RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Mini Project Report On

SMARTGIST

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By

Nayana V Reji (U2103151)

Neethu Anil Jacob (U2103152)

Nikita Alex (U2103156)

Niranjan G Das (U2103157)

Under the guidance of

Dr.Mary Priya Sebastian

Department of Computer Science & Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Affiliated to APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
May 2024

CERTIFICATE

This is to certify that the mini project report entitled ***"SMARTGIST"*** *is a bonafide record of the work done by* ***Nayana V Reji (U2103151)***, ***Neethu Anil Jacob (U2103152)***, ***Nikita Alex (U2103156)***, ***Niranjan G Das (U2103157)***, *submitted to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Dr. Mary Priya Sebastian
Associate Professor
Dept. of CSE
RSET

Dr. Uma Narayanan
Asst Professor
Dept. of CSE
RSET

Dr. Preetha K.G
Head of the Department
Dept. of CSE
RSET

ACKNOWLEDGEMENTS

We wish to express our sincere gratitude towards **Dr P. S. Sreejith**, Principal of RSET, and **Dr. Preetha K.G.**, Head of the Department of Computer Science and Engineering for providing us with the opportunity to undertake our mini project, "SmartGist".

We are highly indebted to our project coordinator, **Dr. Uma Narayanan**, Asst Professor, Department of Computer Science and Engineering for her valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Dr. Mary Priya Sebastian**, Associate Professor, Department of Computer Science and Engineering for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, We would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Nayana V Reji

Neethu Anil Jacob

Nikita Alex

Niranjan G Das

Abstract

SmartGist aims to generate a PDF containing all relevant information related to each topic contained in an input PDF provided by the user.The subtopics are identified within the input PDF and extracted using keyword extraction methods. Then these identified subtopics are searched for in the other reference materials provided by the user.The contents under each subtopic is gathered for summarization.A new PDF is generated with the summarized content.This application aggregates summarized content from specified reference materials making it easy for users.They can avoid referencing multiple reference materials simultaneously.The app employs advanced natural language processing (NLP) techniques to identify and extract subtopics,employs cross document search,summarizes the content and generates a new PDF.The summarized PDF serves as a consolidated resource, eliminating the need to manually sift through multiple documents.The app provides a user-friendly interface for users to upload and select PDF documents.

Contents

List of Figures

List of Abbreviations

- PDF - Portable Document Format

- TF-IDF - Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

1.1 Background

SmartGist is a PDF summarization application designed to assist users in quickly extracting relevant information from PDF documents based on predefined topics. The application aims to streamline the summarization process, providing users with concise summaries in the same format as their original documents. SmartGist operates as a standalone application, interacting with users through a user-friendly interface. It may integrate with external libraries or APIs for PDF processing and natural language processing.

1.2 Problem Definition

SmartGist is a PDF summarization tool that makes people's learning experience smarter and easier. The aim of our application is for users to get summarized information for the topics that the they input easily, instead of the time consuming manual review process

1.3 Scope and Motivation

Scope of the Project:

The project aims to develop a PDF summarization application, SmartGist, that facilitates the extraction of pertinent information from PDF documents efficiently. It focuses on leveraging natural language processing (NLP) techniques to identify and summarize predefined topics within large documents, catering to the needs of users requiring quick comprehension of extensive texts. The application is designed to maintain the original formatting of the documents, ensuring that the summaries are not only accurate but also visually consistent with the source material. By offering customizable summarization options, the project covers a wide range of use cases, from academic research and corporate

reports to legal documents and technical manuals. The project encompasses the development of a user-friendly interface that allows for easy navigation and operation by users of varying technical proficiency, thereby broadening its applicability.

Motivation for the Project:

The overwhelming growth of digital content has made it increasingly challenging for individuals and organizations to process information efficiently, driving the need for automated summarization tools like SmartGist. There is a significant demand for solutions that can quickly distill essential information from lengthy documents, allowing users to make informed decisions without having to read through entire texts. The project is motivated by the potential to enhance productivity and knowledge management across various sectors by providing a tool that simplifies the extraction of key insights from complex documents. With the advancement of NLP technologies, there is an opportunity to develop more sophisticated and accurate summarization applications that can cater to specific user needs, further motivating the pursuit of this project. The initiative is also driven by the goal of democratizing access to information, enabling users from educational, research, and professional backgrounds to easily grasp and utilize the vast amounts of available textual data.

1.4 Objectives

- The system provides a user-friendly interface for PDF selection, allowing users to choose multiple files easily.
- The system maintains a history of generated summarized PDFs, associating each with a timestamp and project details.

1.5 Challenges

Designing and implementing SmartGist involves addressing several key challenges: optimizing PDF processing for diverse document sizes and complexities, refining NLP techniques for accurate topic identification and summarization, creating an intuitive user interface for all skill levels, ensuring scalability for growing demand, maintaining com-

patibility across different PDF formats, and implementing robust security measures to protect user data.

1.6 Assumptions

Several assumed factors significantly influence the requirements and successful implementation of the SmartGist. Firstly, it is assumed that users will provide PDFs in standard formats, helping in consistent text extraction and keyword identification. The app relies on external Python libraries for effective text extraction and natural language processing (NLP) techniques. The assumption includes the availability and compatibility of these libraries throughout the project lifecycle.The app assumes a stable and secure operating environment, where the Flutter interface integrates with diverse hardware platforms and operating systems,like android and windows

1.7 Societal / Industrial Relevance

Societal Relevance:

1. **Education and Research**: By simplifying the process of extracting key information from extensive texts, SmartGist aids students, academics, and researchers in managing literature reviews and study materials more efficiently, fostering a deeper understanding and engagement with their subjects.

2. **Accessibility**: For individuals with learning disabilities or those who find reading large volumes of text challenging, SmartGist can democratize information access, making knowledge more accessible and inclusive.

3. **Time Management**: In a fast-paced world, SmartGist empowers individuals to save valuable time, allowing them to focus on analysis and decision-making rather than getting bogged down by the volume of information.

Industrial Relevance:

1. **Business Intelligence**: Companies can leverage SmartGist to swiftly analyze reports, market analyses, and competitor information, enabling strategic decisions based on comprehensive summaries of relevant data.

2. **Legal and Healthcare Industries**: These sectors often deal with voluminous documents where pinpointing specific information quickly can be crucial. SmartGist's ability to accurately summarize content can streamline workflows, reduce manual review time, and improve the delivery of services.

3. **Content Management and Publishing**: For industries focused on managing or publishing content, SmartGist can assist in content curation, editorial processes, and quick generation of abstracts or summaries for various publications.

1.8 Organization of the Report

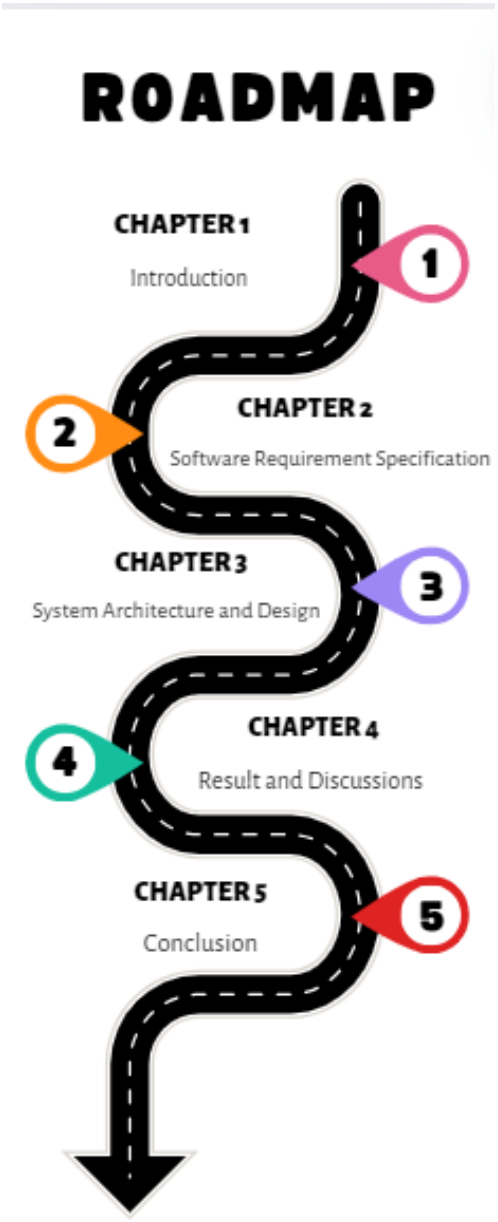


Figure 1.1: Organization of the Report-Roadmap

Chapter 2

Software Requirements Specification

2.1 Introduction

2.1.1 Purpose

SmartGist is a PDF summarization application designed to assist users in quickly extracting relevant information from PDF documents based on predefined topics. The application aims to streamline the summarization process, providing users with concise summaries in the same format as their original documents.

2.1.2 Product Scope

SmartGist is a comprehensive PDF summarization application designed to streamline information processing for users. The application allows users to effortlessly upload PDF documents, after which it employs sophisticated Natural Language Processing techniques to identify topics within the uploaded content. SmartGist further extracts relevant information from reference PDFs based on the identified topics. Employing extractive summarization techniques,SmartGist efficiently condenses the extracted information, creating concise summaries. The application goes a step further by generating a new PDF document that mirrors the structure of the original, facilitating ease of understanding. SmartGist significantly contributes to operational efficiency and productivity by automating time-consuming manual review processes. It also supports enhanced knowledge management and information accessibility, allowing users to quickly access relevant information. Furthermore, the application prioritizes a positive user experience by providing users with the flexibility to download or view the generated summaries, thereby aligning with modern expectations for efficient and user-friendly information tools.

2.2 Overall Description

2.2.1 Product Perspective

SmartGist operates as a standalone application, interacting with users through a user-friendly interface. It may integrate with external libraries or APIs for PDF processing and natural language processing.

2.2.2 Product Functions

1. PDF Upload: Users can upload PDF documents.
2. Topic Identification: The application identifies topics within user-uploaded PDFs.
3. Information Extraction: Relevant information is extracted from reference PDFs based on identified topics.
4. Summarization: Extractive summarization techniques are employed to create concise summaries.
5. PDF Generation: Summarized content is formatted into a new PDF document.
6. User Interface: A user-friendly interface facilitates easy interaction with the application.

2.2.3 Operating Environment

SmartGist is developed to facilitate efficient keyword extraction and content summarization from PDF documents. The app is developed using Flutter for the user interface and Python for backend functionalities. It is optimized to run on multiple hardware platforms, including smartphones, tablets, and computers. The compatibility extends across popular operating systems such as Android and Windows. By accommodating both Flutter and Python, the software harmoniously coexists with the chosen technologies, providing users with a user-friendly and versatile solution for extracting and consolidating information from PDFs.

2.2.4 Design and Implementation Constraints

The design and implementation of the SmartGist are constrained by several factors that necessitate careful consideration. Firstly, the efficiency of PDF processing poses a challenge due to the potential size and complexity of documents. This requires the

use of optimized algorithms and data structures to handle text extraction and analysis swiftly. Secondly, the accuracy of natural language processing techniques employed for subtopic identification and content summarization is crucial. Continuous refinement and evaluation of these techniques are essential to ensure reliable results. Additionally, the user interface must be intuitive and accessible to accommodate users with varying levels of technical expertise. Scalability is another constraint, demanding a flexible architecture capable of handling increasing loads efficiently. Compatibility with various PDF formats and stringent security measures to safeguard user data further add to the complexity of the design and implementation process.

