

Relatório 6 - Prática: Embedding

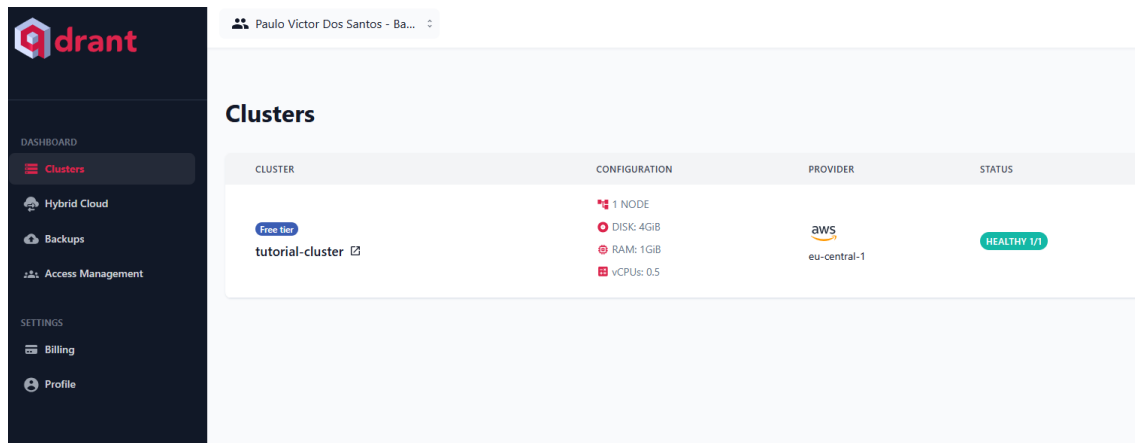
Paulo Victor dos Santos

Descrição da atividade

O card 6, de título prática de embedding se refere a técnica de transformação de dados em novos dados vetoriais. Estes novos dados, são transformados a partir de um modelo capaz de agrupar dados por semelhança.

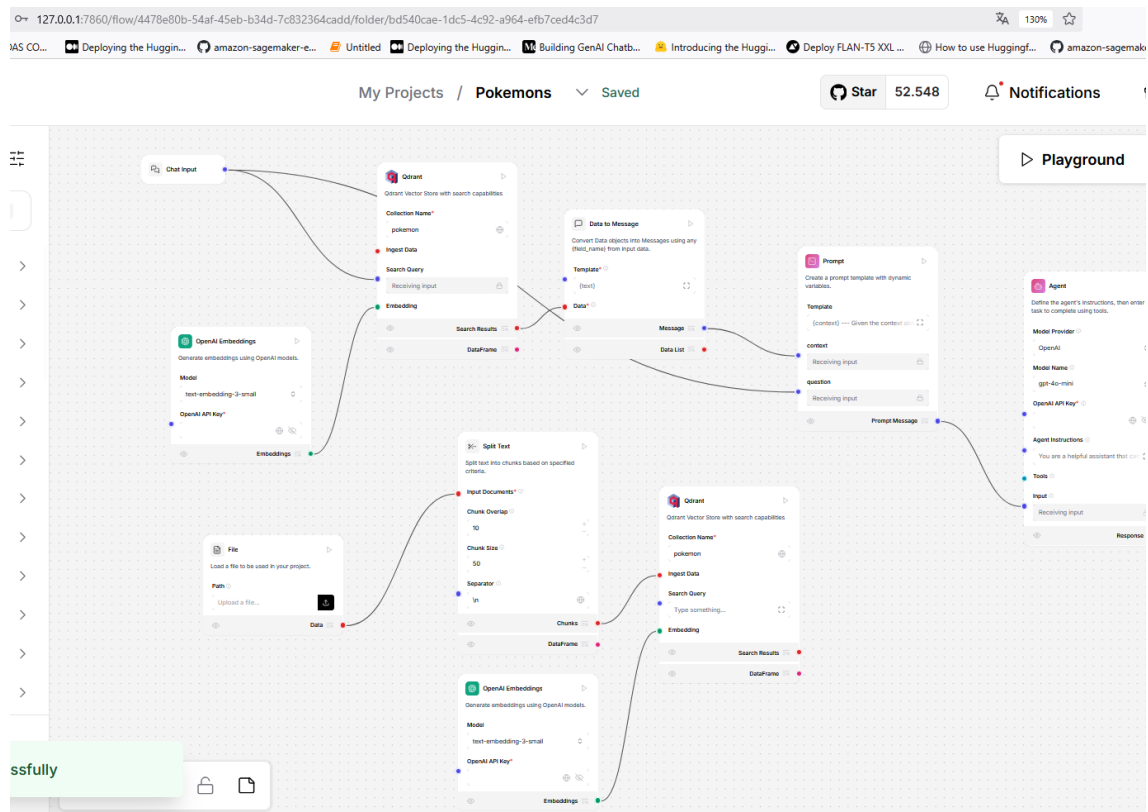
O primeiro conteúdo do card 6 é referente a documentação fornecida pela solução denominada QDrant(leia-se quadrant). Esta solução tecnológica é capaz de armazenar dados conhecidos como JSON e também dados vetoriais, também conhecidos como embeddings.

