Desenvolvimento de um Assistente RAG Temático Star Wars com Ollama e Eel

Vinícius Santos Monteiro

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Carlos – SP – Brazil

vini.mon@usp.br

Abstract. *This study describes the development of a prototype Retrieval-Augmented Generation (RAG) assistant focused on the Star Wars universe. The objective was to create a chatbot capable of answering user questions based on a curated collection of texts about the saga. The methodology involved preparing a dataset of 18 documents, generating semantic embeddings using a Sentence-BERT model (‘paraphrase-multilingual-MiniLM-L12-v2‘), and indexing these embeddings for efficient retrieval. The query pipeline utilizes the Ollama platform to run the ‘mistral-nemo‘ large language model locally, orchestrated by the Langchain framework, to generate context-aware answers. A simple graphical user interface was developed using the Eel library, allowing interaction via a web browser. Preliminary results demonstrate the system's ability to retrieve relevant information and generate coherent answers based on the provided documents.*

Resumo. *Este estudo descreve o desenvolvimento de um protótipo de assistente baseado em Geração Aumentada por Recuperação (RAG) focado no universo Star Wars. O objetivo foi criar um chatbot capaz de responder perguntas de usuários com base em uma coleção curada de textos sobre a saga. A metodologia envolveu a preparação de um conjunto de dados com 18 documentos, a geração de embeddings semânticos utilizando um modelo Sentence-BERT (‘paraphrase-multilingual-MiniLM-L12-v2‘), e a indexação destes para recuperação eficiente. O pipeline de consulta utiliza a plataforma Ollama para executar localmente o modelo de linguagem grande ‘mistral-nemo‘, orquestrado pelo framework Langchain, para gerar respostas contextualmente fundamentadas. Uma interface gráfica simples foi desenvolvida com a biblioteca Eel, permitindo a interação via navegador web. Resultados preliminares demonstram a capacidade do sistema em recuperar informações pertinentes e gerar respostas coerentes com base nos documentos fornecidos.*

1. Introdução

A capacidade de extrair informações específicas e responder perguntas complexas a partir de grandes volumes de texto é um desafio central na inteligência artificial. Sistemas de Geração Aumentada por Recuperação (RAG) surgem como uma abordagem promissora, combinando a recuperação de informações relevantes de uma base de conhecimento com a capacidade de geração de texto de grandes modelos de linguagem (LLMs) [1].

Este relatório detalha o desenvolvimento de um protótipo de assistente RAG temático, focado no rico e vasto universo de Star Wars. O objetivo foi construir um

sistema capaz de responder perguntas factuais sobre personagens, eventos e conceitos da saga, baseando suas respostas em um conjunto pré-definido de documentos textuais. A motivação para este projeto reside na exploração da viabilidade de construir sistemas RAG funcionais utilizando ferramentas acessíveis e modelos executados localmente.

A arquitetura implementada utiliza o modelo de embedding 'paraphrase-multilingual-MiniLM-L12-v2' para vetorização dos textos, o framework Langchain para orquestrar o pipeline de RAG, e o LLM 'mistral-nemo' executado localmente através da plataforma Ollama. Adicionalmente, uma interface gráfica simples foi criada com a biblioteca Eel para facilitar a interação do usuário. Este trabalho visa demonstrar a construção de um sistema RAG ponta-a-ponta, desde a preparação dos dados até a interface final.

O código-fonte e os experimentos completos podem ser encontrados no repositório GitHub e no Google Colab: [1]. [2].

2. Objetivos

O principal objetivo deste estudo é projetar, implementar e demonstrar um protótipo funcional de assistente RAG para responder perguntas sobre o universo Star Wars. Os objetivos específicos incluem:

- Preparar uma coleção de documentos textuais relevantes sobre Star Wars.
- Implementar um pipeline para gerar e indexar embeddings semânticos desses documentos.
- Configurar um fluxo de RAG utilizando Langchain e um LLM local (Ollama com 'mistral-nemo') para gerar respostas baseadas nos documentos recuperados.
- Desenvolver uma interface de usuário simples com Eel para interação com o assistente.
- Avaliar qualitativamente a capacidade do sistema de fornecer respostas relevantes e coerentes.

3. Abordagem Proposta

A solução foi implementada como um pipeline RAG, compreendendo as etapas de preparação de dados, indexação e consulta.

3.1. Preparação dos Dados

O conjunto de dados fonte foi construído a partir de 18 textos extraídos de fontes online (como a Star Wars Wiki Fandom [3]) e livros oficiais da saga. Cada texto possui entre 300 e 800 palavras, abordando diferentes aspectos do universo Star Wars (ex: Império Galáctico, Aliança Rebelde, Estrela da Morte). Os textos foram organizados em um arquivo CSV simples, contendo uma única coluna denominada 'text', onde cada linha representa o conteúdo completo de um documento. Nenhum pré-processamento complexo foi aplicado além da estruturação no formato CSV.

[1] https://github.com/vini-mon/RAG-FAQ-Star_Wars

[2] https://colab.research.google.com/drive/1AScSjRKifsJ72MU5JLrsZemPfh312-tr?usp=sharing

3.2. Indexação e Extração de Características

A etapa de indexação converte os documentos textuais em representações vetoriais (embeddings) para permitir a busca por similaridade semântica.