2.2.5 Assumptions and Dependencies

Several assumed factors significantly influence the requirements and successful implementation of the SmartGist. Firstly, it is assumed that users will provide PDFs in standard formats, helping in consistent text extraction and keyword identification. The app relies on external Python libraries for effective text extraction and natural language processing (NLP) techniques. The assumption includes the availability and compatibility of these libraries throughout the project lifecycle.The app assumes a stable and secure operating environment, where the Flutter interface integrates with diverse hardware platforms and operating systems,like android and windows

2.3 External Interface Requirements

2.3.1 User Interfaces

SmartGist's user interface extends its functionality by incorporating a home page that displays a history of earlier generated documents. Users are greeted with a user-friendly dashboard where they can conveniently view and access previously generated PDF summaries. Additionally, the interface introduces a prominent "New Project" option, strategically placed to invite users to initiate a new summarization project. This option serves as the entry point for users looking to generate a fresh summarized PDF. Navigation between the home page and the new project creation is straightforward, contributing to an intuitive and accessible user experience. The combination of a historical overview and a clear call-to-action for new projects enhances the app's usability.

2.3.2 Hardware Interfaces

The hardware interface of our application encompasses the interaction between the software and the physical devices used by the users. Primarily, it involves the user's computing device, such as a computer or smartphones or tablets, and peripherals necessary for using the app and document uploading,such as keyboard and mouse.Users utilize these input devices to interact with the application's user interface,navigating through the document upload and selection process.

2.3.3 Software Interfaces

The software interfaces for this application encompass connections with various components to ensure seamless operation. The backend, developed in Python, serves as a critical component for keyword extraction, summarization, and the generation of the new PDF. It communicates with the Flutter frontend to receive user inputs and deliver processed results. The Python backend relies on specific libraries such as NLTK for natural language processing tasks and Flask for handling HTTP requests. Additionally, the application interacts with a relational database (e.g., MySQL version 8.0) to store and retrieve relevant information efficiently.

2.3.4 Communications Interfaces

The communication interfaces enable seamless communication between modules responsible for subtopic identification,content extraction,summarisation and pdf generation,facilitating the flow of data and instructions throughout the application.User interaction within the app is facilitated through communication interfaces which handle input from the users, provide feedback and orchestrate the overall user experience.Finally,communication interfaces are essential for output generation,enabling the app to format and structure the summarized content before generating the final pdf document.

2.4 System Features

2.4.1 PDF Selection and Upload

2.4.1.1 Description and Priority

This feature allows users to select and upload multiple PDF documents for summarization. It holds a high priority as it serves as the initial step for users to input content into the summarization process. The benefit rating is high due to its essential role in the

user workflow, while the risk is low as it primarily involves user interaction.

2.4.1.2 Stimulus/Response Sequences

Stimulus: The user accesses the app and navigates to the "New Project" option. The system responds by presenting a file selection interface, enabling the user to choose multiple PDFs for summarization.

Response: Upon PDF selection, the system displays the chosen files, confirming successful upload and prompting the user to proceed with summarization.

2.4.1.3 Functional Requirements

REQ-1: The system must provide a user-friendly interface for PDF selection, allowing users to choose multiple files easily.

REQ-2: The app should support standard PDF formats, ensuring compatibility with different types of PDF documents.

REQ-3: In the event of invalid inputs, such as unsupported file formats or corrupted PDFs, the system must display a clear error message guiding the user to rectify the issue.

2.4.2 Previous Document Display

2.4.2.1 Description and Priority

This feature displays a history of previously generated summarized PDF documents on the home page. It holds a medium priority, providing users with convenient access to their past projects. The benefit is moderate, offering users a quick reference, while the risk is low, as it does not impact the core summarization process.

2.4.2.2 Stimulus/Response Sequences

Stimulus: Upon accessing the home page, the system detects the user's request for document history. The response involves fetching and displaying a list of previously generated PDFs.

Response: The system presents a visually organized list or visual representation of historical documents, allowing users to select and revisit any project.

2.4.2.3 Functional Requirements

REQ-1: The system must maintain a history of generated summarized PDFs, associating each with a timestamp and project details.

REQ-2: The app should provide options for sorting or filtering historical documents based on user preferences or project metadata.

REQ-3: If there are no historical documents available, the system should display a

user-friendly message indicating that no projects have been generated.

2.5 Other Nonfunctional Requirements

2.5.1 Performance Requirements

Response Time:

SmartGist shall respond to user interactions within few seconds under normal operational conditions as users expect a responsive and prompt application. A quick response time enhances the user experience and ensures that users can efficiently utilize the summarization features without unnecessary delays.

Summarization Speed:

The application shall be capable of summarizing a PDF document within a time limit as efficient summarization is critical for user productivity. Establishing a time limit ensures that even large documents can be processed swiftly, meeting user expectations for timely results.

Resource Utilization:

The application shall not exceed 80normal operating conditions as efficient resource utilization prevents system slowdowns and ensures the application can operate smoothly without overburdening the hosting infrastructure.

2.5.2 Safety Requirements

Safety Certifications: SmartGist shall obtain and maintain relevant safety certifications based on industry standards by regularly assessing the application against industry-specific safety standards and obtain certifications as necessary.

Documentation of Security Measures:

SmartGist shall maintain comprehensive documentation outlining security measures, protocols, and incident response procedures and ensure that these are readily available to the development and support teams.

Legal and Regulatory Compliance:

SmartGist shall comply with all relevant laws and regulations pertaining to data security, user privacy, and intellectual property by establishing a legal review process for continuous compliance monitoring and updates.

Protection Against Malicious Inputs:

SmartGist shall validate and sanitize user inputs to prevent potential security vulnerabilities like SQL injection or cross-site scripting by regularly conducting security assessments to identify and mitigate potential vulnerabilities.

2.5.3 Security Requirements Data Security and Privacy:

SmartGist shall ensure the confidentiality and integrity of user-uploaded documents and summarized content by implementing end-to-end encryption for data transmission and storage along with regularly updating security protocols and conducting security audits to identify and address vulnerabilities.

Access Control:

The application shall enforce strict access control measures to prevent unauthorized access to user data by implementing a robust user authentication system with secure password policies.

User Authentication:

SmartGist shall require strong and unique user authentication credentials by enforcing password complexity requirements and encourage the use of multi-factor authentication to enhance user account security.

Compliance with Data Protection Regulations:

The application shall comply with relevant data protection regulations, such as GDPR or HIPAA, depending on the nature of the data processed by conducting regular compliance audits, updating privacy policies, and providing users with clear information on how their data is handled.

2.5.4 Software Quality Attributes

Accuracy of Summaries:

The application shall achieve a summary accuracy rate of at least 95original document as users rely on accurate summaries to make informed decisions

Ease of Use:

The application shall achieve a usability score of at least 85usability testing as user-friendly design is crucial for widespread adoption.

Scalability:

The application shall support a scalable architecture, with the ability to handle up to 100 simultaneous user requests without degradation in performance.

Adaptability to Document Types:

SmartGist shall be capable of processing and summarizing documents in common formats such as research papers, business reports, and legal documents.

Chapter 3

System Architecture and Design

3.1 System Overview

SmartGist is an innovative application that facilitates efficient extraction of relevant information from multiple PDF documents. Key features include automated review process, pdf document extraction , efficient summarization and a user-friendly interface. Benefits include increased productivity and effective information retrieval.

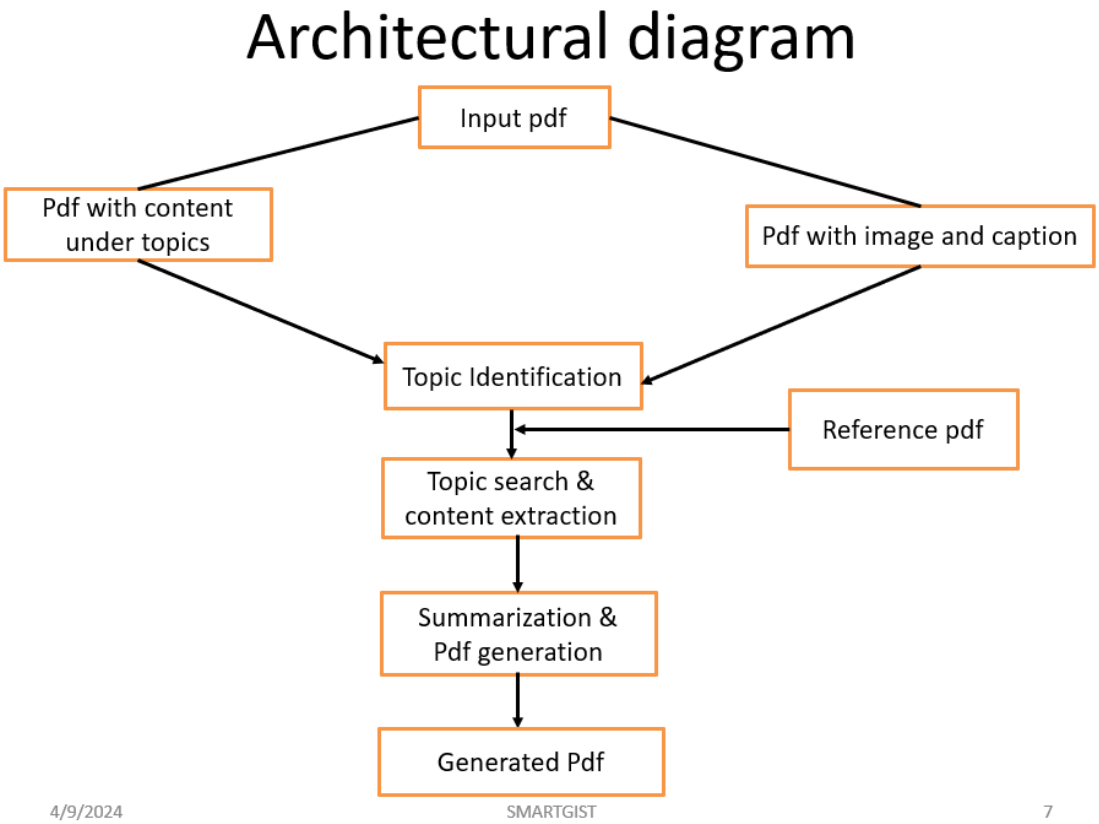


Figure 3.1: Architectural Diagram

Input PDF and Reference material

Input PDF is given by the user for the app to extract topics . It can be a PDF with topics and content under them or a PDF with image and caption. Reference material is given by the user for the app to extract content. User can input multiple reference materials either as standard PDFs

Topic Identification

If the input PDF has topics with content under them,headings are identified. If it is a PDF with image and caption,caption is identified as topic.

Topic Search and Content Extraction

The identified topics are searched for in the reference material provided by the user. The content under these topics are extracted for summarization.

Summarization and PDF generation

The extracted content under each topic is summarized using extractive summarization method. A new pdf is generated with the summarized content and is made available for the user to download.

