alura

LangChain: Criando chatbots inteligentes com RAG

Hybrid Search & Técnicas Avançadas

MERGULHE EM TECNOLOGIA_

Instrutor(a): Leonardo Pena

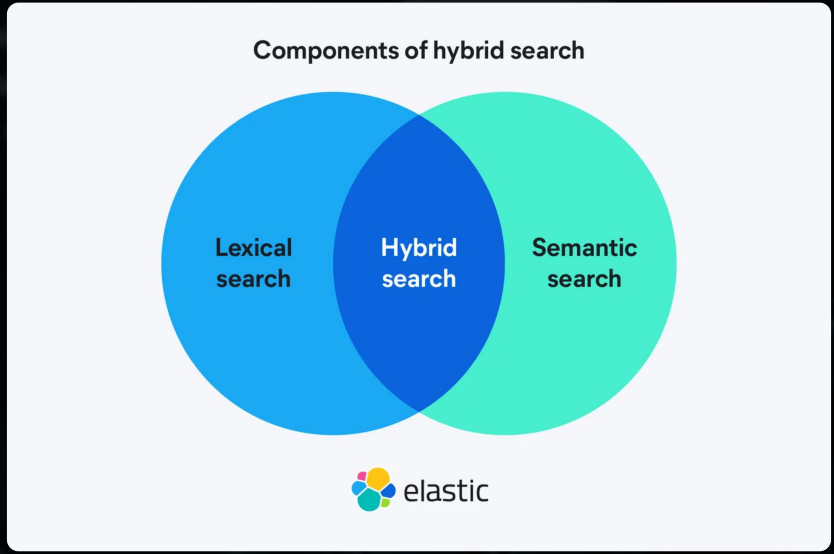
Seja bem vindo (a)!

Aula

Hybrid Search & Técnicas Avançadas

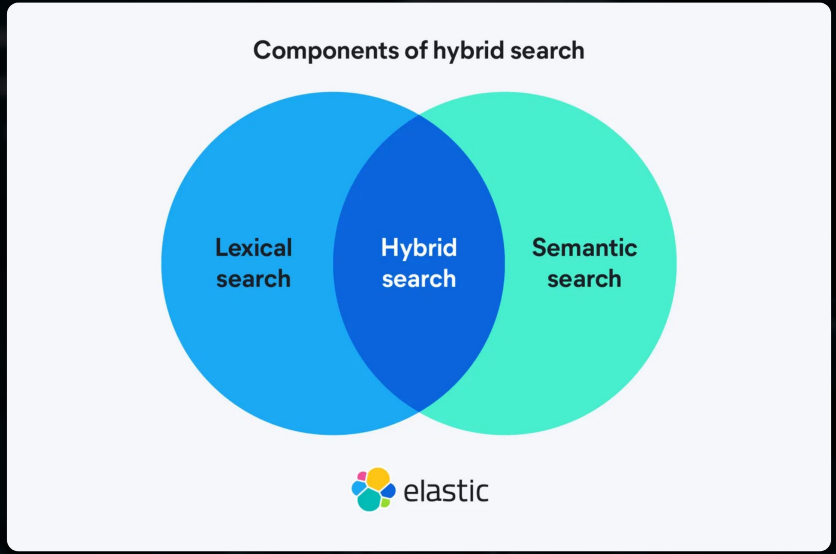
Hybrid Search (Busca Híbrida)

Hybrid Search é uma técnica de recuperação que combina a busca vetorial (semântica) com a busca por palavras-chave (lexical, ex: BM25).



Hybrid Search (Busca Híbrida)

- Busca semântica captura o significado e contexto, mesmo com palavras diferentes
- Busca lexical (BM25) encontra correspondências exatas de termos específicos
- A combinação supera as limitações de cada abordagem individual
- Resulta em recuperação mais robusta, precisa e completa

