Integrando Modelos de Recuperação e Geração: Uma Abordagem de RAG para Respostas em Perguntas do Usuário usando Llama 3.2 e Neo4j

Por: Emanuel de Jesus
Orientador: Eduardo Palhares

DEFESAS PÚBLICAS DE PÓS-GRADUAÇÃO DO PROJETO ARANOUÁ

INSTITUTO FEDERAL Amazonas · SAMSUNG · FAEPI · ARANOUÁ

SEJAM BEM VINDOS!

Nesta apresentação, exploraremos a integração de modelos de recuperação e geração. Abordaremos o RAG para respostas, tanto no prompt, quanto no chat, em perguntas usando Llama 3.2:1b e Neo4j. Descubra como essa combinação resulta em respostas mais precisas dadas pelo chat em relação ao prompt do Llama.

Introdução

Nos últimos anos, a inteligência artificial (IA) tem experimentado avanços significativos, especialmente no campo dos modelos de linguagem, como os transformadores. Uma das áreas que tem se destacado é a de Retrieval-Augmented Generation (RAG), que combina técnicas de recuperação de informações com geração de texto. Esses modelos são projetados para responder de forma mais precisa e contextualizada a perguntas complexas, uma capacidade crucial para a evolução das interações humano-máquina. No entanto, a implementação de RAG em sistemas de respostas a perguntas ainda enfrenta uma série de desafios significativos, que são essenciais para garantir a relevância, a precisão e a confiança nas respostas fornecidas {Proenca2024}.

O que é Geração Aumentada por Recuperação (RAG)?

Conceito

RAG combina recuperação de informações com geração de texto. Ele melhora as respostas com conhecimento externo.

Etapas

Primeiro, recupera informações relevantes. Depois, gera respostas com base nessas informações.

Vantagens

Oferece conhecimento atualizado e respostas contextuais. Garante precisão e relevância.

