Pontifícia Universidade Católica do Rio de Janeiro
Departamento de Informática

Projeto final de programação

Chatbot para recuperação de informação específica de domínio

Marina Condé Araújo

Documentação do projeto final de programação com orientação do professor Marcos Kalinowski

Rio de Janeiro

Dezembro de 2024

Sumário

1 Breve descrição

1.1 Problema

Em um mundo com crescente produção científica, a busca por informações específicas de domínio apresenta desafios, principalmente em contextos onde o conteúdo é altamente técnico, especializado ou denso, como em artigos acadêmicos, manuais de engenharia ou regulamentos jurídicos. Tradicionalmente, esses documentos são acessados por meio de buscas simples baseadas em palavras-chave, que frequentemente falham em capturar o contexto necessário para fornecer respostas relevantes. Além disso, usuários não especialistas podem ter dificuldade em formular perguntas precisas, agravando a frustração. Em áreas específicas, a necessidade de interpretar termos técnicos, conceitos complexos e inter-relações entre tópicos torna a recuperação manual de informações ineficiente e suscetível a erros. Assim, existe uma lacuna para ferramentas que combinem agilidade e precisão na extração de dados em domínios especializados.

1.2 Justificativa

A importância de um chatbot focado na recuperação de informações específicas de domínio reside na necessidade de otimizar o acesso a conteúdos técnicos e complexos. Profissionais e acadêmicos frequentemente enfrentam prazos apertados e grandes volumes de informação, tornando ferramentas tradicionais de busca inadequadas. Além disso, em domínios especializados, como medicina, direito ou tecnologia, a precisão e o contexto são fundamentais para evitar interpretações incorretas que podem levar a decisões equivocadas. Com o avanço dos modelos de linguagem (LLMs) e o uso de técnicas como a Recuperação Aumentada de Geração (RAG), é possível oferecer respostas contextuais baseadas em documentos carregados, transformando o modo como usuários acessam informações críticas. Essa ferramenta não apenas melhora a produtividade, mas também democratiza o acesso a conhecimentos especializados.

1.3 Objetivo

O principal objetivo do programa é fornecer uma solução inteligente para processar, compreender e recuperar informações específicas de domínio, permitindo aos usuários localizar e compreender informações específicas de maneira rápida, precisa e contextualizada, por meio de um chatbot integrado a uma interface gráfica. Este objetivo abrange:

- Facilitar a extração de informações específicas em documentos PDF de maneira amigável;

- Oferecer respostas interpretadas e sintetizadas com base no conteúdo enviado, atendendo às necessidades de pesquisadores, estudantes e profissionais;
- Melhorar a experiência de consulta e análise de documentos complexos, reduzindo o tempo gasto em leituras manuais extensas.

1.4 Principais funcionalidades

- Carregamento do Documento: O usuário faz o upload de um arquivo PDF diretamente na interface gráfica. O programa suporta documentos técnicos, acadêmicos, jurídicos e outros formatos de texto estruturado;
- Processamento e Indexação: O conteúdo do documento é dividido em pequenos trechos (chunks), que são indexados utilizando algoritmos de similaridade semântica (baseados em embeddings gerados por modelos de aprendizado profundo). Essa indexação torna a busca rápida e precisa, mesmo em documentos longos;
- Interação com o Usuário: O usuário digita uma pergunta em linguagem natural na interface. Por exemplo: "Quais são os métodos mais utilizados no conjunto de artigos fornecidos?";
- Recuperação Aumentada: O sistema identifica os trechos mais relevantes do documento que podem conter a resposta, utilizando técnicas avançadas como o método MMR (Maximal Marginal Relevance), que equilibra relevância e diversidade dos resultados;
- Geração da Resposta: Após identificar os trechos relevantes, o programa utiliza um modelo de linguagem (como GPT-4 ou LLaMA) para construir uma resposta que seja direta, concisa e compreensível;
- Apresentação e Transparência: Além de apresentar a resposta gerada, o programa exibe os trechos exatos do documento que embasaram a resposta, permitindo que o usuário verifique a fonte diretamente.

1.5 Público-alvo

O programa foi desenvolvido para atender diferentes tipos de usuários que necessitam acessar e interpretar informações específicas em documentos específicos e extensos.

