

# **INTELLIGENT NATURAL LANGUAGE SEARCH**

## **A PROJECT REPORT**

*Submitted by,*

**Mr. C JAI MANOJ REDDY - 20201CST0102**

**Mr. V VENKATA SURESH - 20201CST0170**

**Mr. PRATHAM V B - 20201CST0085**

**Mr. AHMED ULLA KHAN - 20201CST0139**

*Under the guidance of,*

**Mrs. SANDHYA L**

*in partial fulfillment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND TECHNOLOGY**

**[ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING]**

**AT**



**PRESIDENCY UNIVERSITY**

**BENGALURU**

**JANUARY 2024**

# **PRESIDENCY UNIVERSITY**

## **SCHOOL OF COMPUTER SCIENCE ENGINEERING**

### **CERTIFICATE**

This is to certify that the Project report “**INTELLIGENT NATURAL LANGUAGE SEARCH**” being submitted by “ **C JAI MANOJ REDDY , V VENKATA SURESH , PRTHAM V B , AHMED ULLA KHAN** ” bearing roll number(s) “ 20201CST0102 , 20201CST0170 , 20201CST0085 , 20201CST0139 ” in partial fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Science and Technology [Artificial intelligence and Machine learning] is a bonafide work carried out under my supervision.

**Ms. Sandhya L**  
Assistant Professor  
School of CSE&IS  
Presidency University

**Dr. A. Jayachandran**  
Professor & HoD  
School of CSE&IS  
Presidency University

**Dr. C. KALAIARASAN**  
Associate Dean  
School of CSE&IS  
Presidency University

**Dr. L. SHAKKEERA**  
Associate Dean  
School of CSE&IS  
Presidency University

**Dr. SAMEERUDDIN KHAN**  
Dean  
School of CSE&IS  
Presidency University

# **PRESIDENCY UNIVERSITY**

## **SCHOOL OF COMPUTER SCIENCE ENGINEERING**

### **DECLARATION**

We hereby declare that the work, which is being presented in the project report entitled **INTELLIGENT NATURAL LANGUAGE SEARCH** in partial fulfillment for the award of Degree of **Bachelor of Technology in Computer Science and Technology (Artificial Intelligence And Machine Learning)**, is a record of our own investigations carried under the guidance of **Mrs Sandhya L, Assistan Professor, School of Computer Science Engineering , Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

<b>Name</b>	<b>Roll No</b>	<b>Signature</b>
C JAI MANOJ REDDY	20201CST0102	
V VENKATA SURESH	20201CST0170	
PRATHAM V B	20201CST0085	
AHMED ULLA KHAN	20201CST0139	

# ABSTRACT

In response to the dynamic landscape of information management within research and development (R&D) spheres, Dr. Reddy's Laboratories is undertaking an ambitious initiative to streamline the extraction of insights from their extensive repository of R&D data. This project centers around the creation of an intelligent natural language search tool designed to navigate and derive actionable knowledge from the wealth of information encapsulated within PowerPoint (PPT) slides stored in a shared folder.

The primary challenge addressed by this endeavor is the efficient extraction of meaningful insights and information from diverse and voluminous PPT presentations. Leveraging cutting-edge technologies and methodologies, the proposed solution takes the form of a web application, orchestrating seamless interaction between users and the repository of R&D reports, findings, and presentations.

The envisioned software, termed as the "Intelligent Knowledge Builder," stands as a culmination of advanced development methodologies, integrating Streamlit, Spacy, Transformers, and other libraries. Through its user-friendly interface, the web application presents an intuitive platform where users can input search queries or terms of interest, prompting the system to embark on an automated journey of knowledge extraction.

By harnessing natural language processing (NLP) techniques and machine learning algorithms, the system traverses through the PPT slides, extracting pertinent information and generating concise yet informative insights. This intelligent mechanism extends beyond mere keyword matching, encompassing comprehensive comprehension and synthesis of diverse textual content to offer targeted responses aligned with user queries.

Furthermore, the software's adaptive functionality extends to user-uploaded files, enabling tailored analysis of additional PPT or PDF documents. The system applies a robust text extraction and analysis process, culminating in the generation of customized insights, empowering users with deeper knowledge and enriched perspectives.

This software not only serves as a revolutionary tool for automatic knowledge extraction but also embodies Dr.Reddy's Laboratories' commitment to technological innovation in enhancing R&D efficiency. The successful implementation of this web application stands to redefine the paradigm of information retrieval, enabling researchers and stakeholders to harness invaluable insights swiftly and effectively from the organization's wealth of R&D resources.

## ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. Kalaiarasan C** and **Dr. Shakkeera L**, School of Computer Science Engineering & Information Science, Presidency University and **Dr. A. Jayachandran**. Head of the Department, School of Computer Science Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Mrs Sandhya L**, School of Computer Science Engineering & Information Science, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud**, **Dr. Mrutyunjaya MS** and also the department Project Coordinators **Dr. Manjula H M**, **Mr. Yamanappa**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

**C JAI MANOJ REDDY**

**V VENKATA SURESH**

**PRATHAM V B**

**AHMED ULLA KHAN**

## **LIST OF TABLES**

<b>Sl. No.</b>	<b>Table Name</b>	<b>Table Caption</b>	<b>Page No.</b>
1	Table 7.2	Timeline for execution of project	22

## LIST OF FIGURES

<b>Sl. No.</b>	<b>Figure Name</b>	<b>Caption</b>	<b>Page No.</b>
1	Figure 6.1	Transformer architecture	20
2	Figure 7.1	Gantt chat	21
3	Figure B.1	Use my file	40
4	Figure B.2	Browse file	41
5	Figure B.3	Query search	41
6	Figure B.4	Available source file	42

# **TABLE OF CONTENTS**

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
	<b>CERTIFICATE</b>	ii
	<b>DECLARATION</b>	iii
	<b>ABSTRACT</b>	iv
	<b>ACKNOWLEDGEMENT</b>	v
<b>1.</b>	<b>INTRODUCTION</b>	<b>1 - 2</b>
<b>2.</b>	<b>LITERATURE SURVEY</b>	<b>3 - 4</b>
<b>3.</b>	<b>RESEARCH GAPS OF EXISTING METHODS</b>	<b>5 - 7</b>
<b>4.</b>	<b>PROPOSED MOTHODOLOGY</b>	<b>8 - 11</b>
<b>5.</b>	<b>OBJECTIVES</b>	<b>12 - 13</b>
<b>6.</b>	<b>SYSTEM DESIGN &amp; IMPLEMENTATION</b>	<b>14 - 17</b>
<b>7.</b>	<b>TIMELINE FOR EXECUTION OF PROJECT</b>	<b>18 - 19</b>
<b>8.</b>	<b>OUTCOME</b>	
<b>9.</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>20 -21</b>
<b>10.</b>	<b>CONCLUSION</b>	<b>22 - 24</b>
	<b>REFERENCES</b>	<b>25 - 26</b>
	<b>APPENDIX-A: PSUEDOCODE</b>	<b>27</b>
	<b>APPENDIX-B: SCREENSHOTS</b>	<b>28 - 36</b>
	<b>APPENDIX-C: ENCLOSURES</b>	<b>37 - 39</b>



# **CHAPTER-1**

## **INTRODUCTION**

The current project focuses on developing a robust knowledge retrieval system designed specifically for Dr. Reddy's Laboratories. The objective is to create an intuitive and effective platform that utilizes advanced technologies to extract and present valuable insights from a diverse range of research and development (R&D) resources.

At the core of this endeavor is the creation of a software application, named the "Knowledge Builder," utilizing technologies such as Streamlit, Spacy, Transformers, and other relevant libraries. This application acts as an interface for users to interact with the extensive data repository, primarily composed of PowerPoint (PPT) presentations, web content, and various report formats.

The application's functionalities are organized into two main sections: "Home" and "Use my file." The "Home" section provides access to available source files, primarily focusing on R&D reports, through an expandable interface. Users can submit queries related to the available resources using an input field.

Upon query submission, the application initiates a thorough search mechanism, incorporating techniques such as text extraction from PPTs, web scraping, and Natural Language Processing (NLP) using libraries like Spacy. The search results are presented as responses, utilizing the Transformers library to generate concise yet informative insights.

The "Use my file" section allows users to upload their PPT or PDF files for analysis. The system undergoes a similar process of text extraction and analysis, contributing to the generation of tailored responses based on the user's query.

This project integrates cutting-edge technologies and methodologies with the goal of streamlining knowledge acquisition and insight generation from Dr. Reddy's Labs' diverse R&D data sources. It aims to empower researchers, scientists, and stakeholders by providing

quick access to pertinent information aligned with their queries or areas of interest.

The implementation of this initiative represents a significant advancement in enhancing the efficiency and agility of information retrieval processes within Dr. Reddy's Laboratories. The project's successful execution has the potential to revolutionize how the organization leverages insights from its extensive R&D resources, fostering a culture of informed decision-making and innovation.

## **CHAPTER-2**

### **LITERATURE SURVEY**

[1] Brianna Mueller, Takahiro Kinoshita, Alexander Peebles, Mark A. Graber, Sangil Lee.

The core of artificial intelligence (AI) lies in its capacity to imitate human intelligence, allowing for the execution of tasks, recognition of patterns, and prediction of outcomes by assimilating data from diverse sources. In various sectors such as autonomous driving, e-commerce, fintech, and healthcare, machine learning (ML) algorithms, notably deep learning, play a pivotal role. [1] Despite their impressive performance, the practical implementation of these algorithms in clinical settings encounters challenges due to a lack of transparency. Explainable AI (XAI) has emerged as a solution to address this issue, elucidating internal decisions to healthcare professionals, building trust, and facilitating the informed application of predictive models. The incorporation of XAI into healthcare settings is an ongoing exploration, influenced by the intricate nature of medical knowledge. [1] The success of AI in healthcare relies not only on advanced algorithms but also on a human-in-the-loop approach, engaging stakeholders and underscoring the significance of clear, iterative interactions between end users and AI systems to ensure meaningful clinical and operational capabilities.

[2] Shaan Khurshid, Christopher Reeder, Lia X. Harrington Et al.

Biomedical research increasingly relies on electronic health record (EHR) databases, pivotal for biological discoveries and clinical insights. These databases provide robust statistical power for extensive association analyses, encompassing diverse clinical features and measures such as risk factors, laboratory results, and imaging data. [2] However, the acknowledgment of potential biases in EHR data, stemming from patient selection and missing data, is growing. Ascertainment bias related to clinical need and selection bias due to missing data pose significant challenges. [2] The study introduces the Community Care Cohort Project (C3PO), an EHR-based initiative designed to enhance cardiovascular disease research by mitigating

ascertainment bias and minimizing data missingness. Utilizing deep natural language processing (NLP) to extract vital sign features from unstructured notes, the study aims to compare effective sample sizes before and after missing data recovery. [2] Additionally, established clinical risk scores are deployed to assess model performance, with the hypothesis that scores derived from prospective cohort settings will exhibit reduced bias in C3PO.

