RAG no Diário Oficial: Um Estudo sobre os Limites da Busca Semântica

Análise de Performance e Desafios da Recuperação de Informação em Documentos Não Estruturados

Gustavo Almeida Valentim

Julho de 2025

Resumo

Este trabalho aborda o desafio de extrair informações de Diários Oficiais, documentos públicos de notória complexidade e difícil acesso. Para democratizar esta informação, foi investigada a aplicação de uma arquitetura de Geração Aumentada por Recuperação (RAG). Uma arquitetura de linha de base, utilizando busca vetorial simples, foi sistematicamente avaliada contra um dataset verificado para estabelecer um benchmark de performance. Os resultados revelaram um teto de performance com acurácia máxima de 34.38%, evidenciando que a falha crítica do sistema não reside na capacidade de geração do *Large Language Model (LLM)*, mas na ineficácia da etapa de recuperação de informação (*retrieval*) em encontrar os contextos corretos em meio ao ruído semântico do jargão jurídico. O estudo conclui que a busca vetorial pura é insuficiente para este domínio e prova a necessidade de arquiteturas mais sofisticadas, como Busca Híbrida e *Re-ranking*, como o caminho para a construção de ferramentas de IA verdadeiramente eficazes para a promoção da transparência pública.

Sumário

1 Introdução

A transparência dos atos governamentais é um pilar essencial para a democracia, e o Diário Oficial é a principal ferramenta para garantir esse direito. No entanto, existe uma contradição fundamental: apesar de ser pública, a informação contida nesses documentos é, na prática, de difícil acesso para a maioria das pessoas.

O problema está no formato em que esses dados são apresentados. Publicados como arquivos de formato `PDF` (Portable Document Format), eles são frequentemente longos, não estruturados e escritos em uma linguagem técnica e burocrática. Encontrar uma informação específica, como o valor de um contrato ou os detalhes de uma nomeação, exige um esforço manual enorme. Isso afasta o cidadão comum e limita o acesso à informação a um pequeno grupo de especialistas, o que vai contra o princípio da transparência.

Os Modelos de Linguagem de Grande Porte, ou *Large Language Models (LLMs)*, revolucionaram a forma como interagimos com a tecnologia por sua capacidade de entender e gerar texto. Contudo, seu conhecimento é limitado aos dados de seu treinamento e eles não conhecem o conteúdo de documentos específicos, como os do Diário Oficial. Para resolver essa limitação, surge a arquitetura de Geração Aumentada por Recuperação, ou *Retrieval-Augmented Generation (RAG)* [1]. O RAG atua como uma ponte, conectando o poder de raciocínio de um LLM a uma base de documentos específica, permitindo que ele responda perguntas sobre informações que não viu durante seu treinamento.

O objetivo deste trabalho é documentar a implementação e, mais importante, a análise metodológica de um sistema RAG de linha de base, projetado para o corpus do Diário Oficial do Distrito Federal (DODF). A hipótese central não é apenas a de que o sistema pode funcionar, mas que o processo de análise de uma arquitetura simples revelará suas limitações. Entender essas falhas é o que justifica, de forma científica, a necessidade de evoluir para técnicas mais avançadas, buscando um nível de precisão confiável para o público geral.

Este relatório está estruturado da seguinte forma: a Seção 2 detalha a metodologia empregada na construção do sistema base e no desenho dos experimentos. A Seção 3 apresenta os resultados quantitativos obtidos. A Seção 4 discute as implicações desses resultados, analisando as causas das falhas e o desempenho do sistema. Finalmente, a Seção 5 conclui o trabalho e aponta os caminhos para trabalhos futuros, baseados nas evidências coletadas.

2 Metodologia

Para avaliar sistematicamente a eficácia de um sistema RAG no domínio de documentos jurídicos, foi definida uma metodologia robusta, abrangendo a preparação dos dados, a arquitetura do sistema de linha de base e um processo de avaliação multifacetado.

2.1 Corpus de Dados

A base de conhecimento para este estudo é composta por um corpus de 57 documentos em formato `PDF`, extraídos de diversas edições do Diário Oficial do Distrito Federal (DODF). Este corpus é caracterizado por sua alta heterogeneidade, contendo uma mistura de atos de naturezas distintas, como extratos de contrato, termos aditivos e portarias de nomeação. Os documentos são inerentemente não estruturados, com grande variação de

layout e extensão, o que representa um desafio significativo para a extração de informação automatizada.