- Pesquisadores acadêmicos: Pesquisadores de áreas como ciências naturais, engenharia, saúde e outras disciplinas técnicas enfrentam o desafio de analisar grandes

volumes de artigos, relatórios e dados experimentais. Muitas vezes, localizar informações específicas, como metodologias detalhadas, resultados de experimentos ou conclusões-chave, é uma tarefa demorada e trabalhosa. O chatbot de recuperação de informação específica de domínio apoia essa necessidade ao permitir que pesquisadores realizem buscas eficientes e contextuais em seus materiais. Essa funcionalidade reduz significativamente o tempo de revisão bibliográfica, otimiza o processo de coleta de dados e possibilita maior foco em análises críticas e na produção de conhecimento científico;

- Estudantes de graduação e pós-graduação: Estudantes de cursos intensivos, como engenharia, direito, medicina e ciências da computação, muitas vezes enfrentam dificuldades para localizar conceitos fundamentais, explicações técnicas e metodologias específicas em materiais didáticos extensos. Essas barreiras podem prejudicar o aprendizado e aumentar o tempo gasto em atividades acadêmicas. Com o chatbot, os estudantes podem realizar perguntas direcionadas em linguagem natural e obter respostas rápidas e objetivas, facilitando a compreensão de conteúdos complexos. O programa apoia o aprendizado direcionado, a preparação para provas e o desenvolvimento de projetos acadêmicos, tornando o estudo mais eficiente e organizado;

- Profissionais especializados: Profissionais de áreas como direito, engenharia, saúde e tecnologia frequentemente consultam documentos técnicos ou jurídicos para obter informações críticas, como cláusulas contratuais, especificações técnicas, diretrizes regulatórias ou protocolos médicos. Esses documentos muitas vezes são extensos e complexos, tornando a busca manual por informações uma tarefa demorada e suscetível a erros. O chatbot apoia esse público oferecendo uma interface que permite localizar dados relevantes de forma rápida e precisa, utilizando linguagem natural. Isso pode melhorar a produtividade, reduzir o tempo necessário para análise de documentos e aumentar a precisão nas atividades profissionais, promovendo maior eficiência no dia a dia.;

1.6 A natureza do programa

O programa foi concebido como uma solução para facilitar a recuperação de informações específicas de domínio em documentos textuais, integrando tecnologias avançadas de busca contextual e geração de respostas precisas. Ele foi projetado como um sistema funcional e adaptável, permitindo que evolua conforme novas necessidades ou tecnologias sejam incorporadas. Essa flexibilidade garante sua aplicabilidade em cenários reais, como suporte a atividades acadêmicas, técnicas e profissionais em áreas como engenharia, saúde e direito. O foco está em resolver desafios enfrentados por quem precisa acessar dados

críticos ou conhecimentos específicos em textos extensos, demonstrando como a tecnologia pode simplificar e agilizar esses processos complexos.

1.7 Ressalvas

Embora o programa ofereça uma solução para o problema proposto, é importante que os usuários considerem algumas limitações práticas. Essas considerações não diminuem o impacto do programa, mas reforçam a importância de entender suas características e possibilidades para tirar o máximo proveito da solução apresentada.

- Ele é projetado para documentos em formato PDF com texto legível; arquivos baseados em imagens podem exigir processamento adicional antes do uso;
- Documentos muito extensos ou perguntas extremamente complexas podem exigir mais tempo para processamento, mas o sistema busca oferecer respostas claras mesmo nesses casos;
- O desempenho depende do ambiente técnico onde o programa é executado, como disponibilidade de recursos computacionais e acesso a modelos de linguagem configurados.

