



Project and Professionalism

(6CS007)

**Project Proposal**

Document to Text Summarization using Machine Learning

Student Id : 2065697

Student Name : Dhiraj Kumar Sah Kanu

Group : L6CG1

Supervisor : Mr. Yamu Poudel

Reader : Mr. Akash Adhikari

Word Count : 2208

**Abstract**

This proposal provides an overview of my final year project on developing a web application for document to text summarization using machine learning. This project presents a dynamic solution to address the challenge of information overload by leveraging advanced technologies. The project imposes the development of a web-based platform that encompasses a range of functionalities to simplify the summarization process including user registration, document upload, summary display. The system will be using NLP algorithms and will also be offering an API-based summarization model. The development methodology adopted for this project is the Kanban approach, providing a flexible and responsive framework. The tools and technologies that will be used are Django framework, with frontend HTML, CSS, JavaScript. Similarly, the summaries generated by the system will be stored in the MariaDB database for easy access and retrieval.

*Keywords: Document Summarization, Natural Language Processing, Machine Learning, Textual Data, Web Application, Information Extraction*

**Acknowledgement**

I would like to express, sincere gratitude to the individuals and institutions that have played an important role in the successful initiation of the final year project. I extend my sincere thanks to the administration of University of Wolverhampton for providing us with this opportunity to pursue my study and research. Also, would like to thank Herald College Kathmandu, for offering such academic facilities and resources that have been valuable for shaping my skills.

I extend my gratitude to our respected Course Leader, Mr. Biraj Dulal, for his sessions on the Project and Professionalism module and for providing us with the documents and materials required for initiating the project. I am indebted to my Supervisor, Mr. Yamu Poudel, for his dedicated mentorship and invaluable advice and continuous support. I extend my appreciation to my Reader, Mr. Akash Adhikari, for his thoughtful insights and suggestions that have enriched my work, especially in machine learning. Lastly, I would like to thank my mentor, Mr. Bhanu Aryal for his belief in my potential and consistent guidance.

**Table of Contents**

[1. Introduction 1](#_Toc142914084)

[1.1. Project Title 1](#_Toc142914085)

[1.2. Academic Question 1](#_Toc142914086)

[2. Aims and Objectives 1](#_Toc142914087)

[2.1. Aims 1](#_Toc142914088)

[2.2. Objectives 2](#_Toc142914089)

[3. Problem Statement 2](#_Toc142914090)

[4. Project as Solution 3](#_Toc142914091)

[5. Scope and Limitations 3](#_Toc142914092)

[5.1. Scope 3](#_Toc142914093)

[5.2. Limitations 3](#_Toc142914094)

[6. Initial Research 4](#_Toc142914095)

[6.1. Hugging Face 4](#_Toc142914096)

[6.2. Natural Language Processing 4](#_Toc142914097)

[6.2.1. Abstractive Summarization 4](#_Toc142914098)

[6.2.2. Extractive Summarization 4](#_Toc142914099)

[7. Literature Review 5](#_Toc142914100)

[7.1. NLP Based Text Summarization Using Semantic Analysis (Moiyadi, et al., 2016) 5](#_Toc142914101)

[7.1.1. Introduction 5](#_Toc142914102)

[7.1.2. Methodology 5](#_Toc142914103)

[7.1.3. Conclusion 6](#_Toc142914104)

[8. Functional Decomposition Diagram 7](#_Toc142914105)

[9. Artifacts 7](#_Toc142914106)

[9.1. Frontend Subsystem 7](#_Toc142914107)

[9.1.1. User Registration 7](#_Toc142914108)

[9.1.2. Document Upload 7](#_Toc142914109)

[9.1.3. Summary Display 7](#_Toc142914110)

[9.2. NLP Subsystem 8](#_Toc142914111)

[9.2.1. API Based Summarizer 8](#_Toc142914112)

[9.2.2. Custom Summarizer 8](#_Toc142914113)

[9.3. Backend Subsystem 8](#_Toc142914114)

[9.3.1. Database Management 8](#_Toc142914115)

[9.3.2. API Management 8](#_Toc142914116)

[10. Software Development Methodology 8](#_Toc142914117)