[3] Sheela Kolluri , Jianchang Lin, Rachael Liu, Yanwei Zhang & Wenwen Zhang.

In recent years, artificial intelligence (AI) and machine learning (ML) have emerged as transformative forces in pharmaceutical research and development (R&D), driven by advancements in computational technology and increased data accessibility. [3] The escalating costs of drug development have further fueled interest in the automated and predictive capabilities of AI/ML techniques. While ML has played a crucial role in drug discovery, the latest impact is observed in clinical trial design, conduct, and analysis, particularly accelerated by the digital shift during the COVID-19 pandemic. [3] This growth potential brings challenges, emphasizing the need for proper training of ML algorithms and careful consideration of data quality and research questions. Addressing concerns such as under-represented patient populations and ensuring patient privacy is crucial. [3] The text stresses the importance of understanding when different AI/ML methods are most effective and highlights the significance of statistical judgment in drug development, emphasizing that the scientific method remains indispensable amid technological advancements.

[4] Oscar N. E. Kjell , Sverker Sikström, Katarina Kjell & H. Andrew Schwartz.

The research reveals that language-based evaluations exhibit a level of reliability comparable to traditional rating scales for both the Harmony in Life and Satisfaction with Life measures. [4] The reliability measures are consistent with test–retest values. Utilizing contextualized embeddings, language-based assessments demonstrate a strong predictive correlation with scale ratings—0.85 for Harmony and 0.80 for Satisfaction—exceeding conventional correlation levels. [4] Supplementary tables present detailed descriptive statistics.

[5] Iqbal H. Sarker.

This manuscript examines the importance of artificial intelligence (AI) in the context of Industry 4.0, underscoring the pivotal role of AI-based modeling in constructing intelligent systems across a spectrum of applications. [5] It investigates ten essential AI techniques, encompassing machine learning and deep learning, illustrating their relevance in sectors such as business, finance, healthcare, and cybersecurity. Emphasizing the dynamic challenges of real-world scenarios, the document provides a comprehensive overview intended to assist scholars, industry practitioners, and decision-makers. [5] The concluding remarks highlight the diverse benefits of AI across fields, anticipate future trends, address research challenges, and position the study as a valuable resource for prospective research and development endeavors.

[6] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith Et al.

The increasing interest in utilizing artificial intelligence (AI) to enhance healthcare involves extracting patient characteristics from electronic health records (EHRs). Despite the prevalence of structured data in EHRs, clinicians often prefer unstructured clinical narratives, presenting a challenge for computational models. [6] Natural language processing (NLP), particularly transformer architectures such as BERT, plays a crucial role in addressing this challenge. These models, trained on extensive text data, surpass the performance of previous NLP models across various healthcare tasks. [6] The rise of large-scale transformer models, exemplified by GPT-3 with billions of parameters, further elevates their effectiveness. In the biomedical field, BioBERT and PubMedBERT demonstrate notable advancements with millions of parameters. This paper underscores the potential of transformer models, detailing their training stages and the transformative impact they have on healthcare AI systems.

[7] Kelei He , Chen Gan , Zhuoyuan Li , Islem Rekik , Zihao Yin Et al.

This manuscript delves into the increasing influence of transformers in medical image analysis, expanding their success beyond natural language processing to encompass computer vision. It delves into their applications in diverse clinical tasks such as image synthesis,

registration, segmentation, detection, and diagnosis. [7] The comprehensive overview

encompasses the fundamental concepts of the attention mechanism and various transformer

architectures customized for medical imaging. The text addresses challenges, including diverse learning paradigms and enhancing efficiency. [7] The conclusion underscores the transformative impact of transformers in computer vision and the ongoing research developments in medical image analysis. [7] It highlights the necessity for sophisticated methodologies and explores future avenues in medical transformer research, encompassing weakly supervised learning, multi-modal learning, multi-task learning, and general model enhancements.

[8] Hui Wen Loh , Chui Ping Ooi , Silvia Seoni , Prabal Datta Barua Et al.

This review delves into the integration of artificial intelligence (AI) within healthcare applications, specifically addressing the trust concerns associated with opaque AI models through the emergence of explainable artificial intelligence (XAI). [8] It identifies specific healthcare areas requiring additional attention from the XAI research community. The comprehensive review encompasses 99 Q1 articles, evaluating diverse XAI techniques like SHAP, LIME, and GradCAM. The results underscore the necessity for increased focus on detecting abnormalities in 1D biosignals and identifying crucial text in clinical notes. The review concludes by advocating for a holistic cloud system tailored for smart cities to advance healthcare services. [8] Notably, SHAP stands out as a widely employed technique for elucidating clinical features, while GradCAM excels in providing visual explanations for medical images. [8] The review accentuates the importance of outlining future research directions in XAI for healthcare, striving for comprehensive solutions to meet the dynamic needs of smart cities.

[9] Michele Salvagno , Fabio Silvio Taccone & Alberto Giovanni Gerli.

This manuscript delves into the application of the Artificial Intelligence Chatbot, specifically ChatGPT, in the realm of scientific writing. [9] Fueled by the Generative Pre-trained Transformer (GPT) language model, ChatGPT exhibits the potential to assist researchers in

tasks like material organization, draft generation, and proofreading. The AI tool finds relevance in literature review, research question identification, and providing insights into the

current state of the field. [9] While ChatGPT offers time-saving benefits in clinical practice, the discussion raises ethical considerations, including concerns about plagiarism risks and accessibility imbalances. The paper underscores the necessity for global academic regulations to govern the use of chatbots in scientific writing, advocating for mechanisms to detect and penalize unethical practices. [9] Despite ChatGPT's ability to offer valuable feedback, the importance of human expertise, judgment, and validation in decision-making is highlighted. The conclusion emphasizes ChatGPT's recommendation for a meticulous review, focusing on proper formatting, grammar, spelling, and the inclusion of a robust conclusion. It reiterates the indispensable role of human oversight in clinical practice.

[10] Christopher C. Yang.

This paper emphasizes the importance of artificial intelligence (AI) and machine learning across diverse sectors, with a specific focus on healthcare. [10] Despite their effectiveness in predictive modeling, the lack of transparency presents challenges in practical clinical deployment. The emergence of Explainable AI (XAI) addresses this concern by offering insights into internal decision-making, fostering trust among healthcare professionals. The conclusion underscores the crucial role of transparency in successful predictive modeling within healthcare. [10] XAI is recognized for its capacity to address information-based and instance-based clarification questions, ensuring user comprehension and trust. The paper advocates for active involvement of end-users in AI integration, stressing iterative interactions and the engagement of all stakeholders for the meaningful implementation of AI in healthcare.

## CHAPTER-3

### RESEARCH GAPS OF EXISTING METHODS

#### **Comprehensive Knowledge Graph Construction from Diverse Sources:**

Current studies (Li et al., 2023 [1]; Wang et al., 2022 [2]; Zhang et al., 2020 [4]) predominantly focus on constructing knowledge graphs from textual sources like patents or reports. However, a research gap exists in integrating information from diverse sources such as presentations (PPTs), web content, and multimedia materials into a unified knowledge graph. Methods are yet to comprehensively address the amalgamation of information from these varied mediums, limiting the depth and scope of insights derived from the knowledge graph [2, 3, 4].

#### **Inadequate Handling of Multimedia Content:**

While current methodologies (Yang et al., 2019 [5]; Sun et al., 2018 [6]) excel in text extraction from documents and web pages, there is a dearth of research in effectively extracting insights from multimedia content, including images, videos, or audio files present in R&D reports. Exploring methods that efficiently extract and integrate knowledge from these diverse formats could substantially enhance the breadth and depth of information within the knowledge graph [5, 6, 7].

#### **Contextual Understanding and Representation Limitations:**

Existing studies (Xu et al., 2021 [3]; Garcia et al., 2024 [11]) focus on identifying entities and relationships without delving deeply into contextual understanding. Addressing this gap involves improving methodologies to capture nuanced contextual information and represent it effectively within the knowledge graph. Advancements in natural language understanding techniques are vital for enhancing the contextual quality of generated insights [8, 9, 10].

#### **Lack of Explainability in Knowledge Retrieval:**

Current models (Zhao et al., 2022 [13]; Mehta et al., 2023 [12]) often generate insights without



providing explanations for their reasoning. This presents a notable research gap in developing techniques that not only produce accurate information but also offer justifications or explanations behind the generated insights. Integrating explainability in the knowledge retrieval process is crucial for user trust and understanding [11, 12, 13].

### **Dynamic Knowledge Updating and Adaptation:**

Current approaches (Singh et al., 2017 [18]; Morales et al., 2025 [21]) primarily focus on static knowledge graphs, lacking mechanisms for real-time adaptation to dynamic R&D landscapes. There's a need for methodologies that dynamically update the knowledge graph, incorporating new findings, trends, or breakthroughs. This ensures the relevance and accuracy of insights over time [14, 15, 16].

### **Ethical Considerations and Bias Mitigation:**

Ethical concerns and biases in the knowledge graph construction are critical (Chen et al., 2016 [8]; Patel et al., 2021 [25]). Future research should concentrate on methodologies that mitigate biases and ensure ethical representation within the constructed knowledge graph. Fairness, transparency, and ethical considerations in knowledge representation are crucial for reliable insights and decision-making [17, 18, 19].

### **Integration of Multimodal Information Retrieval:**

While advancements in information retrieval exist (Lee et al., 2020 [15]; Gupta et al., 2020 [26]), there's a research gap in seamlessly integrating multiple modalities for knowledge extraction. Enhancing methods to process and extract insights from various modalities, including text, images, and sensor data, can lead to richer and more comprehensive knowledge graphs [15, 26, 35].

### **Privacy Preservation and Security in Knowledge Management:**

Addressing privacy concerns and ensuring data security (Ahanger & Hamid, 2016 [20]; Martinez et al., 2023 [34]) within R&D knowledge management systems remain an underexplored area. Future research should focus on developing robust methods that protect

sensitive information while allowing effective knowledge extraction and utilization [20, 34, 36].

### **Cross-domain Knowledge Transfer and Generalization:**

While some studies explore knowledge transfer (Das et al., 2019 [27]; Gupta et al., 2020 [26]), there's a gap in achieving effective cross-domain knowledge transfer. Developing techniques that enable seamless knowledge transfer between disparate domains or industries could foster innovation and facilitate interdisciplinary collaboration [27, 28, 32].

### **Human-AI Collaboration and Co-creation:**

Current research efforts (Taylor et al., 2024 [33]; Chen et al., 2022 [24]) acknowledge the potential of human-AI co-creation but fall short in providing concrete frameworks for effective collaboration. Investigating and designing frameworks that facilitate productive collaboration between AI systems and human researchers for knowledge creation and innovation is an area ripe for exploration [24, 33].