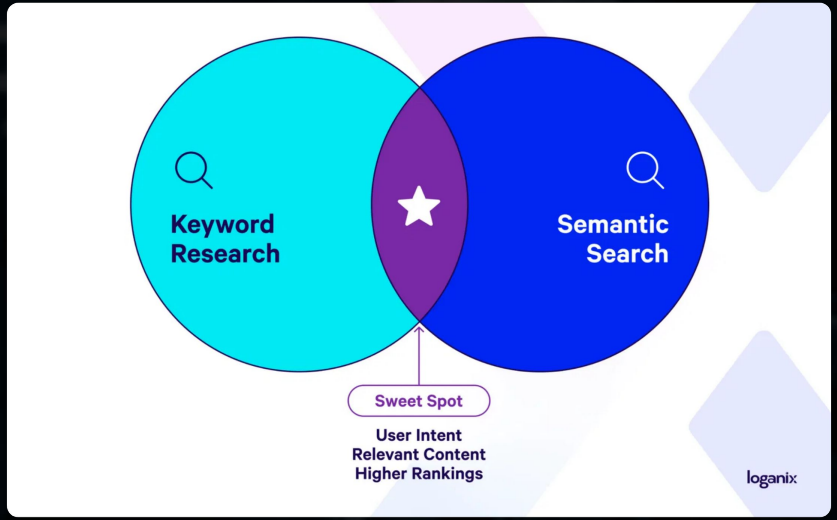


O objetivo é aproveitar **melhor dos dois mundos**: A compreensão contextual da busca semântica e a precisão da busca lexical.

Vantagens da Busca Híbrida

A busca híbrida combina o melhor de duas abordagens para superar suas limitaçoes individuais

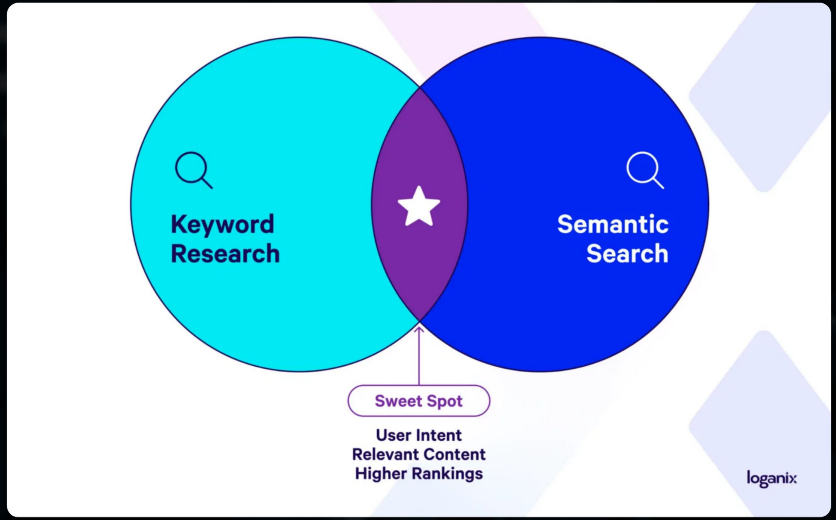
Tipo de Busca	Pontos Fortes	Limitações
Semântica	Entende contexto e significado	Pode perder termos específicos
Lexical (BM25)	Excelente em termos exatos	Não entende contexto ou sinônimos
Híbrida	Combina ambas as forças	Complexidade de implementação



Vantagens da Busca Híbrida

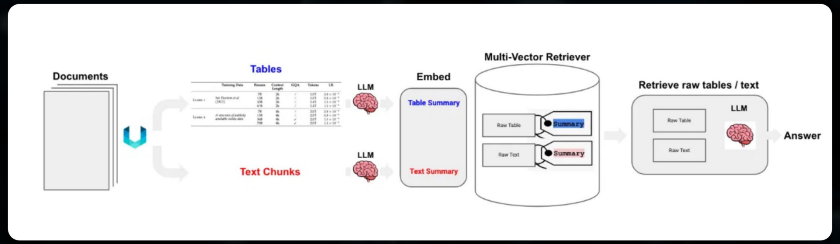
Casos de uso ideais para busca híbrida

- Documentação técnica com terminologia específica
- Catálogos de produtos com códigos e descrições
- Bases de conhecimento com conceitos complexos



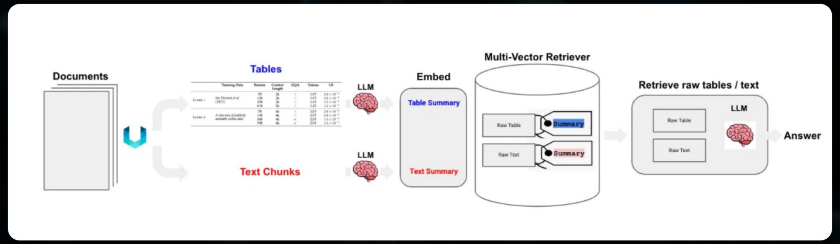
Multi-Vector RAG

Multi-Vector RAG é uma técnica avançada onde diferentes aspectos de um único documento são representados por múltiplos vetores.



