## Project Report: Semantic Search System for Policy Documents

### Overview

This project aims to develop a sophisticated semantic search system specifically designed for policy documents. It integrates advanced techniques for document processing, vector embeddings, and coherent answer generation. The system is structured into three key layers: embedding, searching, and generation, each optimized for superior performance.

### Project Structure

#### Embedding Layer

- \*\*Document Processing and Chunking\*\*: Explore various strategies for processing and chunking PDF documents.

- \*\*Vector Representations\*\*: Choose between OpenAI's embedding model and SentenceTransformers for generating vector representations of the text.

#### Search Layer

- \*\*Query Design\*\*: Develop three distinct queries that reflect potential user questions about policy documents.

- \*\*Vector Database Searches\*\*: Implement searches against ChromaDB using a vector database, incorporating a cache mechanism to optimize performance.

- \*\*Result Enhancement\*\*: Enhance search results with a re-ranking block using cross-encoding models from HuggingFace.

#### Generation Layer

- \*\*Prompt Design\*\*: Create a detailed and instructive prompt for the Language Model (LM) to ensure coherent and relevant answer generation.

- \*\*Few-Shot Learning\*\*: Include a few-shot example in the prompt to improve the output quality of the LM.

### Performance Evaluation

- \*\*Impact Assessment\*\*: Conduct a thorough evaluation of the impact of different strategies, models, and components on the system's performance.

- \*\*Scalability\*\*: Address scalability concerns by considering potential increases in the number of documents or user queries.

### Getting Started

1. \*\*Install Required Libraries\*\*:

```sh

pip install pdfplumber tiktoken openai chromaDB sentence-transformers -q

```

2. \*\*Mount Google Drive\*\*:

```python

from google.colab import drive

drive.mount('/content/drive', force\_remount=True)

```

3. \*\*Implementation\*\*: Follow the step-by-step instructions in the codebase to implement the Embedding, Search, and Generation layers.

---

By structuring the system in this way, we aim to create an efficient and scalable semantic search solution tailored for policy documents. This project not only addresses the immediate needs for advanced search capabilities but also lays the groundwork for future enhancements and expansions.