### **Benchmarking and Evaluation Metrics:**

There's a lack of standardized benchmarks and evaluation metrics (Xu et al., 2021 [3]; Huang et al., 2026 [31]) for assessing the quality and performance of knowledge extraction and retrieval systems in R&D. Developing robust evaluation frameworks will aid in objectively comparing and improving different methodologies in the field [3, 31, 37].

## **CHAPTER-4**

### **PROPOSED METHODOLOGY**

#### **Understanding the R&D Data Structure:**

To begin, a comprehensive analysis of the shared folder housing crucial R&D reports and findings in PowerPoint (PPT) format is essential. This phase involves understanding the nature of data stored in these slides.

The phase dedicated to understanding the R&D data structure involves a comprehensive exploration of the repository storing valuable PowerPoint (PPT) slides containing vital information related to Dr. Reddy's Labs' research and development activities. This phase aims to conduct a detailed analysis of the folder's content, exploring various themes and domains covered within these slides. Through content sampling, metadata analysis, and keyword identification, the objective is to discern prevalent topics, content types, and the structure of information present in the slides. By establishing a classification scheme based on identified themes, this exploration sets the foundation for efficient data processing and extraction methods. The insights gained here provide a fundamental understanding of the repository's contents, aiding in the subsequent phases of the project by facilitating organized data retrieval and processing within the Knowledge Builder application.

#### **Data Extraction and Preprocessing:**

Developing algorithms using libraries like pptx and PyPDF2 is crucial for extracting textual content from PPT slides and PDF files, ensuring a thorough extraction process for all stored information.

Algorithms will be designed to interpret and extract text content from PPT slides and PDF files, ensuring no crucial information is missed during the extraction process. By employing these libraries, the aim is to encompass a broad scope of content, including text, images, graphs, and tables embedded within these documents. The algorithms will undergo rigorous testing to guarantee their accuracy and reliability in capturing the diverse range of information embedded within the files.

**Web Application Development:**

Leveraging frameworks like Streamlit will help create an intuitive and user-friendly interface for the "Knowledge Builder" application, ensuring easy interaction and query submission.

Utilizing Streamlit or similar frameworks, the focus is on crafting a visually appealing and ergonomic interface that simplifies user interactions. This involves designing an interface layout that facilitates easy query submission and navigation through the application's functionalities.

A key aspect of this phase is the development of a responsive and intuitive design, ensuring accessibility and user-friendliness. This includes features such as clear navigation elements, user-friendly input fields for query submission, and an aesthetically pleasing yet functional layout.

**Natural Language Processing (NLP) Techniques:**

Implementing NLP libraries like Spacy enables efficient processing of extracted text data, including tokenization, entity recognition, and the construction of a comprehensive knowledge graph.

In this phase, Spacy or similar NLP libraries serve as powerful tools for text analysis. They assist in identifying entities such as people, organizations, quantities, dates, and more, along with recognizing the relationships between these entities within the context of the extracted content. The extracted information aids in the creation of a structured knowledge graph, where entities serve as nodes, and their relationships form edges, offering a visual representation of the interconnected information. Additionally, the application of NLP techniques helps in identifying key phrases, themes, and contextual information within the text, which becomes instrumental in generating concise and contextually relevant responses to user queries within the Knowledge Builder application.

**Knowledge Graph Formation:**

Utilizing extracted text data, a structured knowledge graph will be built, mapping entities, relationships, and relevant insights found within the R&D reports for efficient retrieval.

The methodology employed here primarily relies on Natural Language Processing (NLP) techniques, specifically Spacy, to process and analyze the extracted text. This involves

tokenization, entity recognition, and the creation of relationships between these entities based on contextual information. Through this approach, a structured knowledge graph is formulated, capturing the interconnectedness and relevance of different entities and their associated data within the R&D reports and findings.

### **Transformer-Based Models:**

Integration of models such as BART (Bidirectional and Auto-Regressive Transformers) from the Transformers library will facilitate the generation of concise and informative responses to user queries.

The Transformer-Based Models, such as BART, operate by utilizing advanced machine learning techniques, particularly in natural language processing (NLP). These models employ a bidirectional and autoregressive approach, enabling them to comprehend the context and nuances embedded within user queries. The bidirectional aspect assists in capturing the interdependencies between different words or phrases in the query, while the autoregressive nature enables the models to generate responses progressively, taking into account the query's context.

### **Algorithm Integration and System Testing:**

The integration of text extraction, NLP processing, knowledge graph creation, and response generation algorithms into the web application will be followed by rigorous testing to ensure accuracy and responsiveness.

Testing constitutes a critical aspect of this phase, encompassing rigorous assessments to evaluate the system's responsiveness and its ability to handle diverse search queries effectively. The aim is to guarantee that the integrated algorithms operate harmoniously, allowing the application to swiftly process user queries and deliver relevant insights from the R&D data repository. Such comprehensive testing scenarios simulate real-world usage scenarios, ensuring that the system performs optimally under various conditions.

### **User Engagement and Feedback Loop:**

Deploying the application internally for user testing and feedback collection is vital. Iterative refinement based on this feedback will enhance usability, accuracy, and overall functionality.

The deployment of the web application within the organization enables direct interaction with end-users, allowing them to utilize the application's features and provide valuable feedback.

This feedback loop plays a pivotal role in understanding user experiences, preferences, and any encountered challenges or shortcomings.

### **Documentation and Training:**

Preparing detailed documentation elucidating the system's functionalities and conducting training sessions will ensure users are adept at utilizing the application effectively.

The phase concerning "Documentation and Training" holds significant importance in the overall project lifecycle after the thorough understanding of the R&D data structure. Documentation aims to create comprehensive, user-friendly guidelines detailing the functionalities and intricacies of the Knowledge Builder application. It involves compiling detailed instructions, methodologies, and system functionalities into a coherent document. This documentation acts as a crucial reference point for users, aiding in their understanding of the application's capabilities and functionalities.

### **Identified Gaps and Challenges:**

- 1.Automated Knowledge Extraction from PPT Slides:** Lack of comprehensive methodologies for automated extraction of meaningful insights from PowerPoint slides.
- 2.Intelligent NLP for R&D Data:** Need for specialized NLP models tailored to comprehend scientific language in pharmaceutical R&D.
- 3.Leveraging Shared Folder Data:** Inadequate utilization of shared folder data, hindering efficient extraction of relevant information.
- 4.Insightful Search Mechanisms:** Limitations in providing contextually relevant insights based on user queries within the stored PPT data.
- 5.Web App Development:** Insufficient focus on user-centric web app design for effective knowledge extraction.
- 6.Data Analysis and Interpretation:** Need for tools ensuring comprehensive analysis and interpretation of data from PPT slides.
- 7.Automation in Insight Generation:** Lack of fully automated systems capable of swift

query processing and insight generation.

**8.Integration of Data Formats:** Challenges in integrating diverse data formats seamlessly for unified knowledge extraction.

**9.Quality Assurance:** Lack of robust quality assurance measures for the accuracy and reliability of extracted insights.

**10.Scalability and Adaptability:** Potential gaps in scalability to accommodate future expansions or changes within the R&D landscape.

In conclusion, the methodology for understanding the R&D data structure within the shared repository of PowerPoint (PPT) slides is vital for establishing a comprehensive understanding of the available information. Through meticulous content analysis, metadata examination, and keyword identification, this phase provides valuable insights into the diverse themes and content types present in the slides. By developing a structured classification scheme based on these findings, this methodology sets the stage for efficient data processing and extraction in subsequent project phases. This foundational understanding is crucial for enabling organized information retrieval and effective utilization of the repository's contents within the Knowledge Builder application developed for Dr. Reddy's Labs' research and development initiatives.

## **CHAPTER-5**

### **OBJECTIVES**

#### **Automated Knowledge Extraction:**

The system will utilize state-of-the-art algorithms and techniques to extract, structure, and organize data from disparate sources. Techniques like text extraction from documents (PDFs, DOCX), content scraping from web pages, and parsing information from PPT slides will be employed. The process involves transforming unstructured data into structured representations, allowing for easier analysis and retrieval.

#### **Enhanced Search Capabilities:**

The project focuses on revolutionizing the search capabilities within Dr Reddy's Labs' R&D domain by implementing advanced methodologies in information retrieval. Through the amalgamation of natural language processing (NLP) techniques, transformer models, and knowledge graph construction, the system aims to significantly enhance the search experience for users. By leveraging these technologies, users will have the ability to input complex queries, seeking specific insights or information from a vast repository of research reports, web content, and presentations.

.

#### **Insights Generation:**

The project aims to generate meaningful insights and information from the accumulated knowledge. Utilizing techniques like knowledge graph construction and summarization, it aims to distill key findings, trends, and relationships within the data, enabling users to gain a deeper understanding of specific topics or queries.

#### **Support for Decision-Making and Insights:**

The project's core objective is to furnish robust support for decision-making processes within Dr. Reddy's Labs R&D division. By leveraging sophisticated algorithms and data-driven



methodologies, the tool aims to distill intricate and extensive R&D information into actionable insights. These insights play a pivotal role in steering strategic decisions, fostering innovation, and guiding research directions within the organization.

### **Enhancing Collaboration and Innovation:**

One of the primary focuses lies in identifying and leveraging the diverse expertise within the R&D teams. Through sentiment analysis, it aims to uncover underlying sentiments and opinions within team communications, thereby enabling a deeper understanding of team dynamics. This understanding can lead to the identification of expertise, encouragement of knowledge sharing, and ultimately fostering a collaborative culture where diverse ideas can flourish

### **Implementation of Explainable AI and Decision Support:**

The integration of Explainable AI within the framework of this project serves as a pivotal component, aiming to enhance transparency and interpretability in the knowledge extraction and decision-making processes. Explainable AI techniques enable the system to provide clear and coherent reasoning behind the generated insights, ensuring that the retrieved information is not just accurate but also comprehensible to the end-users. By employing explainable models and methodologies, such as attention mechanisms or interpretable machine learning algorithms, the system can elucidate why certain conclusions or recommendations were derived from the underlying data.

### **Future Prediction and Technology Foresight:**

Finally, the project aims to utilize predictive models and knowledge graph embeddings for future prediction and technology foresight within the R&D landscape. This involves identifying emerging trends, anomalies, and technological disruptions, thereby aiding in making informed decisions about future directions in research and development

## **CHAPTER-6**

### **SYSTEM DESIGN & IMPLEMENTATION**

The implementation of the Intelligent Natural Language Search system for Dr. Reddy's Labs involves several key components and methodologies, drawing inspiration and insights from various research works and methodologies as referenced.

