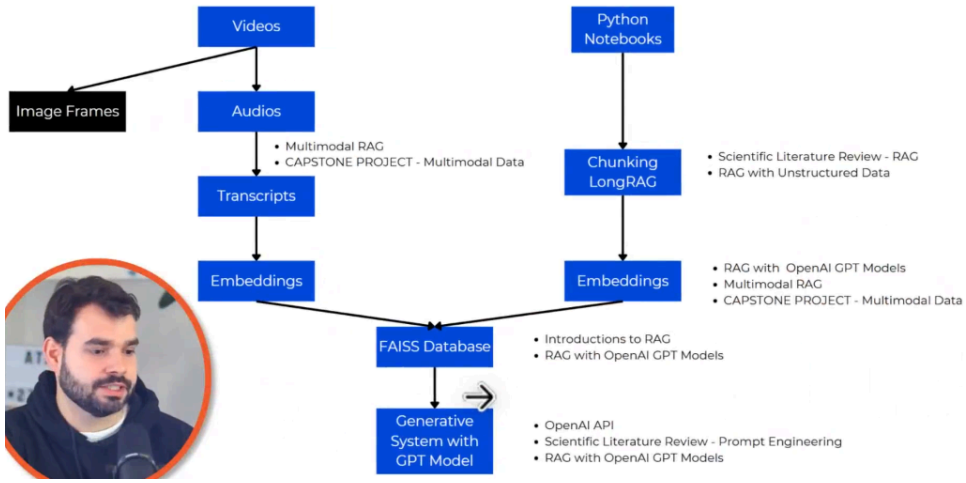
ZTM RAG: Fundamentos

segunda-feira, 16 de junho de 2025 15:12

O que é RAG:
RAG é uma forma de fazer com que os LLM (Modelos de linguagens como GPT) usem apenas os dados que você fornece para responder as perguntas.

Projeto de RAG criado pelo cara do curso: https://rubber-ducky.streamlit.app/

Arquitetura para chegar no RAG e onde será aprendido no curso:



Informações Relevantes / Key Concepts

O que é indexing / indexação
Indexing organiza dados para busca fácil, como criando um roadmap para os retrieval systems achar rapidamente informações relevantes. Sem indexing, as buscas seriam lentas.
Como indexing funciona?
Tokenização > Normalização > Inversão

Exemplo de index:
Doc1: A Croácia é um país da Europa.
Doc2: O rio Danúbio não passa na Croácia.
Doc3: A população estimada da Croácia é de 4 milhões de habitantes.

index
"Croácia":[1, 2, 3]
"país":[1]
"Europa":[1]
"rio":[2]
"Danúbio":[2]
"passa":[2]
"população"[3]
"estimada"[3]
"milhões"[3]
"habitantes"[3]

O que é querying

Querying é procurar dados indexados para achar documentos relevantes baseados no input do usuário.

Tipos de querying:
Keyboard queries (procura a palavra digitada)
Boolean queries (usa "and", "or" e "not")
Phrase Queries (procura a exata sequência de termos)
Natural Language Queries (usa processamento de linguagem natural para procurar)

Como querying funciona:
Input processing > Index lookup > Result compilation

O que é ranking

Ranking ordena os resultados por relevância, mostrando os mais relevantes primeiro.
Métodos comuns de ranking:
TF-IDF, Cosine Similarity, PageRank (o Google usa esse), Machine Learning Models.

Como ranking funciona:
Calculo de score > Sorting / Ordenação (maior score primeiro) > Apresentação

ReAct

ReAct faz com que LLMs alternem entre pensar e agir, como pensar "Eu preciso de informações" e querying uma API, ajustando baseado no resultado.

Chain-of-Thought (CoT)

É uma técnica de engenharia de prompts que visa melhorar o desempenho de modelos de linguagem em tarefas complexas, solicitando que o modelo gere uma explicação passo a passo antes de fornecer a resposta final.

Como usar CoT

- **Divida a questão**
  Ex:
  ❌ Qual o resultado de 15 + 9 – 4?
  ✅ Primeiro, some 15 com 9. Depois subtraia 4. Qual o resultado?
- **Encoraje a explicação**
  Ex: "Explique o seu raciocínio", "Explique passo a passo".
- **Use demonstrações antes da questão**
  Ex: "Dado a pergunta 'tal' é esperado que você responda assim 'passo 1, passo 2, passo 3...'. Agora responda essa pergunta:"