20MCA246 MAIN PROJECT

SYNOPSIS

Name of Student : Nikhitha Babu

Roll No : 39

Title : AI-Based Text Summarization System with OCR Integration

Synopsis approved : Yes / No

Name and Signature of Guide with date : Prof. Swapna K J

Any Other Remarks :

Name and Signature of Coordinators with date:

Dr. Reena Murali / Dr. Sangeetha Jose

Name and Signature of HOD with date:

Dr. Reena Murali

20MCA246 MAIN PROJECT

SYNOPSIS

AI-Based Text Summarization System with OCR Integration

Introduction

Text summarisation is the technique used to distill lengthy documents into concise, coherent, and informative summaries. Provides properly phrased information by filtering huge number of paragraphs into valuable modules upon user's preference.Finalize a solution to deep reading and condense, which is time consuming for users to compress the required contents without any grammatical errors. Text summarization encompasses the creation of a concise, accurate, and coherent summary derived from a longer textual document. This task is a crucial component of natural language processing (NLP), aiming to condense extensive text while maintaining its essence to convey the core message effectively. The escalating volume of online text data has heightened the necessity for automated text summarization techniques, facilitating the swift discovery and consumption of pertinent information. Entire system is fully trainable by giving diverse level of datasets which can improve the overall text mapping accuracy. With a focus on efficiency and precision, it is possible to speed up the entire process of text condensing with the aid of an AI model.

Existing System

Existing text summarization tools predominantly employ methods such as extractive summarization to condense textual content. Extractive summarization selects and combines important sentences or phrases from the original text, while abstractive summarization generates a summary by paraphrasing and rephrasing the content. However, these methods often struggle with maintaining coherence, generating human-like summaries, and accurately capturing the essence of the text. Extractive methods may result in disjointed summaries, while abstractive methods may struggle with generating contextually relevant and grammatically correct summaries. Additionally, both approaches may encounter difficulties in handling complex language nuances, leading to summaries that lack clarity or convey inaccurate information. Despite offering notable advantages, automatic text summarization also presents significant drawbacks. Its implementation demands considerable resources, often requiring specialized tools and computational power. Furthermore, machines currently lack the nuanced summarization capabilities of humans, particularly in understanding context, tone, and emphasis in a text. This limitation is further exacerbated when dealing with complex materials such as scientific papers or legal documents, which pose challenges for accurate and effective summarization due to their intricate nature and domain-specific terminologies. It is important to note that existing text summarization systems do not offer an option for summarizing text from images, PDFs, or text files, highlighting a gap in their functionality.

Proposed System

The proposed AI text summarization system is a highly advanced solution aimed at transforming lengthy texts into concise, coherent, and meaningful summaries using state-of-the-art techniques. The system is designed to be accessible and adaptable, featuring an intuitive interface and seamless integration with various platforms, making it versatile enough for a broad range of applications. These include academic research, corporate workflows, and personal use cases where quick and accurate content understanding is essential. By supporting multiple input formats and maintaining a focus on inclusivity, the system ensures that diverse user needs are met, fostering a broader appeal among different demographic and professional groups.

The development process emphasizes robustness and precision through meticulous data collection and preprocessing methods, ensuring the system can handle diverse text types, including complex, domain-specific content. Extensive testing frameworks are employed to validate performance, accuracy, and consistency, simulating real-world scenarios to guarantee reliability. The iterative design process ensures computational efficiency without compromising on the quality of the summaries, allowing real-time processing for high-volume tasks.

The system also incorporates advanced functionalities to extend its capabilities. It supports text extraction from various sources, including scanned documents and multimedia formats, enabling users to work seamlessly with diverse content types. Multilingual support further enhances its global usability, enabling users to generate and translate summaries in multiple languages, broadening its accessibility to non-native speakers.

User-centric features, such as feedback mechanisms and customizable settings, provide an enhanced experience by allowing users to tailor the summarization outputs to their specific needs. These features foster continuous refinement and ensure that the system evolves to meet changing user requirements over time. Designed for efficiency and adaptability, the system is equipped to cater to professional, academic, and personal environments, establishing itself as a reliable, high-performance tool for handling large-scale text summarization tasks with precision and ease.

Objectives

1. The primary objective of the AI text summarization project, named ”Concisely,” is to develop a robust and user-friendly system that can efficiently summarize text from various sources. The project aims to achieve efficient text summarization by developing an AI model. This model should accurately summarize text from PDFs, images, and user input.

2. Translation feature translate summaries into multiple languages, enhancing the system's usability for a global audience. To enable versatile input processing, the project will incorporate Optical Character Recognition (OCR) functionality for extracting text from images and a PDF reader for extracting text from PDF files. This will allow the system to process a wide range of input sources as well.

Scope and Relevance

The model processes images, PDFs, and text files using OCR and abstractive summarization to generate concise summaries, aiding in content curation, information retrieval, and document analysis. It addresses the growing need to manage and utilize vast textual data efficiently, enhancing accessibility and comprehension across diverse content types.

References

1. Paulus Setiawan Suryadjaja, Rila Mandala., "Improving the Performance of the Extractive Text Summarization by a Novel Topic Modeling and Sentence Embedding Technique using SBERT." (2021)

2. Mirza Alim Mutasodirin, Radityo Eko Prasojo., "Investigating Text Shortening Strategy in BERT: Truncation vs Summarization." (2021)

Keywords : :Text Summarization, Abstractive Summarization, Optical Character Recognition (OCR)

Nikhitha Babu

Roll No:39