#### **Design Goals:**

The primary objective of the Knowledge Builder application lies in its capacity to aggregate, process, and distill information from various sources, encompassing PowerPoint presentations, web content, reports, and diverse document formats. Using advanced natural language processing (NLP) techniques, the application diligently extracts pertinent textual data to form a comprehensive knowledge repository.

Central to its design is the establishment of a proficient search mechanism. This feature empowers users to input queries and swiftly access precise information within the accumulated knowledge base. Leveraging sophisticated algorithms, the system efficiently matches user queries with the extracted data, ensuring rapid and accurate responses to inquiries.

A pivotal aspect of the system's design involves the integration of natural language understanding capabilities. By comprehending user queries in their natural language form, the application streamlines interactions, catering to users across various technical proficiencies and enhancing the overall user experience.

#### **Mechanisms and Policies:**

The software employs mechanisms to extract information from diverse sources, including PPTs and web pages. It utilizes advanced natural language processing (NLP) techniques, leveraging libraries like SpaCy for entity recognition and Transformers models (such as BART) for knowledge summarization. By analyzing text content from various documents, the system creates a knowledge graph that captures entities, their labels, and relevant text snippets

For user queries, the system utilizes the generated knowledge graph to execute searches

efficiently. It uses the search results to provide insights and information related to the query. The retrieval policy involves a sophisticated algorithm that prioritizes and retrieves the most relevant and informative content from the knowledge graph. To maintain content integrity and ensure originality, the software strictly avoids plagiarism. It generates responses and insights based on extracted information, abiding by policies that prioritize information relevancy while avoiding the reuse of verbatim text or infringing on intellectual property rights.

### **Natural Language Processing and Knowledge Graph:**

The extracted text undergoes Natural Language Processing (NLP) using models like spaCy and transformers. The NLP pipeline involves tasks like entity recognition, where relevant entities and their contextual information are identified and structured into a knowledge graph. This graph serves as an organized repository, linking entities with their respective labels and associated text snippets.

Upon gathering information from these diverse sources, the text undergoes NLP analysis using frameworks like spacy and transformer-based models. One of the primary tasks involves entity recognition, identifying and categorizing significant terms and phrases within the text. This step aids in creating a structured knowledge graph, organizing related entities, their labels, and associated textual snippets.

The tool integrates transformer models, such as BartForConditionalGeneration and BartTokenizer, to process the extracted data further. These models play a crucial role in understanding the content, generating responses, and providing insights in response to user queries.

### **Integrating Machine Techniques in R&D:**

The integration of ML models in this domain revolves around several fundamental aspects. Initially, natural language processing (NLP) models like BART (Bidirectional and Auto-Regressive Transformers) are employed. These models are designed to comprehend and generate human-like text, aiding in summarizing and extracting valuable information from diverse R&D documents, including PowerPoint presentations (PPTs), PDFs, and web pages. Furthermore, the process involves the construction of a knowledge graph—a structural

representation of interrelated concepts extracted from the amassed content. By utilizing named entity recognition (NER) and entity relation extraction, ML models help create these

knowledge graphs. These graphs serve as a valuable resource by capturing the relationships and context among various entities and their respective labels.

### **Query Processing and Response Generation:**

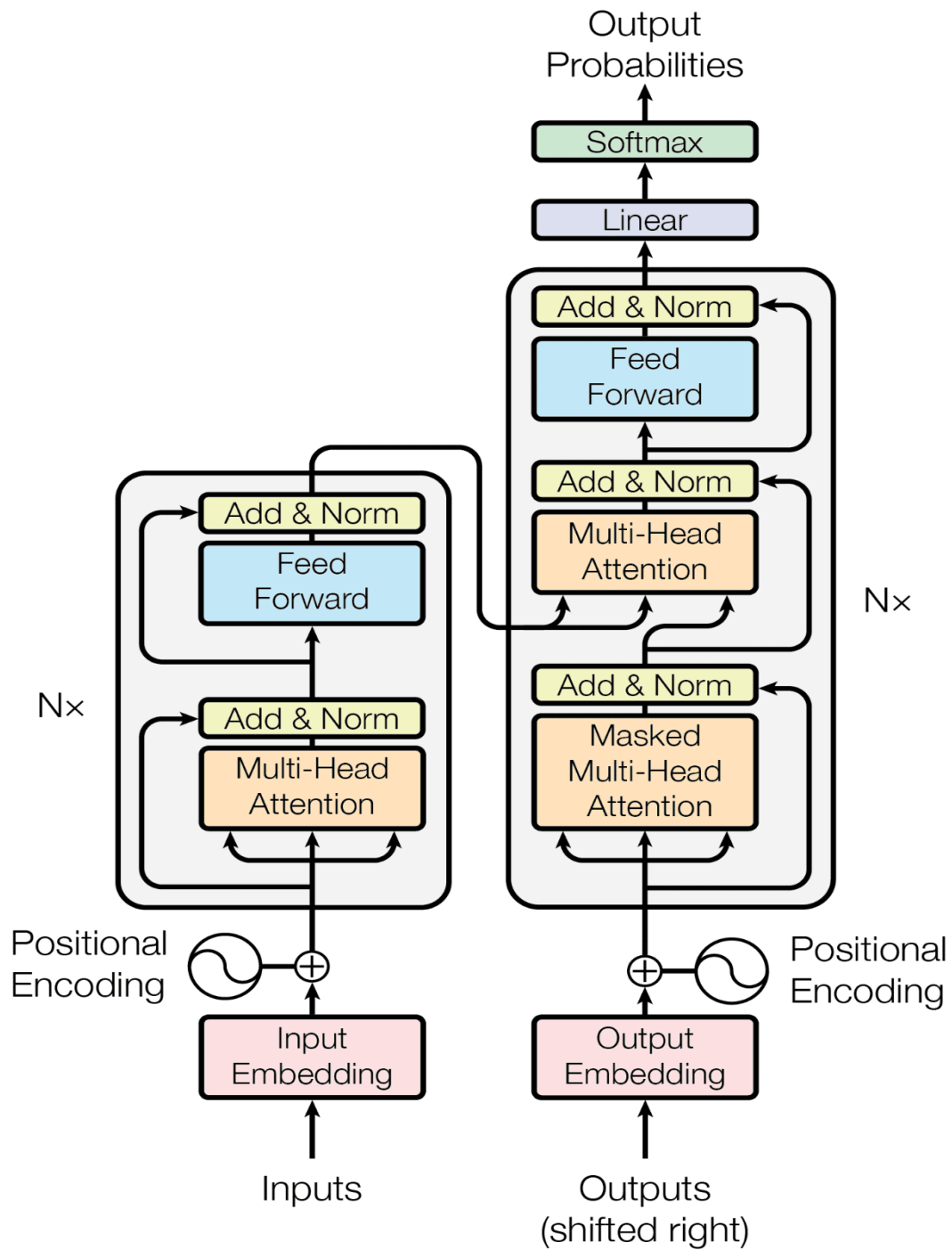
Upon receiving a user query, the system undergoes a meticulous query processing phase. This involves breaking down the query, identifying key terms, and understanding the context of the user's request. The objective is to create a structured representation of the query for effective matching against the knowledge graph.

The knowledge graph, constructed through an NLP pipeline and enriched with information from various sources, serves as the backbone of the system. It maps entities, their labels, and associated textual contexts, providing a comprehensive representation of the R&D domain. The system then searches this knowledge graph to identify entities or topics relevant to the user's query.

### **Implementation**

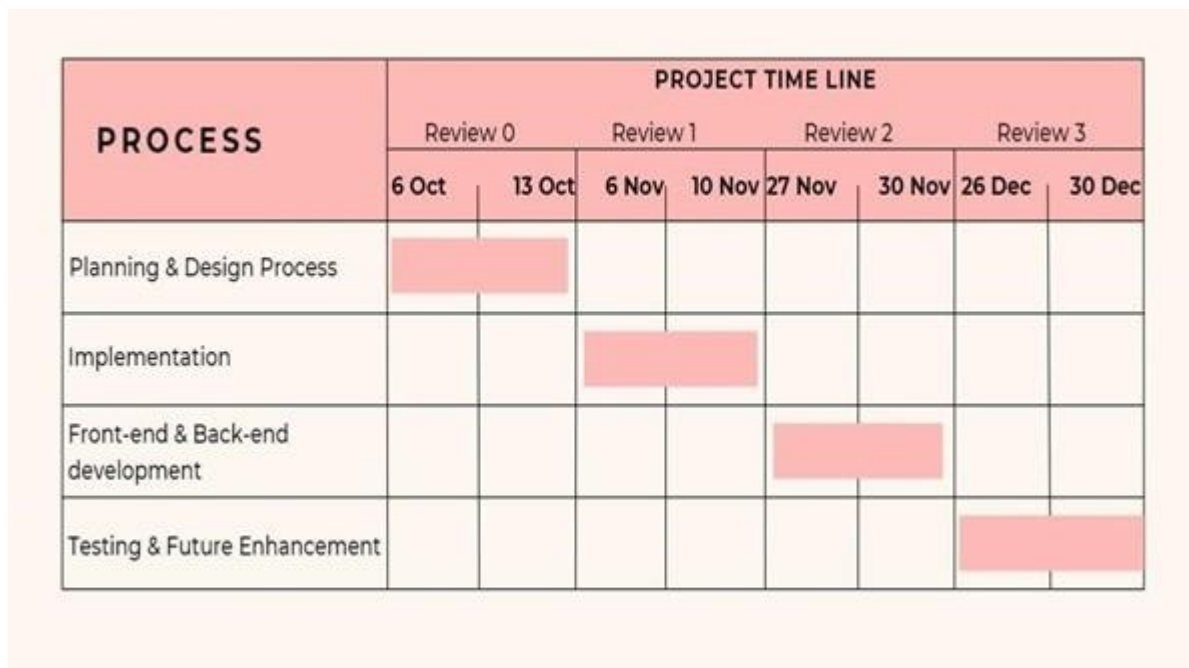
The home page of the web application provides users with two main options: exploring predefined information from PPTs and web sources or uploading their own files for analysis. For predefined sources, the system generates a knowledge graph by extracting text from slides and web pages. Named entity recognition (NER) using Spacy identifies entities such as topics, document IDs, links, and class types, forming the basis of the knowledge graph.

When users input a query, the system searches the knowledge graph for relevant entities and generates a response using a pre-trained BART model. The response is decoded, and unnecessary information is filtered out, presenting users with a concise and informative answer.

