

Inteligência Artificial 2

Aula 10 (Parte 2): Engenharia de Recuperação Métricas IR, Limites dos Embeddings e Busca Híbrida

Prof. Aléssio Miranda Júnior
alessio@cefetmg.br

CEFET-MG - Campus Timóteo
Dep. Engenharia de Computação

Setembro - 2026/2

Onde Paramos?

Na sessão anterior, vocês construíram o motor semântico artesanal:

- **Limite da busca literal:** A busca por caracteres não modela sinonímia ou polissemia por si só.
- **Similaridade de Cosseno:** $\cos \theta = \frac{A \cdot B}{\|A\| \|B\|}$ mede ângulo, ignorando magnitude de texto.
- **Vetorização com spaCy:** No modelo usado no laboratório, frases viram vetores densos em $\mathbb{R}^{300}$; outros modelos podem ter outra dimensionalidade.
- **Ranking Top-K:** O buscador varreu o corpus e trouxe “taxa de juros” para “economia e dinheiro”.

A Pergunta de Engenharia para Hoje

Uma demonstração bonita no terminal prova que o buscador é bom?
Como avaliar formalmente a recuperação em escala industrial?

- **A Ilusão da PoC:** Por que um ranking isolado não serve para homologar um produto de IA.
- **Consultas Rotuladas (Ground Truth):** A fundação para medir qualidade de Information Retrieval (IR).
- **Métricas Formais:** Compreender e calcular Precision@K e Recall@K.
- **Limites a investigar:** Avaliar consultas com identificadores exatos e códigos técnicos (ERR-504).
- **Busca Híbrida (Hybrid Search):** Fusão ponderada de sinais léxicos e semânticos.
- **RAG Responsável:** Mecanismo de abstenção com threshold e proveniência auditável.

A Ilusão da Prova de Conceito (PoC)

O Erro Clássico do Engenheiro Júnior

Testar 2 ou 3 queries convenientes, ver resultados coerentes no terminal e declarar:
“Pronto, o buscador semântico está perfeito!”

O Risco Corporativo

- E se o usuário buscar uma sigla jurídica?
- E se buscar um número de protocolo fiscal?
- E se fizer uma pergunta fora do domínio?

Demo Visual (Terminal)

3 queries escolhidas a dedo

↓ quebra

Produção Real

Milhares de queries com ruído

↑ sustenta

Avaliação Sistemática

Benchmark com Ground Truth

- Para medir a qualidade de recuperação, precisamos de um **padrão-ouro (Ground Truth)**.
- Para cada consulta de teste q_i , especialistas determinam previamente o conjunto R_{q_i} de documentos que **realmente respondem** àquela pergunta.

Exemplo de Mapeamento Ground Truth no Nosso Corpus

ID Query	Texto da Consulta	Documentos Relevantes (R)
Q_1	“economia e dinheiro”	Doc 1 (Juros), Doc 8 (Dólar)
Q_2	“problemas de saúde”	Doc 4 (Cardíaco), Doc 5 (Vacina UBS)
Q_3	“animais domésticos”	Doc 0 (Gato no sofá), Doc 7 (Cachorro parque)
Q_4	“código técnico ERR-504”	Doc 10 (Incidente de infraestrutura)

Métrica 1: Precision@K (Precisão no Top-K)

Seja R o conjunto de itens relevantes e L_K a lista dos primeiros K resultados retornados:

$$\text{Precision@K} = \frac{|R \cap L_K|}{K} \quad (1)$$

- Responde: **“Dos K itens que o buscador trouxe, quantos prestam?”**
- Mede a **taxa de ruído** entregue ao usuário ou ao LLM.
- Se o Top-2 traz 1 documento certo e 1 errado: $\text{Precision@2} = 1/2 = 50\%$.

Top-K Retornado (L_K)

Relevante ($R \cap L_K$)

Relevante ($R \cap L_K$)

Ruído (Irrelevante)

Métrica 2: Recall@K (Revocação no Top-K)

$$\text{Recall@K} = \frac{|R \cap L_K|}{|R|} \quad (2)$$

- Responde: **“De tudo o que existia de relevante no corpus, quanto conseguimos resgatar no Top-K?”**
- Mede a **cobertura da evidência**.
- Crítico para RAG: se o documento que continha a resposta ficou na posição #15 e o Top-K é 3, o Recall é baixo e o LLM não terá como responder.

O Dilema Precision vs. Recall

- Para um ranking fixo, aumentar K **não reduz o Recall** (há mais chances de incluir a resposta).
- Mas aumentar K costuma **reduzir a Precision** (injeta mais ruído e gasta tokens da janela de contexto).

