Documentação do Programa em Python

Visão Geral

O programa é uma aplicação Python que utiliza a biblioteca langchain para realizar um sistema de perguntas e respostas baseado em recuperação, empregando um modelo de linguagem e um mecanismo de busca vetorial.

Dependências

Certifique-se de ter as seguintes bibliotecas instaladas:

!pip install langchain sentence-transformers faiss

Fluxo de Execução

Carregamento de Documentos:

Utiliza o DirectoryLoader para carregar documentos de um diretório específico.

Divisão de Texto em Trechos:

Utiliza o RecursiveCharacterTextSplitter para dividir os documentos em trechos de texto.

Geração de Vetores de Embedding:

Utiliza o HuggingFaceEmbeddings para gerar vetores de embedding com base em um modelo pré-treinado.

Armazenamento em Vetores:

Utiliza o FAISS para armazenar os vetores de embedding em um índice eficiente para busca.

Configuração do Modelo de Linguagem:

Carrega um modelo de linguagem usando o CTransformers da biblioteca langchain.

Template de Pergunta e Resposta:

Define um template de pergunta e resposta usando o PromptTemplate.

Configuração do Módulo de Recuperação de Perguntas e Respostas:

Configura o módulo de recuperação de perguntas e respostas (RetrievalQA) usando o modelo de linguagem, o vetorizador e o template de pergunta.

Interação com o Usuário:
Inicia um loop de interação com o usuário, permitindo que eles forneçam perguntas.

Busca por Similaridade e Resposta:
A cada iteração do loop, o programa realiza uma busca por similaridade nas respostas armazenadas e fornece uma resposta com base na pergunta do usuário.

Uso
Execute o programa e insira perguntas quando solicitado. Para sair, digite "exit".

Roadmap para Escala do Programa:

Otimização de Desempenho:
Avaliar e otimize o desempenho do código para garantir uma execução eficiente, especialmente ao lidar com grandes conjuntos de dados.

Paralelização do Processamento:
Implementar técnicas de processamento paralelo para acelerar operações intensivas, como a geração de embeddings e a busca por similaridade.

Integração com Serviços de Nuvem:
Considerar a migração para serviços de nuvem para facilitar a escalabilidade automática e garantir recursos suficientes conforme a carga aumenta.

Gestão de Grandes Volumes de Dados:
Implementar estratégias eficazes para lidar com grandes volumes de dados, como carregamento de documentos sob demanda e técnicas de armazenamento eficientes.

Melhorias na Interface do Usuário:
Aprimorar a interface do usuário para facilitar a interação e compreensão do sistema, considerando interfaces gráficas ou interfaces de linha de comando mais amigáveis.

Aprimoramento do Modelo de Linguagem:
Avaliar modelos de linguagem mais avançados e treinados para lidar com uma variedade maior de perguntas e contextos.

Implementação de Aprendizado Contínuo:
Desenvolver um mecanismo de aprendizado contínuo que permita ao programa aprimorar seu desempenho com base no feedback do usuário e em novos dados disponíveis.

Suporte a Múltiplos Idiomas:
Estender o programa para suportar múltiplos idiomas, proporcionando uma experiência mais global.

Integração com APIs de Plataformas de Mensagens:
Integrar o programa com APIs de plataformas de mensagens populares para permitir a interação através de aplicativos de mensagens, chatbots ou interfaces de voz.

Implementação de Cache Inteligente:
Introduzir um sistema de cache inteligente para armazenar resultados frequentemente acessados, reduzindo a necessidade de recalculos e melhorando a velocidade de resposta.

Monitoramento e Logging:
Implementar ferramentas de monitoramento e logging para rastrear o desempenho, identificar gargalos e fornecer informações valiosas sobre o uso do sistema.

Documentação Abrangente:
Elaborar uma documentação abrangente que explique detalhadamente a configuração, os requisitos e os procedimentos de escalabilidade para facilitar futuras implementações e manutenções.

Ao seguir este roadmap, o programa estará mais bem equipado para lidar com volumes crescentes de dados e interações de usuários, garantindo uma experiência escalável e eficiente.