Versionamento :: [major.minor.patch] Major :: Conclusões de cada section Minor :: Conclusões de sub sections Patch :: Revisões do professor orientador

TODO Agradecimentos DONE Introdução DONE Motivação DONE Objetivos DONE Metodologia TODO Fundamentação TODO Conceitos Fundamentais TODO Referencia Cientifica para LLMs TODO Explicar o conceito de RAG TODO Modelagem TODO Modelo Conceitual TODO Tecnologias Utilizadas TODO Explicar Python TODO Explicar virtual environment TODO Explicar principais bibliotecas (Docling, LamaIndex, etc.) TODO Explicar Banco Vetorial (Elasticsearch) TODO Explicar streamlit TODO Solução Desenvolvida TODO Figuras da solução desenvolvida TODO Resultados concretos (respostas das perguntas) TODO Explicar Funcionalidades TODO Conclusão TODO Considerações Finais TODO Limitações TODO Referências

Universidade Federal do Estado do Rio de Janeiro
Centro de Ciências Exatas e Tecnológicas
Escola de Informática Aplicada

Retrieval Augmented Generation Aplicada à Bibliotecas

Breno Costa da Silva Filgueiras

Rio de Janeiro, RJ – Brasil
Dezembro, 2024

Retrieval Augmented Generation Aplicada à Bibliotecas

Breno Costa da Silva Filgueiras

Projeto de graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) como cumprimento de requerimento parcial para obtenção título de Bacharel em Sistemas de Informação.

Approved by:

Supervisor, D.Sc. – UNIRIO

Supervisor 2, D.Sc. – UNIRIO

Supervisor 3, D.Sc. – XXXX

Rio de Janeiro, RJ – Brasil

Dezembro, 2024

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Resumo

Em uma parceria entre Seagate e a International Data Corporation (IDC) foi realizado o estudo “The Digitization of the World From Edge to Core”, nele a IDC fala sobre diversos aspectos referentes aos dados presentes no mundo digital e um dos tópicos abordados no estudo é “Mankind is on a quest to digitize the world” e neste mesmo tópico eles explicam que os dados que geramos no dia a dia está em constante crescimento, ou seja, estamos gradualmente produzindo mais dados.

Com um volume cada vez maior de dados, uma busca por informação otimizada é essencial, dado que são necessárias ferramentas que nos garantam confiança e precisão da informação adquirida. Com isso em mente, este trabalho visa o desenvolvimento de um sistema capaz de ler, processar e armazenar documentos diversos de determinada biblioteca (conjunto de documentos) para que possamos utilizar um Large Language Model (LLM) para responder perguntas que os usuários possam ter acerca dos documentos.

A ideia é conseguir processar documentos de diferentes épocas, temas, formatos e conseguir responder o maior número possível de perguntas dos usuários com a melhor confiança possível.

Palavras-chave: retrieval, augmented, generation, inteligência, artificial.

Abstract

In a partnership between Seagate and the International Data Corporation (IDC), the study "The Digitization of the World From Edge to Core" was conducted. In it, IDC discusses various aspects related to data present in the digital world and one of the topics covered in the study is "Mankind is on a quest to digitize the world". In this same topic, they explain that the data we generate on a daily basis is constantly growing, that is, we are gradually producing more data.

With an ever-increasing volume of data, an optimized search for information is essential, given that tools are needed that guarantee reliability and accuracy of the information acquired. With this in mind, this work aims to develop a system capable of reading, processing and storing various documents from a given library (set of documents) so that we can use a Large Language Model (LLM) to answer questions that users may have about the documents.

The idea is to be able to process documents from different periods, themes and formats and to be able to answer as many user questions as possible with the greatest possible confidence.

Keywords: retrieval, augmented, generation, artifical, inteligence.