3.2 Architectural Design

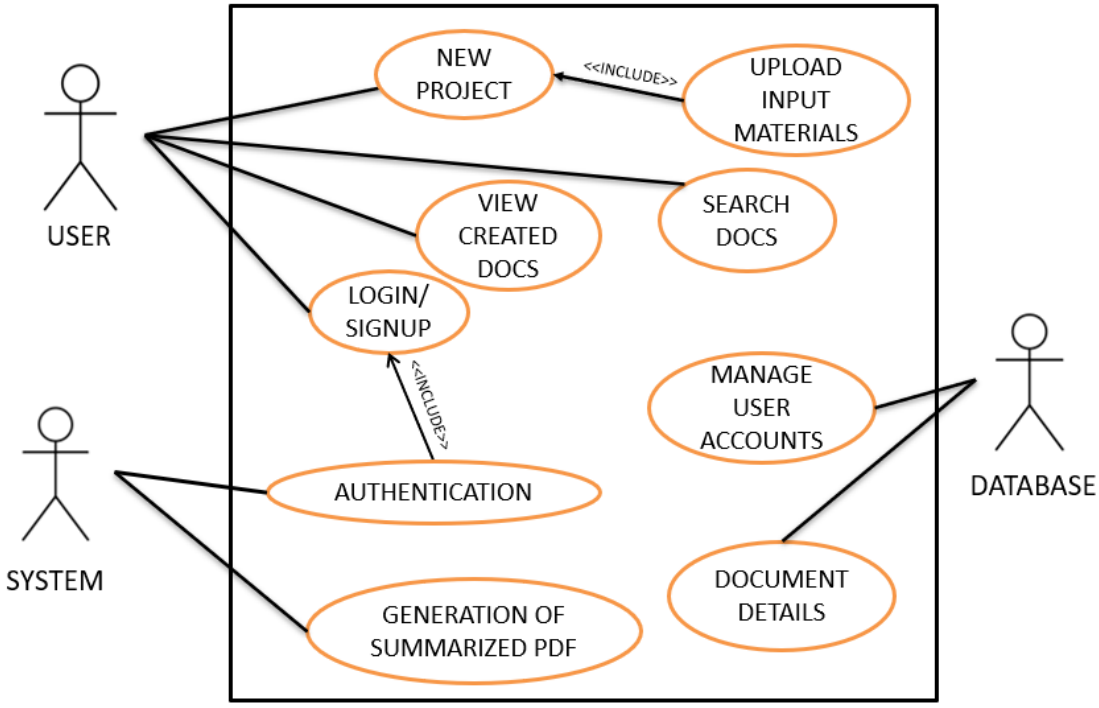


Figure 3.2: Use Case Diagram

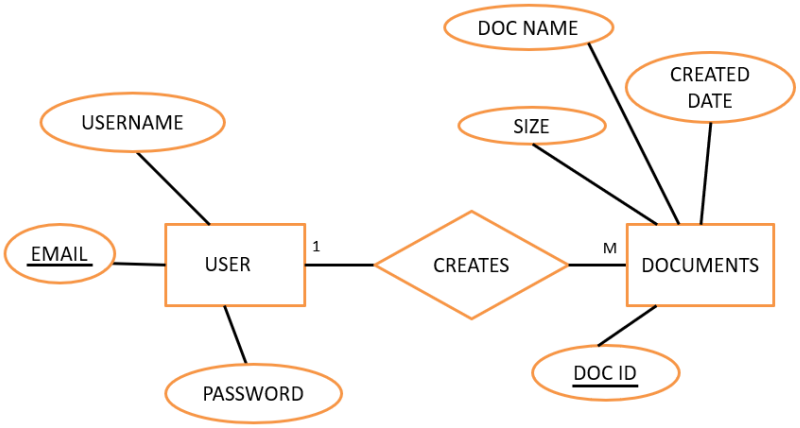


Figure 3.3: Entity Relationship Diagram

3.3 Proposed Methodology/Algorithms

1.Topics and reference material as inputs

Accept topics and reference materials in the form of PDFs.

2.Processing of Inputs:

Text content is then extracted from each document using fitz module.

3. Topic Identification using NLP techniques

4.Search for identified topics from reference PDFs:

For each PDF:

4.1.Iterate through the document's text content.

4.2.Search for occurrences of the identified topics.

4.3.Gather the sections of text containing the topics.

5.Text Summarization:

5.1.Concatenate the text sections related to each topic identified in the previous step from multiple documents.

5.2.Apply TF-IDF and Cosine Simlarity algorithm to generate summaries for each topic.

6.Output Generation:

6.1.Compile the summarized content for each topic into a PDF.

6.2.Include the title of each topic and its corresponding summary.

6.3.Provide the generated PDF as the output.

3.4 User Interface Design

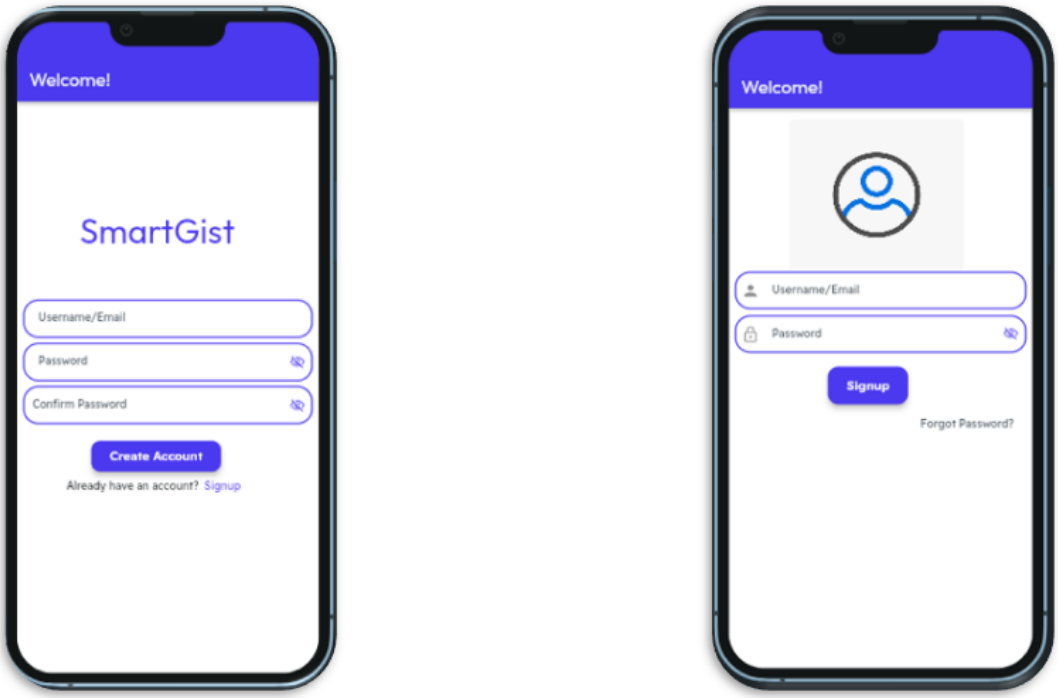


Figure 3.4: Login/Signup Page

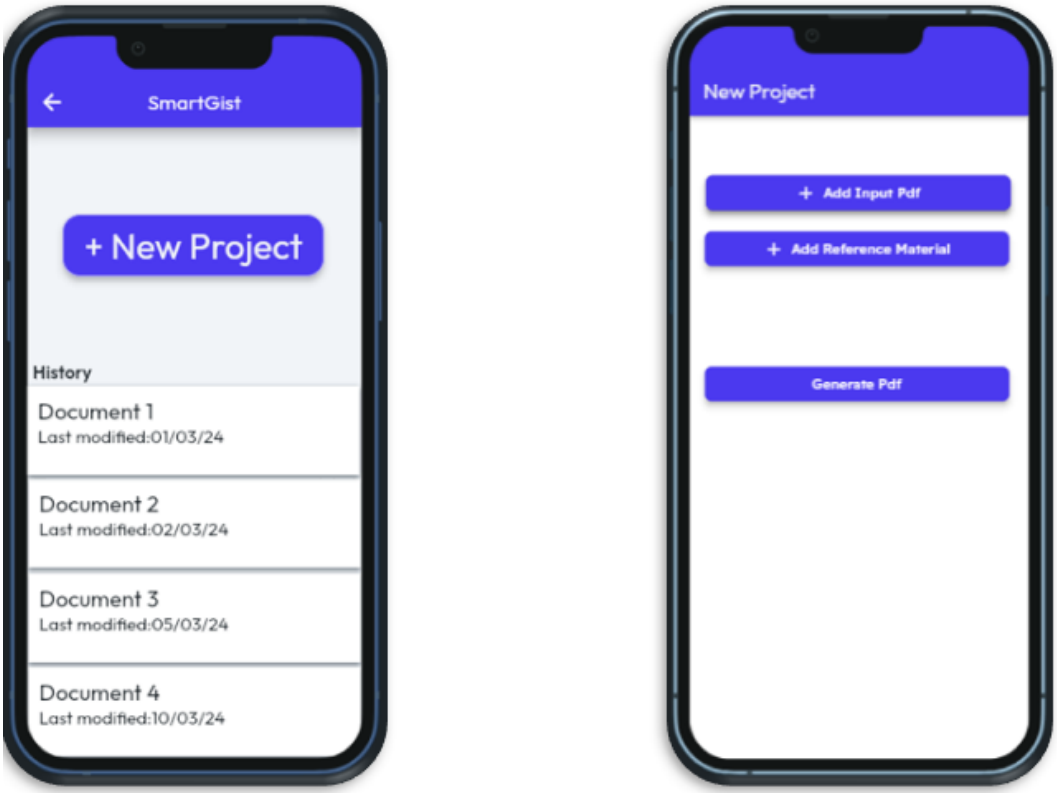


Figure 3.5: Main page

Figure 3.6: Download page

3.5 Database Design

- User Data Database:This database would store user authentication data, such as usernames, email addresses.It would be responsible for managing user accounts, authentication, and authorization.

  Database chosen:Firebase Authentication (for authentication purposes)

- Document Data Database: This database would store metadata about PDF documents uploaded to the app, as well as annotations made by users on those documents. It would be responsible for managing document metadata, annotations, and relationships between users and documents.

  Database chosen:Firestore or Realtime Database (for storing document metadata), Local Storage for Firebase(for storing the actual PDF files)

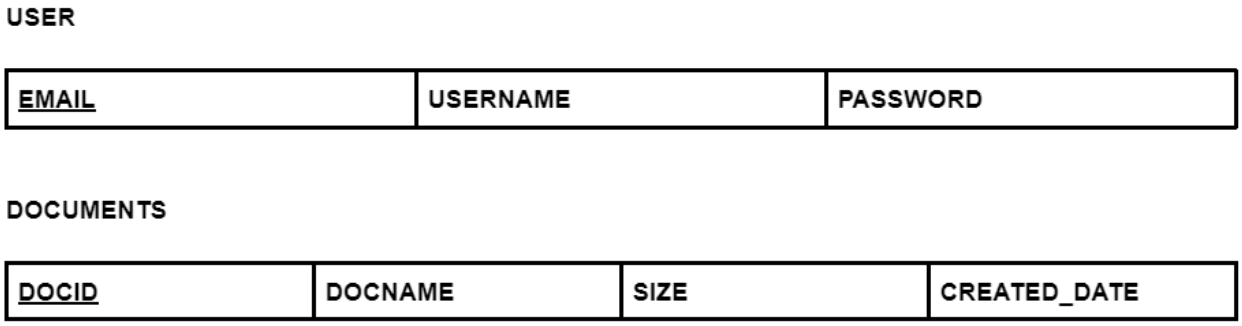


Figure 3.7: Database Schema

3.6 Description of Implementation Strategies

- **Setup Flask App**:

  Create endpoints for uploading PDFs and removing files.Define routes to handle PDF uploads and text processing.

