BCG X Challenge

Alavancando o poder de GenAI para transformar o planejamento climático.

Equipe: Hugo, Joel, Juliana, Mikhail

EcoDocs A.I

O Desafio - objetivos:

- Dar suporte a gestores públicos com modelos de GenAI construídos a partir de documentos.
- Desenvolver a capacidade de extrair dados que tornem o LLM preciso.
- Definir o ranking de documentos similares usando transformers e embeddings.
- Construir prompts que proporcionasse melhores respostas para o gestor.
- Criar um chatbot assistente com uma boa interface e de fácil interação.
- Integrar a pipeline de inteligência artificial, usando RAG com o chatbot assistente.

EcoDocs A.I

Pré-processamento dos dados

- Os oito arquivos são salvos no banco de dados assim que a aplicação inicia, caso eles ainda não tenha sido salvos.
- O Unstructured é crucial para pipelines RAG, pois permite a extração eficiente de dados de diversos formatos. Ele ajuda a quebrar dados de forma inteligente, melhora o pré-processamento, focando na relevância das informações e na recuperação e geração de respostas.
- Uma vez que os arquivos em PDF são carregados, o texto é limpo com a função clean_pdf. Esta função normaliza o texto (removendo acentos e caracteres especiais), remove as stopwords em portugués, e converte todas as palavras em letras minúsculas.

Embedding e vector database

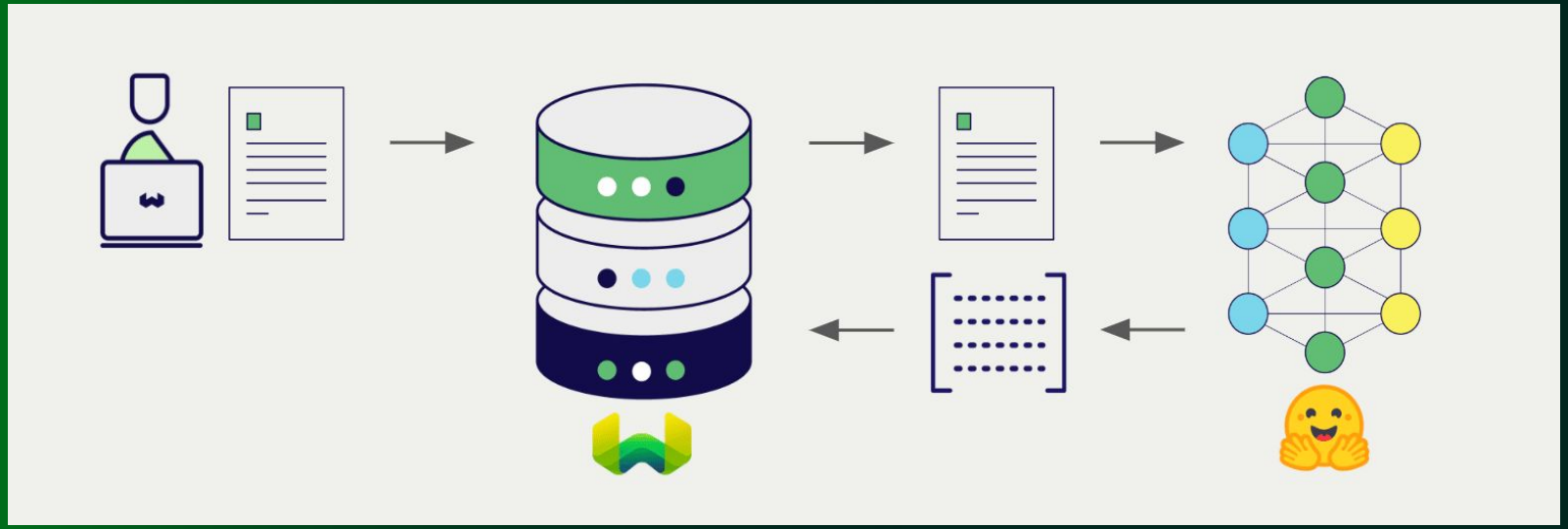
A base de conhecimento é responsável por armazenar documentos de texto usando representação numérica, que são os dados vetoriais.

- O Weaviate é um banco vetorial nativo para AI onde ele consegue além de armazenar dados, também realizar embedding localmente, sem usar recursos da Open AI Embeddings.

Os documentos relevantes são recuperado a partir da questão do gestor, onde o contexto que for similar, é adicionado ao prompt antes de enviar para o modelo generativo. O modelo sentence-transformers do huggingfaces mapeia sentenças para vetores de 384 dimensões, otimizado para busca semântica.