2.2 Datasets de Avaliação

Para conduzir uma análise completa, foram utilizados um conjunto de dados de avaliação:

2.2.1 Dataset Sintético para Avaliação Quantitativa

O conjunto consiste em um dataset de larga escala, gerado programaticamente. A partir de um arquivo de dados pré-existente contendo pares de `objeto-valor` de contratos, um *Large Language Model (LLM)* foi empregado com a técnica de *few-shot prompting* para gerar variações de perguntas em linguagem natural para cada objeto. Este método permitiu a criação de um volume significativo de dados de teste, ideal para uma avaliação quantitativa e automatizada, focada primariamente na métrica de acurácia por correspondência exata de valores. Reconhece-se, contudo, que este dataset pode conter um viés de "modelo-no-loop"(*model-in-the-loop bias*), dado que as perguntas foram formuladas por uma IA.

2.3 Arquitetura do Sistema de Linha de Base

A arquitetura inicial foi deliberadamente simplificada para servir como um ponto de partida controlável, permitindo uma análise clara de suas capacidades e limitações. Os componentes são detalhados a seguir.

2.3.1 Processamento e Divisão de Texto (Chunking)

Os documentos do corpus foram processados com a biblioteca LangChain [2]. Para a segmentação do texto, foi empregado o `RecursiveCharacterTextSplitter`, uma técnica que tenta preservar a integridade semântica ao dividir o texto priorizando quebras em parágrafos e sentenças. Os hiperparâmetros `CHUNK_SIZE` e `CHUNK_OVERLAP` foram sistematicamente variados durante a fase de testes para avaliar seu impacto na performance do sistema.

2.3.2 Vetorização e Armazenamento Vetorial

Para a conversão dos chunks de texto em representações numéricas (vetores), foi utilizado o modelo de embedding `mxbai-embed-large:latest`, servido através de uma instância local da plataforma Ollama [4]. Os vetores gerados foram indexados em uma base de dados vetorial utilizando a biblioteca `FAISS` (Facebook AI Similarity Search)[3]. O FAISS otimiza a busca por similaridade euclidiana, sendo um componente essencial para a eficiência da etapa de recuperação do RAG.

2.3.3 Geração de Resposta e Cadeia RAG

Na etapa final, o LLM `llama4:latest` (67GB), servido por uma instância remota do Ollama, foi empregado como o componente gerador de respostas. A orquestração do fluxo foi realizada pela montagem de uma cadeia `RetrievalQA` simples. Esta cadeia implementa o fluxo RAG básico: a pergunta do usuário é vetorizada, o retriever busca

no índice FAISS os k chunks mais similares, e estes, junto com a pergunta original, são inseridos em um prompt e entregues ao LLM para que ele sintetize a resposta final.

2.4 Procedimento e Métricas de Avaliação

A avaliação do sistema foi conduzida utilizando o datasets descrito:

1. **Avaliação Quantitativa:** Utilizando o dataset sintético, a performance foi medida pela Acurácia de Correspondência Exata (Exact Match Accuracy). Um script customizado (`evaluate_rag.py`) foi desenvolvido para extrair os valores monetários da resposta gerada e compará-los com o gabarito, com 192 perguntas.

3 Resultados e Análise

3.1 Resultados dos Testes de Hiperparâmetros

Com o objetivo de otimizar a arquitetura de linha de base, foi conduzido um experimento sistemático para avaliar o impacto de diferentes hiperparâmetros de processamento de texto e recuperação de informação. A performance de cada configuração foi medida utilizando a métrica de Acurácia de Correspondência Exata sobre o *Golden Dataset Verificado*, composto por 192 perguntas. Os resultados mais significativos do experimento são apresentados na Tabela 1.

Tabela 1: Resultados de Acurácia vs. Configuração de Chunking e Retrieval

CHUNK_SIZE	CHUNK_OVERLAP	RETRIEVER_K	Acurácia (%)
512	100	10	**34.38%**
256	64	10	21.35%

A melhor performance observada foi uma acurácia de 34.38%, obtida com a configuração de `CHUNK_SIZE=[512]` e `RETRIEVER_K=[10]`. Este resultado, embora represente um avanço significativo em relação a configurações menos otimizadas, também estabelece um teto de performance claro para a arquitetura de busca vetorial simples, indicando que mais de 65% das perguntas, mesmo com a resposta presente no corpus, não foram respondidas corretamente.