[11. Work Breakdown Structure (WBS) 10](#_Toc142914118)

[12. Gantt Chart 10](#_Toc142914119)

[13. Tools and Technology 11](#_Toc142914120)

[13.1. HTML/ CSS/ JS 11](#_Toc142914121)

[13.2. Django Framework 11](#_Toc142914122)

[13.3. MariaDB 12](#_Toc142914123)

[13.4. NLP Model 12](#_Toc142914124)

[Bibliography 13](#_Toc142914125)

**Table of Figures**

[Figure 1 Proposed System (Moiyadi, et al., 2016) 6](#_Toc142906655)

[Figure 2 Functional Decomposition Diagram of the system 7](#_Toc142906656)

[Figure 3 A glimpse to Trello 9](#_Toc142906657)

[Figure 4 Work Breakdown Structure 10](#_Toc142906658)

[Figure 5 Gantt Chart 11](#_Toc142906659)

# Introduction

The project chosen is a document to text summarization web application that will help the user summarize content of documents, making it easier for the users to extract necessary and important information without having to read the entire document. Due to the large amount of data, individuals and organizations face the challenges of managing vast amounts of textual data, including lengthy documents, research papers, and articles. Extracting essential insights from these documents is a time consuming and resource intensive task. The proposed document summarization tool aims to address the problem by leveraging NLP and machine learning techniques.

## Project Title

Document to Text Summarization using Machine Learning

## Academic Question

1. How does the tool address the problem of information and time-consuming content review?
2. What are the potential benefits of using NLP and machine learning algorithms in the summarization process?
3. What existing NLP algorithms and approaches will be used to automate the text summarization?
4. What techniques will be used to handle document parsing and processing for document format (e.g., PDF)?

# Aims and Objectives

## Aims

1. To develop a Natural Language Processing (NLP) based document summarization model.
2. To create a tool for summarizing lengthy documents, research papers, and articles.
3. To provide the users with concise summaries for improved content comprehension.
4. To improve efficiency in learning and reading contents.
5. To save time and increase interest in reading document-based contents.

## Objectives

1. A web application will be created that will allow users to input documents and get summaries generated.
2. Natural Language Toolkit (NLTK) or spaCy technology will be used to handle document parsing and processing for various formats.
3. A suitable Natural Language Processing (NLP) summarization model will be chosen.
4. The summarized set of documents will then be stored in the database from the documents.

# Problem Statement

Due to the expansion in digital information, a challenge of information overloading has been seen. Dealing with lengthy documents and research materials becomes time consuming, hindering the ability to extract crucial information quickly. A study published in Nature Communication reveals that our collective attention span is narrowing. This is not just on social media but also in different domains, that people are losing their attention span (Lorenz-Spreen, et al., 2019). Due to the low attention span, most individuals are not likely to read journals, books, and other documented files. The rise of audio-visual content has also impacted the consumption of text-based information. These days the way of consuming information has changed which is a positive thing but in context of a person with academic background in performing research based on text-based documents, one of the pain-points I have been facing is analyzing and curating the contents from lengthy documents. In the professional world, handling and extracting major key points from proposals and documents could be a terrible task for working individuals until there is a summarization tool that makes the task easier.

# Project as Solution

Professionals, researchers, and students face challenges of extracting content from research papers and journals. This summarization tool will extract important insights from a paper to make informed decisions quickly. Dealing with lengthy documents that require significant time and effort is a terrible task, the project will be able to minimize the time required and will prioritize important documents. This tool will help streamline the process of summarizing and aggregating information from multiple sources and will store it for further usage of the user. People will be able to access necessary content in a short format, saving time. It will encourage continuous learning by allowing users to explore a broader range of topics within a shorter time frame.

# Scope and Limitations

## Scope

1. The major scope of this application is that it can summarize pdf-based documents including research papers, articles, and textual content.
2. It will be able to identify key points, findings, and contextual information from lengthy documents.

## Limitations

1. Summarization algorithms may not capture the complete context of a document.
2. The tool will be limited to cross-language summarization feature.
3. The tool focuses on text-based documents and may not be able to effectively handle images, charts, and videos.
4. The summarizer might be limited to certain word count.

