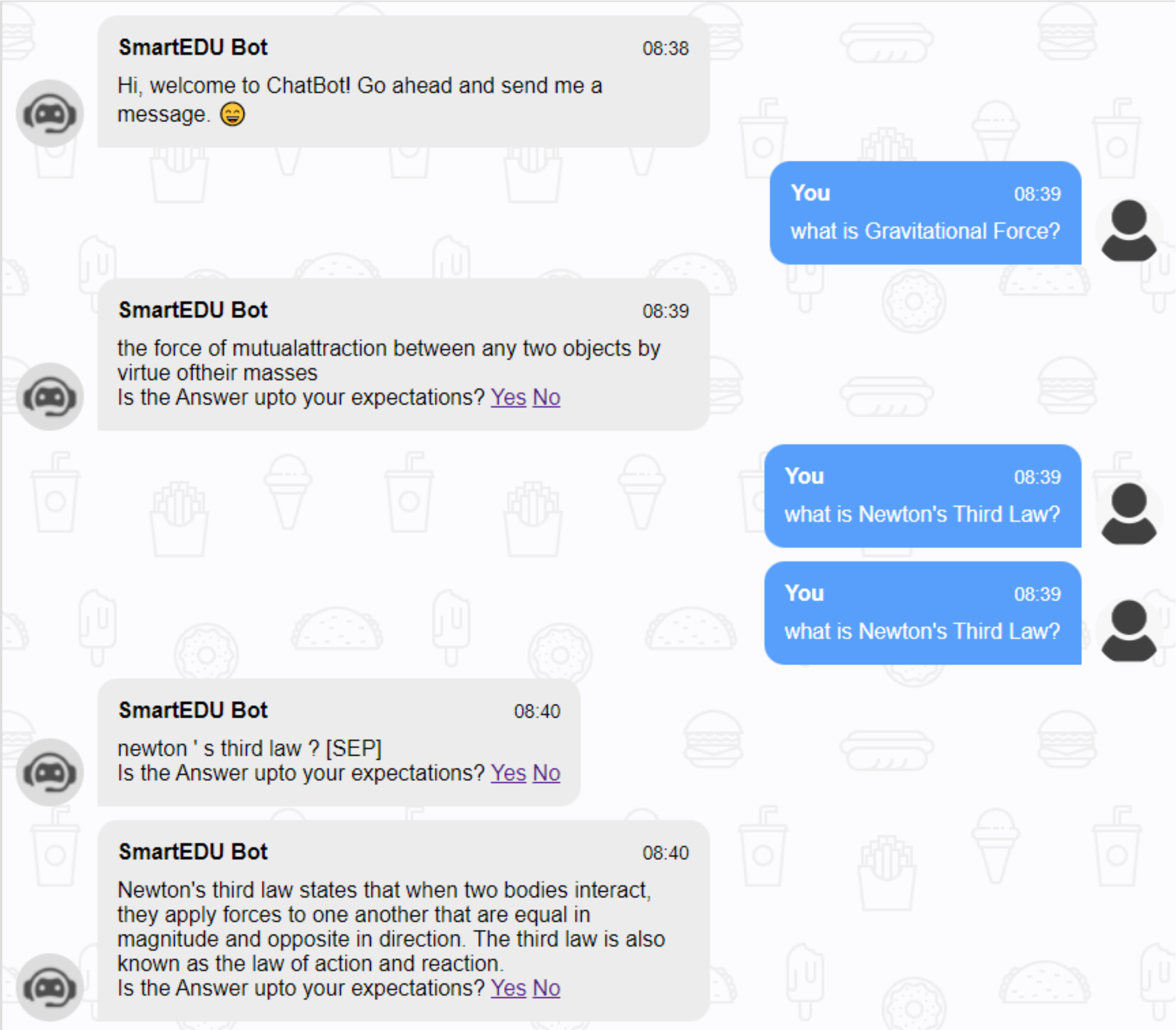


SmartEDU Bot 08:31
Hi, welcome to ChatBot! Go ahead and send me a message. 😄

You 08:34
what is Gravitational Force?

SmartEDU Bot 08:34
Please Select an appropriate Department for your queries
Is the Answer upto your expectations? Yes No

Fig. 9.1 Outputs



SmartEDU Bot 08:38
Hi, welcome to ChatBot! Go ahead and send me a message. 😄

You 08:39
what is Gravitational Force?

SmartEDU Bot 08:39
the force of mutualattraction between any two objects by virtue oftheir masses
Is the Answer upto your expectations? Yes No

You 08:39
what is Newton's Third Law?

You 08:39
what is Newton's Third Law?