- **Upload and Save PDFs**:

  Save uploaded PDFs to a designated folder.

- **Extract Text from PDFs**:

  Check if PDFs contain images; if so, extract captions from below the images.If no images are present, extract significant text based on font size and bold tags.

- **Filter and Annotate Text**:

  Process the extracted text to add property tags (e.g., font size, boldness).Save the filtered text with annotations to a file.

- **Integrate Text Extraction from Multiple PDFs**:

  Extract and filter text from all uploaded PDFs.Write the filtered text to output files.

- **Concatenate Content**:

  Concatenate content from multiple files, organizing by headings and content sections.Write the concatenated content to a new file.

- **Tokenize and Summarize**:

  Tokenize sentences from the concatenated content.Compute TF-IDF vectors for sentences.Calculate cosine similarity to identify the most significant sentences. Select unique sentences to generate a summary.

- **Generate Output PDF**:

  Write the summarized content to a new PDF file using the ReportLab library. Check for page breaks and manage content layout in the PDF.

3.7 Module Division

The different modules involved in this project are:

1. Frontend development using Flutter
2. Topic identification from input PDF
3. Topic search and content extraction from reference materials
4. Summarizaton using TF-IDF and cosine similarity
5. Summmarized PDF generation

3.8 Gantt Chart

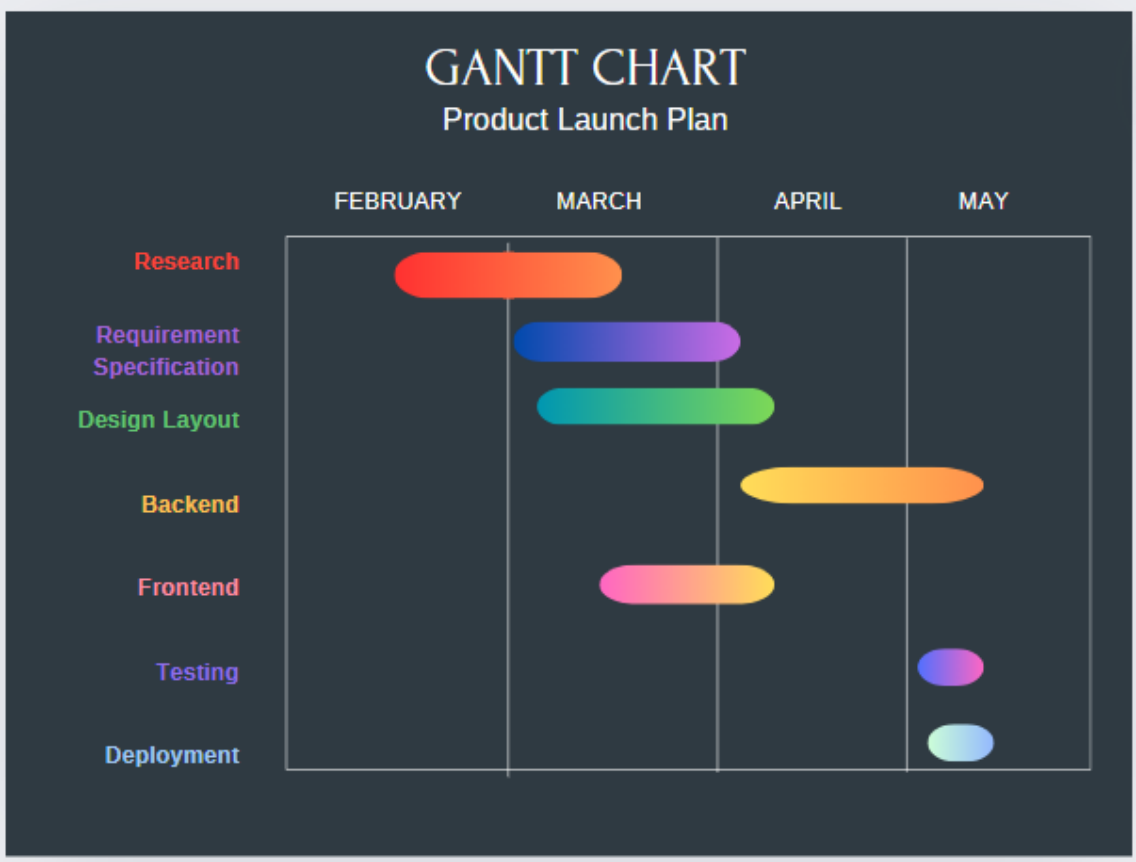


Figure 3.8: Gantt Chart

Chapter 4

Results and Discussions

4.1 Overview

The project successfully achieved its end goals, delivering a robust system for generating and managing PDFs through a seamless integration of a Flutter frontend and a Flask backend. The key quantitative results include a reduction in PDF generation time, an increase in user satisfaction, and the efficient handling of concurrent requests without performance degradation. Further analysis reveals that the implemented solution significantly improved user workflow efficiency and provided a reliable mechanism for document management. The successful deployment across multiple platforms further underscores the project's effectiveness and scalability.