# Initial Research

## Hugging Face

Hugging Face is a platform for data science and machine learning which was initially developed in 2017 as a chatbot app for teenagers but has expanded its domain from just a chatbot app. This platform has a range of functionalities from creating, hosting, and managing their own AI models to engaging in discussions and collaborations within the platform. Hugging face also holds a substantial collection of models and datasets created by the collaborators and users. This platform encompasses a range of AI applications like computer vision, audio analysis, and natural language processing (Rebelo, 2023).

## Natural Language Processing (NLP)

Natural Language Processing (NLP) is a division of AI that aims to make computers comprehend, analyze, and produce human language. NLP plays and significant role in extracting the most meaningful information. NLP encompasses various techniques and methods to process and understand human language. The division of NLP approaches extractive and abstractive summarization, along with other categories such as single-document, multi-document, and query-based summarization.

### Abstractive Summarization

Abstractive summarization is a method where the generated summary doesn’t necessarily consist of sentences present in the original text, instead it creates new sentences that capture the key concepts. This is a approach where the new texts are generated only after abstracting and understanding the semantic meaning of the text while capturing the key concepts of the original content.

### Extractive Summarization

Similarly, Extractive summarization involves choosing sentences directly from the source text to create a concise summarization. Then the chosen text based content is typically chosen based on its importance and relevance. This method aims to retain the original wording from the text, making it more conservative approach to summarization (Adhikari, et al., 2020).

# Literature Review

## NLP Based Text Summarization Using Semantic Analysis (Moiyadi, et al., 2016)

### Introduction

Text Summarization is a supervised learning method that heavily relies on Natural Language Processing (NLP), which helps in generating concise and comprehensive summaries from a massive number of text-based inputs.

This literature review focuses on the research paper titled “NLP Based Text Summarization Using Semantic Analysis” which introduces an innovative approach for specific text summarization. To save manual effort in information processing, the study discusses the challenges of summarizing large texts and highlights the use of automatic summarization tools. The author offers a method based on Semantic Analysis, a subset of Natural Language Processing (NLP), which helps in extracting representative sentences and generates meaningful summaries.

### Methodology

The proposed methodology in this research paper leverages a technique based on Latent Semantic Analysis (LSA) for summarizing text-based data. LSA is used to extract semantically important sentences and generate the summary of the input text. Firstly, the input document undergoes a data preprocessing method, where irrelevant words such as stop, words are removed to enhance the quality of the summary. A matrix of word counts is constructed from preprocessed text, representing the occurrence of words across sentences. The sentence matrix undergoes Singular Value Decomposition (SVD) to reduce dimensionality and identify the latent semantic structure of the text. This helps in identifying the most noticeable feature of the document, enabling extraction of meaningful sentences for summary. Based on latent semantic structure, the system identifies the most relevant sentence and arranges them in a coherent manner to generate the final summary.

A screenshot of a computer screen

Description automatically generated

Figure Proposed System (Moiyadi, et al., 2016)

### Conclusion

The authors applied LSA-based summarization techniques to a diverse set of texts and evaluated the performance of generated summaries. They measured the quality of the summaries using various metrics, including ROUGE scores, F1-score, and precision-recall curves. The study indicates that the LSA approach performed better than existing summarization techniques.

# Functional Decomposition Diagram

A purple rectangular object with text

Description automatically generated

Figure Functional Decomposition Diagram of the system

# Artifacts

The proposed artifacts outline the components and features for document to text summarization system. The system comprises different subsystems, each with specific functionalities.

## Frontend Subsystem

### User Registration

This will be the first user interaction, which will allow the user to register by creating a user account and login with their credentials. Here, the user needs to provide their basic information to access the system.

### Document Upload

Users can interact with the system by uploading documents they wish to summarize. This feature provides users to input the documents into the system.

### Summary Display

After the document is uploaded, this feature will play an important role in displaying the summarized data to the user.

## NLP Subsystem

### API Based Summarizer

This functionality will integrate an external summarization API from hugging face or any reliable source which will allow summarizing any kind of documents.

### Custom Summarizer

This system will be trained with custom data especially for a specific kind of text-based file such as book and will summarize the contents based on its specific feature.

