Hope to Skill Free AI Basic Course Hackathon

Final Hackathon Coding Project

Project Name: DocumentGPT

Presented to: Xeven Solutions

Instructors: Irfan Malik, Dr. Sheraz Naseer, Muhammad Haris Tariq

Submitted By: Shaheryar Yousaf

DocumentGPT: Comprehensive Documentation

Executive Summary

DocumentGPT revolutionizes the way users interact with documents. By integrating advanced artificial intelligence with a user-friendly interface, it transforms static documents into dynamic conversations. This detailed documentation aims to provide a thorough understanding of DocumentGPT, its unique features, the technology that powers it, and its potential applications, making it an essential guide for both users and potential investors.

Introduction

Concept and Vision

DocumentGPT is not just a software application; it's a paradigm shift in document management and interaction. The vision behind DocumentGPT is to create an intuitive platform where documents are no longer passive entities but active participants in a conversational exchange. This approach opens up new possibilities in data accessibility, knowledge extraction, and user engagement.

Transform Your Documents into Conversational Partners!

In-Depth Feature Analysis

User Authentication and API Integration

- **API Key Authentication**: The application requires users to authenticate using their OpenAI API key. This key is pivotal for accessing the full range of features.
- **Error Handling Mechanism**: In case of invalid API keys, users are not left in the dark. The application provides clear error messaging, guiding users to rectify the issue.

Advanced File Handling Capabilities

- **Multiple File Formats**: Emphasizing its versatility, DocumentGPT supports a variety of formats including PDF, DOCX, and ZIP, catering to diverse user needs.
- **Intelligent File Processing**:
    - **PDF & DOCX Processing**: These files undergo a systematic process of storage, text extraction, and chunking, ensuring data is manageable and easily retrievable.

    - **ZIP File Processing**: ZIP files receive special attention where their contents are first verified and then processed similarly to individual files.

Data Storage and Vectorization

- **Qdrant DB Integration**: The choice of Qdrant DB stands out for its efficiency in handling large datasets, essential for the app's performance.
- **Vector Store Creation**: This step is critical in transforming text data into a format that's conducive for quick retrieval and analysis, leveraging advanced AI techniques.

Interactive User Interface

- **Dynamic Query Input**: Post-processing, the application presents a user-friendly interface for querying the document. This interactive feature is the heart of DocumentGPT, where documents come to life through conversation.

Comprehensive Technology Stack Overview

Core Frameworks and Libraries

- **Streamlit**: Offers a sleek and interactive web app experience, making DocumentGPT accessible and easy to navigate.
- **LangChain and Lanchain-Community**: These packages enhance the app's language understanding capabilities, crucial for processing and responding to user queries.
- **Transformers by Hugging Face**: A cornerstone for implementing cutting-edge NLP models, enabling sophisticated text analysis and response generation.

Document Processing Technologies

- **Docx2txtLoader & PyPDFLoader**: These libraries are integral in extracting text from DOCX and PDF files, ensuring that all document content is accurately captured and processed.

Database and Data Handling

- **Qdrant-Client**: Manages interactions with the Qdrant database, a pivotal component in storing and retrieving processed document data.
- **HuggingfaceEmbeddings**: Used for creating vector embeddings, a process that enhances the efficiency and accuracy of the chatbot's responses.

Unfortunately, I wasn’t able to use Pinecone and Contextual compression Retrievers.

Analysis of Text Splitting Techniques in DocumentGPT

The analysis involved applying both text splitting techniques to the same set of legal clauses and comparing their outputs. The clauses chosen for this analysis were "Exclusive Remedies" and "Severability" from a legal agreement. Each technique was tasked with providing an explanation for these clauses.

Results

Inputs and Outputs

1. **Exclusive Remedies Clause:**
    - **Original Input:** "The Buyer and Seller Parties acknowledge and agree that their sole and exclusive remedy with respect to any and all claims... for any breach of any representation, warranty, covenant, agreement, or obligation set forth herein..."
    - **Text Character Split Output:** "The exclusive remedies provision states that the buyer and seller can only seek indemnification for any claims related to the agreement, except for fraud, criminal activity, or intentional misconduct."
    - **Recursive Text Split Output:** "Exclusive Remedies refers to the sole and exclusive recourse available to the Buyer and Seller Parties for any claims related to breaches of representations, warranties, covenants, agreements, or obligations in the Agreement, except for fraud, criminal activity, or willful misconduct."
2. **Severability Clause:**
    - **Original Input:** "Whenever possible, each provision of this Agreement shall be interpreted... but if any provision of this Agreement is held to be prohibited by or invalid under any applicable law, such provision will be ineffective only to the extent of such prohibition or invalidity..."
    - **Text Character Split Output:** "Severability refers to the principle that if any provision of an agreement is deemed invalid or prohibited by law, it will only be ineffective to the extent of the invalidity or prohibition."
    - **Recursive Text Split Output:** "Severability refers to the principle that if any provision in an agreement is found to be prohibited or invalid under applicable law, that specific provision will be ineffective to the extent of the prohibition or invalidity."

Analysis

- **Text Character Split Technique:**
  - Provided concise and straightforward summaries.
  - Focused on key aspects but lacked specific details.
  - Ideal for users seeking quick, general insights.
- **Recursive Text Split Technique:**
  - Delivered more detailed and nuanced information.
  - Closely adhered to the original text's intent and details.
  - Better suited for in-depth analysis and legal interpretations.

In the analysis of text splitting techniques, it was found that the Text Character Split method excels in providing concise and straightforward summaries, making it ideal for quick insights. On the other hand, the Recursive Text Split technique delivers more detailed and nuanced information, suitable for in-depth analysis and precise legal interpretations. The choice between these techniques should be based on specific user needs and document complexity within DocumentGPT.

Potential Applications and Use Cases

DocumentGPT has a broad range of applications across various sectors:

- **Educational Sector**: Assisting students and educators in interacting with academic material in an engaging and interactive manner.
- **Corporate Sector**: Enhancing business decision-making by providing quick access to insights from reports and analytics.
- **Healthcare Industry**: Facilitating the review of medical documents, research papers, and patient records for healthcare professionals.
- **Legal Sector**: Streamlining the process of legal document analysis and information retrieval for lawyers and legal researchers.
- **Publishing Industry**: Assisting editors and writers in navigating and referencing large manuscripts or literary works.
- **Government Agencies**: Enabling efficient document handling and information extraction from policy documents, legal texts, and administrative paperwork.

Conclusion

DocumentGPT represents a significant innovation in how we interact with and extract value from documents. Its unique blend of advanced AI, user-friendly design, and versatile functionality positions it as a frontrunner in the field of document management and interaction. As you prepare for your investor presentation, remember that DocumentGPT is more than a product; it's a solution that redefines the boundaries of document engagement, promising not just to meet but exceed the expectations of a diverse range of users and industries.