Relatório 6 - Prática: Embedding (II)

Guilherme Mileib

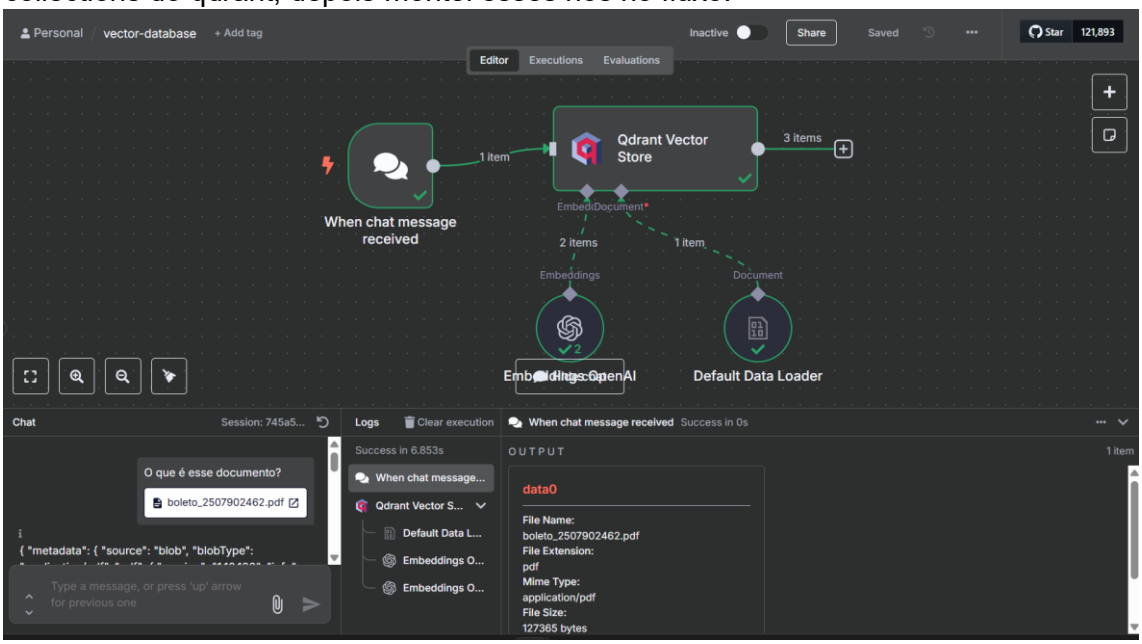
Descrição da atividade

Durante essa atividade fizemos um fluxo no n8n com o ‘qdrant’, que é um banco de dados vetorial. Ele funciona de forma diferente dos bancos de dados convencionais, o qual armazena os dados estruturados e como uma busca literal. Enquanto os bancos de dados vetoriais não entendem texto ou imagem, pois convertem para uma representação numérica chamada vetor (*embedding*). Esse vetor captura o significado semântico e o contexto do dado original.

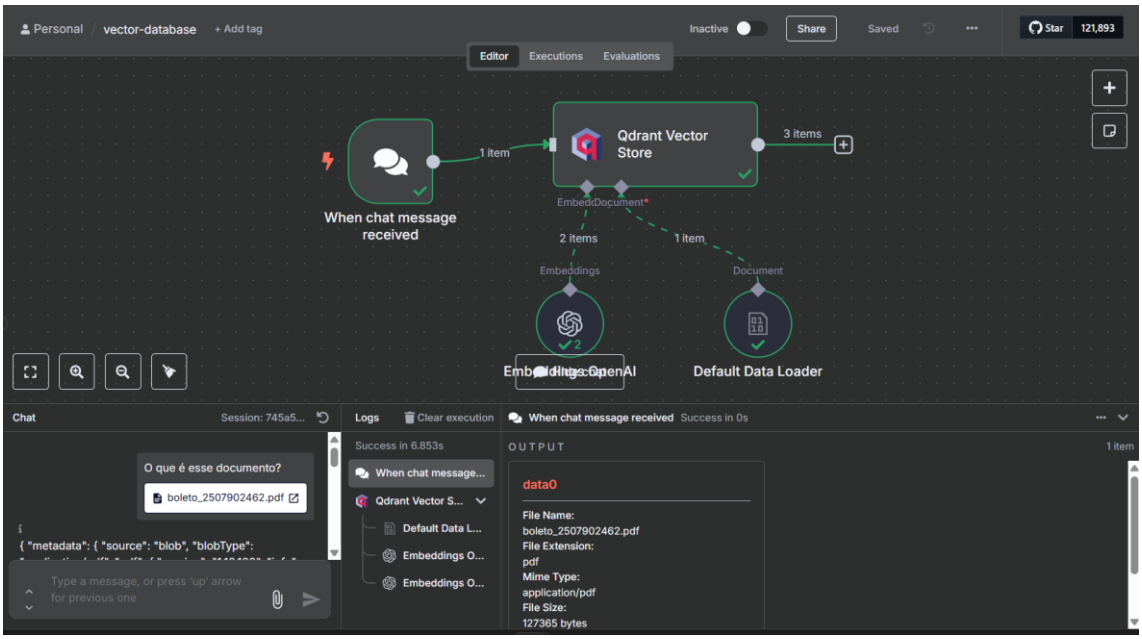
A função do baco de dados é uma ferramenta otimizada para encontrar os vetores mais próximos ou semelhantes a um vetor de consulta.

