Scheme Research Tool

Introduction

The Scheme Research Tool is a web-based application designed to streamline the process of researching and summarizing government schemes. Developed using Python and Streamlit, the tool enables users to input scheme URLs or upload text files containing URLs, processes the content using advanced natural language processing techniques, and generates summaries and answers to user queries. It leverages OpenAI embeddings and FAISS for efficient information retrieval.

Features

Input Handling

- Users can input URLs directly or upload text files containing URLs.
- Validation of URLs ensures that only valid and accessible links are processed.
- Support for PDF and web-based articles through dedicated loaders.

Document Processing

- Uses UnstructuredURLLoader for HTML content and PyPDFLoader for PDFs.
- Downloads PDF content temporarily, processes it, and removes temporary files securely.
- Splits documents into smaller chunks for embedding using the RecursiveCharacterTextSplitter.

Embedding and Indexing

- Generates embeddings for document chunks using OpenAI embeddings.
- Indexes embeddings using FAISS for fast and accurate similarity-based search.
- Stores the FAISS index in a pickle file (faiss_store_openai.pkl) for persistence.

Scheme Summarization

- Constructs a prompt to summarize scheme details into key categories:
    1. Benefits
    2. Application Process
    3. Eligibility
    4. Required Documents
- Utilizes OpenAI's LLM to generate a comprehensive summary of the scheme.

Question-Answering

- Allows users to query processed schemes interactively.
- Implements a ConversationalRetrievalChain to retrieve relevant information and generate responses.
- Displays answers with references to the source URLs.

User Interface

- Built with Streamlit, providing a simple, responsive, and interactive interface.
- Sidebar for input handling and buttons for initiating processing.
- Main content area for displaying scheme summaries and answers to user queries.

File Structure

- **main.py**: Main application script containing all the functionalities.
- **.config**: Configuration file for storing the OpenAI API key.
- **requirements.txt**: Dependency file listing required Python libraries.
- **faiss_store_openai.pkl**: Serialized FAISS index for storing embeddings.

Libraries Used

- **Streamlit**: For building the web-based interface.
- **LangChain**: For document loaders, text splitting, and LLM interaction.
- **OpenAI**: For generating embeddings and answering questions.
- **FAISS**: For indexing and similarity-based search.
- **NLTK**: For downloading required linguistic models.
- **Validators**: For validating input URLs.
- **Requests**: For downloading PDF content.

Usage Instructions

1. Place the OpenAI API key in .config as:

```
{
  "OPENAI_API_KEY": "your-api-key"
}
```

2. Install dependencies:

pip install -r requirements.txt

3. Run the application: streamlit run main.py
4. Input URLs via the sidebar or upload a file with URLs.
5. Click **Process URLs** to generate summaries and enable querying.
6. Enter a question in the main content area to receive answers.

Future Work

- **Caching**: Cache processed URLs to avoid reprocessing.
- **Multilingual Support**: Extend support to schemes in regional languages.
- **Advanced Validation**: Include additional checks for content accessibility.
- **Incremental Updates**: Enable updating the FAISS index without overwriting existing embeddings.
- **Enhanced Security**: Encrypt sensitive files like .config to protect API keys.

Conclusion

The Scheme Research Tool is a robust application that simplifies accessing and summarizing government scheme information. By integrating modern NLP techniques with a user-friendly interface, it empowers users to retrieve actionable insights efficiently. Future enhancements can further extend its capabilities and utility.