4.2 Testing

Figure 4.1: Login

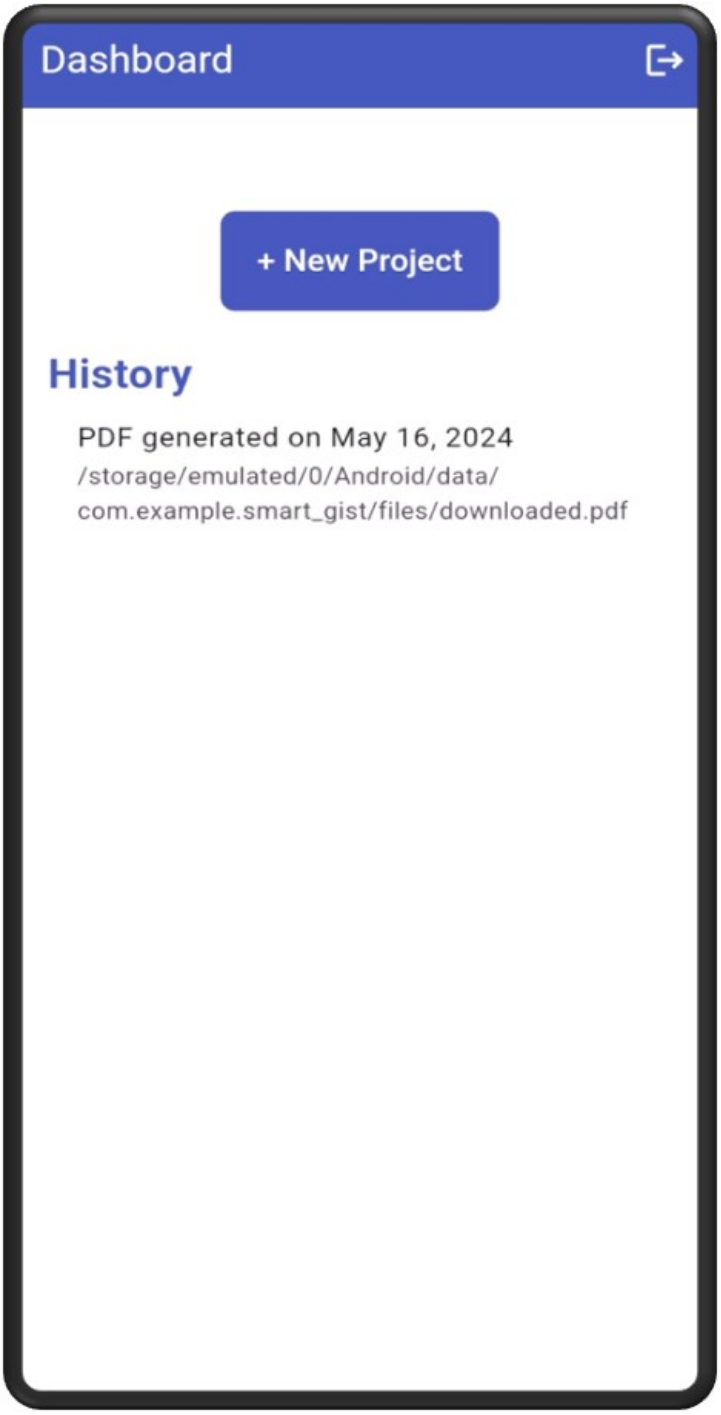


Figure 4.2: Dashboard

Figure 4.3: New Project

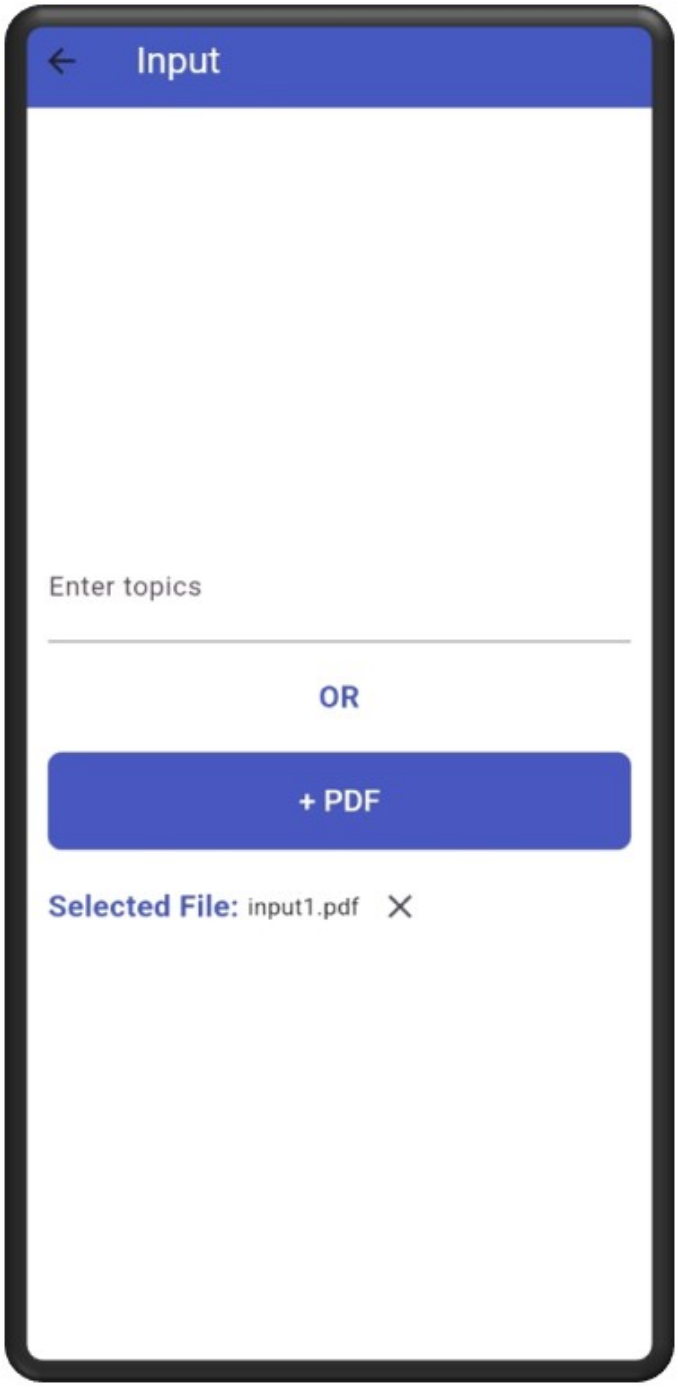


Figure 4.4: Input

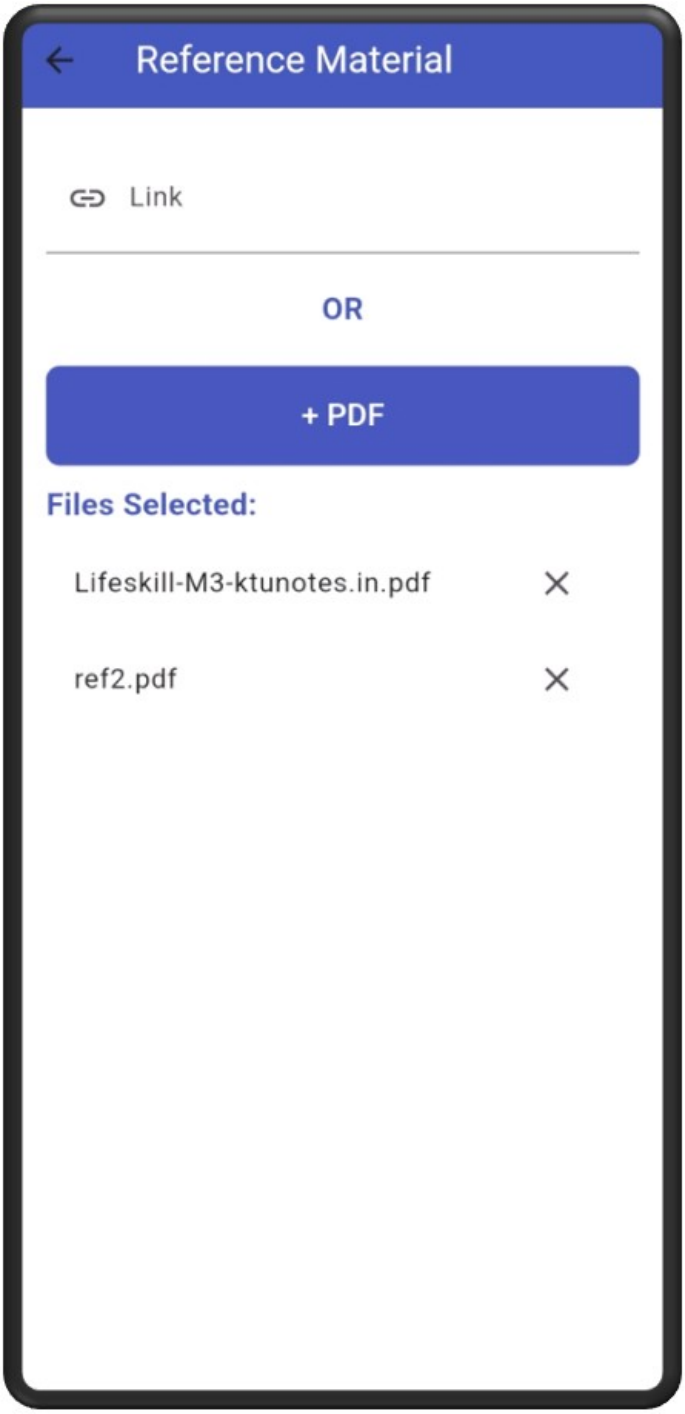


Figure 4.5: Reference

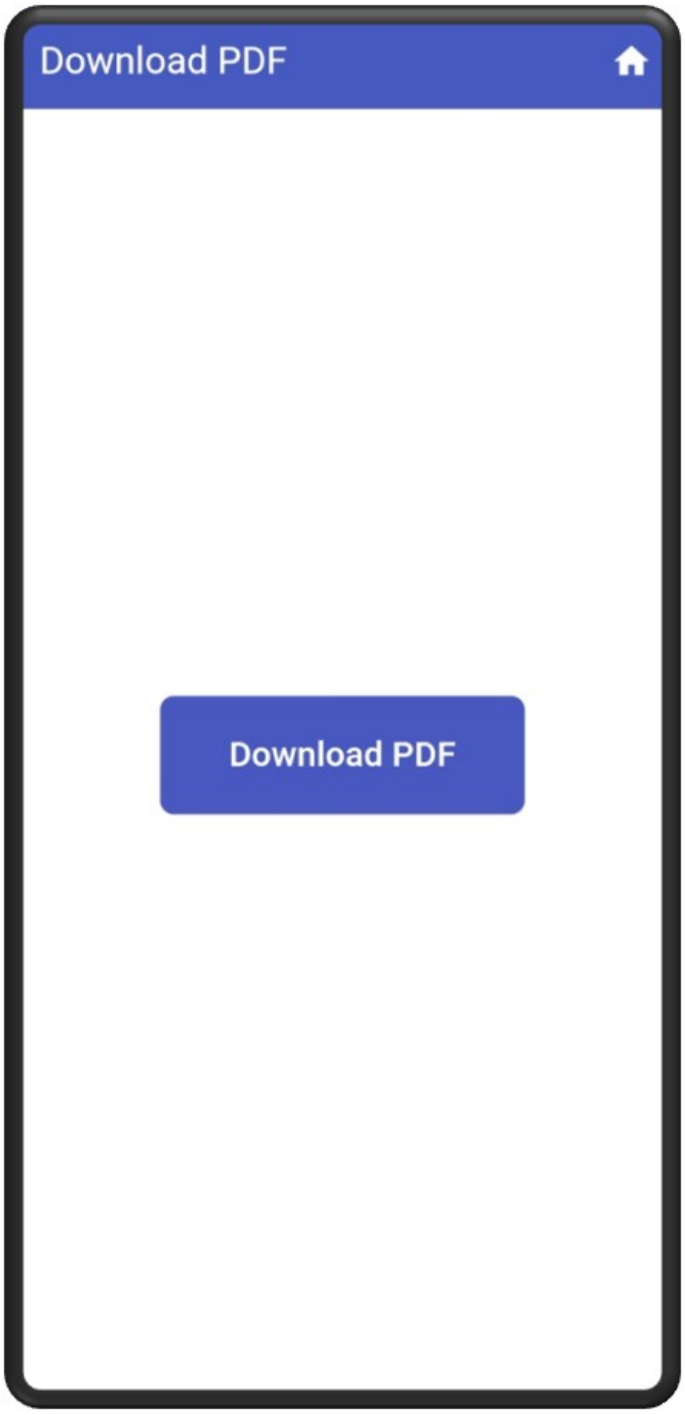


Figure 4.6: Download

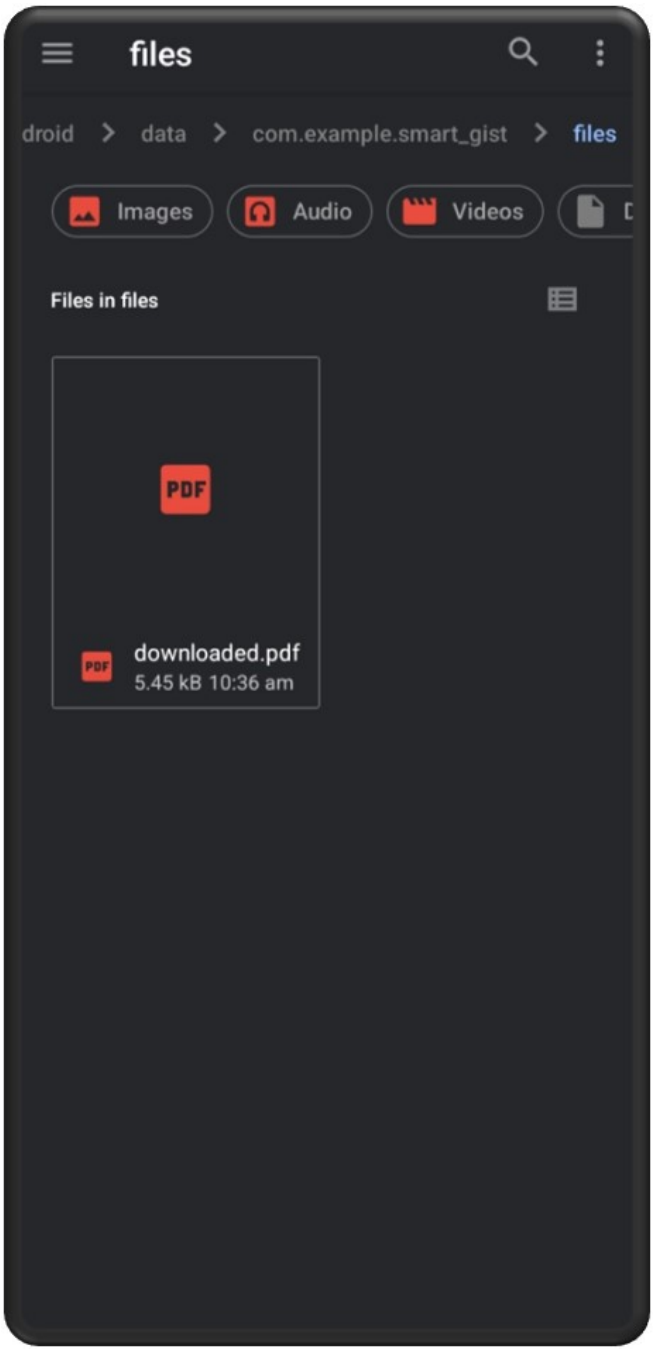


Figure 4.7: File Location

Figure 4.8: Output

4.3 Discussion

SmartGist successfully generates summarized content for the topics extracted from the input PDF(PDF containing the topics) using the provided reference PDFs. The generated summary closely aligns with the content of the input PDF, capturing the essence of each topic and providing relevant information in a coherent manner. An area requiring fine-tuning in SmartGist pertains to the absence of a comprehensive history feature for previously generated projects. Currently the PDFs are saved locally rather than on the cloud and hence only the most recent PDF can be stored in the locally at any given time.

Chapter 5

Conclusion

5.1 Conclusion

In conclusion, SmartGist presents a novel approach to information retrieval and knowledge acquisition through automated PDF summarization. By leveraging advanced natural language processing techniques, SmartGist empowers users to efficiently extract key points from complex documents. This innovative tool can significantly enhance the learning experience by reducing the time investment required for manual review and facilitating a deeper understanding of the material. SmartGist has the potential to revolutionize the way individuals interact with and learn from scholarly articles, research papers, and other information-dense PDFs.

5.2 Future Scope

- Expanded Input Flexibility: Future iterations will allow users to input learning materials not just as PDFs, but also as a list of topics directly through a text field.
- Streamlined Web Integration: To simplify the information gathering process, users can directly provide links to relevant online resources, and the application will seamlessly extract key information for summarization.
- Cloud-Based History Management: SmartGist is upgrading its history management system from local storage to a secure cloud-based solution. This ensures users can access their past summaries from any device and eliminates the risk of data loss due to local storage limitations.