2 Visão do projeto

Os cenários que apresentados têm como objetivo principal ilustrar como o programa pode ser utilizado em diferentes contextos. Eles servem para orientar o desenvolvimento do software, garantindo que ele permaneça alinhado às necessidades do público-alvo, e para demonstrar ao usuário como ele pode aproveitar as funcionalidades do programa de maneira eficaz. Além disso, os cenários também auxiliam na evolução do programa, permitindo que colaboradores e desenvolvedores identifiquem oportunidades de melhorias e novos usos para a ferramenta. Eles proporcionam uma visão ampla sobre o que o programa faz, como pode ser usado e quais limitações precisam ser consideradas, contribuindo para a consolidação do projeto como uma solução eficiente e confiável para os desafios apresentados. A seguir, são apresentados os cenários divididos entre positivos, que ilustram interações bem-sucedidas e esperadas, e negativos, que expõem limitações do programa e sugerem ajustes ou soluções para esses casos. Esses exemplos foram elaborados com foco em situações específicas, destacando o impacto da ferramenta em contextos reais.

2.1 Cenário Positivo 1: Aluno de Mestrado Fazendo sua Dissertação

Paulo é aluno de mestrado em engenharia elétrica e está trabalhando em sua dissertação sobre fontes renováveis de energia, com foco em painéis solares. Ele precisa realizar uma revisão bibliográfica sobre a eficiência energética em climas tropicais, mas tem pouco tempo para analisar todos os artigos disponíveis. Paulo carrega sua coleção de documentos no chatbot e pergunta: "Quais fatores mais impactam a eficiência dos painéis solares em climas tropicais?". O chatbot analisa os documentos e fornece uma resposta detalhada, listando os fatores mais mencionados, como temperatura, umidade e intensidade solar, com trechos destacados. Com isso, Paulo organiza suas referências de forma mais eficiente e avança na escrita de sua dissertação.

2.2 Cenário Positivo 2: Profissional da Saúde Buscando Informação Clínica

Dr. Luís é cardiologista e está se preparando para uma apresentação sobre o uso de medicamentos antiplaquetários em pacientes com risco de infarto. Ele possui vários artigos científicos e guidelines médicos armazenados, mas precisa localizar rapidamente informações específicas sobre as diferenças entre dois medicamentos comumente usados. Dr. Luís carrega os documentos no chatbot e pergunta: "Quais estudos comparam os efeitos do medicamento A com o medicamento B em pacientes com risco de infarto?". O chatbot analisa os materiais carregados e retorna uma resposta consolidada, destacando trechos dos artigos que mencionam resultados comparativos, como eficácia e efeitos colaterais. Com as informações prontamente acessíveis, Dr. Luís organiza sua apresentação de forma eficiente, economizando tempo de pesquisa e garantindo precisão nas informações apresentadas.

2.3 Cenário Negativo 1: Formato de documento incompatível

Clara, uma advogada, está revisando contratos para identificar cláusulas sobre penalidades em caso de atraso na entrega. Ela carrega um contrato no programa e pergunta: "Quais são as cláusulas relacionadas a penalidades por atraso?". No entanto, o arquivo carregado é um PDF em formato de imagem, sem texto editável. O chatbot exibe uma mensagem indicando que não é possível processar o documento devido ao formato incompatível. Clara entende a limitação e utiliza um software de OCR para converter o documento em texto antes de tentar novamente. Apesar do contratempo, ela consegue utilizar o programa após a conversão para encontrar as cláusulas desejadas.

2.4 Cenário Negativo 2: Limitação de conteúdo da base

Mariana, uma nutricionista, está pesquisando os impactos do consumo de alimentos ricos em ômega-3 na saúde cardiovascular. Ela carrega um artigo científico no chatbot e pergunta: "Quais são os principais benefícios do ômega-3 para a saúde do coração?". O programa analisa o documento, mas retorna uma mensagem indicando que o artigo não aborda diretamente os benefícios do ômega-3, apenas fornece dados sobre saúde cardiovascular de forma geral. Mariana percebe que precisará complementar sua base de pesquisa com artigos mais direcionados sobre ômega-3 para obter informações detalhadas. O chatbot ajuda a identificar a lacuna no conteúdo, orientando-a na busca de fontes adicionais.

3 Documentação técnica do projeto

3.1 Especificação dos requisitos funcionais

- Permitir o upload de arquivos PDF para análise e indexação;
- Implementar uma interface de chatbot para interagir com os documentos enviados;
- Fornecer respostas contextuais utilizando um modelo de LLM (*Large Language Models*);
- Integrar diferentes provedores de modelos, como HuggingFace, OpenAI e Ollama;
- Armazenar e recuperar vetores de forma eficiente utilizando FAISS.

