1. Visão Geral O DocMind é um assistente inteligente baseado na arquitetura RAG (Retrieval-Augmented Generation). Ele foi projetado para ler documentos PDF técnicos e responder perguntas com alta precisão, evitando alucinações comuns em modelos de linguagem puros.

2. Especificações Técnicas (Stack) O sistema foi construído utilizando as tecnologias mais modernas de 2025:

- **Framework Web:** Next.js 16 (App Router) garantindo renderização rápida e rotas de API server-side.
- **Linguagem:** TypeScript para tipagem estática e segurança de código.
- **Estilização:** Tailwind CSS para uma interface limpa e responsiva (Dark Mode nativo).

3. Motor de Inteligência Artificial O "cérebro" do DocMind opera em duas etapas distintas:

- **Geração de Texto (LLM):** O sistema utiliza o modelo **Llama 3.3-70b-versatile**, provido pela plataforma **Groq**. A escolha do Groq se deve à sua velocidade infernal de inferência (LPU), permitindo respostas quase instantâneas.
- **Embeddings (Vetores):** Para converter texto em números, utilizamos o modelo **sentence-transformers/all-mpnet-base-v2**, processado via API da **Hugging Face**.

4. Armazenamento de Conhecimento (Memória) Os fragmentos de texto (chunks) e seus respectivos vetores são armazenados no **Pinecone**, um banco de dados vetorial especializado.

- **Processo de Ingestão:** O sistema lê arquivos PDF usando a biblioteca `pdf-parse` (versão 1.1.1 para compatibilidade).
- **Chunking:** Os textos são divididos em blocos de aproximadamente 1000 caracteres com sobreposição de 200 caracteres, garantindo que o contexto não se perca entre os cortes.

5. Fluxo de Operação (RAG) Quando um usuário faz uma pergunta:

1. A pergunta é convertida em vetor (Embedding).
2. O Pinecone busca os 5 trechos de texto mais similares matematicamente à pergunta.
3. Esses trechos são injetados no prompt do sistema como "Contexto".
4. O Llama 3.3 gera a resposta final baseada estritamente nesse contexto.

6. Solução de Problemas (FAQ)

- **Erro "DOMMatrix is not defined":** Ocorre ao usar versões do `pdf-parse` superiores a 1.1.1 em ambiente Node.js. Solução: Downgrade para v1.1.1.
- **Respostas Lentas:** Verifique a latência da API do Hugging Face ou do Pinecone. O Groq geralmente é estável.
- **Alucinações:** Se o DocMind inventar informações, verifique se o parâmetro `topK` no Pinecone está configurado corretamente (recomendado: 5).