Conteúdo

Lista de Figuras

Lista de Tabelas

List of Algorithms

1 Introdução

1.1 Motivação

Recentemente precisei implementar uma solução de onboarding para funcionários de uma determinada empresa. Por ser uma empresa grande, diversas regras e normas estavam distribuidas em inúmeros documentos (documentos de formatos distintos e sem uma padronização específica), o que gerava uma dor para funcionários recém contratados, que nem sempre sabiam qual documento consultar.

Para implementar esta solução, foi escolhida uma abordagem que utilizasse o conceito de Retrieval Augmented Generation (RAG) para trazer aos funcionários a informação buscada no menor tempo possível e com confiança de que a informação é válida. Ao longo da implementação da solução, lembrei de todas as vezes que precisei ler um artigo, livro ou até slides só por conta de um determinado tópico ou assunto e o tempo que gastei procurando uma informação que nem sempre seria útil, seja por falta de referência ou a própria informação.

Com isso em mente, busquei implementar um projeto que seria capaz de processar uma biblioteca de documentos e implementar o conceito de RAG para que um usuário seja capaz de fazer perguntas em uma interface de chat simples e com base nos dados processados da biblioteca de documentos, um Large Language Model (LLM) irá retornar uma resposta humanizada contendo uma resposta e a referência, de qual documento veio a resposta, ao usuário.

O tema de busca por informação é importante, pois na internet ainda encontramos diversos dados sem referência ou representados de maneiras distintas (como em um gráfico e em um texto descritivo, ambos com a mesma informação), alguns dados também podem estar alocados em aplicações pouco intuitivas o que acaba aumentando o tempo de busca por informação, seja um site difícil de navegar ou um portal com o mecanismo de busca ruim. Todo esse tempo investido na busca por uma informação, nem sempre é vantajoso para estudantes e pesquisadores, o que pode dificultar pesquisas e trabalhos a longo prazo.

1.2 Objetivos

O principal objetivo deste trabalho é provar que é possível implementar uma solução RAG para bibliotecas de documentos específicos, neste caso os Trabalhos de Conclusão de Curso (TCCs), disponíveis na biblioteca de publicações do curso de Bacharelado em Sistemas de Informação (BSI) da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) [tccs_unirio<empty citation>].

Mais especificamente, estarei usando como biblioteca todos os TCCs de 2023 disponíveis na biblioteca do curso. Embora sejam do mesmo curso e ano, nem todos os trabalhos possuem a mesma estrutura, os alunos são livres para escrever seus trabalhos de diversas formas e não há como garantir uma padronização comum à eles, tornando o processamento dos documentos mais complexo. No entanto, todos os documentos disponíveis na bibliotecas são do tipo Portable Document Format (PDF) com extensão .pdf, isso vai contribuir durante a etapa de implementação dos extratores mais a frente no projeto.

1.3 Metodologia

Neste projeto irei utilizar uma abordagem de Design Science Research (DSR) para que ao final do projeto o artefato modelado esteja implementado e funcionando como planejado.

O DSR tem suas raízes na engenharia e nas ciências do artificial [simon_1996<empty citation>
É uma metodologia fundamental para a resolução de problemas, buscando aprimorar o conhecimento humano com a criação de artefatos inovadores e a geração de conhecimento de design por meio de soluções para problemas do mundo real [design_science<empty citation>].

Deste modo, ao utilizar o DSR, ao final do projeto haverá um artefato que foi produzido com base na aplicação de tudo estudado e discutido nas próximas seções deste trabalho.

2 Fundamentação

2.1 Conceitos Fundamentais

No livro RAG-Driven Generative AI, do Denis Rothman, ele diz que "Mesmo o modelo mais avançado de Inteligência Artificial (IA) generativa é limitado a responder sobre dados nos quais ela foi treinada." rothman<empty citation> o que nos chama a atenção a um probelma em especial, como fazer para que uma IA saiba responder perguntas referentes a um conjunto específico de dados?