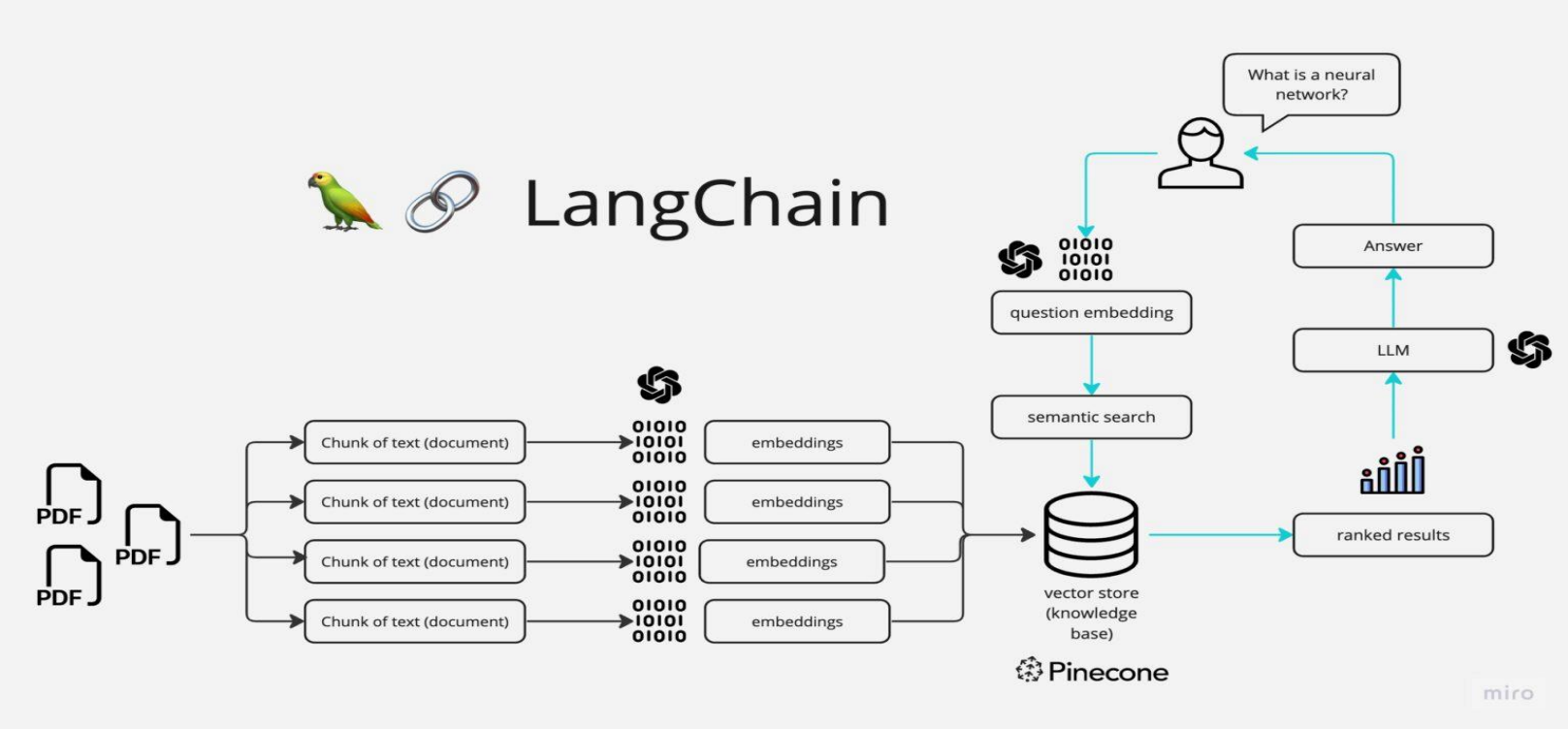
Metodologia

A construção do sistema proposto baseia-se na combinação de diferentes tecnologias que permitem integrar mecanismos de recuperação de informações com modelos generativos de linguagem. A estrutura central apoia-se na arquitetura RAG, implementada com o uso do **framework LangChain(que facilita a criação de aplicações com modelos de linguagem (como o ChatGPT) que interagem com dados externos, APIs, bancos de dados**) e que orquestra a comunicação entre os componentes. O armazenamento e a consulta de informações são realizados por meio do **banco de dados em grafos Neo4j**, enquanto a busca semântica é viabilizada por embeddings gerados pela **plataforma Ollama**. As respostas finais são produzidas pelo **modelo Llama 3.2 (1B),** ajustado para lidar com interações em linguagem natural e enriquecer as respostas com base nos documentos recuperados.

Arquitetura RAG

A arquitetura RAG (Geração Aumentada por Recuperação) usando LangChain que é uma estrutura de software ajuda a criar aplicações baseadas em LLM's e melhora as respostas em modelos de IA ao integrar informações externas. Estrutura essa que facilita a implementação do RAG, permitindo a conexão com bancos de dados, vetores de embeddings e documentos estruturados, o que aumenta a precisão e confiabilidade das respostas geradas{Vidivelli2023}, como demonstrada na imagem abaixo.

Arquitetura do Modelo RAG.



Banco de dados orientado a grafos

O Neo4j é um banco de dados orientado a grafos, projetado para armazenar e consultar dados que podem ser representados como grafos. Ele foi criado para lidar com relações complexas entre dados, sendo ideal para aplicações que exigem interações dinâmicas entre entidades, como redes sociais, sistemas de recomendação e detecção de fraudes{Khan2023}.

Um breve manuseio de Banco de dados Orientado a Grafos

Neo4j: Banco de Dados de Grafos para Recuperação Eficiente

Grafos

Neo4j usa grafos para representar o conhecimento. Ele armazena relações e entidades.

Armazenamento

Neo4j gerencia informações relevantes para RAG. As informações são armazenadas com eficiência.

Cypher

Consultas Cypher recuperam informações rapidamente. As consultas são precisas e eficientes.

Arquitetura do Sistema RAG com Llama 3.2 e Neo4j

1

Pergunta

Usuário faz uma pergunta.

2

Recuperação

Neo4j recupera informações relevantes.

3

Geração

Llama 3.2 gera a resposta final.