MRR (Mean Reciprocal Rank)

- Avalia a posição do **primeiro item relevante** encontrado:

$$RR = \frac{1}{\text{rank}_{\text{primeiro}}}$$

- 1º lugar: $RR = 1.0$
- 2º lugar: $RR = 0.5$
- 4º lugar: $RR = 0.25$
- Ideal para assistentes de resposta direta onde o usuário só lê a 1ª sugestão.

nDCG (Normalized DCG)

- Trata relevância com **pesos graduados** (ex: 0 = irrelevante, 1 = parcial, 2 = perfeito).
- Aplica desconto logarítmico pela posição:

$$DCG@K = \sum_{i=1}^K \frac{rel_i}{\log_2(i+1)}$$

- Penaliza pesadamente colocar o melhor documento no fim da lista.

O Calcanhar de Aquiles dos Embeddings

Embeddings Não Resolvem Todo Tipo de Consulta

Embeddings projetam textos em um espaço geométrico contínuo. Eles costumam ajudar com **conceitos difusos**, mas podem perder sinal quando a busca exige **exatidão literal**; isso depende do modelo, do corpus e do protocolo de consulta.

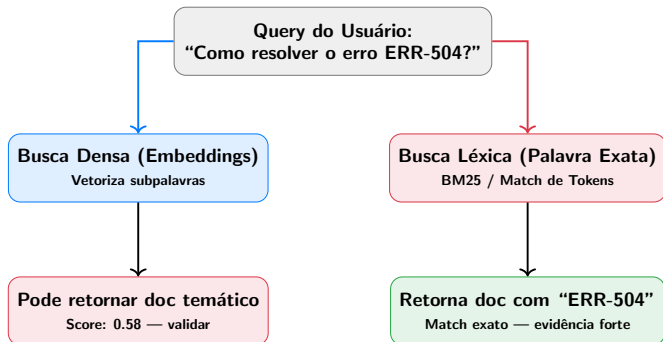
Onde Embeddings Brilham

- Sinônimos (carro $\approx$ automóvel).
- Paráfrases e linguagem natural.
- Consultas conceituais abertas.
- Busca semântica multilíngue.

Onde Embeddings Falham

- Códigos de erro (ex: ERR-504) e outros identificadores exatos.
- Nomes próprios incomuns e siglas.
- Números e identificadores de lei.
- Negações lógicas (“sem glúten”).

Estudo de Caso Prático: O Código de Erro



Diagnóstico

Um modelo de embedding pode não representar ERR-504 de modo suficientemente distinto no corpus. Por isso, o resultado deve ser medido: a busca densa pode recuperar contexto relacionado ou pode perder o identificador, enquanto a busca léxica preserva o sinal exato.

Tabela Comparativa: Quem Vence Onde?

Cenário de Consulta	Busca Léxica (BM25)	Busca Densa (Cosseno)	Busca
Paráfrase (“dinheiro e juros”)	Pode perder sinônimos	Pode ajudar	Avaliar n
Código exato (ERR-504)	Forte para match exato	Pode perder o identificador	Avaliar n
Sinonímia médica (“dor de cabeça”)	Pode perder paráfrases	Pode ajudar	Avaliar n
Número de contrato / CPF	Forte para token exato	Pode perder o número	Avaliar n
Palavra rara fora do vocabulário	Depende da indexação	Depende do modelo	Avaliar n

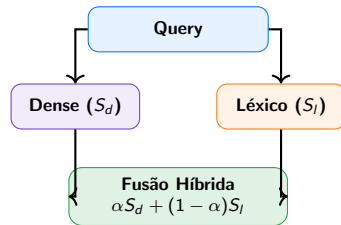
Conclusão Industrial

A escolha entre busca densa, léxica ou híbrida deve ser guiada por métricas e análise de erros do domínio.

Busca Híbrida Como Estratégia de Fusão

- Executamos simultaneamente a **Busca Léxica** e a **Busca Densa**.
- Transformamos os scores para uma escala comparável; isso não garante calibração por si só.
- Aplicamos uma combinação linear por um hiperparâmetro $\alpha \in [0, 1]$, escolhido na validação:

$$S_{\text{híbrido}} = \alpha \cdot S_{\text{denso}} + (1 - \alpha) \cdot S_{\text{léxico}} \quad (3)$$



Fusão Alternativa: Reciprocal Rank Fusion (RRF)

- Somar scores requer que as escalas estejam calibradas (o que nem sempre é verdade).
- O **RRF (Reciprocal Rank Fusion)** ignora os scores absolutos e olha apenas as **posições nos rankings**:

$$\text{Score}_{\text{RRF}}(d) = \sum_{m \in \{\text{léxico}, \text{denso}\}} \frac{1}{k + r_m(d)} \quad (4)$$

Propriedades Elegantes do RRF

- $r_m(d)$ é a posição (rank 1, 2, ...) do documento no método m .
- A constante k (frequentemente iniciada em 60) reduz o domínio de uma posição isolada e deve ser validada no domínio.
- Um documento bem colocado em ambos os métodos pode subir no ranking final, sem garantia de ocupar a primeira posição.