SmartEDU Bot 08:40
newton ' s third law ? [SEP]
Is the Answer upto your expectations? Yes No

SmartEDU Bot 08:40
Newton's third law states that when two bodies interact, they apply forces to one another that are equal in magnitude and opposite in direction. The third law is also known as the law of action and reaction.
Is the Answer upto your expectations? Yes No

Fig.9.2 Output

The output is generated as follows:

1. The user query, “What is Gravitational Force?” is fetched from the website using GET Method in Python Flask;
2. The user query is then put into Keyword Extraction Method, to extract the necessary keywords for context identification;
3. The word “Gravitational Force” is extracted;
4. The keyword is then passed through the parsed PDF contents of the selected department, in this case, PHY11;
5. The necessary context is then extracted;
6. The context as well as the user query is then passed onto BERT Algorithm which returns the appropriate answer;
7. Lets ask another question, this time the question is not available in the PDF
8. It returns an answer which is not according to user standards, the user presses the “No” button
9. The previous question is then passed on to the web scrapper to get the answer.

Test Cases and Accuracy:

Test Cases and Accuracy of our Chatbot are stated below:

User Asked: “What is Gravitational Force?”
Expected Answer: “The gravitational force is the force of mutual attraction between any two objects by virtue of their masses”
Calculated Answer: “the force of mutualattraction between any two objects by virtue of their masses.”

User Asked: “What is Newton’s Third Law?”
Expected Answer: “To every action, there is always an equal and opposite reaction.”
Calculated Answer: “newton ‘ s third law ?”
Web scrapped Answer: “Newton's third law states that when two bodies interact, they apply forces to one another that are equal in magnitude and opposite in direction. The third law is also known as the law of action and reaction.”

User Asked: “What classes come under Phylum?”
Expected Answer: “Classes comprising animals like fishes, amphibians, reptiles, birds along with mammals constitute the next higher category called Phylum.”
Calculated Answer: “fishes , amphibians , reptiles , birds”

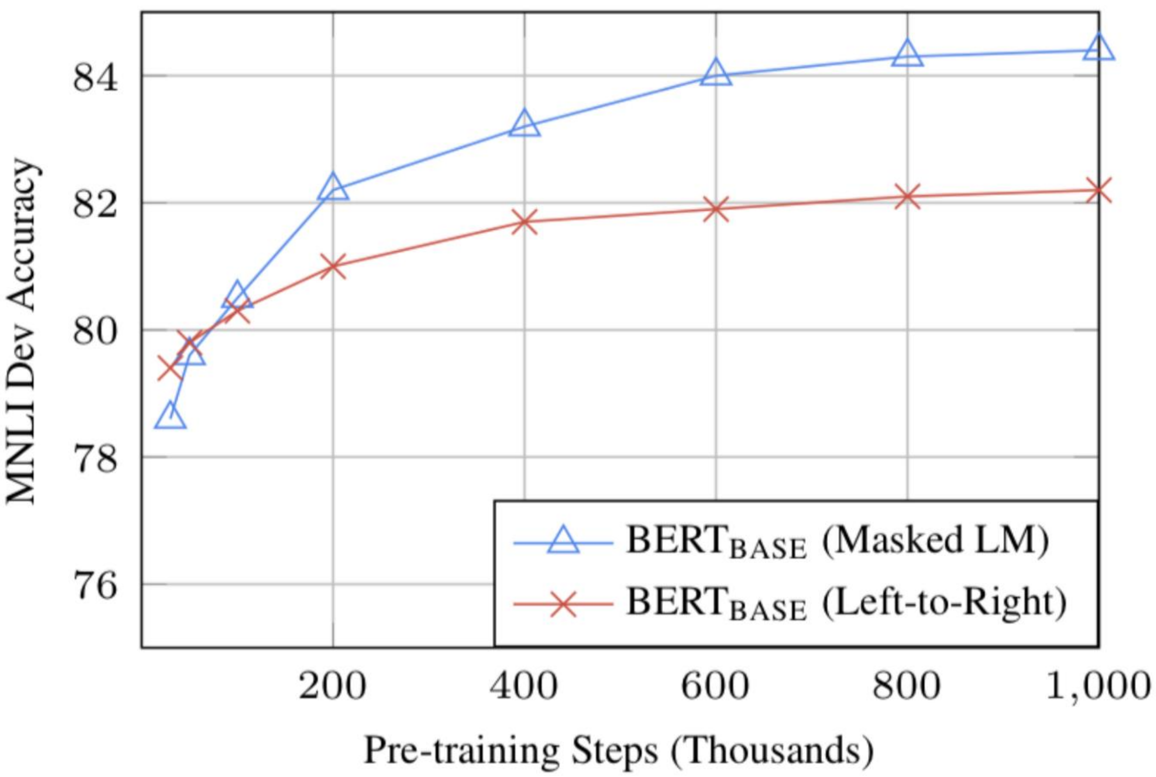
User Asked: “What is Genus?”
Expected Answer: “Genus comprises a group of related species which has more characters in common in comparison to species of other genera.”
Calculated Answer: “a group of related species”