3.2 Especificação dos requisitos não funcionais

- O sistema deve ser responsivo e fácil de usar;
- Garantir baixa latência nas respostas do chatbot;
- Manter compatibilidade com diferentes modelos de linguagem e frameworks;
- Documentação clara para manutenção e extensibilidade.

3.3 Arquitetura do projeto (modelos utilizados)

- Upload de arquivos PDF através da interface do usuário (Streamlit);

- Divisão dos documentos em chunks utilizando a técnica de divisão recursiva (RecursiveCharacterTextSplitter);
- Criação de embeddings com o modelo HuggingFace;
- Indexação dos embeddings em um vetor FAISS para busca eficiente;
- Configuração de um pipeline de RAG (Retrieval-Augmented Generation) para gerar respostas contextuais.

3.4 Descrição funcional do software

- Entrada: Arquivos PDF enviados pelo usuário
- Processamento:
  - Divisão dos documentos em partes menores;
  - Geração de embeddings vetoriais;
  - Indexação e recuperação de documentos com FAISS;
  - Geração de respostas usando modelos LLM.
- Saída: Respostas contextuais baseadas no conteúdo dos documentos enviados.

3.5 Linguagem e frameworks utilizados no projeto

- Python: Linguagem principal;
- Streamlit: Interface gráfica;
- LangChain: Orquestração de NLP e pipeline RAG;
- FAISS: Armazenamento e recuperação de vetores;
- HuggingFace: Geração de embeddings e LLMs;
- OpenAI: Integração com GPT.

4 Manual de Utilização para Usuários Contemplados

Este manual foi desenvolvido para orientar todos os tipos de usuários sobre como utilizar o programa de forma eficaz. O sistema é uma aplicação baseada em Streamlit, destinada a análise e interação com documentos PDF utilizando modelos de linguagem natural (LLMs).

4.1 Funcionalidades

4.1.1 Upload de arquivos PDF

- Objetivo: Carregar documentos para análise, conforme a figura 1;
- Passos:
  1. Na barra lateral, clique em "Enviar arquivos";
  2. Selecione arquivos PDF;
  3. Confirme o upload e aguarde o processamento.

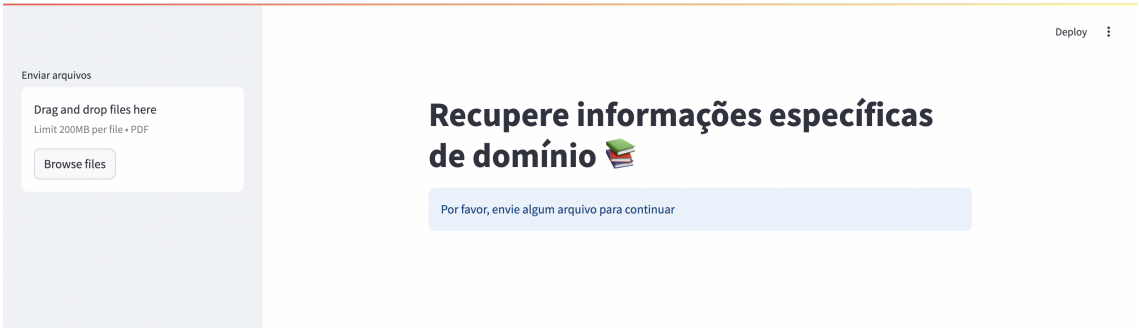

Figura 1: Visualização inicial do chatbot

4.1.2 Interação com o Chatbot

- Objetivo: Fazer perguntas sobre o conteúdo carregado, conforme a figura 2;
- Passos:
  1. Digite a pergunta no campo "Digite sua mensagem aqui...";
  2. Pressione Enter;
  3. A resposta para a pergunta será exibida diretamente na interface do chatbot, utilizando linguagem clara e objetiva. Caso a informação específica solicitada pelo usuário não seja encontrada nos documentos carregados, o sistema notificará o usuário, indicando que os dados necessários para responder à pergunta não estão disponíveis na base fornecida.

     Além disso, o chatbot fornecerá uma funcionalidade adicional: os trechos exatos dos documentos nos quais a resposta foi baseada serão destacados e apresentados ao usuário.