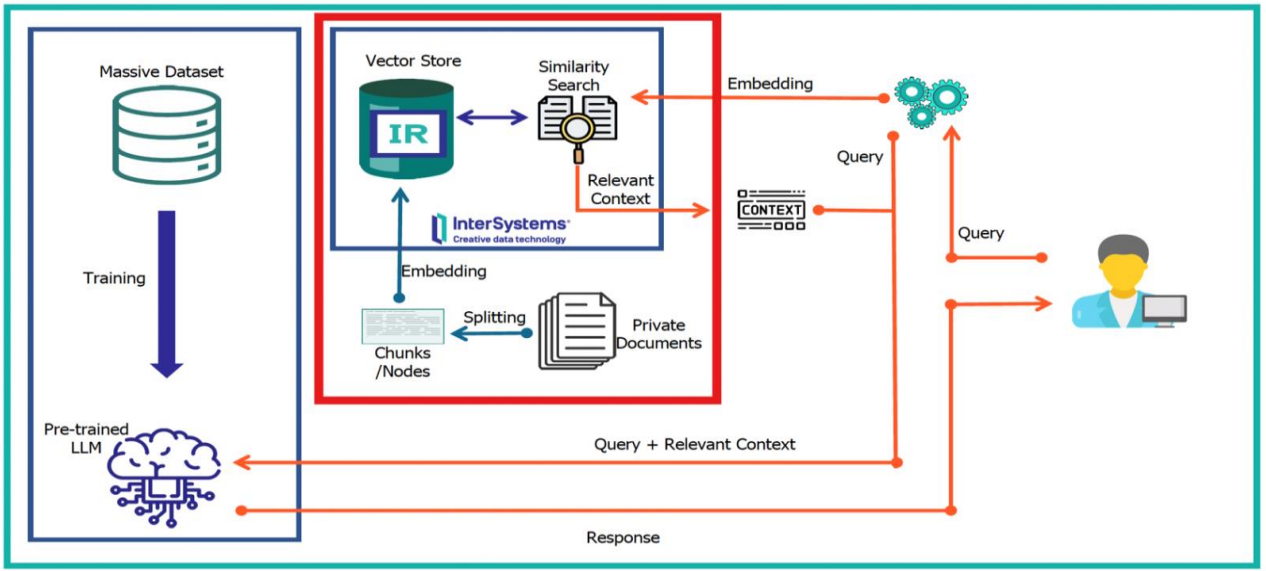
Embeddings Semânticos com Ollama

A recuperação de documentos relevantes neste trabalho é baseada em representações vetoriais de textos, conhecidas como embeddings semânticos. Esses vetores numéricos capturam relações de similaridade contextual entre termos, permitindo que consultas textuais sejam comparadas com documentos de forma mais precisa do que abordagens puramente lexicais. No contexto da arquitetura RAG, os embeddings são fundamentais para encontrar os documentos mais relevantes que servirão de base para a geração da resposta final pelo modelo de linguagem.

O sistema implementado segue um fluxo estruturado: o usuário insere uma pergunta por meio de uma interface web, e essa entrada é transmitida para a função \textit{chat_with_bot}. Essa função aciona o método \textit{query_neo4j}, responsável por gerar o embedding da consulta, calcular sua similaridade com os documentos vetorizados previamente no Neo4j e recuperar os conteúdos mais relevantes. O resultado desse processo são os documentos com maior proximidade semântica em relação à consulta, que serão usados como contexto para a geração da resposta final.A métrica utilizada para comparar os vetores da consulta e dos documentos é a similaridade de cosseno, definida pela fórmula:

$$silimaridade\,(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Após o cálculo das similaridades, o sistema seleciona os cinco documentos mais relevantes com base na pontuação obtida. Esses documentos são então encaminhados ao modelo gerador (Llama 3.2 (1B)), que os utiliza como contexto para elaborar uma resposta. Essa combinação entre recuperação baseada em embeddings e geração com LLMs é o que caracteriza a arquitetura RAG.



Métricas de Avaliação das Respostas Geradas

A qualidade das respostas geradas pelo sistema é avaliada por meio de duas métricas amplamente utilizadas na literatura de processamento de linguagem natural: ROUGE-L e BERTScore. Ambas têm como objetivo medir o grau de similaridade entre a resposta gerada pelo modelo e uma resposta de referência, porém adotam abordagens distintas e complementares.

A métrica ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) baseia-se na identificação da maior subsequência comum (Longest Common Subsequence, LCS) entre as palavras da resposta gerada e da referência. Essa subsequência considera apenas a ordem das palavras, desconsiderando interrupções intermediárias. A partir da LCS, são calculadas três medidas:

- Precisão: proporção da LCS em relação ao número de palavras da resposta gerada.
- Recall: proporção da LCS em relação ao número de palavras da resposta de referência
- F1-score: média harmônica entre precisão e recall.

O BERTScore, por sua vez, utiliza modelos pré-treinados como o BERT para gerar embeddings semânticos das palavras da resposta gerada e da referência. Com base nesses embeddings, calcula-se a similaridade semântica entre os textos. Assim como o ROUGE-L, o BERTScore fornece medidas de precisão, recall e F1-score, mas com a vantagem de reconhecer sinônimos, reescritas e reformulações estruturais.