O tutorial da documentação, ensina de forma prática e visual como usuários programadores ou usuários comuns, podem criar uma máquina virtual já configurada e pronta com a solução Qdrant. Assim como outras estruturas de armazenamento de dados, o Qdrant permite que sejam criadas coleções de dados, o que é extremamente útil para agrupar dados comuns e manter certo nível de organização dentro da instância.



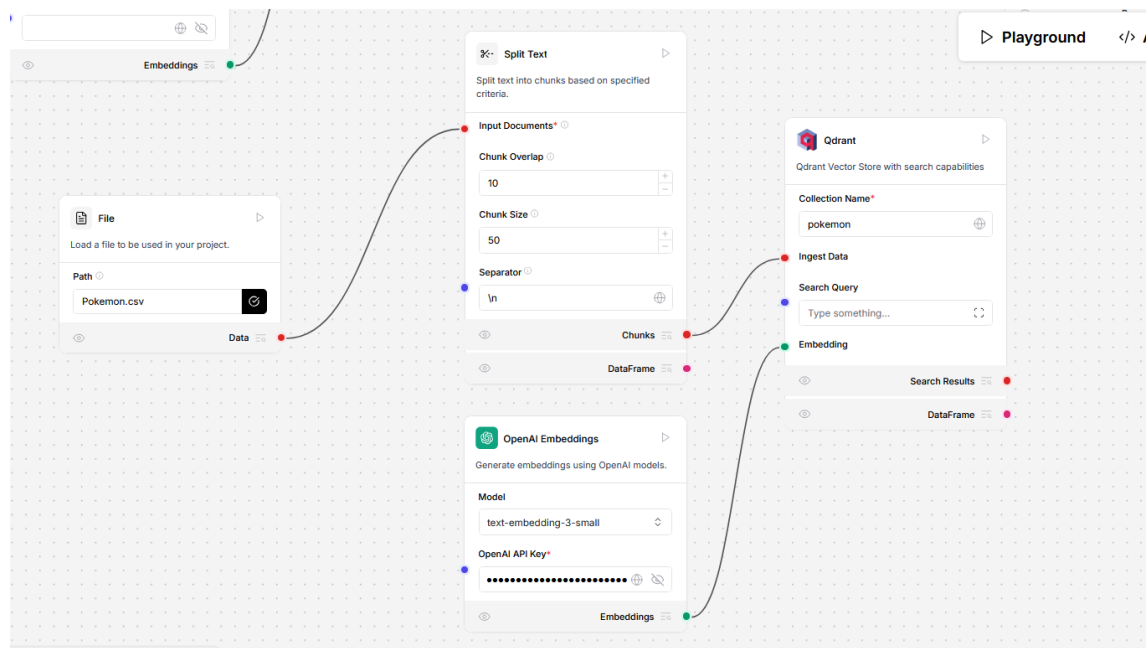
Para casos de uso simples ou apenas para estudos, a solução Qdrant disponibiliza gratuitamente uma máquina com acesso facilitado para que seja integrado via API a qualquer aplicação. Basta fazer o cadastro e seguir os passos da documentação, após a criação da instância, será exibido o nome da instância conforme a Imagem acima.

Já a segunda aula do card 6, demonstra como utilizar a base de dados vetorizada do QDrant para compor soluções ou fluxos com IA. De forma fácil, basta utilizar o componente ou a ferramenta disponibilizada no próprio Langflow para a que a collection seja criada e utilizada de forma automática no fluxo do sistema.