- **Geração de Embeddings:** Utilizou-se o modelo 'paraphrase-multilingual-MiniLM-L12-v2' [2] da biblioteca Sentence-Transformers. Este modelo foi escolhido por seu bom equilíbrio entre performance e custo computacional, gerando vetores de 384 dimensões para cada documento na coleção.
- **Indexação:** Os embeddings gerados foram armazenados juntamente com os textos originais (ou FAQs, se aplicável) em uma estrutura que permite a busca rápida por vetores similares durante a fase de consulta.

Esta etapa de indexação é computacionalmente mais intensiva, mas precisa ser executada apenas uma vez (ou sempre que a coleção de documentos for atualizada).

3.3. Pipeline de Consulta RAG

Quando um usuário submete uma pergunta através da interface Eel, o seguinte pipeline é acionado:

1. **Embedding da Pergunta:** A pergunta do usuário é convertida em um vetor de embedding usando o mesmo modelo Sentence-BERT utilizado na indexação.
2. **Recuperação (Retrieval):** O vetor da pergunta é comparado com os vetores indexados dos documentos. Os k documentos (configurado para $k = 2$ neste projeto) cujos embeddings são mais similares (geralmente por similaridade de cosseno) ao embedding da pergunta são recuperados. Estes documentos formam o "contexto" para a geração da resposta.
3. **Geração (Generation):** Um prompt é construído combinando um template pré-definido, o contexto recuperado (os trechos de texto dos documentos mais similares) e a pergunta original do usuário. Este prompt completo é então enviado para o LLM 'mistral-nemo' (via Ollama). O LLM utiliza as informações do contexto para gerar uma resposta textual à pergunta do usuário. A orquestração deste fluxo (embedding, recuperação, formatação do prompt, chamada ao LLM) é gerenciada pelo framework Langchain.
4. **Interface:** A resposta gerada pelo LLM é enviada de volta para a interface Eel e exibida ao usuário no navegador.

4. Resultados

A avaliação do protótipo foi realizada qualitativamente através de exemplos de interação. O sistema demonstrou capacidade de processar perguntas em linguagem natural e gerar respostas relevantes baseadas nos documentos indexados. A seguir, alguns exemplos de perguntas e as respectivas respostas geradas pelo assistente:

Pergunta 1: *O que é a Estrela da Morte?*

Resposta Gerada: *A Estrela da Morte era uma estação de batalha móvel construída pelo Império Galáctico como uma arma psicológica de terror em escala planetária. Era concebida para erradicar a dissidência através da ameaça de aniquilação instantânea e absoluta, servindo como um instrumento de controle através do medo inquestionável.*

Pergunta 2: *O que é a Aliança Rebelde?*

Resposta Gerada: *A Aliança Rebelde, também conhecida como a Aliança para Restauração da República, foi uma organização de resistência formada durante o período do Império Galáctico. Ela surgiu como uma resposta à tirania e à opressão do Império, após a diplomacia falhar. A Aliança era composta por diferentes facções de resistência espalhadas pela galáxia, incluindo grupos de guerrilha em Onderon, estrategistas Mon Cala que forneciam frota e inteligência, pilotos talentosos de mundos agrícolas como Tatooine, e financiadores secretos em mundos centrais. Um dos maiores feitos da liderança da Aliança Rebelde foi unir essas facções diferentes sob um comando centralizado, o Alto Comando Rebelde. A Aliança não tinha um planeta natal e operava a partir de bases móveis e secretas em mundos remotos para evitar aniquilação.*

Pergunta 3: *Quem comandava o império?*

Resposta Gerada: *Com base no contexto fornecido, o Imperador Galáctico era na verdade Darth Sidious, que também era conhecido como Sheev Palpatine. Portanto, Darth Sidious comandava o Império Galáctico.*

Discussão: Os exemplos demonstram que o assistente RAG consegue extrair informações presentes nos documentos de origem e sintetizá-las em respostas diretas. A qualidade da resposta depende fortemente da presença da informação relevante nos documentos recuperados durante a etapa de *Retrieval*. A interface desenvolvida com Eel provou ser uma forma eficaz e de baixo custo para interagir com o backend RAG implementado em Python. Observou-se que perguntas cujas respostas exigem a combinação de informações de múltiplos documentos ou inferências mais complexas podem apresentar desafios para o modelo atual.

5. Conclusão

Este estudo demonstrou com sucesso a construção de um protótipo de assistente RAG temático para o universo Star Wars, utilizando ferramentas acessíveis como Ollama, Langchain, Sentence Transformers e Eel. A arquitetura implementada, separando a indexação baseada em embeddings da geração de respostas com um LLM local, mostrou-se funcional para responder perguntas factuais com base em uma coleção limitada de documentos.

O uso de um LLM executado localmente (Ollama) e um modelo de embedding eficiente (‘paraphrase-multilingual-MiniLM-L12-v2‘) contribui para a viabilidade da solução em ambientes com recursos computacionais limitados, embora a etapa inicial de indexação ainda demande algum processamento. A interface Eel proporcionou um meio simples e eficaz para a demonstração interativa do sistema.