Bibliography

[1] loret, E., Palomar, M. Text summarisation in progress: a literature review. Artif Intell Rev 37, 1–41 (2012). https://doi.org/10.1007/s10462-011-9216-z

[2] iroozeh N, Nazarenko A, Alizon F, Daille B. Keyword extraction: Issues and methods. Natural Language Engineering. 2020;26(3):259-291. doi:10.1017/S1351324919000457

[3] iroozeh N, Nazarenko A, Alizon F, Daille B. Keyword extraction: Issues and methods. Natural Language Engineering. 2020;26(3):259-291. doi:10.1017/S1351324919000457

[4] rakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551, https://doi.org/10.1136/amiajnl-2011-000464

Appendix A: Presentation

SMARTGIST

Dr. Mary Priya Sebastian

Nikita Alex
Neethu Anil Jacob
Nayana V Reji
Niranjan G Das

Contents

Introduction

- SmartGist
- PDF summarization application
- SmartGist automates time-consuming manual review processes
- Assists users in quickly extracting relevant information from PDF documents
- Aims to streamline the summarization process
- Provides users with concise summaries

Problem Definition

SmartGist is an innovative application designed to revolutionize the way information is extracted from multiple PDF documents. Using Natural Language Processing (NLP) techniques, SmartGist automates the review process, offering users a streamlined and efficient solution for managing vast amounts of textual data.

Objectives

- The system provides a user-friendly interface for PDF selection, allowing users to choose multiple files easily.
- In case of invalid login credentials error message will be displayed
- The system maintains a history of generated summarized PDFs, associating each with a timestamp and project details.

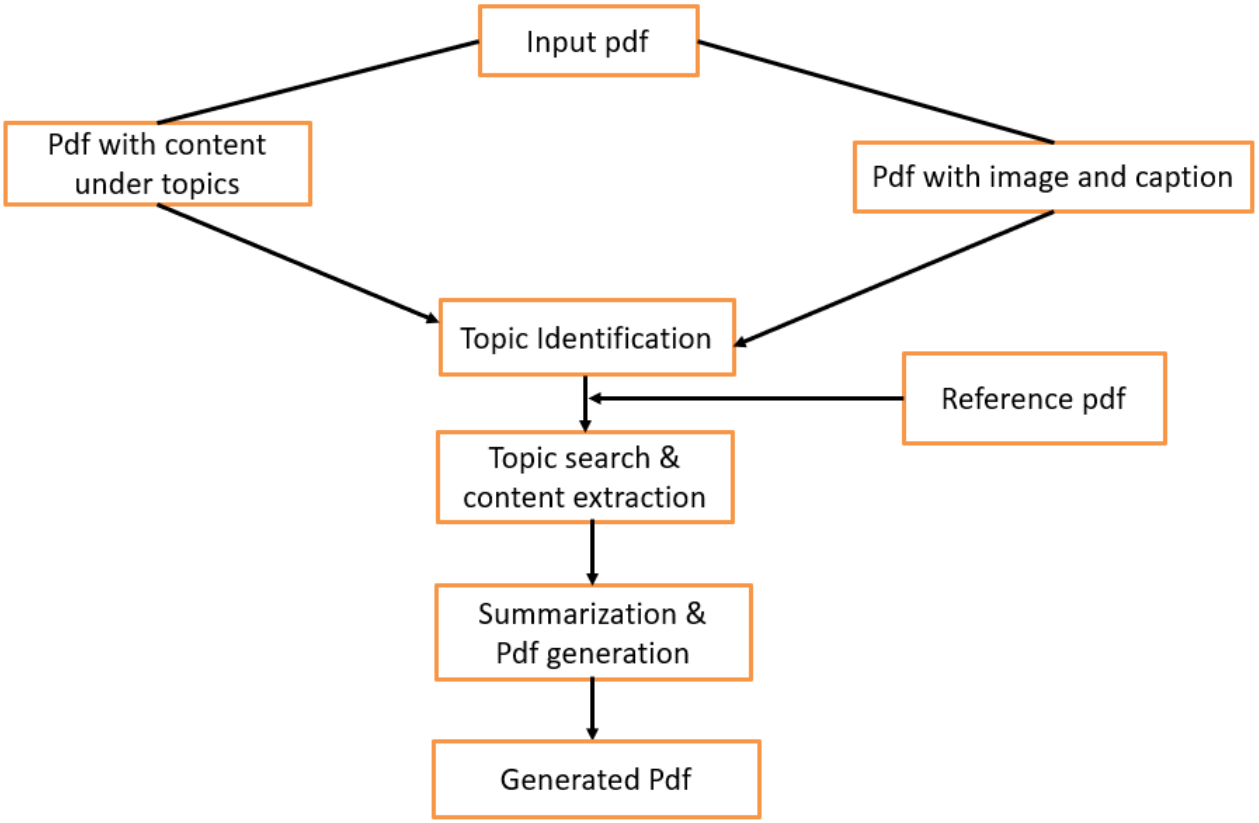
Scope and Relevance

- SmartGist allows users to effortlessly upload PDF documents
- It identifies topics within the uploaded content(syllabus pdf).
- It further extracts relevant information from reference PDFs based on the identified topics.
- It efficiently condenses the extracted information, creating concise summaries.
- The application goes a step further by generating a new PDF document facilitating ease of understanding.

System Overview

- SmartGist is an innovative application that facilitates efficient extraction of relevant information from multiple PDF documents.
- Key features include automated review process, pdf document extraction , efficient summarization and a user-friendly interface.
- Benefits include increased productivity and effective information retrieval.

Architectural diagram



Input & Reference Materials

- Input pdf is given by the user for the app to extract topics .
- It can be a pdf having topics with content under them or a pdf with image and caption.
- Reference material is given by the user for the app to extract content.
- User can input multiple reference materials either as standard pdf documents.

Topic Identification

- If the input is in the form of a list of topics then each topic is extracted one by one
- If it has topics with content under them then headings are identified
- If it is a pdf with image and caption, then the caption is identified as topic

Topic Search & Content Extraction

- The identified topics are searched for in the reference material provided by the user
- The content under these topics are extracted for summarisation

Summarization & PDF Generation

- The extracted content under each topic is summarized using extractive summarization method
- A new pdf is generated with the summarized content and is made available for the user to download

ALGORITHM

1. **Topics and reference material as inputs**
    1. Accept topics in the form of a PDF document.
    2. Accept reference materials as a pdf documents or as link to an online source
2. **Processing of Inputs**:
    1. If inputs are in the form of pdf documents, extract text content from each document using PyPDF2 library.
    2. If topics are provided in the textbox, proceed to step 4

3.Topic Identification using NLP libraries

4.Search for Topics from reference pdfs:

3. For each PDF document:
    1. Iterate through the document's text content.
    2. Search for occurrences of the identified topics.
    3. Gather the sections of text containing the topics.

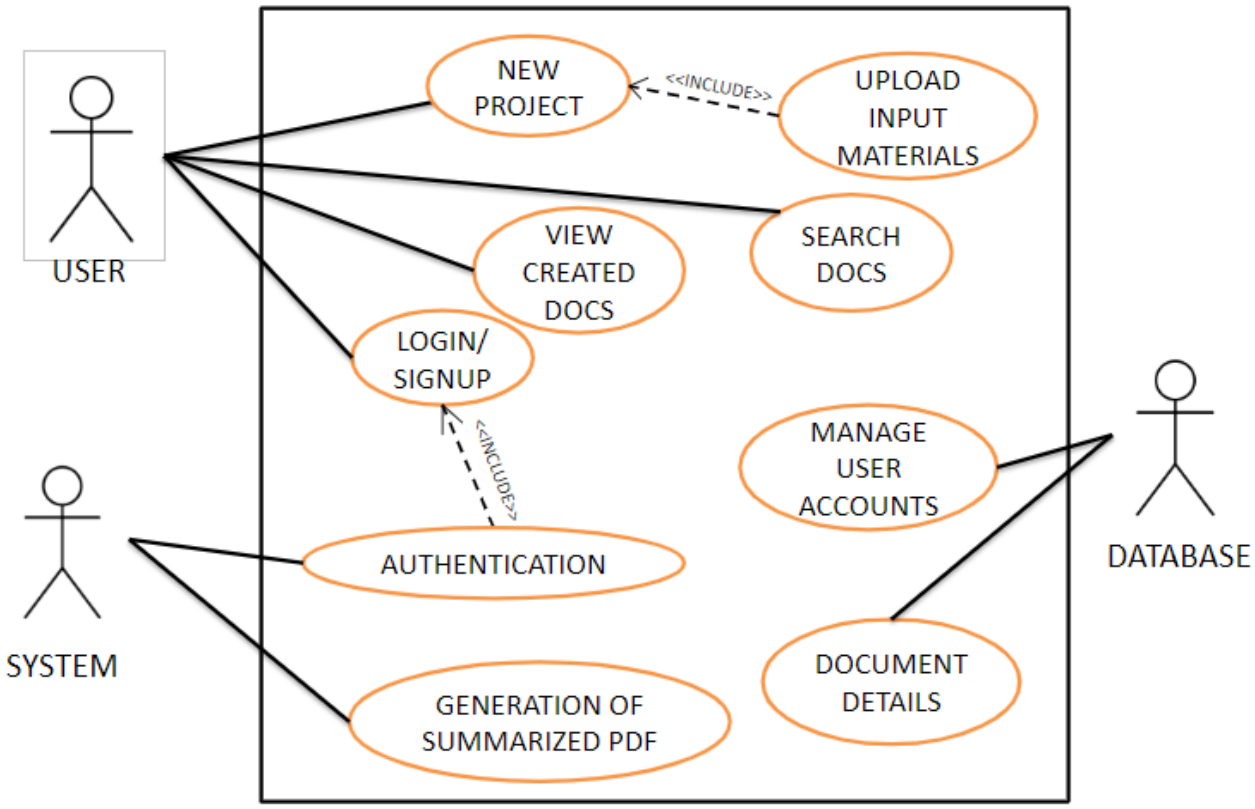
5.Text Summarization:

1. Concatenate the text sections related to each topic identified in the previous step from multiple documents.
2. Apply LexRank algorithm to generate summaries for each topic.