De fato, uma IA não tem como saber o que ela não sabe, não existe conhecimento além dos dados nos quais ela foi treinada. Perguntas fora do contexto do treinamento de uma IA geralmente leva a halucinações, viés ou pura besteira. Para isso, foi implementado um framework, ou estrutura, que combina abordagens baseadas em recuperação com modelos generativos, esta estrutura é a Retrieval Augmented Generation (RAG).

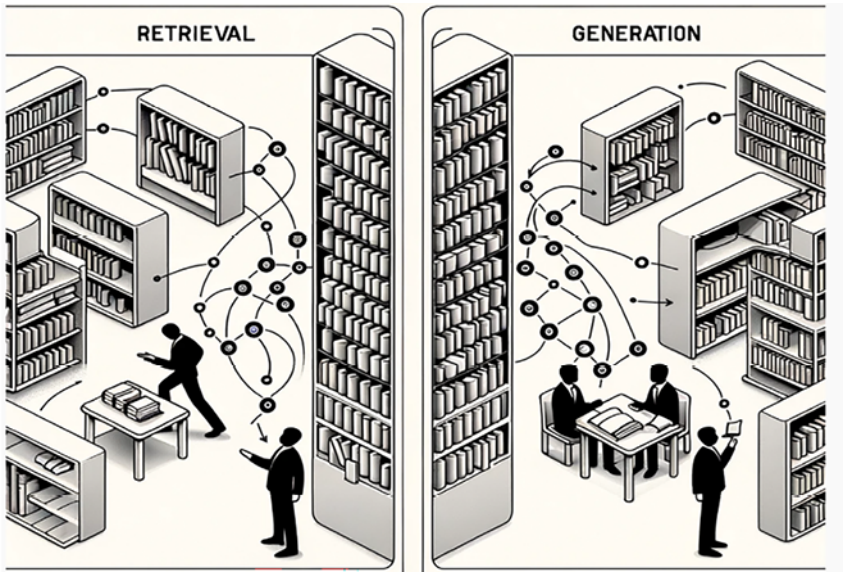
A RAG recupera dados relevantes de fontes externas em tempo real e usa esses dados para gerar respostas contextualmente relevantes. Uma de suas principais vantagens é a adaptabiolidade, tendo em vista que a estrutura pode ser aplicada independente do tipo de dado abordado na solução, seja texto, imagens, áudios ou documentos diversos.

2.2 RAG

Quando um modelo de IA generativa não sabe responder determinada pergunta com precisão, diz-se que ele está alucinando ou apresentando viés, mas, na prática, está apenas gerando respostas sem sentido. Isso ocorre porque o modelo não foi treinado com as informações solicitadas ou por conta de limitações em sua configuração, resultando em sequências prováveis, mas não precisas necessariamente.

A RAG começa onde a IA generativa termina, fornecendo informações que um modelo de LLM não possui para responder com precisão. A RAG otimiza tarefas de recuperação de informações e adiciona os dados recuperados durante a entrada (seja consulta do usuário ou um prompt automatizado), gerando uma saída melhorada e

mais amigável ao usuário. O funcionamento geral do RAG pode ser resumido na figura a seguir:



3 Modelagem

3.1 Modelo Conceitual

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

3.2 Tecnologias

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus.

Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

4 Solução Desenvolvida

4.1 Visão Geral

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

4.2 Funcionalidades

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus.

Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

5 Conclusão

5.1 Considerações Finais

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5.2 Limitações

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus.

Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

5.3 Trabalhos Futuros

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Referências

Instruções de bibliografia a seguir foram retiradas do manual de referência da Sociedade Brasileira de Computação [sbc]:

As referências bibliográficas devem ser de entendimento único e uniformes. Nós recomendamos dar ao autor nomes de referências em colchete, e.g. [knuth], [smith]; ou datas nos parênteses, knuth<empty citation>, smith<empty citation>.

As referências devem ser listadas usando o tamanho de fonte de 12 pontos, com 6 pontos do espaço antes de cada referência. A primeira linha de cada referência não deve ser recuada, quando a subseqüente dever ser recuada 0.5 cm.