EcoDocs A.I

Embedding e vector database



EcoDocs A.I

Arquitetura RAG

EcoDocs A.I

- Uso de técnica RAG de aprimoramento do LLM através de informações contextuais.
- Sintetizar com precisão a capacidade de busca usando mecanismos do RAG.
- Problemas resolvidos:
    - **Recuperação**: O RAG, proporciona ao realizar upload de novos documentos atualizar a memória do LLM, pode-se usar a busca por similares.
    - **Aumento**: essa forma dinâmica de recuperação da informação reduz a necessidade de retreinamento do LLM.
    - **Geração**: Com o gerador pode-se mitigar alucinações, que são respostas falsas ou inventadas, as informações passam a ser de um contexto verdadeiro e mais precisas.

Engenharia De Prompt

A engenharia de prompt aqui utilizada na aplicação visa construir um fluxo de mensagens que oriente o modelo a:

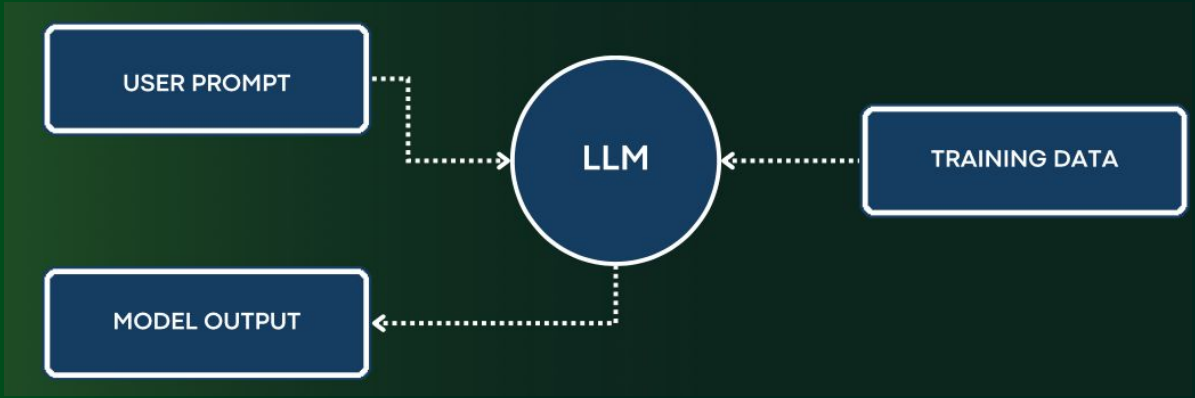
- Focar nas informações dos documentos fornecidos.
- Ser preciso e transparente na atribuição de fontes.
- Evitar respostas sem base em evidências.
- Prover um valor adicional à resposta com uma análise aplicada ao contexto de gestão pública, pois é o escopo do desafio.

Isso ajuda a garantir que as respostas sejam fundamentadas, confiáveis e relevantes para o cenário dos documentos.

EcoDocs A.I

Aplicativo QnA

- Frontend com Streamlit: apresenta uma interface de chat para entrada de perguntas
- Backend com FastAPI: gerencia consultas e uploads de arquivos PDF de forma eficiente.
- Serviço de Consulta e Upload: permite recuperar respostas de uma banco de dados, além de alimentar esse banco com uploads do usuário.



EcoDocs A.I

Upload de documentos

A aplicação permite fazer o upload de documentos em formato PDF para construir o modelo LLM. Dessa forma, a aplicação permite construir um modelo LLM com qualquer documento em formato PDF.

- O upload de documentos pode ser feito através da UI ou da API da aplicação web.
- Garantia de compatibilidade e estruturação dos dados para processamento eficiente

EcoDocs A.I

Solução Final

- Usar os oito documentos disponibilizados para criar uma assistência para gestores públicos, além de permitir que técnicos e usuários monitore os resultados, com o objetivo de melhorar as respostas do chatbot através de métricas estatísticas.
- Alguns adicionais permitem que o usuário realize upload de mais documentos que ajude o LLM cumprir seu papel de gerar informações cada vez mais precisas.
- A ferramenta também consegue separar além das informações geradas a partir da questão do gestor, cria um resumo de como realizar uma gestão mais efetiva.

EcoDocs A.I

Conheça a Equipe:

Hugo Angulo
Data Engineer

Joel Maykon
Data Scientist

Juliana Gonçalves
Data Scientist

Mikhail Futorny
Software Engineer

EcoDocs A.I