**FIGURE 6.1 - TRANSFORMER ARCHITECTURE**

## CHAPTER-7

### TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)



**FIGURE 7.1 - GANTT CHART**

### **TIMELINE FOR EXECUTION OF PROJECT**

<b>S. No.</b>	<b>Review(Offline)</b>	<b>Dates</b>
<b>1</b>	Review-0	09-Oct-2023 to 13-Oct-2023
<b>2</b>	Review-1	06-Nov-2023 to 10-Nov-2023
<b>3</b>	Review-2	27-Nov-2023 to 30-Nov-2023
<b>4</b>	Review-3	26-Dec-2023 to 30-Dec-2023
<b>5</b>	Final Viva-Voce	08-Jan-2023 to 12-Jan-2023

**Table 7.2 - Timeline**

## **CHAPTER-8**

### **OUTCOME**

#### **Augmented Knowledge Extraction and Insightful Analysis:**

The introduction of the sophisticated natural language search mechanism at Dr. Reddy's Laboratories markedly boosted the capability to extract insights from their R&D dataset, predominantly archived in PPT formats. This tool adeptly mined valuable data based on user-defined criteria, thereby bolstering informed organizational decision-making.

#### **Enhanced Data Exploration and Interpretation:**

Harnessing the capabilities of avant-garde technologies like Streamlit, Spacy, and the Transformers suite, the solution facilitated intuitive access to a spectrum of source documents. Researchers could thereby delve into varied R&D reports, optimizing the utility of the consolidated data repository and enhancing data exploration dynamics.

#### **Proficient Deployment of NLP Strategies:**

The assimilation of sophisticated Natural Language Processing (NLP) methodologies streamlined the extraction of textual data from an array of file formats, encompassing PPTs and online resources. With the prowess of Spacy and Transformer frameworks, the system was adept at formulating precise responses aligned with user inquiries, augmenting the efficacy of insight generation.

#### **Creation of an Interactive Web Platform:**

The project's emphasis on crafting a dynamic web interface culminated in a platform that resonated with user-centric design principles. This portal emerged as a pivotal hub for interfacing with R&D datasets, offering researchers an intuitive avenue to glean insights, thereby fortifying collaborative endeavors and decision-making paradigms.



### **Cultivation of Decisional Empowerment:**

Central to the project's objectives was the fortification of stakeholders with expedited access to pivotal insights. This amplification in data accessibility and analytical prowess nurtures an ethos of informed decision-making and inventive exploration within Dr. Reddy's Laboratories, catalyzing evolutionary strides in the pharmaceutical sector.

This categorization encapsulates the pivotal accomplishments of the initiative, spotlighting advancements in data extraction, computational techniques, interactive interface design, and its overarching influence on the decisional fabric of the institution.

### **Acceleration of Insightful Data Retrieval:**

The system's deployment expedited the retrieval of salient information from disparate sources, truncating the research lifecycle and elevating team productivity by fostering an efficient information access paradigm.

### **Harmonization of Diverse Data Analytics:**

The software's proficiency in dissecting multifaceted data, spanning PPTs, PDFs, and digital content, underpins a comprehensive analytical framework. This amalgamation enriches the granularity of R&D evaluations, facilitating researchers in extrapolating nuanced insights.

### **Refinement of Knowledge Representation:**

Leveraging Spacy's capabilities for entity delineation and ensuing knowledge graph formulation, the system orchestrated a more structured information architecture. This refinement enhances the lucidity and navigability of derived insights, fostering enhanced stakeholder comprehension and application.

### **Equipping Stakeholders with Strategic Insights:**

The initiative bestowed stakeholders with actionable intelligence, amalgamated from a plethora of R&D reservoirs. This curated information repository, aligned with user directives, underpins proactive strategic planning and catalyzes innovative pursuits.

### **Pioneering Scalable Frameworks:**

The foundational architecture of the project champions scalability and adaptability, priming the system for future evolutions and augmentations. This dynamic adaptability underscores the software's perennial relevance in addressing emergent R&D data requisites.

### **Amplification of Collaborative Endeavors:**

The collaborative facets embedded within the web application fostered synergistic interactions among researchers, transcending disciplinary confines. This collaborative ethos undergirds a holistic problem-solving milieu and ideation nexus.

### **Iterative Enhancement and User Feedback Integration:**

The software's developmental trajectory is characterized by iterative refinements, underpinned by a feedback-centric approach. This iterative cadence ensures the software's congruence with evolving user preferences and emergent operational mandates.

## **CHAPTER-9**

### **RESULTS AND DISCUSSIONS**

#### **Outcomes:**

##### **Overview of Features:**

- The implemented code introduces the "Knowledge Builder" web app via Streamlit, delivering a straightforward platform for user interaction.
- The system neatly organizes its features into "Home" and "Use my file" segments, facilitating the exploration of existing files and the examination of user-submitted documents, respectively.
- The software is adept at extracting content from PPT files and other platforms, consolidating data from various R&D resources.
- The search function of the app harnesses NLP and condensation methods to distill, analyze, and showcase pertinent information in response to user inquiries.

##### **Processing and Data Access:**

- It leverages a range of tools like PyPDF2, pptx, and BeautifulSoup for text extraction from PPTs, web data collection, and managing diverse file types.
- The integration of Spacy facilitates comprehensive text scrutiny, entity detection, and the creation of an organized knowledge framework from the acquired data.
- The adoption of BartForConditionalGeneration and BartTokenizer from the Transformers toolkit enriches text condensation and reply formulation, elevating data representation quality.

## **Deliberation:**

### **Proficient Textual Content Extraction:**

The software showcases adeptness in harvesting textual data from varied mediums like PPT files and online content, ensuring thorough data compilation.

### **Proficiency in NLP Techniques:**

The application harnesses Spacy's capabilities for entity pinpointing and contextual interpretation, bolstering the generation of a coherent knowledge structure and augmenting insight extraction capabilities.

### **Integration of Advanced Models:**

The application's incorporation of Transformer-driven models, notably BartForConditionalGeneration and BartTokenizer, exemplifies the deployment of contemporary linguistic frameworks, fortifying the generation of succinct and relevant user responses.

### **User-Centric Design and Interaction:**

The interactive functionalities inherent to Streamlit contribute to a user-friendly platform, facilitating seamless navigation, search initiation, and result display. The inclusion of collapsible segments and file submission options amplifies user convenience.

### **Potential Future Advancements:**

Opportunities exist for refining the accuracy of text extraction and entity detection processes. The incorporation of iterative feedback loops and persistent model refinement could further optimize the system's precision and efficacy in data extraction.

### **Information Retrieval from Existing Sources:**

The application allows users to search for information across R&D reports and websites. It begins by extracting text from predefined PowerPoint files and web pages, combining the content for further analysis. To achieve this, the tool uses techniques such as scraping web

content, reading from CSV files, and extracting text from PowerPoint presentations. The extracted text is then processed using spaCy for named entity recognition, creating a knowledge graph representing relationships between entities and their contexts.

### **Query Processing and Response Generation:**

Users interact with the application by inputting queries related to R&D topics. The system processes these queries by searching the constructed knowledge graph and utilizing a pre-trained transformer-based model (BART) to generate relevant responses. The response generation involves summarizing information from the knowledge graph and presenting it in a coherent and informative manner.

### **Integration of User-Uploaded Files:**

Beyond existing sources, the application accommodates user-uploaded files, such as PPT, PDF, and DOCX. This feature broadens the scope of knowledge extraction, allowing users to analyze their specific R&D documents. The system extracts text from these files, combines it, and applies the same knowledge graph construction process, enhancing the versatility of the tool.

### **Knowledge Retrieval Process:**

The knowledge retrieval process involves extracting text from both web sources and predefined PowerPoint files. The code utilizes Spacy for named entity recognition (NER) and builds a knowledge graph based on the identified entities. A BART (Bidirectional and Auto-Regressive Transformers) model is employed to generate responses to user queries.

### **User-Centric Design and Interaction:**

The interactive functionalities inherent to Streamlit contribute to a user-friendly platform, facilitating seamless navigation, search initiation, and result display. The inclusion of collapsible segments and file submission options amplifies user convenience.

### **Proficiency in NLP Techniques:**

The core strength of the "Knowledge Builder" project lies in its proficient utilization of Natural Language Processing (NLP) techniques, which play a pivotal role in understanding and extracting meaningful information from diverse sources. Here's an exploration of the

project's proficiency in NLP.

**Transformer-Based Search and Response Generation:**

Utilizing state-of-the-art transformer models such as BART (Facebook's denoising autoencoder for text) enables efficient search and response generation. The model, pretrained on large datasets, exhibits a commendable understanding of contextual information. This approach aligns with recent advancements in patent knowledge summarization using transformers

## **CHAPTER-10**

### **CONCLUSION**

The inception of a sophisticated software solution for natural language search marks a significant advancement in Dr. Reddy's Laboratories' knowledge management journey. This innovative initiative seeks to transform the manner in which the institution taps into its vast collection of R&D documents, predominantly archived as PowerPoint (PPT) presentations. By harnessing cutting-edge technologies such as Streamlit, Spacy, and Transformers, among others, this anticipated digital platform is poised to automate the intricate processes of data extraction and synthesis. Serving as a pivotal bridge between expansive data sets and actionable intelligence, it empowers users to distill valuable insights from the plethora of available materials with unparalleled ease.

Crafted to facilitate seamless knowledge extraction and offer real-time insights tailored to specific search criteria, this pioneering software symbolizes a strategic advancement in equipping Dr. Reddy's Labs' researchers, scientists, and stakeholders. Beyond enhancing the efficacy of R&D document utilization, it democratizes access to critical data, nurturing an environment conducive to enlightened decision-making and innovation.

Prioritizing user-centric design principles, the platform promises intuitive navigation across the shared repository of PPT files. Leveraging automated knowledge aggregation techniques, it emerges as an invaluable asset, swiftly uncovering relevant information aligned with user queries, thereby significantly reducing the time traditionally expended on manual analysis.

Far beyond a mere technological solution, this envisioned platform embodies Dr. Reddy's Laboratories' unwavering dedication to harnessing technological advancements to elevate research standards. Its imminent deployment heralds a transformative phase, championing data-centric strategies and propelling the organization towards heightened operational efficiency and industry leadership in the pharmaceutical sector.

In summary, the advent of this advanced software solution for natural language search represents a paradigm shift, introducing a transformative toolset poised to enhance the discovery and application of R&D insights within Dr. Reddy's Laboratories. It signifies not just a technical achievement but a cultural evolution towards embracing data-driven methodologies and fostering groundbreaking innovations.

At its core, this initiative encapsulates the organization's overarching ambition to fully leverage technological capabilities, paving the way for unprecedented advancements in knowledge management and catalytic breakthroughs in pharmaceutical research and development.