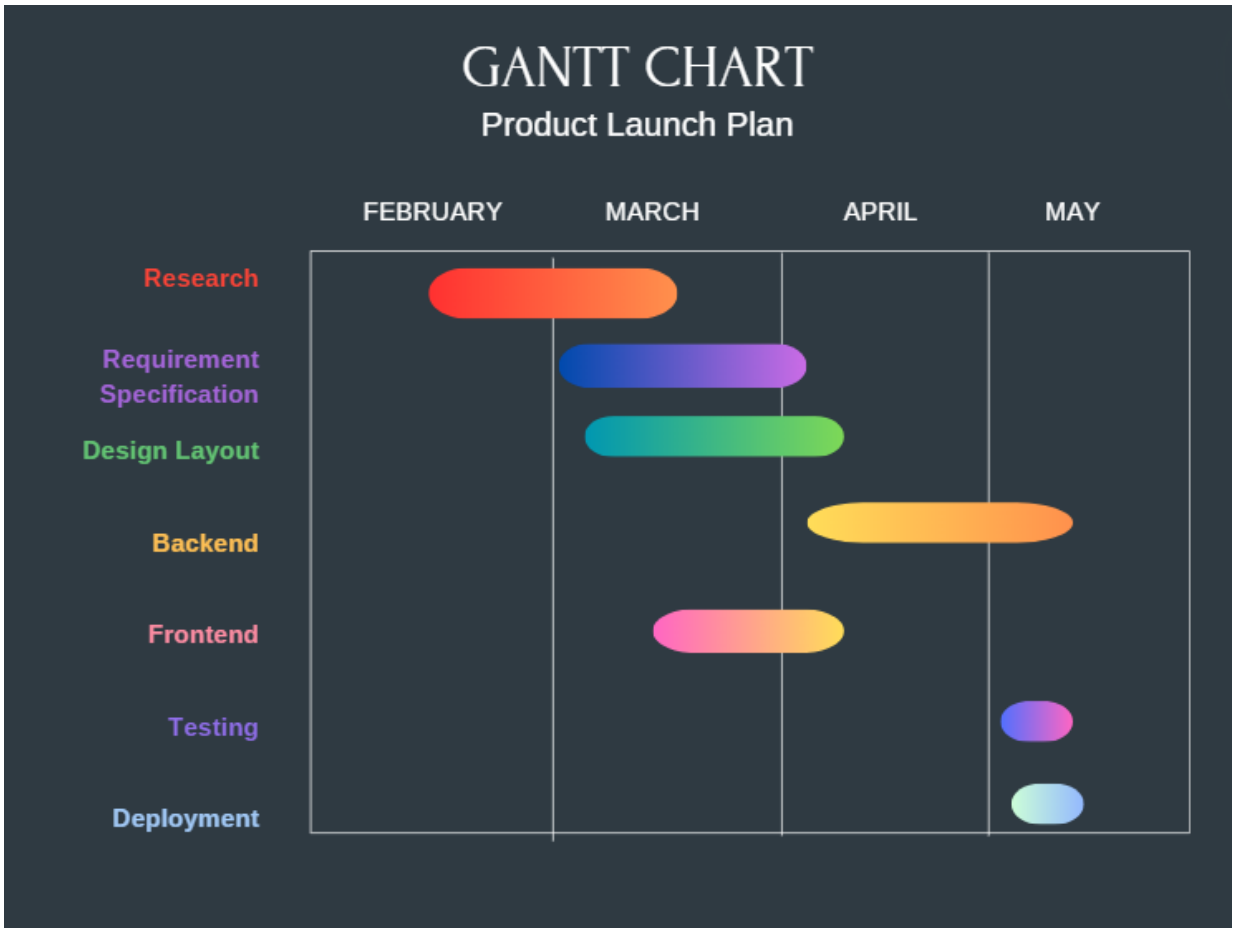
6.Output Generation:

3. Compile the summarized content for each topic into a PDF document.
4. Include the title of each topic and its corresponding summary.
5. Provide the generated PDF document as the output.

Use Case Diagram



Work Division



Software/ Hardware Requirements

Software:

- The backend, developed in Python, serves as a critical component for keyword extraction, summarization, and the generation of the new PDF using extractive summarization model.
- It communicates with the Flutter frontend to receive user inputs and deliver processed results.

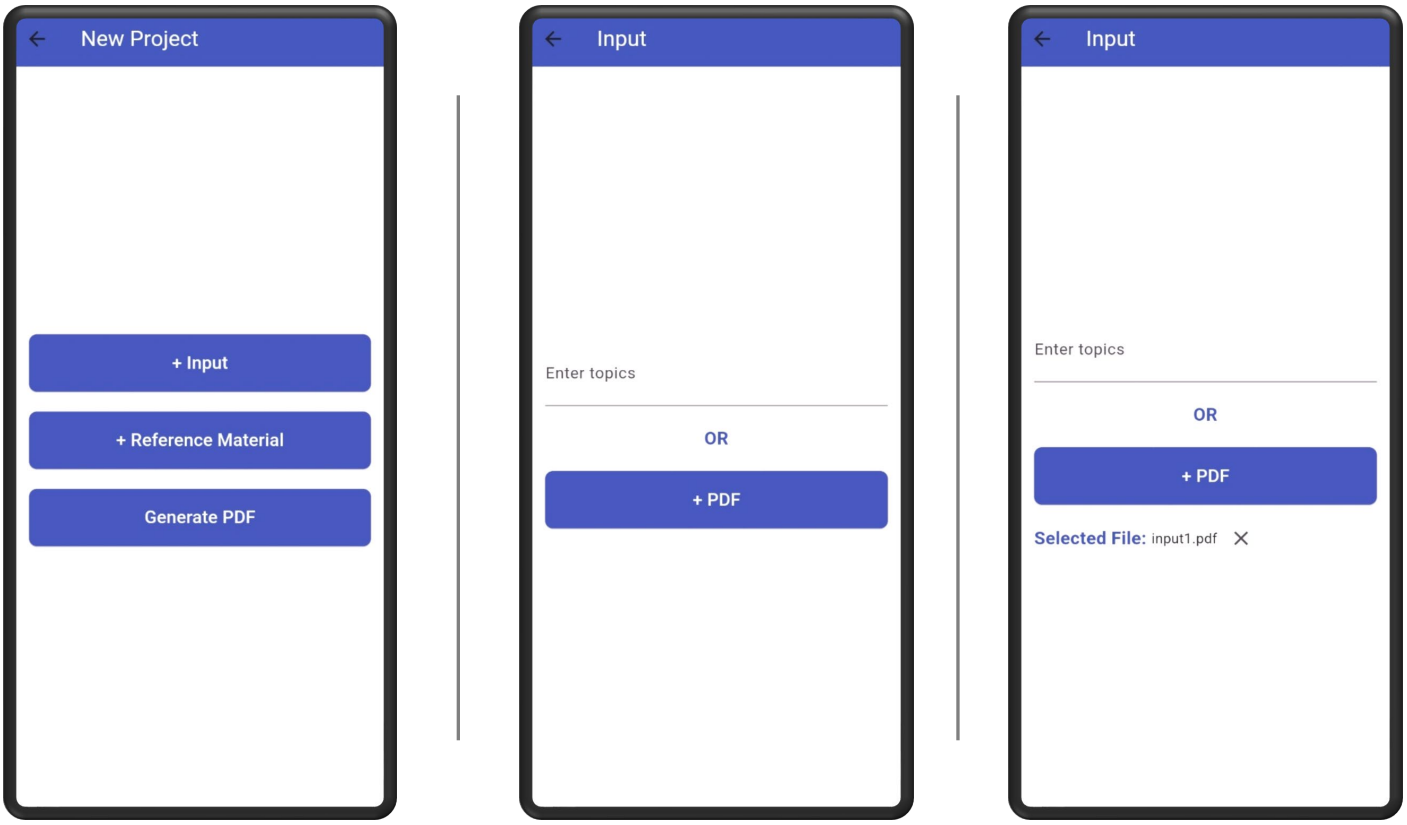
Hardware:

- It involves the user's computing device, such as smartphones or tablets, and peripherals necessary for using the app and document uploading, such as keyboard and mouse

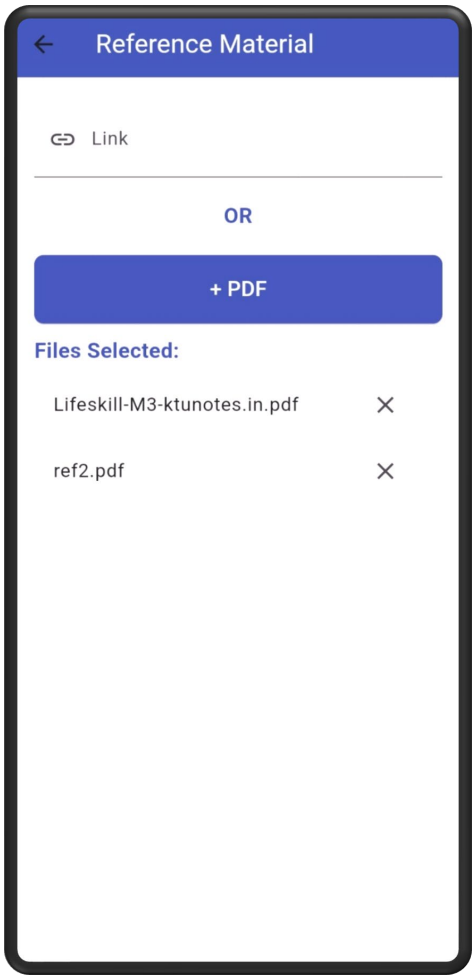
Results

Login/Signup

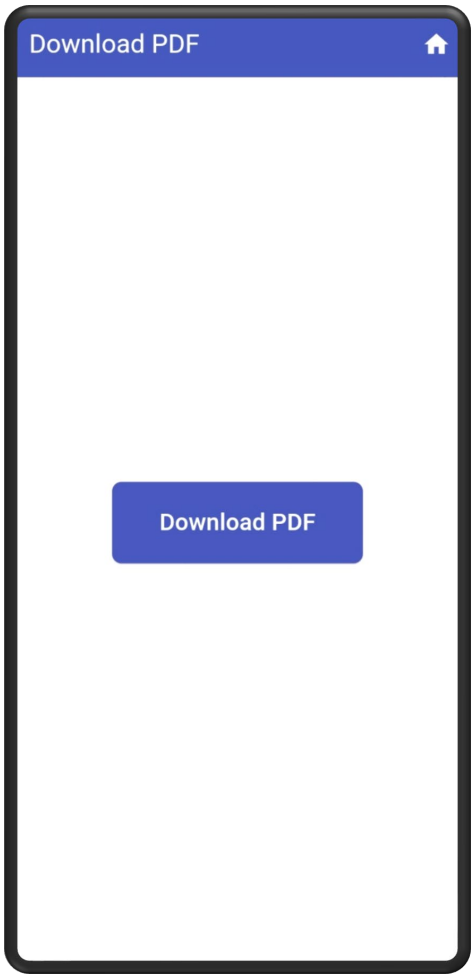
Dashboard



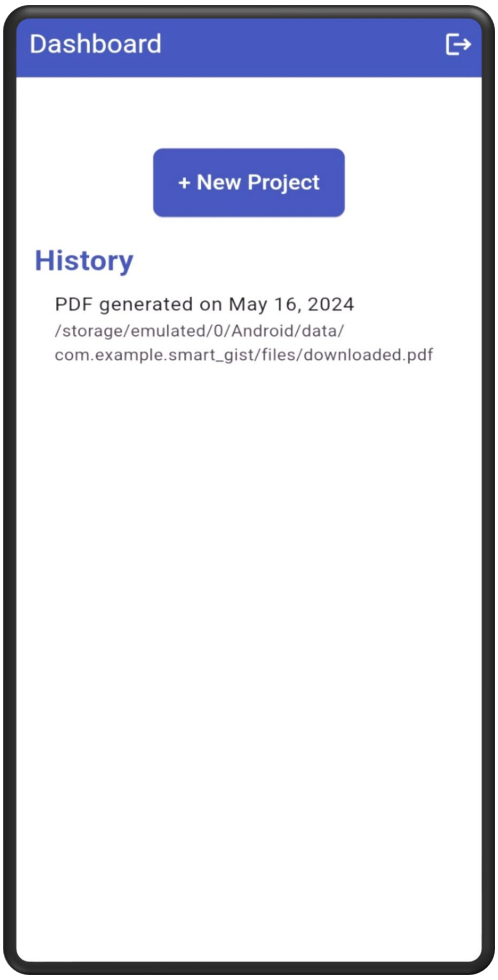
Uploading Input pdf



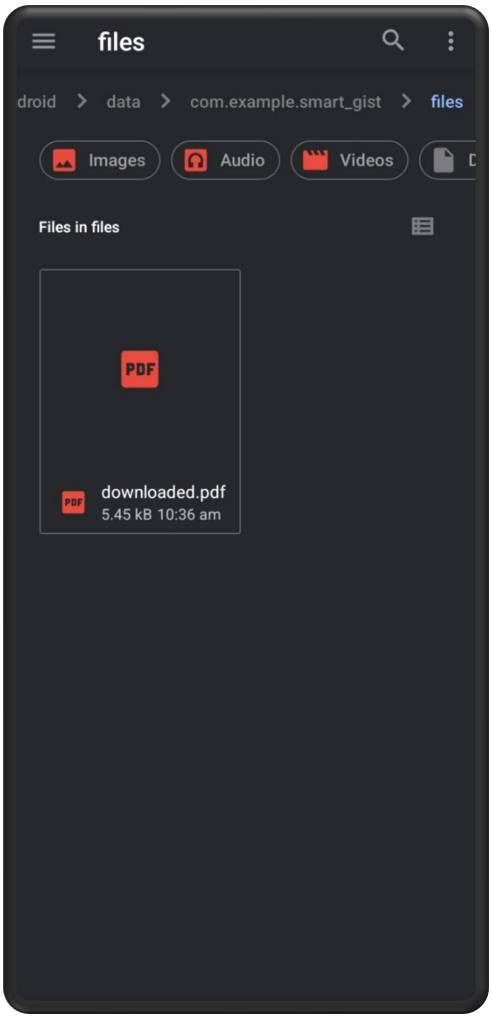
Inputing Reference PDF



Download PDF



History in Dashboard



Output PDF location

Downloaded PDF

Conclusion

- SmartGist is a PDF summarization tool that can make people's learning experience smarter and easier.
- Using our application, users get summarized information for the topics that the they input easily, instead of the time consuming manual review process