Multi-Vector RAG

- Cada documento é dividido em diferentes representações vetoriais
- Permite recuperar informações com granularidade mais apropriada
- Lida melhor com conteúdo complexo e heterogêneo
- Aumenta a precisão da recuperação para consultas específicas



Query Transformation

Query Transformation consiste em usar um LLM para reescrever ou expandir a pergunta original do usuário antes de enviá-la ao sistema de recuperação

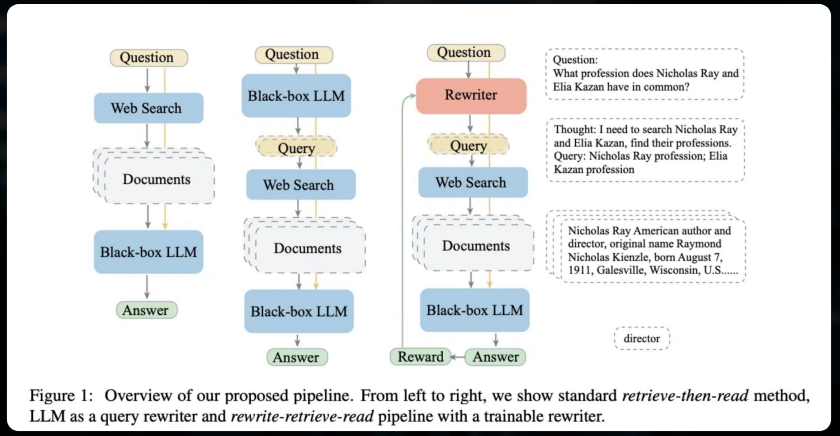


Figure 1: Overview of our proposed pipeline. From left to right, we show standard *retrieve-then-read* method, LLM as a query rewriter and *rewrite-retrieve-read* pipeline with a trainable rewriter.

Query Transformation

Expansão de Consulta
Adiciona termos relacionados, sinônimos ou contexto à consulta original.
Exemplo: "carros elétricos" → "carros elétricos veículos EV Tesla
bateria autonomia"

Contextualização de Diálogo
Incorpora o histórico da conversa para entender consultas ambíguas.
Exemplo: "Qual é o preço?" → "Qual é o preço do Tesla Model 3
2023 mencionado anteriormente?"

Geração de Múltiplas Consultas
Cria várias versões da mesma pergunta para capturar diferentes aspectos.
Exemplo: "Como funciona IA?" → ["O que é inteligência artificial?", "Como os modelos de IA são treinados?", "Aplicações práticas de IA"]

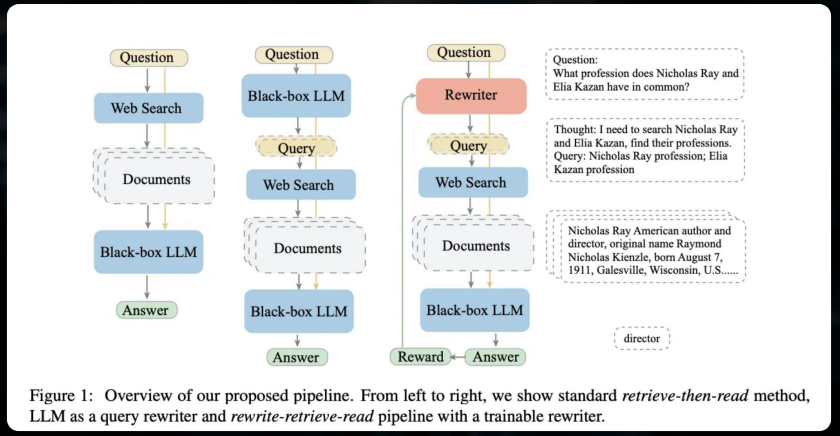


Figure 1: Overview of our proposed pipeline. From left to right, we show standard *retrieve-then-read* method, LLM as a query rewriter and *rewrite-retrieve-read* pipeline with a trainable rewriter.

Esta técnica melhora significativamente qualidade da recuperação, especialmente para consultas ambíguas, curtas ou com contexto implícito