Essa métrica é especialmente adequada para avaliar respostas geradas por modelos de linguagem, já que leva em consideração o significado das palavras, e não apenas sua ordem ou exatidão superficial. Por esse motivo, é comum que os valores de F1 obtidos pelo BERTScore sejam significativamente mais altos do que os do ROUGE-L, mesmo quando ambos avaliam as mesmas respostas. Essa diferença será retomada e discutida na seção de resultados.

Resultados e discussão

Esta seção apresenta os resultados obtidos a partir da aplicação do modelo Llama 3.2 (1B) em dois cenários distintos: com e sem a utilização da arquitetura RAG. O objetivo é comparar a qualidade das respostas geradas em cada abordagem, com base em perguntas elaboradas a partir do conteúdo do livro A Riqueza das Nações, de Adam Smith.

Ao final, os resultados são quantificados por meio de métricas objetivas ROUGE-L e BERTScore, permitindo avaliar, de forma mais precisa, o impacto da técnica RAG na geração das respostas. A seção termina com uma discussão interpretativa, que analisa as principais tendências observadas e suas implicações para o uso prático de modelos LLMs com RAG.

Resposta com RAG

Pergunta 1 — Qual o conceito de valor?

Resposta com RAG

O conceito de valor, segundo Adam Smith, está ligado ao trabalho necessário para produzir um bem. Ele distingue entre valor de uso, que é a utilidade de um bem, e valor de troca, que é a capacidade de ser trocado por outros bens no mercado. O trabalho é a medida real do valor de troca.

Pergunta 2 — De onde surge o conceito de valor?

Resposta com RAG

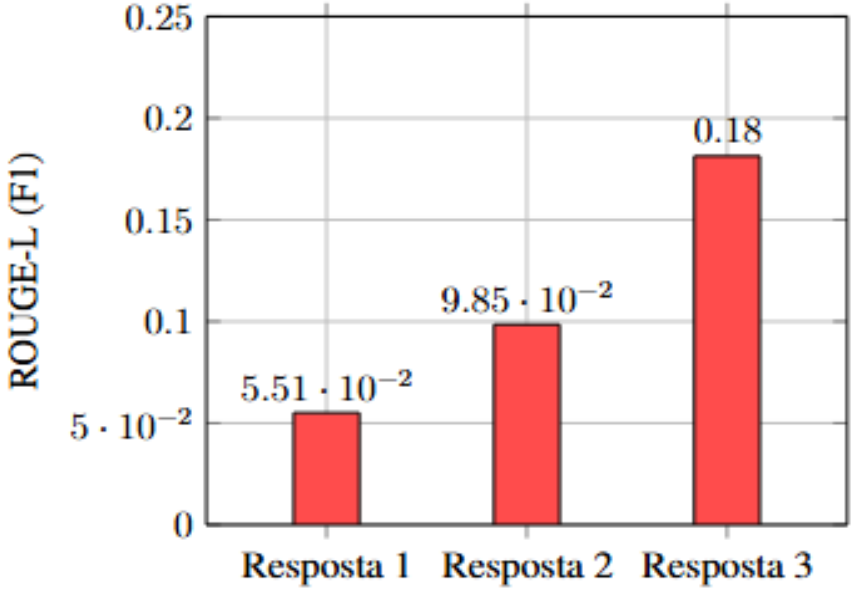
Segundo Adam Smith, o valor dos bens surge da quantidade de trabalho empregada em sua produção. O trabalho é a medida real do valor de troca de qualquer mercadoria.

Pergunta 3 — Quais os fatores de produção que geram o conceito de valor?