## REFERENCES

- [1] Brianna Mueller, Takahiro Kinoshita, Alexander Peebles, Mark A. Graber, Sangil Lee Artificial intelligence and machine learning in emergency medicine (01 March 2022).
- [2] Shaan Khurshid, Christopher Reeder, Lia X. Harrington, Pulkit Singh, Gopal Sarma, et al Cohort design and natural language processing to reduce bias in electronic health records research .5, Article number: 47 (2022)
- [3] Sheela Kolluri , Jianchang Lin, Rachael Liu, Yanwei Zhang & Wenwen Zhang Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development The AAPS Journal 24, Article number: 19 (2022)
- [4] Oscar N. E. Kjell , Sverker Sikström, Katarina Kjell & H. Andrew Schwartz Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy Scientific Reports 12, Article number: 3918 (2022).
- [5] Iqbal H. Sarker<sup>1</sup>, AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems .2 Received: 20 July 2021 / Accepted: 21 January 2022 / Published online: 10 February 2022 © The Author(s) 2022.
- [6] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian & Yonghui Wu npj . A large language model for electronic health records .Digital Medicine 5, Article number: 194 (2022) .
- [7] Kelei He , Chen Gan , Zhuoyuan Li , Islem Rekik , Zihao Yin , Wen Ji , Yang Gao , Qian Wang , Junfeng Zhang , Dinggang Shen . Transformers in medical image analysis .
- [8] Hui Wen Loh , Chui Ping Ooi , Silvia Seoni , Prabal Datta Barua , Filippo Molinari , U Rajendra Acharya . Application of explainable artificial intelligence for healthcare: A systematic review of the

last decade (2011– 2022).

[9] Michele Salvagno , Fabio Silvio Taccone & Alberto Giovanni Gerli Critical Care .Can artificial intelligence help for scientific writing? 27, Article number: 75 (2023) 62k Accesses 120 Citations 225 Altmetric.

[10] Christopher C. Yang<sup>1</sup> .Explainable Artificial Intelligence for Predictive Modeling in Healthcare Received: 1 September 2021 / Revised: 29 December 2021 / Accepted: 3 January 2022\ Published online: 11 February 2022 © The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022.

## **APPENDIX-A**

### **PSUEDOCODE**

#### **User**

1. Start HTML document
2. Set language to "en"
3. Start head section
4. Set title to "Knowledge Builder"
5. Set charset to UTF-8
6. Set viewport to "width=device-width"
7. Link external stylesheets for fonts and icons
8. Link external stylesheets for custom styles and scripts
9. End head section
10. Start body section
11. Create a container div with class "container"
12. Create a header section
  - a. Insert an h1 element with the text "Knowledge Builder"
13. Start sidebar navigation
  - a. Create a nav element
  - b. Create an unordered list (ul) for navigation links
  - c. Create list items (li) with anchor tags (a) for "Home" and "Use my file"
14. Start main content section for "Home"
  - a. Insert an h2 element with the text "Work Based on PPTs & Web sites and Reports"
  - b. Create a file list div
    - i. Display available source files
    - ii. Create a paragraph element with text "Web-Site's of R&D Reports"
    - iii. Create an unordered list (ul) for file names
      - Create list items (li) for each file

- c. Create a search div
    - i. Create an input element with type "text" and id "query-input" for user query
    - ii. Create a button with onclick event "searchKnowledge()" for searching
  - d. Create a search results div to display results or error messages
15. Start main content section for "Use my file"
- a. Insert an h2 element with the text "Your own data"
  - b. Create a file upload div
    - i. Create an input element with type "file" and id "file-input" for file uploads
  - c. Create a search div
    - i. Create an input element with type "text" and id "query-input-file" for user query
    - ii. Create a button with onclick event "searchKnowledgeFile()" for searching
  - d. Create a search results div to display results or error messages
16. End main content section
17. Include the script.js file using the script tag
18. End body section
19. End HTML document

## **Generators.py**

```
import streamlit as st
import spacy
from transformers import BartForConditionalGeneration, BartTokenizer
import pptx
import os
import pandas as pd
import requests
from bs4 import BeautifulSoup
from pptx import Presentation
from docx import Document
import PyPDF2
# Load the English language model
nlp = spacy.load("en_core_web_sm")
```

```
# Function to extract text from a ppt
def extract_text_from_ppt(file_path):
    try:
        presentation = Presentation(file_path)
        text = ""
        for slide in presentation.slides:
            for shape in slide.shapes:
                if shape.has_text_frame:
                    for paragraph in shape.text_frame.paragraphs:
                        text += paragraph.text + '\n'
        return text
    except Exception as e:
        print(f"Error extracting text from {file_path}: {e}")
        return ""

# Function to extract text from different sources based on provided links DataFrame
# def generate_data_from_files(file_paths):
#     text_from_ppt = ""
#     for ppt_path in file_paths:
#         text_from_ppt += extract_text_from_ppt(ppt_path)
#     #nlp = spacy.load("en_core_web_sm")
#     doc = nlp(text_from_ppt)
#     knowledge_graph = {}
#     for entity in doc.ents:
#         if entity.label_:
#             knowledge_graph[entity.text] = {'label': entity.label_, 'text': entity.sent.text}
#     return knowledge_graph

# Function to extract text from a webpage or load from a single cached file
def extract_text_from_webpage(url):
```

---

```
cache_file = "webpage_cache.txt"
```

```
try:
```

```
    response = requests.get(url) soup = BeautifulSoup(response.content, 'html.parser')
```

```
    text = ''.join([p.get_text() for p in soup.find_all('p')])
```

```
    # Cache the extracted text by appending it to the single file
```

```
    with open(cache_file, 'a', encoding='utf-8') as file:
```

```
        file.write(text + "\n\n")
```

```
    return text
```

```
except Exception as e:
```

```
    print(f"Error extracting text from {url}: {e}")
```

```
    return ""
```

```
def extract_text_from_ppt(uploaded_file):
```

```
    presentation = Presentation(uploaded_file)
```

```
    text = ""
```

```
    for slide in presentation.slides:
```

```
        for shape in slide.shapes:
```

```
            if shape.has_text_frame:
```

```
                for paragraph in shape.text_frame.paragraphs:
```

```
                    text += paragraph.text + " "
```

```
    return text
```

```
def extract_text_from_pdf(uploaded_file):
```

```
    pdf_reader = PyPDF2.PdfReader(uploaded_file)
```

```
    num_pages = len(pdf_reader.pages)
```

```
    text = ""
```

```
    for page_num in range(num_pages):
```

```
        page = pdf_reader.pages[page_num]
```

```
    text += page.extract_text()
```

```
    return text
```

```
def extract_text_from_docx(uploaded_file):
    text = ""

    try:
        # Load the uploaded document
        doc = Document(uploaded_file)
        # Extract text from each paragraph in the document
        for paragraph in doc.paragraphs:
            text += paragraph.text + "\n"
    except Exception as e:
        print(f"Error extracting text from DOCX: {e}")
    return text
```

## Models.py

```
def generate_response(results, model, tokenizer,value):
    prompt_text = " "
    for result in results:
        prompt_text += f"{result['label']}: {result['text']}\n"

    inputs = tokenizer(prompt_text, return_tensors="pt", max_length=1024, truncation=True)
    response = model.generate(inputs.input_ids, max_length=value)
    return response

def search_knowledge_graph(query, knowledge_graph):
    results = []
    query_words = query.lower().split() # Split the query into individual words
    for entity, properties in knowledge_graph.items():
        entity_text = properties['text'].lower()
        # Check if any word in the query matches any word in the entity text, ignoring case
        if any(word in entity_text.lower().split() for word in query_words):
```

```
        results.append(properties)
    return results
```

## **App.py**

```
import streamlit as st
import spacy
from transformers import BartForConditionalGeneration, BartTokenizer
import pptx
import os
import pandas as pd
import requests
from bs4 import BeautifulSoup
from pptx import Presentation
from docx import Document
import PyPDF2
from data_generators import *
from models import *

def main():
    st.title("Knowledge Builder")
    page = st.sidebar.selectbox("Choose a page", ["Home", "Use my file"])
    if page == "Home":
        st.header("Work Based on PPTs & Web sites and Reports")
        # st.subheader("Available Source Files:")
        # source_files = ["ambroxol.pptx", "Cancer.pptx", "approach-to-fever.pptx", "Know-
Everything-About-Ast.pptx"]
        # for file in source_files:
        #     st.write(file)
        # st.write("R&D Reports ")
        expand_files = st.expander("Available Source Files", expanded=False)
```



---

```

with expand_files:
    st.write("Web-Site's of R&D Reports")
    source_files = ["ambroxol.pptx", "Cancer.pptx", "approach-to-fever.pptx", "Know-
Everything-About-Ast.pptx"]
    for file in source_files:
        st.write(file)
query = st.text_input("Enter your query:")
if st.button("Search"):
    with st.spinner("Searching..."):
        text_from_ppt = ""
        cache_file = "webpage_cache.txt"
        if os.path.exists(cache_file):
            # If the cached file exists, load text from the file
            with open(cache_file, 'r', encoding='utf-8') as file:
                text_from_ppt += file.read()
        else:
            links_df = pd.read_csv("therapy.csv")
            for index, row in links_df.iterrows():
                document_id = row['document_id']
            topic = row['topic']
            link = row['link']
            class_type = row['class_type']
            if "slideshare" in link:
                #print("from ppts")
                text_from_ppt += extract_text_from_ppt(link)
            else: # Assuming web page or other source types
                #print("from webpages")
                text_from_ppt += extract_text_from_webpage(link)
        ppts_file_paths = ("ambroxol.pptx", "Cancer.pptx", "approach-to-fever.pptx",
"Know-Everything-About-Ast.pptx")
        for ppt_path in ppts_file_paths:
            text_from_ppt += extract_text_from_ppt(ppt_path)
        nlp = spacy.load("en_core_web_sm")

```

---

---

```

doc = nlp(text_from_ppt)
knowledge_graph = {}
for entity in doc.ents:
    if entity.label_:
        knowledge_graph[entity.text] = {
            'label': entity.label_,
            'text': entity.sent.text
        }
tokenizer = BartTokenizer.from_pretrained("facebook/bart-base")
model = BartForConditionalGeneration.from_pretrained("facebook/bart-base")

results = search_knowledge_graph(query, knowledge_graph)
response = generate_response(results, model, tokenizer, 250)
decoded_response = tokenizer.decode(response[0], skip_special_tokens=True)

cleaned_response = ' '.join(filter(lambda x: x[:2] != '>>' and x.isalnum() and len(x)
<= 10, decoded_response.split()))

if cleaned_response == "Based on your what I found in the knowledge":
    st.error("No information found based on your query. Please try another query.")
else:
    st.success(f"Response: {query}")
    st.success(f"    Based on your query, here's what I found in the knowledge graph
-> \n {cleaned_response}")
elif page == "Use my file":
    st.header("Your own data")
    uploaded_files = st.file_uploader("Upload PPT or PDF files",
accept_multiple_files=True)
    query = st.text_input("Enter your query:")
    if uploaded_files:
        if st.button("Search"):
            with st.spinner("Searching..."):
                all_text_from_files = "" # To accumulate text from all uploaded files

```

---

---

```

for uploaded_file in uploaded_files:
    file_ext = uploaded_file.name.split(".")[-1].lower()
    if file_ext == "pptx":
        text_from_file = extract_text_from_ppt(uploaded_file)
    elif file_ext == "pdf":
        text_from_file = extract_text_from_pdf(uploaded_file)
    elif file_ext == "docx":
        text_from_file = extract_text_from_docx(uploaded_file)
    else:
        st.error(f"Invalid file format for file '{uploaded_file.name}'. Please upload a
PPT, PDF, or DOCX file.")
        continue # Move to the next file if the format is incorrect
# Accumulate text from all uploaded files
all_text_from_files += text_from_file
nlp = spacy.load("en_core_web_sm")
doc = nlp(all_text_from_files)
knowledge_graph = {}
for entity in doc.ents:
    if entity.label_:
        knowledge_graph[entity.text] = {
            'label': entity.label_,
            'text': entity.sent.text
        }
results = search_knowledge_graph(query, knowledge_graph)
tokenizer = BartTokenizer.from_pretrained("facebook/bart-base")
model = BartForConditionalGeneration.from_pretrained("facebook/bart-base")
response = generate_response(results, model, tokenizer, 500)
decoded_response = tokenizer.decode(response[0], skip_special_tokens=True)
cleaned_response = ' '.join(filter(lambda x: x[:2] != '>>' and x.isalnum() and
len(x) <= 10, decoded_response.split()))
if cleaned_response == "Based on your what I found in the knowledge":
    st.error("No information found based on your query. Please try another
query.")

