



Teste Técnico – Analista de Dados Sênior

Dados Fornecidos

Você receberá uma base de conhecimento contendo arquivos de texto em `.txt` com conteúdo institucional de uma empresa (como políticas internas, produtos, serviços e processos).

https://drive.google.com/drive/folders/1-qkMy5SCGzvHUMsOdJKR0-MtMAsN7QW_?usp=sharing

Tarefas

1. Pré-processamento dos dados

- Normalize e limpe os textos.
- Remova ruídos.
- Separe os textos em **chunks semânticos**, como seções ou parágrafos coesos, com tamanho máximo de 500–1000 tokens.

2. Criação da base vetorial

- Utilize um mecanismo vetorial.
- Faça o embedding dos textos com modelo da OpenAI, HuggingFace, ou outro.
- Salve os dados com metadados úteis (ex: origem do documento, seção, categoria).

3. Implementação do pipeline RAG

- Crie uma função ou endpoint onde o usuário envia uma pergunta e recebe uma resposta.
- A resposta deve:
 - . Buscar os trechos mais relevantes no vector store.
 - . Gerar uma resposta com um modelo LLM (pode ser OpenAI, Mistral, Cohere, etc).
 - . Referenciar os trechos usados (como um sistema de fontes/citações).

4. Estudo de Qualidade e Métricas

- Faça um conjunto de 10 perguntas relevantes.
- Para cada uma:
 - . Armazene os documentos recuperados.

- . Avalie se a resposta está correta, completa e bem referenciada.
- . Calcule métricas como: Recall@3, Precisão percebida, Nível de cobertura semântica.

5. Recomendações e Insights

- Com base nos testes:
 - . Quais tipos de perguntas funcionam melhor?
 - . Há lacunas na base de conhecimento?
 - . O modelo cometeu erros? Se sim, quais tipos?
 - . Sugira melhorias (ex: otimização de chunking, troca de modelo, enriquecimento de metadados, uso de re-ranking, etc).

Entregáveis Obrigatórios

1. **Código funcional** (pode ser Jupyter ou estrutura Python) com instruções de execução.
2. **Relatório técnico (PDF ou Markdown)** com:
 - Descrição das decisões técnicas (chunking, embeddings, vector store).
 - Métricas aplicadas e análise dos resultados.
 - Recomendações futuras.
3. **Tabela com perguntas x respostas x avaliações.**


Diferenciais

1. **Relatórios Gráficos**
 - **Visualização do Espaço Vetorial**
 - . Demonstrar entendimento de como os embeddings estão distribuídos e se agrupam semanticamente.
 - **Heatmap de Similaridade ou Overlap Semântico**
 - . Mostrar graficamente a semelhança entre perguntas e chunks ou entre respostas e fontes.
2. **(Opcional) Demonstração em FastAPI.**


Avaliação

Critério	Peso
Qualidade do tratamento e chunking	20%
Funcionalidade e estrutura do pipeline RAG	25%

Análise crítica das respostas	25%
Clareza do relatório técnico	20%
Diferenciais entregues	10%

 **Ferramentas recomendadas:** Python, LangChain, OpenAI API, FAISS, Pinecone ou Qdrant, Pandas, Matplotlib, Jupyter, e FastAPI.

Sinta-se livre para usar outras ferramentas se justificar no relatório.

 O **prazo** para o envio do projeto é de **7 dias à partir do recebimento**. Enviar o teste para adriel.oliveira@bevioficial.com.br ou para **18 98189-1278 (whatsapp)**