Sobre a parte prática

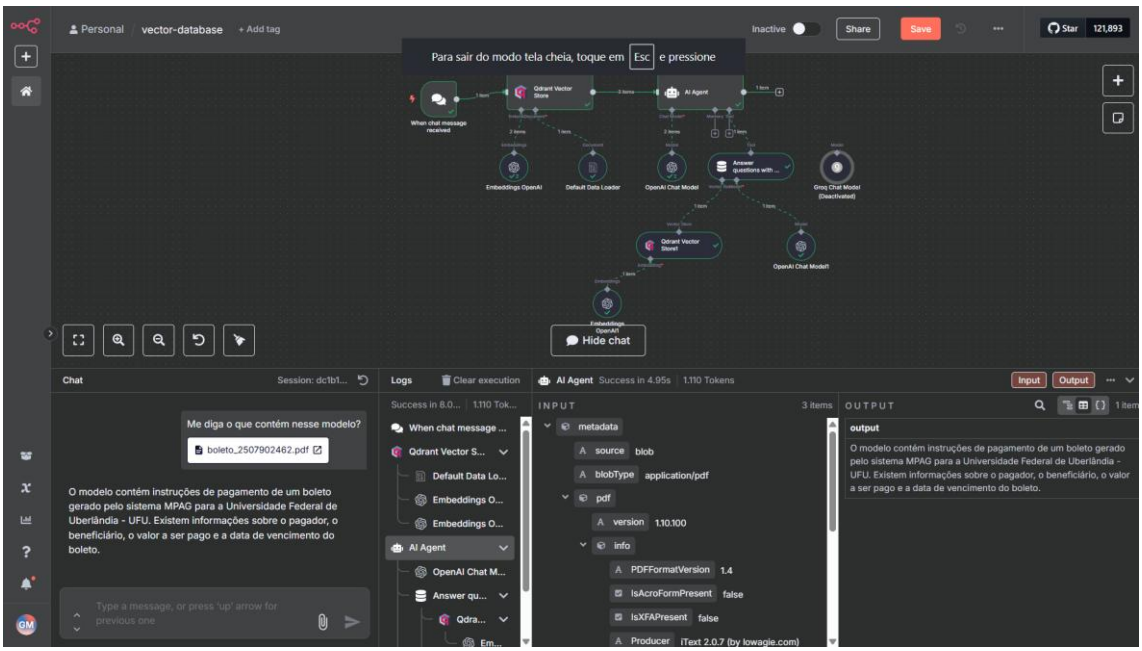
1. Ajustando o qdrant (database vectorial) – criei a Key para ter acesso as collections do qdrant, depois montei esses nós no fluxo:



2. Configurando o input do AI Agent



3. Após configurar o segundo vetor de armazenamento – terminando o workflow:



Esse workflow foi um exemplo de RAG, com buscas de informações a partir de uma base de dados (documentos que eram inseridos no Qdrant). A LLM realiza a busca por trechos no documento que sejam relevantes para a pergunta feita, usando esses trechos para 'contextualizar', formular uma resposta inteligente e precisa.

Dificuldades

Minhas dificuldades foram algumas acerca das diferenças dos nomes apresentados nas aulas devido a versão ser diferente do n8n. Porém, com algumas pesquisas consegui selecionar os 'nodes' corretos. E tudo funcionou muito bem.

Conclusões

Com os conceitos introduzidos nas aulas pude ter contato através de uma pesquisa individual os significados tecnologias que não havia visto antes. Tal como, a criação de um cluster, o qual já havia tido contato com seu significado em matérias da faculdade. O cluster auxilia na escalabilidade e disponibilidade de um sistema, por ser algo flexível, permite que os múltiplos servidores (nós), que operam em um sistema único. Por isso, traz a possibilidade de aumentar distribuição de dados caso haja necessidade, apenas introduzindo um nó ao cluster. Permitindo que o sistema consiga lidar com um volume maior de requisições simultâneas. Ou até mesmo se o sistema cair outros nós podem continuar atendendo a busca.

Referências

- https://www.youtube.com/watch?v=SqQe5jgdg8s (Vídeo de configuração do workflow).
- https://qdrant.tech/documentation/quickstart/ (Documentação Qdrant Local)
- https://www.youtube.com/watch?v=CFBV7gnW0BY&t=254s (Baixando os modelos Ollama)