O Erro Ingênuo do Pipeline RAG

O usuário faz uma pergunta completamente sem sentido ou fora de contexto:

“Qual o segredo para plantar couve na lua?”

O que o buscador ingênuo faz:

- Varre o banco de finanças e medicina.
- Calcula cossenos baixos (ex: 0.28).
- Retorna os Top-3 mesmo assim!
- Injeta os 3 textos no prompt do LLM.

O que o LLM faz:

- Tenta obedecer à instrução a todo custo.
- Conecta juros imobiliários com cultivo lunar.
- **Alucina uma mentira assertiva!**

A Política de Abstenção (Thresholding)

- Um sistema responsável deve oferecer uma política explícita para quando não houver evidência suficiente.
- Definimos um limiar de corte calibrado τ no conjunto de validação; seu valor não é transferível automaticamente entre modelos.
- A decisão combina score, cobertura, qualidade das fontes e domínio. Score baixo pode levar à abstenção, mas não é o único sinal.



Proveniência e Citação Auditável

- Além do texto recuperado, preserve metadados que permitam rastrear a origem da evidência.
- Cada trecho recuperado deve carregar metadados versionados e rastreáveis:

Contrato de Dados da Evidência para RAG

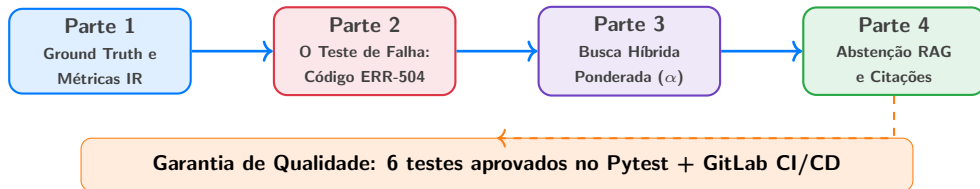
```
{  
  "doc_id": "DOC-004",  
  "score_confianca": 0.8912,  
  "fonte": "manual_tecnico_v2.pdf",  
  "pagina": 42,  
  "trecho": "A anomalia cardiaca exige protocolo imediato de  
    emergencia..."  
}
```

Princípio de Auditoria

Toda afirmação verificável deve apontar para evidência recuperada. A citação permite

1. **Demonstração visual não é métrica:** Homologação exige Ground Truth e métricas de IR.
2. **Precision@K vs Recall@K:** Precision mede ruído; Recall mede cobertura da evidência.
3. **Limites dos embeddings:** Identificadores, números e termos técnicos exatos podem exigir sinal léxico adicional.
4. **Busca Híbrida:** Combina sinais densos e léxicos; pesos, fusão e ganhos precisam ser validados.
5. **RAG Responsável:** Política calibrada de abstenção, proveniência e verificação de suporte para respostas auditáveis.

Roteiro do Laboratório 10 (Parte 2)



Entregáveis da Bancada

- Módulo `src/buscador_cosseno.py` expandido com busca híbrida e métricas.
- Suíte `tests/test_buscador.py` com 6 testes verdes no terminal.

1. Métricas de Recuperação

Implementar `calcular_precision_at_k` e `calcular_recall_at_k` para avaliar o benchmark de 4 consultas rotuladas.

2. Busca Híbrida

Fundir cosseno com contagem normalizada de termos para recuperar ERR-504 em Top-1.

3. Abstenção com Threshold

Criar função `recuperar_para_rag` que retorna status `ABSTENCAO` para queries sem suporte textual.

4. Testes Automatizados

Escrever os 3 novos testes unitários e comitar com git.

Dúvidas?

Email: alessio@cefetmg.br

Aula: Segunda e Quarta-Feira, 7h as 8:40h (Sala 39)

WhatsApp: Dúvidas rápidas

Lembrete

Confira sua Issue no GitLab!

Próxima aula: **Vector DBs e ChromaDB**

100% da Aula 11 Concluída!