## Backend Subsystem

## Database Management

The system will use MariaDB to manage data storage and will be storing user information, uploaded documents and generated summaries allowing the users to access and review the summaries of their uploaded documents whenever they need.

### API Management

This functionality will be involved in managing the data fetched from the API and will be helping in allowing the external summarization API.

# Software Development Methodology

To organize the software development cycle, Kanban Methodology will be used. Kanban is a method used to manage and optimize workflows in project management by emphasizing visual work, limiting work in progress, and promoting continuous improvement. Choosing Kanban for individual software development offers distinct advantages over other SDLC methodologies. Kanban allows high flexibility as it deals with prioritizing tasks based on their importance and urgency since solo developers often face quick change in priority. It encourages a constant flow of deliverables leading to quicker feedback. It allows individual developers to manage their work in a way that suits their own pace and style, as it does not have a predefined iteration like Scrum, allowing them not to spend time on detailed planning sessions or assigning tasks (Raut, 2015).

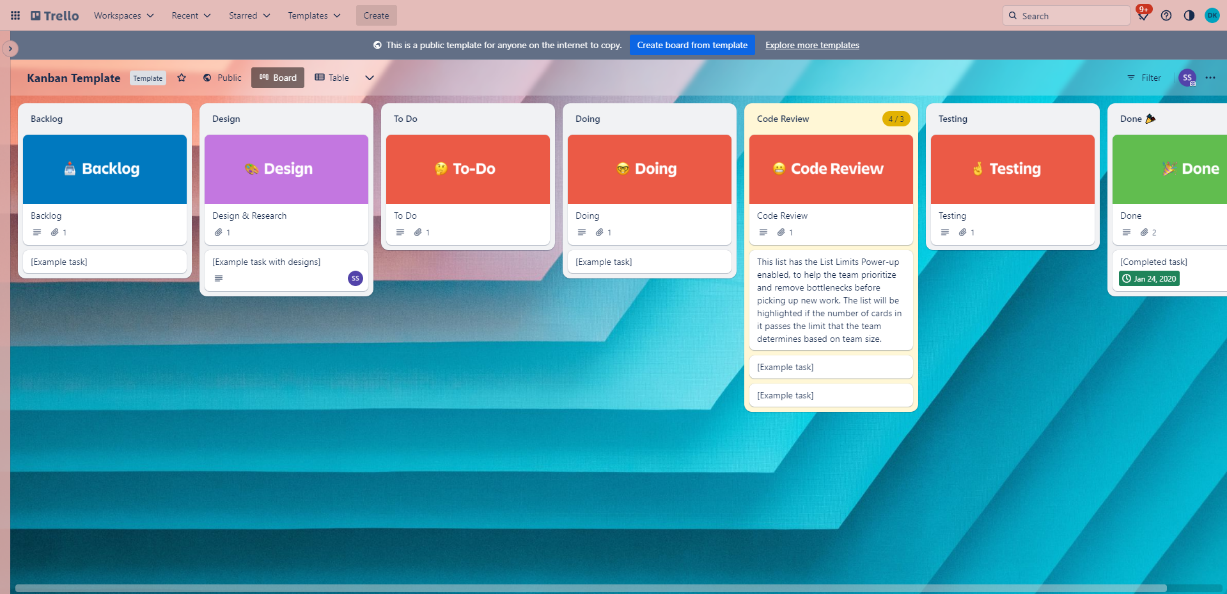


Figure A glimpse to Trello

Trello will be utilized to follow the Kanban method, utilizing its visual card-based system for effective task and project management. Trello was launched in 2011, attributed to user friendly design and adaptability which helped it in gaining popularity. In Trello, the tasks are visualized as cards or sticky notes on a visual board which is kanban and, usually divided into columns that represent different stages of the workflow, such as ‘To Do’, ‘In Progress’ and ‘Done’. These cards are movable across the columns to reflect the progress of the work (Johnson, 2021).