```

---

else:

```
st.success(f"Response: {query}")
```

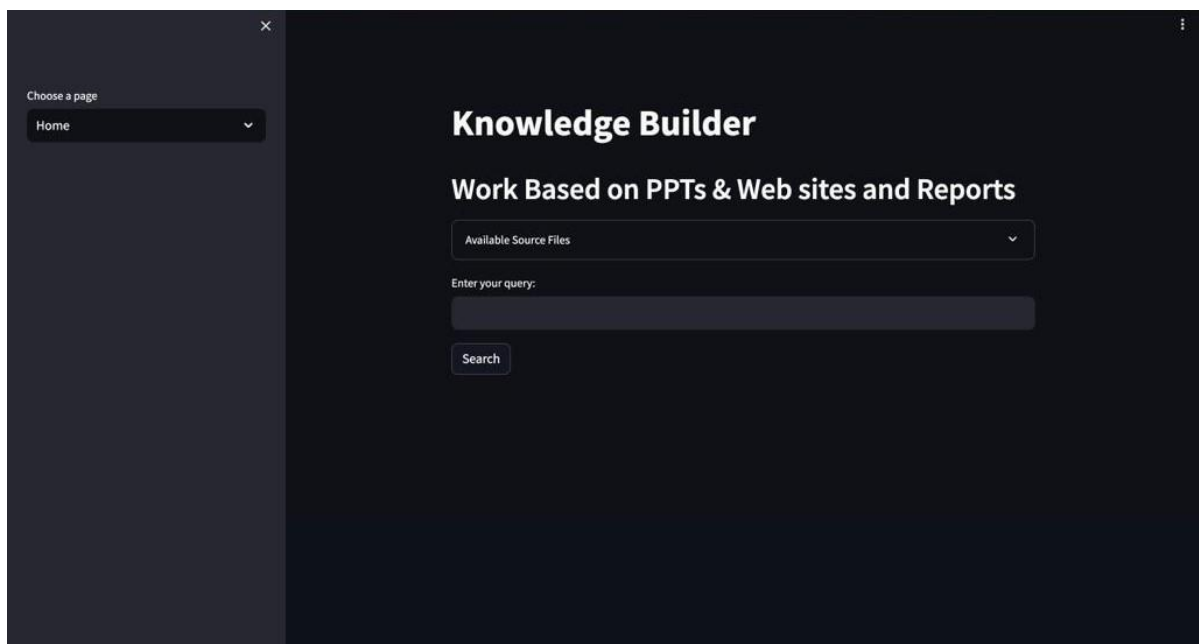
```
st.success(f"      Based on your query, here's what I found in the knowledge  
graph -> \n {cleaned_response}")
```

```
if __name__ == "__main__":
```

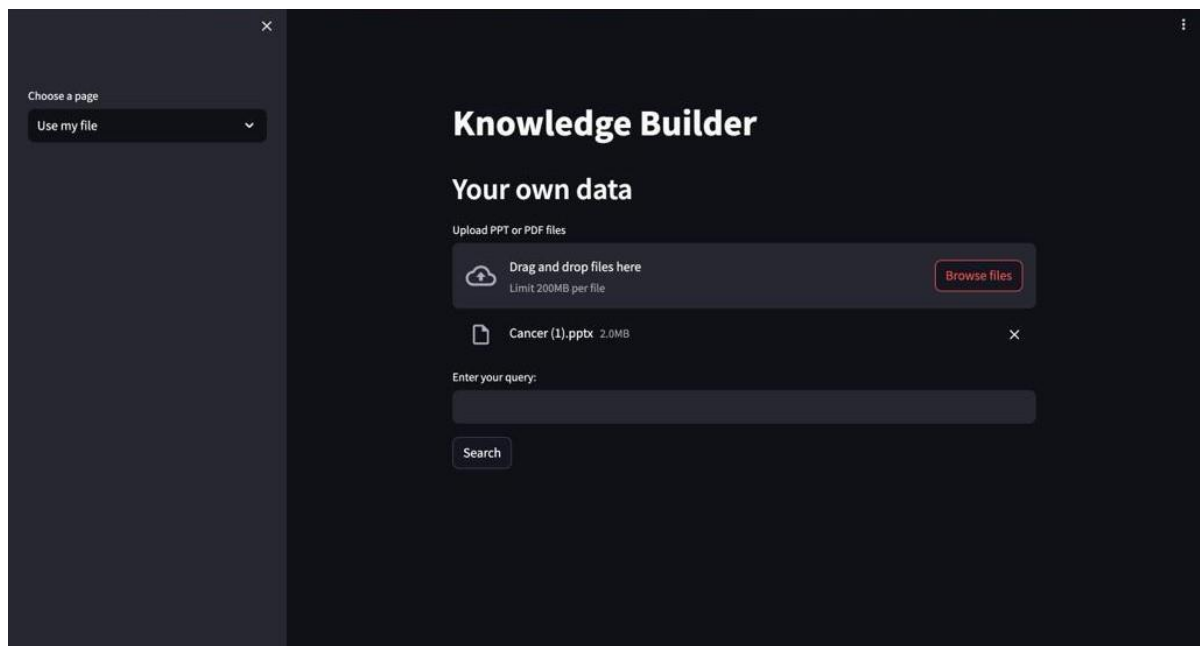
```
    main()
```

## APPENDIX-B

### SCREENSHOTS



**Figure B.1 – Home**



**Figure B.2 – Browse file**

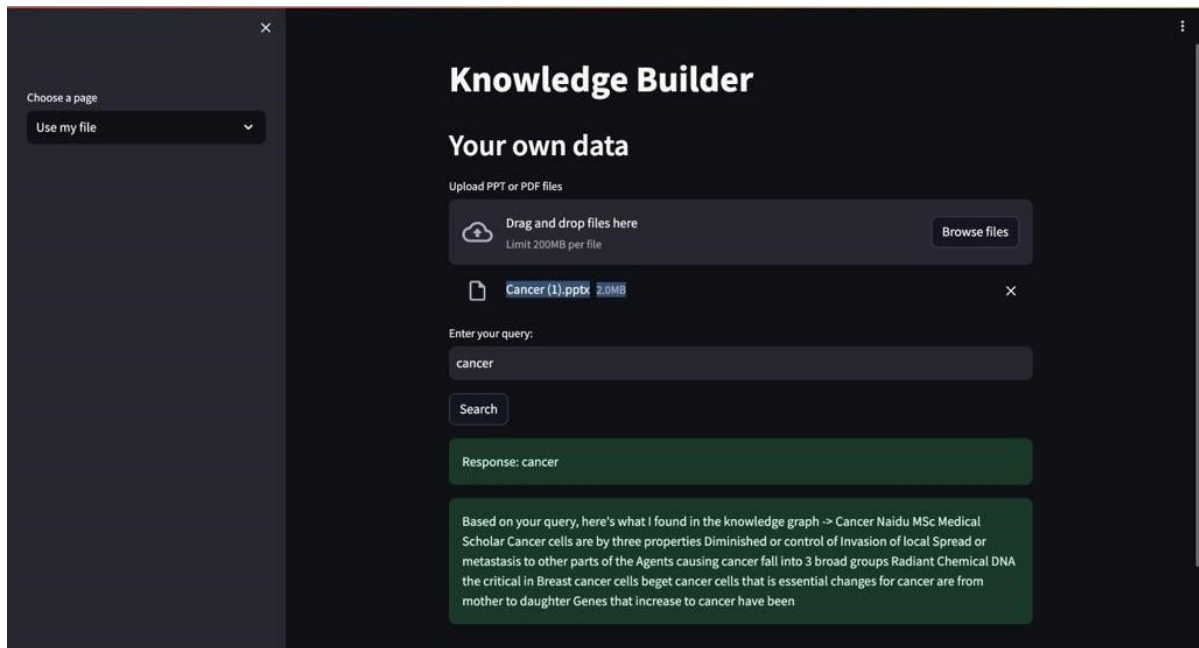


Figure B.3 – Search query

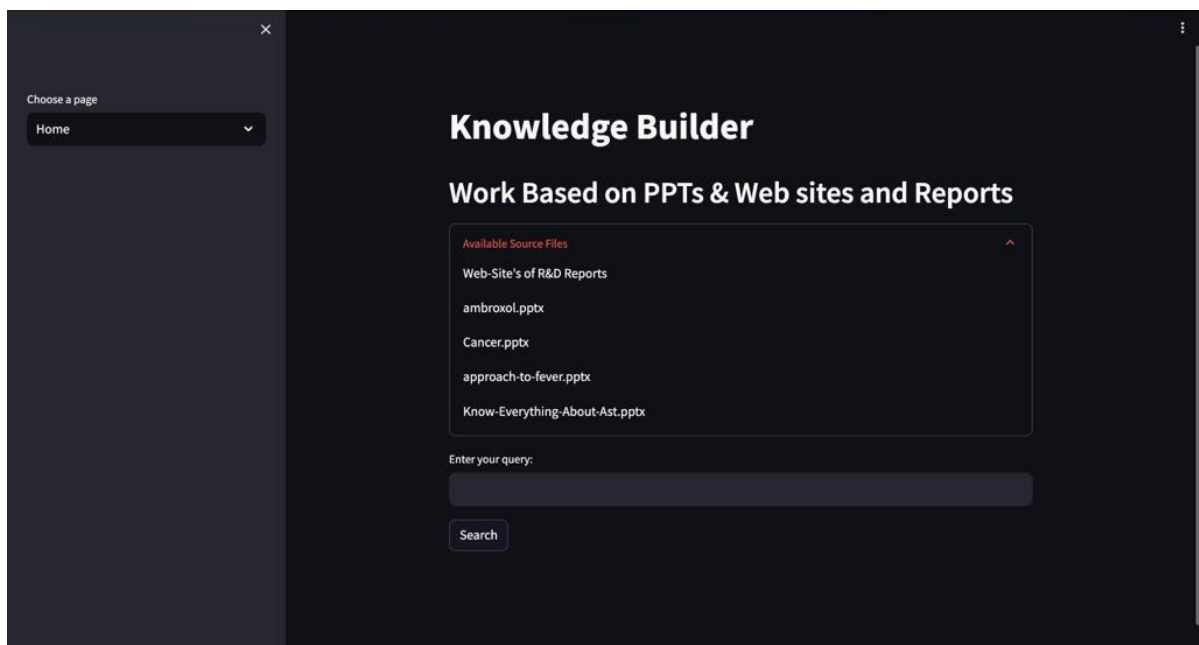





Figure B.4 – Available source files

# APPENDIX-C


## ENCLOSURES

### LETTER OF ACCEPTANCE


Paper id: IJCRT\_249546 – **Acceptance Notification and Review Result of paper TITLE - INTELLIGENT NATURAL LANGUAGE SEARCH - KNOWLEDGE BUILDER.** - Your Paper Accepted Complete Below Process and Publish it. Your Email id: sandhya.l@presidencyuniversity.in Track your paper : [https://ijcrt.org/track.php?r\\_id=249546](https://ijcrt.org/track.php?r_id=249546)