User Asked: “ ”
Expected Answer: “ ”
Calculated Answer: “Error- List Index Out of Range”

As you can see the answers provided by the BERT algorithm may not be according to the University standards but they are to the point, without mention of phrases like “This term Means”, “This term includes”. All the answers comprise of words that co-relate to the user query and are to the point.

Accuracy of BERT Algorithm:
The max number of intents that our Fine-Tuned BERT model can handle is 512. This number can be lower, can be equal to 512 but not greater than it since the model is not able to handle more than 512 intents. BERT Model uses its own Tokenizer, this means it has its own vocabulary. Let’s say you pass the word “playing” as context for BERT Algorithm. The vocabulary of BERT Algorithm contains the word “play” not the word “playing”. It will hence split the word “playing” as “play” and “#ing”. This adds up to 2 intents. Context should be simple enough to be understood and not to complex with words outside the vocabulary.

The calculated accuracy for the BERT Algorithm with 381 intents is 86.11%. This accuracy keeps on decreasing by a logarithmic curve as the number of intents increase. There is no fix accuracy for BERT Algorithm since intents will be different every time and two sentences can always never have the same meaning.


MNLI Dev Accuracy
76
78
80
82
84
200
400
600
800
1,000
Pre-training Steps (Thousands)
BERT$_{\text{BASE}}$ (Masked LM)
BERT$_{\text{BASE}}$ (Left-to-Right)

Chapter 8 - Conclusion

From the research conducted above we aim towards creating a perfect Smart Student Assistant. We are using advanced data parsing libraries such as PyPDF2 to extract information from various PDFs and send it to our backend. We are also using Wikipedia API and some Educational APIs that shall also provide data to our backend. By using this Bot, we assure that the student will be able to find both short and brief answers for the asked question.

Future Scope:

We aim to implement image processing so that images can be exchanged by the bot as well as the user. This will enable them to get appropriate representation for their problem.

We also aim to implement a dynamic PDF upload option so that you, the user can upload your own PDF and start asking questions related to the topics present in the PDF right away.

Chapter 9 - References

M. Farhan, I. M. Munwar, M. Aslam, A. M. Martinez Enriquez, A. Farooq, S. Tanveer, and P. A. Mejia “Automated reply to students’ queries in e-learning environment using Web-BOT,” Eleventh Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence and Applications, Special Session - Revised Paper, 2012.

S. Satu, H. Parvez, and Shamim-AI-Mamun “Review of integrated applications with AIML based chatbot,” First International Conference on Computer and Information Engineering (ICCIE), 2015.

S. Satu, H. Parvez, and Shamim-AI-Mamun “Review of integrated applications with AIML based chatbot,” First International Conference on Computer and Information Engineering (ICCIE), 2015.

Garcia Brustenga, G., Fuertes-Alpiste, M., Molas-Castells, N. (2018). Briefing paper: “Chatbots in Education.” Barcelona: eLearn Center. Universitat Oberta de Catalunya. ISBN: 978-84-09-03944-9

Eric Hsiao-Kuang Wu, Chun-Han Lin, Yu-Yen Ou, Chen-Zhong Liu, Wei-Kai Wang, Chi-Yun Chao “Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot” Ministry of Science, Taiwan, Grant ID 108-2221-E-008-034.

Chinedu Wilfred Okonkwo, Abejide Ade-Ibijola “Computers and Education: Artificial Intelligence”

Rana, Muhammad, “EagleBot: A chatbot based on Multi-Tier Question Answering Systems for Retrieving Answers from Heterogenous Sources Using BERT” (2019). Electronic Theses and Dissertations.

Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, and Antonio Piertrosanto, “Chatbot for E-learning: A Case of Study”, DIIN University of Salerno, Fisciano (SA), Italy

K. Jwala G.N.V.G Sirisha, G.V. Padma Raju, “Developing a Chatbot using Machine Learning”, Electronics Computer Technology (ICECT) 2011 3rd International Conference on, Vol. 4, 2011

Ashok K. Goel and Lalith Polepeddi, “Jill Watson: A Virtual Teaching Assistant for Online Education”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-IS3, June 2019.

Sam Schwager John Solitario, “Question and Answering on SQuAd 2.0: BERT is All You Need” 2016 15th International Conference on Department of Computer Science Stanford University