3.2 Análise das Limitações Persistentes

A análise dos casos de falha, mesmo na configuração otimizada, revela que a limitação fundamental do sistema reside na etapa de *retrieval*. O erro não está na capacidade de geração do *Large Language Model (LLM)*, mas na qualidade do contexto que lhe é fornecido. Duas causas principais foram identificadas:

Ruído Semântico e Ambiguidade: A natureza do corpus do Diário Oficial, com seu vocabulário jurídico e contratual repetitivo, cria um cenário de alta ambiguidade para a busca vetorial. Múltiplos chunks, embora tratando de contratos distintos, podem ser semanticamente muito próximos, dificultando para o retriever distinguir o documento precisamente correto de outros plausíveis. O sistema frequentemente identifica a "vizinhança"correta no espaço vetorial, mas falha em pinçar o documento exato.

Sensibilidade à Formulação da Pergunta (*Prompt Sensitivity*): Observou-se que pequenas variações na forma de perguntar sobre a mesma informação podem levar a resultados drasticamente diferentes. Isso indica que a busca por similaridade euclidiana é frágil e pode falhar em capturar a intenção do usuário se não houver uma sobreposição semântica direta, mesmo quando palavras-chave essenciais estão presentes.

Conclui-se, portanto, que, embora a otimização de hiperparâmetros seja uma etapa crucial que pode gerar ganhos de performance expressivos, ela não soluciona a fragilidade inerente da busca vetorial pura para este tipo de domínio. Atingiu-se um platô de performance que só pode ser superado com a introdução de arquiteturas de RAG mais sofisticadas.

4 Discussão e Trabalhos Futuros

4.1 Análise dos Resultados da Linha de Base

Os resultados quantitativos apresentados na seção anterior, que indicam uma acurácia máxima de 34.38% para o sistema de linha de base, exigem uma interpretação cuidadosa. Este número, embora aparentemente baixo, não deve ser visto como um simples fracasso, mas sim como uma métrica informativa que revela a real complexidade do problema proposto. A baixa acurácia é um sintoma direto de um desafio muito mais profundo relacionado à etapa de recuperação da informação (*retrieval*).

A análise qualitativa dos contextos recuperados (`retrieved_contexts`) para as perguntas que falharam é o ponto central desta descoberta. Observou-se que o problema raramente reside na capacidade de geração do *Large Language Model (LLM)*, que se mostrou robusto ao se recusar a responder quando o contexto não continha a informação precisa. A falha fundamental reside no retriever. A busca vetorial pura, baseada em similaridade euclidiana, demonstrou ser uma ferramenta imprecisa para este domínio por duas razões principais.

Primeiramente, a natureza do corpus do Diário Oficial apresenta uma alta densidade de "ruído semântico". Documentos sobre tópicos distintos, como um contrato para compra de pneus e um para aquisição de caminhões, podem compartilhar um vocabulário contratual e jurídico muito similar. Isso cria uma ambiguidade onde múltiplos chunks, embora semanticamente próximos, são contextualmente incorretos. O retriever, portanto, frequentemente retorna o chunk semanticamente mais "próximo", que nem sempre é o factualmente correto.

Em segundo lugar, a busca vetorial falha em atribuir o peso necessário a palavras-chave e entidades nomeadas que são críticas para desambiguar uma pergunta. Termos como "FUNAP"ou um número de processo específico são o sinal mais forte para encontrar a resposta correta, mas em um cálculo de similaridade de um vetor de alta dimensão, o peso desses termos pode ser diluído pelo resto do contexto, levando o sistema a não conseguir distinguir o sinal relevante em meio a um grande volume de ruído semântico.

4.2 Justificativa para Arquiteturas Avançadas e Trabalhos Futuros

As limitações da arquitetura de linha de base, evidenciadas pelos experimentos, justificam cientificamente a necessidade de explorar técnicas de RAG mais sofisticadas é preciso recorrer a métodos mais avançados que estão na fronteira da pesquisa em IA.