Conclui-se que a abordagem RAG é promissora para a criação de assistentes de conversação especializados. O protótipo atende aos objetivos propostos, fornecendo uma base funcional para futuras explorações.

Para trabalhos futuros, seria interessante a expansão da base de documentos, a experimentação com diferentes modelos de embedding e LLMs (incluindo modelos maiores, se os recursos permitirem), e a implementação de métricas mais formais para

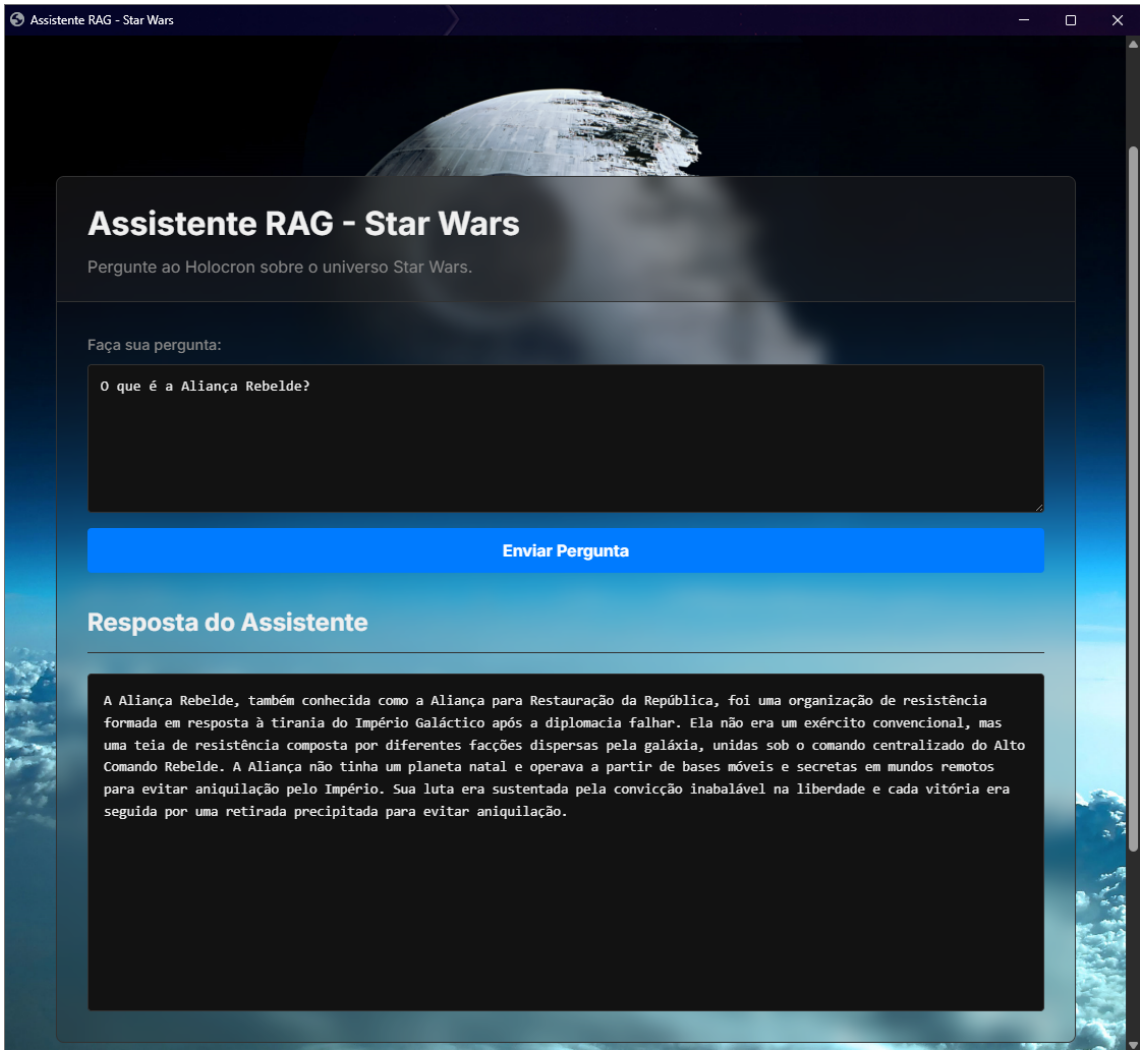

Figure 1. Exemplo da tela da interface gráfica

avaliação da qualidade da recuperação e da geração das respostas (ex: RAGAs, BLEU, ROUGE). A otimização do pipeline para reduzir a latência na resposta também seria um passo importante para melhorar a experiência do usuário.

6. References

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T. and Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, pp.9459–9474.

[2] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[3] Star Wars Wiki PT-BR. Enciclopédia Star Wars em Português. `https://starwars.fandom.com/pt/`. [Accessed: 25/10/2025]

[4] StarWars.com Databank. Official Star Wars Databank. https://www.starwars.com/databank/death-star. [Accessed: 25/10/2025].