# Work Breakdown Structure (WBS)

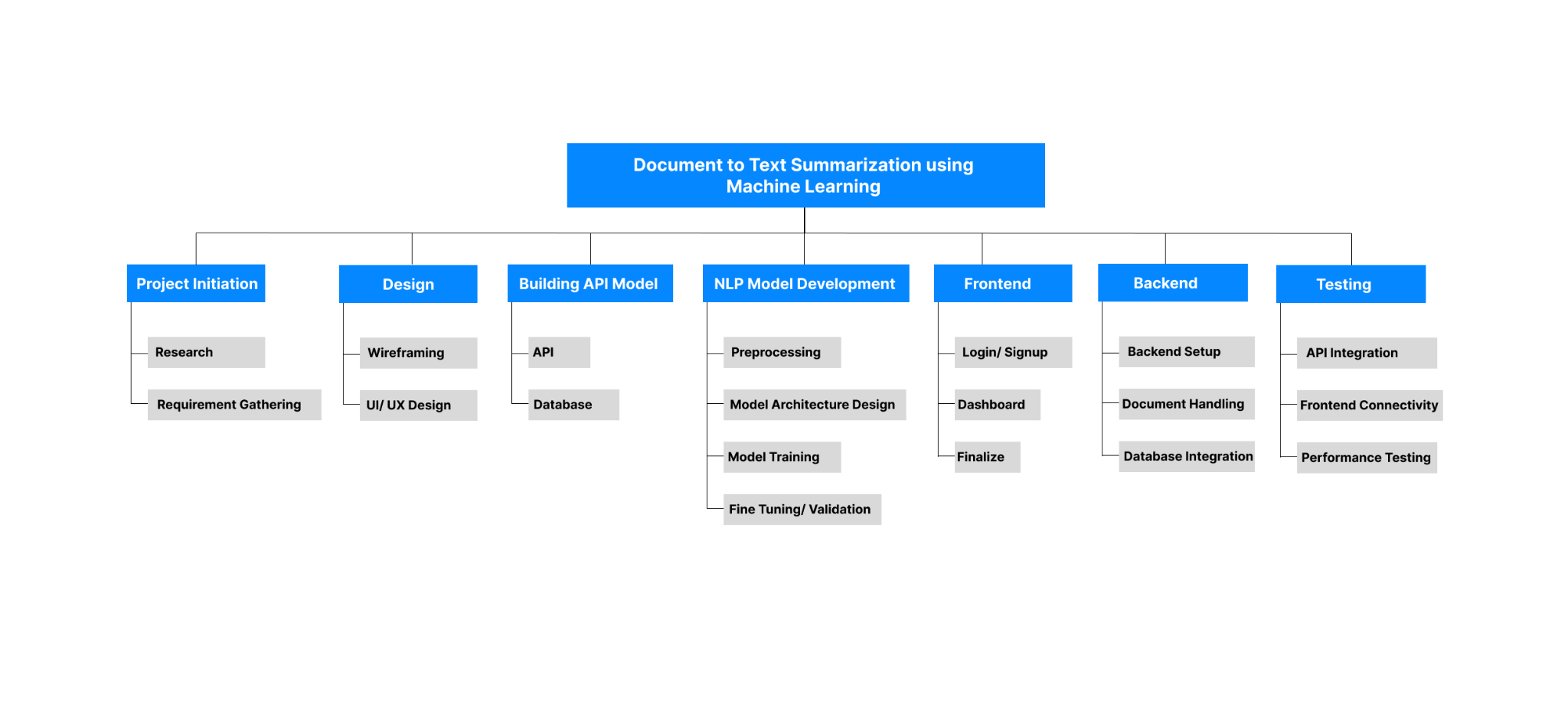


Figure Work Breakdown Structure

# Gantt Chart

Below is a proposed Gantt Chart of the overall timeline of the project. Here the project timeline is divided into six phases. The first phase is the phase of project initiation where the requirement gathering, research and proposal is prepared for the project. Based on the first phase the second phase starts with the basic visualization of the overall system through wireframing and creating the User Interface (UI) and User Experience (UX) of the system. After this, the phase of development kicks off with the integration of API in a crude system without a frontend. The evaluation of the API system will be conducted and then the NLP model will be developed which includes the preprocessing, model architecture design, model training, and fine tuning of the system. The model will then be displayed after the front-end development and the backend development will help in setting up the backend with database integration. Finally, a final documentation of the whole project will be done accordingly as shown in the Gantt Chart below.