O caminho para a melhoria da acurácia, e o foco para a continuação deste trabalho, envolve a implementação e avaliação de duas estratégias principais:

1. **Cross-Encoder Re-ranking:** A primeira melhoria consiste em adicionar uma segunda camada de verificação após a busca inicial. Em vez de entregar os k chunks recuperados diretamente ao LLM, eles seriam primeiro passados por um modelo de Cross-Encoder. Este modelo atuaria como um "juiz", comparando diretamente a pergunta com cada chunk e reordenando-os com base em uma pontuação de relevância muito mais precisa. Isso garante que os exemplos com maior probabilidade de conter a resposta sejam priorizados.

2. **Busca Híbrida:** Para atacar diretamente o problema das palavras-chave, a implementação de uma busca híbrida é o próximo passo lógico. Esta técnica combina os pontos fortes da busca vetorial (que entende o significado) com os da busca lexical tradicional. Ao fundir os resultados de ambas as buscas, o sistema pode recuperar documentos que são tanto semanticamente relevantes quanto lexicalmente precisos, aumentando drasticamente a chance de encontrar o contexto correto.

Em suma, a evolução contínua da área de RAG demonstra que a versão "pura"da técnica é, muitas vezes, apenas o ponto de partida. A verdadeira eficácia para problemas do mundo real, como a democratização do acesso aos Diários Oficiais, reside na combinação inteligente e na orquestração de múltiplas técnicas de recuperação e refinamento.

5 Conclusão

O acesso à informação contida nos Diários Oficiais é um direito do cidadão e um componente essencial para a saúde da democracia. Contudo, como este trabalho demonstrou, a mera disponibilidade desses documentos em formato `PDF` não garante sua acessibilidade. A complexidade inerente à linguagem jurídica e à estrutura não padronizada dos textos representa uma barreira significativa, distanciando a população de informações que deveriam ser transparentes e de fácil consulta. A própria dificuldade encontrada durante a curadoria manual do dataset de avaliação deste estudo evidencia o desafio que um cidadão comum, muitas vezes sem treinamento técnico ou jurídico, enfrenta ao tentar interpretar tais documentos.

Neste contexto, este trabalho investigou a viabilidade de um sistema de Geração Aumentada por Recuperação (RAG) como ferramenta para mitigar essa barreira. Foi implementada e avaliada sistematicamente uma arquitetura de linha de base, utilizando uma busca vetorial simples para recuperar informações de um corpus de Diários Oficiais do Distrito Federal. A avaliação, conduzida revelou os limites práticos desta abordagem fundamental.

A principal descoberta deste estudo é que a eficácia de um sistema RAG, neste domínio, é criticamente dependente da qualidade da sua etapa de recuperação de informação (*retrieval*). Foi provado empiricamente que a busca por similaridade semântica, embora poderosa, é insuficiente para navegar na densidade e na ambiguidade dos textos oficiais, resultando em uma baixa acurácia na localização dos contextos corretos. O gargalo do sistema não reside na capacidade de geração do *Large Language Model (LLM)*, mas sim na sua capacidade de encontrar a "agulha no palheiro"informativa dentro de um grande volume de dados.

A contribuição deste trabalho, portanto, não está na entrega de um sistema com performance otimizada, mas na demonstração clara e metodológica das suas falhas iniciais. Foi provado que, embora um RAG simples já possa gerar informações de valor, ele necessita de arquiteturas mais sofisticadas para se tornar uma ferramenta verdadeiramente confiável. Este estudo estabelece um benchmark sólido e justifica cientificamente a necessidade de trabalhos futuros, focados na implementação de técnicas avançadas como *Re-ranking* e Busca Híbrida, como o caminho para desenvolver sistemas de IA que possam, de fato, fortalecer a democracia através da democratização do acesso à informação pública.

6 Referências

Referências

[1] Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems, 33.

[2] LangChain. (2024). *LangChain Documentation*. Disponível em: https://python.langchain.com/

[3] Johnson, J., Douze, M., & Jégou, H. (2019). *Billion-scale similarity search with GPUs*. IEEE Transactions on Big Data, 7(3).

[4] Ollama. (2024). *Large language models locally*. Disponível em: https://ollama.com/