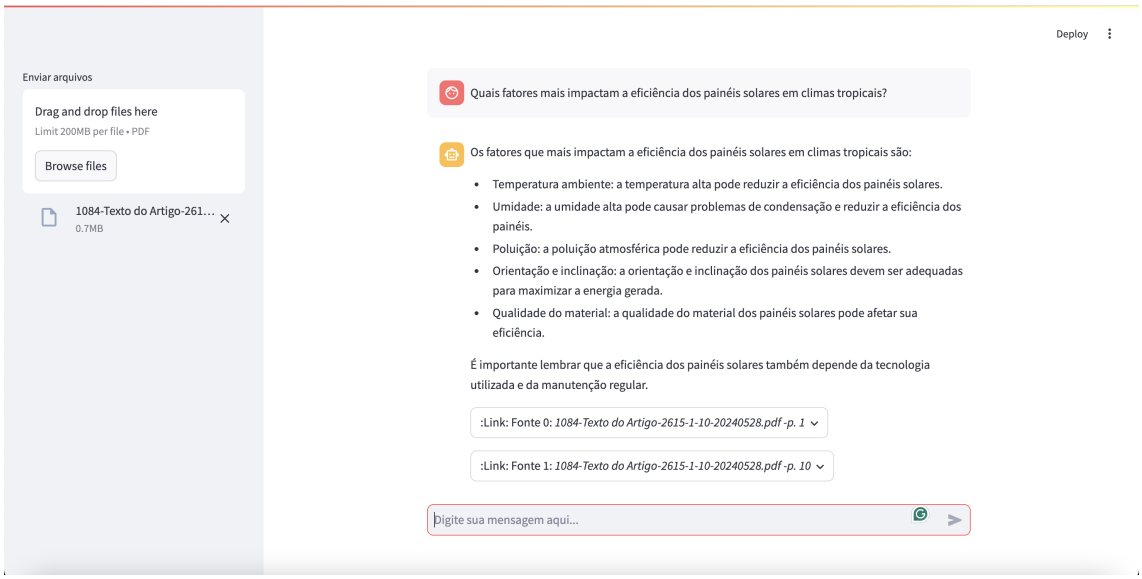

Figura 2: Chatbot retornando a pergunta de domínio

4.1.3 Configurações de modelos LLM

- Objetivo: Alterar o modelo de linguagem utilizado;
- Passos:
  1. No código, ajuste o parâmetro `model_class` para `"hf_hub"`, `"openai"` ou `"ollama"`;
  2. Salve e reinicie o programa.

4.2 Configuração

1. Clone o repositório:

```
git clone https://github.com/seu-repositorio.git
```

2. Instale as dependências:

```
pip install -r requirements.txt
```

3. Execute o programa:

```
streamlit run projetofinal.py
```

5 Conclusão

Este trabalho apresenta uma solução para a interação com documentos textuais, unindo técnicas de recuperação de informações e geração de respostas contextualizadas por meio de modelos de linguagem natural. A proposta de um sistema inteligente, que combina tecnologias como FAISS, LangChain e modelos de linguagem da HuggingFace e OpenAI, demonstra o potencial transformador da Inteligência Artificial no acesso e análise de informações em documentos extensos e complexos.

A solução proposta foi concebida com foco em atender às necessidades de pesquisadores, estudantes e profissionais, permitindo uma interação eficiente com conteúdos técnicos e acadêmicos. Os cenários explorados validaram a aplicabilidade do sistema em situações reais, destacando benefícios como a economia de tempo, a melhoria na qualidade das decisões baseadas em dados e a acessibilidade de informações contextualizadas.

Além disso, as limitações identificadas, como a dependência de documentos legíveis e a exigência de recursos computacionais adequados, não comprometem a relevância do sistema. Pelo contrário, abrem espaço para futuros aprimoramentos, incluindo a integração com tecnologias de OCR e otimizações para ambientes computacionais menos robustos.

Por fim, este trabalho reafirma a importância de soluções tecnológicas para simplificar e otimizar processos complexos, destacando-se como uma iniciativa prática e inovadora na aplicação de inteligência artificial no processamento de documentos textuais.