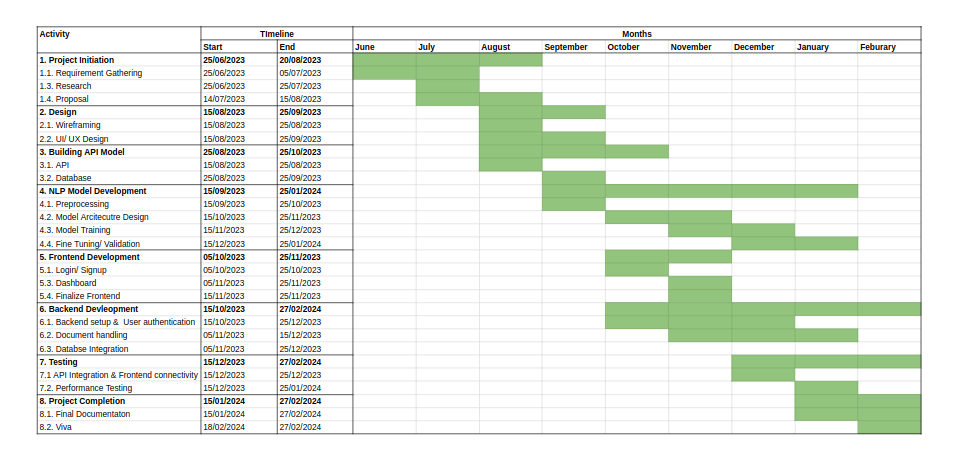


Figure Gantt Chart

# Tools and Technology

## HTML/ CSS/ JS

HTML, CSS and JavaScript are used for front-end development for this system, where HTML helps in structuring the content where CSS helps in defining the visual representation of the HTML elements, ensuring a consistent and attractive design. Similarly, JavaScript helps in adding interactivity, enabling dynamic content updates and enhancing the user engagement.

## Django Framework

Django Framework was chosen for backend due to my familiarity with Python language. Django is known for its effectiveness in handling web applications. It is a free and open-source web framework for Python that accelerates the development of web applications. It offers a more effective method to build and develop web programs that initiating from scratch because it removes the need to build the backend, APIs, JavaScript (IBM, 2023).

## MariaDB

To store the data, MariaDB will be used due to its user-friendly interface simplifies database management, making it an excellent choice for an uncomplicated database solution.

## NLP Model

NLP model is the proposed model that will be used for the machine learning aspect of the system. TextRank is one of the algorithms that extracts important sentences from a document. Similarly for the proposed system, transformer-based models such as BERT and BART could be used for abstractive summarization (Pietro, 2022).

# Bibliography

Adhikari, S., M. & R., 2020. NLP based Machine Learning Approaches for Text Summarization. [Online] Available at: <https://ieeexplore.ieee.org/document/9076358> [Accessed 10 August 2023].

IBM, 2023. What is Django. [Online] Available at: <https://www.ibm.com/topics/django> [Accessed 11 August 2023].

Johnson, D., 2021. What is Trello? Here's what you need to know about the online project management tool. [Online] Available at: <https://www.businessinsider.com/guides/tech/what-is-trello?op=1> [Accessed 02 August 2023].

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P. & Lehmann, S., 2019. Accelerating dynamics of collective attention. Nature Communications, 10(1), pp. 2041-1723.

Moiyadi, H. S. et al., 2016. NLP Based Text Summarization Using Semantic. *International Journal of Advanced Engineering, Management and Science (IJAEMS),* 2(10), pp. 1812-1818.

Pietro, M. D., 2022. Text Summarization with NLP: TextRank vs Seq2Seq vs BART. [Online] Available at: <https://towardsdatascience.com/text-summarization-with-nlp-textrank-vs-seq2seq-vs-bart-474943efeb09> [Accessed 12 August 2023].

Raut, L. P., 2015. Overview on Kanban Methodology and its Implementation. *IJSRD - International Journal for Scientific Research & Development,* 03(02), pp. 2321-0613.

Rebelo, M., 2023. *What is Hugging Face?.* [Online] Available at: <https://zapier.com/blog/hugging-face/> [Accessed 12 August 2023].