Automation on Government Documents: Insights Extraction, Comparison, and Summarization System

1. Introduction

The project aims to develop an automated system for extracting, comparing, and summarizing information from Indian government finance minister budget speeches from 2021 to 2024. By leveraging advanced technology, this system seeks to provide valuable insights to various stakeholders, such as researchers, policymakers, and the public.

2. Problem Statement

The challenge lies in efficiently retrieving data from diverse sources in sansad.in, analysing the retrieved documents to identify common themes, trends, and differences, and summarizing the information into concise, coherent summaries. This necessitates the development of a robust system capable of handling different types of documents and extracting meaningful insights from them.

Develop an automated system for extracting, comparing, and summarizing information from Indian government finance minister budget speeches from 2021 to 2024 .The system should retrieve data from the following URLs:

- https://sansad.in/ls/knowledge-centre/speeches

The system should be capable of:

Information Retrieval: Automatically fetch documents from the provided URLs, extracting text content from Finance minister budget speeches from 2021 to 2024.

Comparison Analysis: Analyse the retrieved documents to identify common themes, trends, and differences between different types of speeches and reports. This could involve sentiment analysis, topic modelling, or any other relevant technique to compare the content.

Text Summarization: Summarize the extracted information into concise, coherent summaries. These summaries should capture the key points and main ideas presented in the documents, enabling users to quickly grasp the essential content without having to read through lengthy texts.

The developed system should be efficient, accurate, and user-friendly, providing valuable insights into the content of Indian government documents for various stakeholders such as researchers, policymakers, and the public.

3. Solution

The proposed solution strategy involves building a proof of concept (POC) to address the requirements. By utilizing tools like Llama Index for efficient data processing and Hugging Face embeddings for text analysis, the system aims to extract key information, identify trends and comparisons, and generate concise summaries. This approach ensures accuracy and lays the foundation for further improvements and customization.

Solution Strategy - Build a POC which should solve the following requirements:

- Extract key information, identify trends and comparisons, and generate concise summaries.
- Provide valuable insights for researchers, policymakers, and the public to understand government communication.

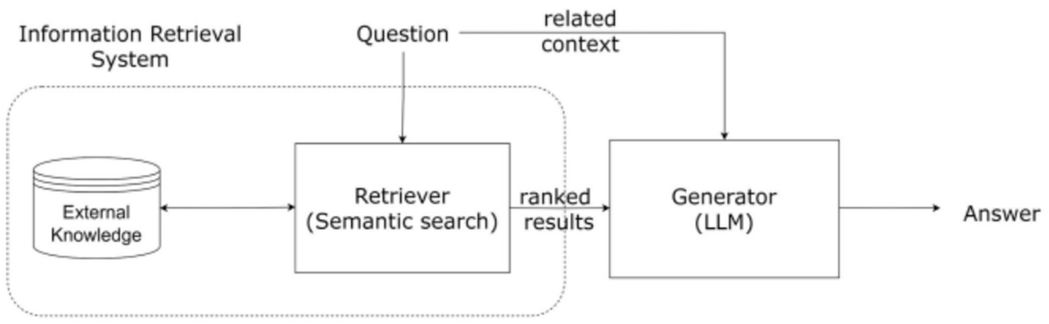
Goal - Solving the above two requirements well in the POC would ensure that the accuracy of the overall model is good and therefore further improvisations and customizations make sense.

Data Used - Indian government budget speeches by Finance Minister from 2021 to 2024

Tools used - LlamaIndex (only for now) has been used due to its powerful query engine, fast data processing using data loaders and directory readers as well as easier and faster implementation using fewer lines of code.

4. Architecture

The system architecture is made up of different parts, such as getting data from specific URLs, parsing and processing documents with tools like Simple Directory Reader and Sentence Splitter, creating embeddings with Hugging Face models, and indexing with Qdrant Vector Store. Query engines are then configured to enable efficient querying and response generation.



5. System Flow

Upon initialization, the system allows users to input questions or queries related to government documents. The query engine processes these queries and uses both retrieval and summarization tools to deliver pertinent responses. The responses are then displayed to the users, allowing them to quickly access the desired information.

6. Challenges Faced

During the development process, challenges such as compatibility issues with dependencies and the integration of different components arose. Additionally, ensuring the accuracy and relevance of the extracted information posed a significant challenge. However, these challenges were addressed through rigorous testing and optimization efforts.

7. Lesson Learned

Throughout the development of this project, several valuable lessons were gleaned. Firstly, the importance of thorough planning and requirement analysis became evident, as it laid a solid foundation for the subsequent stages of development. Secondly, navigating and mitigating dependencies and compatibility issues underscored the significance of robust testing and optimization strategies. Additionally, the iterative nature of problem-solving emphasized the need for adaptability and flexibility in implementing solutions. Lastly, collaboration and communication among team members proved instrumental in overcoming challenges and driving progress. These lessons will inform future projects, guiding decision-making and fostering continuous improvement in development processes.

8. Code URL

https://github.com/PrabakarTS/SemanticSpotter