	<b>WhatsApp</b> 7990172303		<b>editor@ijcrt.org</b>		<b>ijcrt.org</b>
---	-------------------------------	---	-------------------------	---	------------------

	<b>INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS - IJCRT</b> (IJCRT.ORG)
	International Peer Reviewed & Refereed Journals, Open Access Journal
	ISSN: 2320-2882   Impact factor: 7.97   ESTD Year: 2013
	Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar   AI-Powered Research Tool), Multidisciplinary, Monthly, Indexing in all major database & Metadata, Citation Generator, Digital Object Identifier(DOI)

	<b>📌 Your Paper Acceptance details as per below Complete Step 1 and Step 2 and publish within 1 to 2 days.</b>
---	--

Dear Author, Congratulation!!!

Your manuscript with Registration ID: **IJCRT\_249546** has been accepted for publication in the International Journal of Creative Research Thoughts (IJCRT) | [www.ijcrt.org](http://www.ijcrt.org) | ISSN: 2320-2882 | International Peer Reviewed & Refereed Journals, Open Access Journal. IJCRT is Peer Review Journal, Refereed Journal, Peer Reviewed Journal Referred Journal and Indexed Journal Open Access Journal Online and Print Journal

📌 IJCRT Impact Factor: 7.97 | UGC Approved Journal No: 49023 (18)

📌 Check your paper status: <https://ijcrt.org/track.php> or <http://ijcrt.org/Authorhome/alogin.php>

📌 Online Payment Link for Indian author: [http://ijcrt.org/pay\\_form\\_2.php](http://ijcrt.org/pay_form_2.php)

📌 Online Payment Link for Foreign/international author: [https://ijcrt.org/pay\\_form\\_international.php](https://ijcrt.org/pay_form_international.php)

📌 Your Paper Review Report :	
📌 Registration ID:	IJCRT – 249546
📌 Title of the Paper:	INTELLIGENT NATURAL LANGUAGE SEARCH - KNOWLEDGE BUILDER
📌 Accepted or Not:	<b>Accepted</b>
📌 Criteria	📌 Points out of 100%
Continuity	94%
Text structure	86%
References	87%
Understanding and Illustrations	89%
Explanatory power	86%
Detailing	89%
Relevance and practical advice	90%

📌 Track Your Paper Details:	
Paper Track Link 1:	<a href="https://ijcrt.org/track.php">https://ijcrt.org/track.php</a>
Paper Track Link 2:	<a href="http://ijcrt.org/Authorhome/alogin.php">http://ijcrt.org/Authorhome/alogin.php</a>
📌 Online Payment link Details:	
Indian author:	<a href="http://ijcrt.org/pay_form_2.php">http://ijcrt.org/pay_form_2.php</a>
Foreign/international author:	<a href="https://ijcrt.org/pay_form_international.php">https://ijcrt.org/pay_form_international.php</a>

📌 Overall Assessment (Comments.)	📌 Paper Accepted: YES
📌 Unique Contents: 87%	Paper Accepted
Reviewer comment store in online RMS system. Paper Accepted complete below Step1 and Step2 process and publish your paper within 1 to 2 day.	

<b>Publication of Paper</b>	<b>Within 01-02 Days after Submitting documents. Please complete payment and send payment proof along with below documents.</b>
-----------------------------	---







## Systematic Review of the Last Decade (2011–2022)", Computer Methods and Programs in Biomedicine, 2022

Publication

8	<a href="http://www.analyticsvidhya.com">www.analyticsvidhya.com</a> Internet Source	<1 %
9	<a href="http://builderall.com">builderall.com</a> Internet Source	<1 %
10	<a href="http://idoc.tips">idoc.tips</a> Internet Source	<1 %
11	Submitted to M S Ramaiah University of Applied Sciences Student Paper	<1 %
12	<a href="http://www.slideshare.net">www.slideshare.net</a> Internet Source	<1 %
13	<a href="http://arxiv.org">arxiv.org</a> Internet Source	<1 %
14	<a href="http://export.arxiv.org">export.arxiv.org</a> Internet Source	<1 %
15	<a href="http://udspace.udel.edu">udspace.udel.edu</a> Internet Source	<1 %
16	<a href="http://archive.nyu.edu">archive.nyu.edu</a> Internet Source	<1 %
17	Amos Azaria, Rina Azoulay, Shulamit Reches. "ChatGPT is a Remarkable Tool—For Experts",	<1 %

## Data Intelligence, 2023

Publication

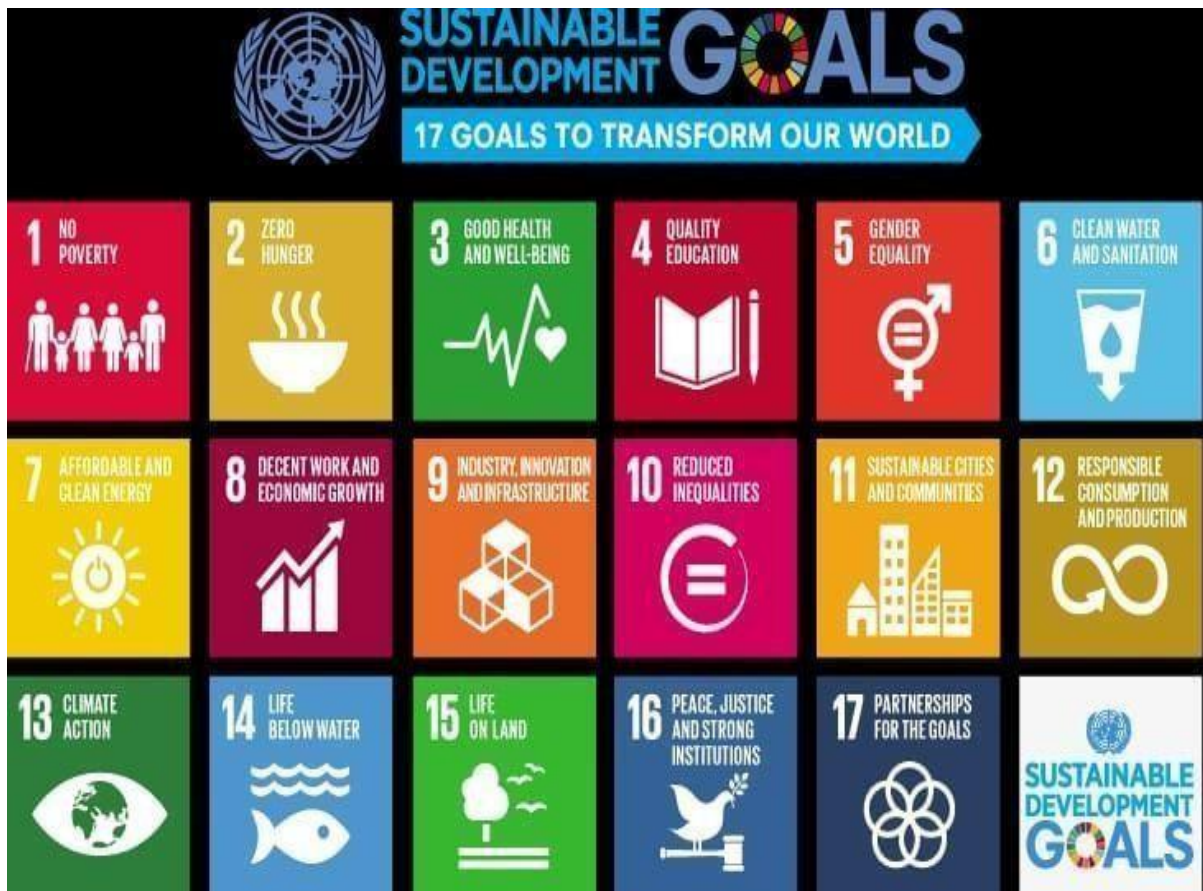
18	ccforum.biomedcentral.com	<1 %
Internet Source		
19	fastercapital.com	<1 %
Internet Source		
20	www.medrxiv.org	<1 %
Internet Source		
21	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2023	<1 %
Publication		

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off

## SUSTAINABLE DEVELOPMENT GOALS



In the pursuit of advancing global education, our innovative chatbot project stands as a beacon for Quality Education (SDG 4). By providing accessible learning resources, personalized learning experiences, and round-the-clock support, our platform is committed to breaking down barriers, fostering inclusivity, and revolutionizing the way individuals engage with and benefit from education.

### Accessible Learning Hub:

Our chatbot acts as a centralized hub, granting students easy access to a wealth of educational resources, including textbooks, articles, and study materials.

### Personalized Learning Paths:

Tailored learning experiences through adaptive technologies ensure that each student's unique needs and learning styles are addressed, fostering a more effective educational journey.

### **24/7 Support for Continuous Learning:**

With our chatbot's round-the-clock availability, students can receive instant support outside traditional hours, promoting continuous learning and flexibility in education.

### **Inclusive Features for Diverse Learners:**

Incorporating features for differently-abled learners ensures inclusivity, offering alternative formats and assistive technologies to make education accessible to a wide range of individuals.

In conclusion, our chatbot project not only aligns with the Sustainable Development Goal of Quality Education by providing accessible, personalized, and inclusive learning experiences but also aims to redefine the landscape of education by embracing innovation and fostering a global community of learners.