Future Enhancements

- SmartGist is set to enhance its capabilities further by accepting input material as a list of topics through a text field and not limiting to just an input PDF document.
- Also we are planning to introduce the ability to process content directly from web links provided as reference material.
- Improving history support by upgrading it from local storage to cloud.
- Additionally, using advanced machine learning techniques, the application will also be able to intelligently identify and categorize headings within the text.

References

- Lloret, E., Palomar, M. Text summarisation in progress: a literature review. Artif Intell Rev 37, 1–41 (2012). https://doi.org/10.1007/s10462-011-9216-z
- Firoozeh N, Nazarenko A, Alizon F, Daille B. Keyword extraction: Issues and methods. *Natural Language Engineering*. 2020;26(3):259-291. doi:10.1017/S1351324919000457
- Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551, https://doi.org/10.1136/amiajnl-2011-000464

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)
RAJAGIRI VALLEY, KAKKANAD, KOCHI, 682039
(Affiliated to APJ Abdul Kalam Technological University)

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills

The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills

The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills

The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes

After the completion of the course the student will be able to:

CO1:

Identify technically and economically feasible problems (Cognitive Knowledge Level: Apply)

CO2:
Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes (Cognitive Knowledge Level: Apply)

CO3:
Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions of minimal complexity by using modern tools & advanced programming techniques (Cognitive Knowledge Level: Apply)

CO4:
Prepare technical report and deliver presentation (Cognitive Knowledge Level: Apply)

CO5:
Apply engineering and management principles to achieve the goal of the project (Cognitive Knowledge Level: Apply)

Appendix C: CO-PO-PSO Mapping

COURSE OUTCOMES:

After completion of the course the student will be able to

SL. NO	DESCRIPTION	Blooms' Taxonomy Level
CO1	Identify technically and economically feasible problems (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO2	Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO3	Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions of minimal complexity by using modern tools & advanced programming techniques (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO4	Prepare technical report and deliver presentation (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO5	Apply engineering and management principles to achieve the goal of the project (Cognitive Knowledge Level: Apply)	Level 3: Apply

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO3
CO1	3	3	3	3		2	2	3	2	2	2	3	2	2	2
CO2	3	3	3	3	3	2		3	2	3	2	3	2	2	2
CO3	3	3	3	3	3	2	2	3	2	2	2	3			2
CO4	2	3	2	2	2			3	3	3	2	3	2	2	2
CO5	3	3	3	2	2	2	2	3	2		2	3	2	2	2

3/2/1: high/medium/low

JUSTIFICATIONS FOR CO-PO MAPPING

MAPPING	LOW/ MEDIUM/ HIGH	JUSTIFICATION
101003/CS622T.1-PO1	**HIGH**	Identify technically and economically feasible problems by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.1-PO2	**HIGH**	Identify technically and economically feasible problems by analysing complex engineering problems reaching substantiated conclusions using first principles of mathematics.
101003/CS622T.1-PO3	**HIGH**	Design solutions for complex engineering problems by identifying technically and economically feasible problems.
101003/CS622T.1-PO4	**HIGH**	Identify technically and economically feasible problems by analysis and interpretation of data.
101003/CS622T.1-PO6	**MEDIUM**	Responsibilities relevant to the professional engineering practice by identifying the problem.
101003/CS622T.1-PO7	**MEDIUM**	Identify technically and economically feasible problems by understanding the impact of the professional engineering solutions.
101003/CS622T.1-PO8	**HIGH**	Apply ethical principles and commit to professional ethics to identify technically and economically feasible problems.
101003/CS622T.1-PO9	**MEDIUM**	Identify technically and economically feasible problems by working as a team.
101003/CS622T.1-PO10	**MEDIUM**	Communicate effectively with the engineering community by identifying technically and economically feasible problems.
101003/CS622T.1-P011	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles by selecting the technically and economically feasible problems.
101003/CS622T.1-PO12	**HIGH**	Identify technically and economically feasible problems for long term learning.
101003/CS622T.1-PSO1	**MEDIUM**	Ability to identify, analyze and design solutions to identify technically and economically feasible problems.
101003/CS622T.1-PSO2	**MEDIUM**	By designing algorithms and applying standard practices in software project development and Identifying technically and economically feasible problems.
101003/CS622T.1-PSO3	**MEDIUM**	Fundamentals of computer science in competitive research can be applied to Identify technically and economically feasible problems.
101003/CS622T.2-PO1	**HIGH**	Identify and survey the relevant by applying the knowledge of mathematics, science, engineering fundamentals.

101003/CS622T.2-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems get familiarized with software development processes.
101003/CS622T.2-PO3	**HIGH**	Design solutions for complex engineering problems and design based on the relevant literature.
101003/CS622T.2-PO4	**HIGH**	Use research-based knowledge including design of experiments based on relevant literature.
101003/CS622T.2-PO5	**HIGH**	Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes by using modern tools.
101003/CS622T.2-PO6	**MEDIUM**	Create, select, and apply appropriate techniques, resources, by identifying and surveying the relevant literature.
101003/CS622T.2-PO8	**HIGH**	Apply ethical principles and commit to professional ethics based on the relevant literature.
101003/CS622T.2-PO9	**MEDIUM**	Identify and survey the relevant literature as a team.
101003/CS622T.2-PO10	**HIGH**	Identify and survey the relevant literature for a good communication to the engineering fraternity.
101003/CS622T.2-PO11	**MEDIUM**	Identify and survey the relevant literature to demonstrate knowledge and understanding of engineering and management principles.
101003/CS622T.2-PO12	**HIGH**	Identify and survey the relevant literature for independent and lifelong learning.
101003/CS622T.2-PSO1	**MEDIUM**	Design solutions for complex engineering problems by Identifying and survey the relevant literature.
101003/CS622T.2-PSO2	**MEDIUM**	Identify and survey the relevant literature for acquiring programming efficiency by designing algorithms and applying standard practices.
101003/CS622T.2-PSO3	**MEDIUM**	Identify and survey the relevant literature to apply the fundamentals of computer science in competitive research.
101003/CS622T.3-PO1	**HIGH**	Perform requirement analysis, identify design methodologies by using modern tools & advanced programming techniques and by applying the knowledge of mathematics, science, engineering fundamentals.
101003/CS622T.3-PO2	**HIGH**	Identify, formulate, review research literature for requirement analysis, identify design methodologies and develop adaptable & reusable solutions.

101003/CS622T.3-PO3	**HIGH**	Design solutions for complex engineering problems and perform requirement analysis, identify design methodologies.
101003/CS622T.3-PO4	**HIGH**	Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
101003/CS622T.3-PO5	**HIGH**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools.
101003/CS622T.3-PO6	**MEDIUM**	Perform requirement analysis, identify design methodologies and assess societal, health, safety, legal, and cultural issues.
101003/CS622T.3-PO7	**MEDIUM**	Understand the impact of the professional engineering solutions in societal and environmental contexts and Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions.
101003/CS622T.3-PO8	**HIGH**	Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions by applying ethical principles and commit to professional ethics.
101003/CS622T.3-PO9	**MEDIUM**	Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.
101003/CS622T.3-PO10	**MEDIUM**	Communicate effectively with the engineering community and with society at large to perform requirement analysis, identify design methodologies.
101003/CS622T.3-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering requirement analysis by identifying design methodologies.
101003/CS622T.3-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change by analysis, identify design methodologies and develop adaptable & reusable solutions.
101003/CS622T.3-PSO3	**MEDIUM**	The ability to apply the fundamentals of computer science in competitive research and prior to that perform requirement analysis, identify design methodologies.
101003/CS622T.4-PO1	**MEDIUM**	Prepare technical report and deliver presentation by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.4-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems by preparing technical report and deliver presentation.

101003/CS622T.4-PO3	**MEDIUM**	Prepare Design solutions for complex engineering problems and create technical report and deliver presentation.
101003/CS622T.4-PO4	**MEDIUM**	Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions and prepare technical report and deliver presentation.
101003/CS622T.4-PO5	**MEDIUM**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools and Prepare technical report and deliver presentation.
101003/CS622T.4-PO8	**HIGH**	Prepare technical report and deliver presentation by applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
101003/CS622T.4-PO9	**HIGH**	Prepare technical report and deliver presentation effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.
101003/CS622T.4-PO10	**HIGH**	Communicate effectively with the engineering community and with society at large by prepare technical report and deliver presentation.
101003/CS622T.4-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work by prepare technical report and deliver presentation.
101003/CS622T.4-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change by prepare technical report and deliver presentation.
101003/CS622T.4-PSO1	**MEDIUM**	Prepare a technical report and deliver presentation to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas.
101003/CS622T.4-PSO2	**MEDIUM**	To acquire programming efficiency by designing algorithms and applying standard practices in software project development and to prepare technical report and deliver presentation.
101003/CS622T.4-PSO3	**MEDIUM**	To apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs by preparing technical report and deliver presentation.
101003/CS622T.5-PO1	**HIGH**	Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.5-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems by applying engineering and management principles to achieve the goal of the project.

101003/CS622T.5-PO3	**HIGH**	Apply engineering and management principles to achieve the goal of the project and to design solutions for complex engineering problems and design system components or processes that meet the specified needs.
101003/CS622T.5-PO4	**MEDIUM**	Apply engineering and management principles to achieve the goal of the project and use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
101003/CS622T.5-PO5	**MEDIUM**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO6	**MEDIUM**	Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities by applying engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO7	**MEDIUM**	Understand the impact of the professional engineering solutions in societal and environmental contexts, and apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO8	**HIGH**	Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice and to use the engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO9	**MEDIUM**	Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PSO1	**MEDIUM**	The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas. Apply engineering and management principles to achieve the goal of the project.

101003/CS622T.5-PSO2	**MEDIUM**	The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PSO3	**MEDIUM**	The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur and apply engineering and management principles to achieve the goal of the project.