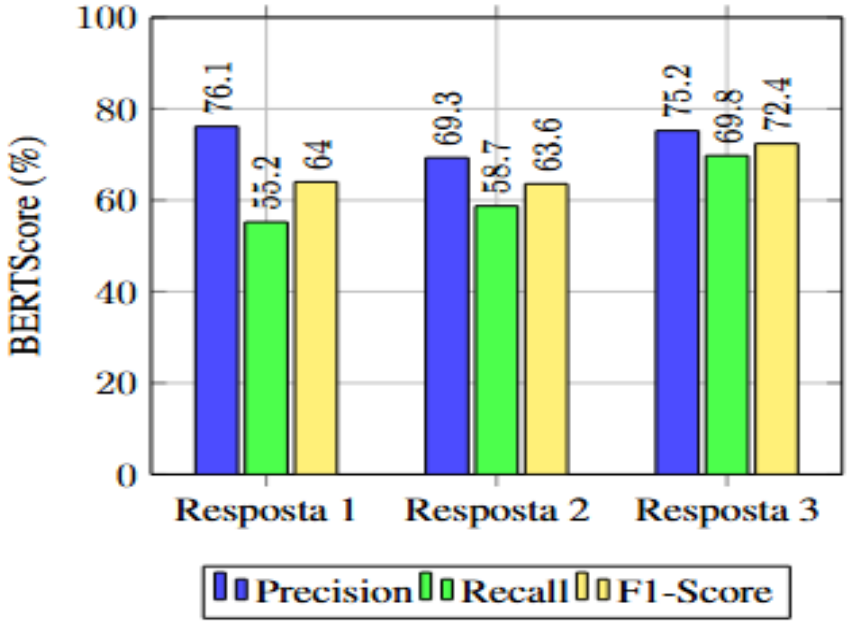
Resposta com RAG

De acordo com Adam Smith, o valor de troca de uma mercadoria é determinado pelo trabalho necessário para produzi-la. O trabalho é o fator fundamental que gera valor, especialmente nas sociedades primitivas, onde o trabalho era o único fator de produção.

Resultados com ROUGE-L



Resultados com BERTScore



A análise das três perguntas revelou diferenças consistentes entre as respostas geradas pelo modelo puro (sem RAG) e aquelas produzidas com o suporte da arquitetura RAG. De modo geral, as respostas com RAG mostraram-se mais precisas, concisas e conceitualmente alinhadas ao conteúdo original da obra de Adam Smith, enquanto as respostas sem RAG tenderam a ser mais genéricas, com traços de interferência de abordagens modernas ou interpretações alternativas.

Do ponto de vista quantitativo, os valores de BERTScore reforçaram essa conclusão: as respostas com RAG obtiveram F1-scores acima de 70%, indicando alta similaridade semântica com as referências. Já os valores de ROUGE-L, embora mais baixos, estão de acordo com o comportamento esperado da métrica, que penaliza reformulações textuais mesmo quando o conteúdo está correto. Isso evidencia a importância de utilizar métricas complementares na avaliação de modelos de linguagem.

Esses resultados indicam que a integração entre mecanismos de recuperação e modelos generativos pode melhorar significativamente a qualidade das respostas em tarefas de pergunta e resposta com base em documentos. No entanto, a eficácia da abordagem RAG depende da qualidade e da cobertura semântica da base de dados utilizada. Em contextos em que o conteúdo de referência é escasso, mal segmentado ou ruidoso, os ganhos observados podem não se repetir.

Conclusão e Trabalhos Futuros

Este trabalho apresentou uma arquitetura baseada em Retrieval-Augmented Generation (RAG) para melhorar a geração de respostas em tarefas de pergunta e resposta a partir de documentos. A proposta integrou a recuperação semântica de informações, por meio de embeddings gerados com a plataforma Ollama, a um modelo de linguagem Llama 3.2 (1B), permitindo que o sistema combinasse busca contextualizada com geração textual em linguagem natural.

Trabalhos futuros podem explorar diversas direções. Uma possibilidade é ampliar a base documental vetorizada, incorporando múltiplas fontes e organizando os dados em estruturas mais robustas, como grafos de conhecimento. Outra linha de investigação envolve a comparação entre diferentes técnicas de vetorização e recuperação semântica, avaliando seu impacto na performance da arquitetura. Também se destaca o potencial de aplicar a abordagem proposta em domínios específicos, como educação, saúde ou direito, avaliando sua robustez em tarefas mais especializadas. Por fim, a incorporação de mecanismos de feedback humano para refinar as respostas geradas representa uma perspectiva promissora para aplicações práticas.

AGRADECIMENTOS

Os autores agradecem à Samsung Eletrônica da Amazônia Ltda., por meio do Projeto Aranouá, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro por meio do Programa de Excelência Acadêmica (PROEX). Este trabalho é resultado do projeto de Pesquisa e Desenvolvimento (P\&D) 001/2021, firmado com o Instituto Federal do Amazonas e a FAEPI, com financiamento da Samsung.