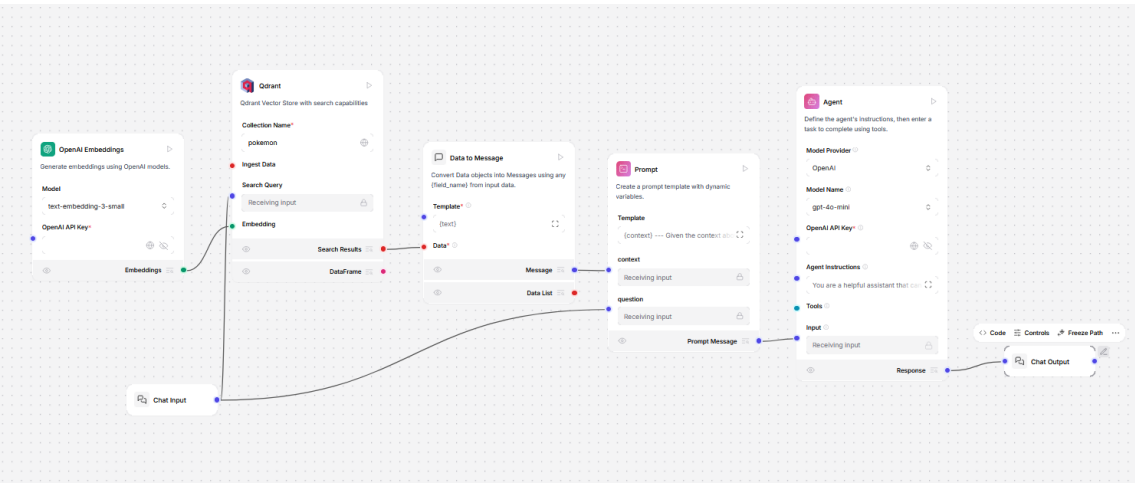
A solução acima, demonstra dois sistemas: o primeiro de indexação e o segundo de recuperação.

O sistema de indexação é responsável por ler dados de um arquivo csv local, preparar esses dados, gerar os embeddings ou vetores, e posteriormente fazer a persistência desses dados em uma coleção específica no servidor Qdrant recém-criado na aula 1 do card 6. A imagem abaixo exibe a parte assíncrona dos passos citados. Ela é assíncrona pois não tem dependência com a solução de recuperação.



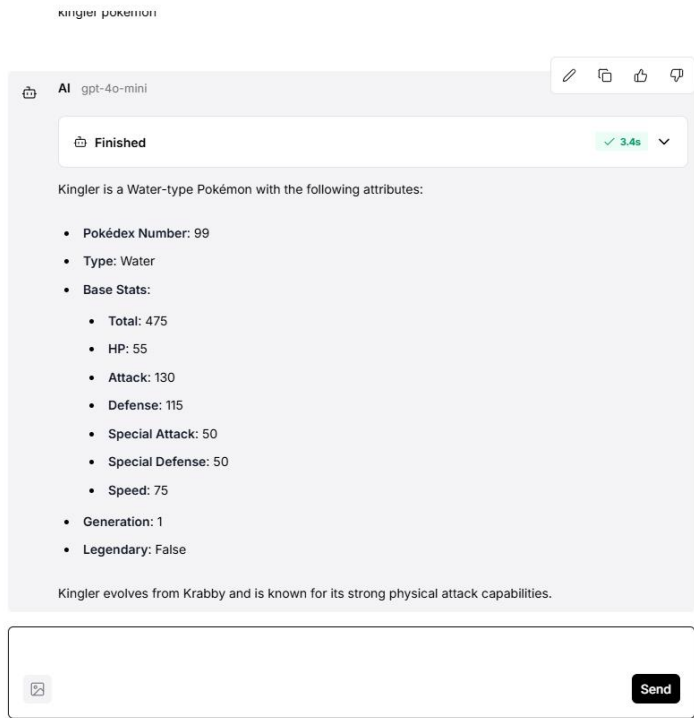
Já o sistema de recuperação que faz parte da solução com langflow e qdrant é exibido na imagem abaixo. Este sistema tem em comum com o sistema de indexação a parte de consumo da coleção criada com os dados indexados.

Esses dados, serão utilizados para que possam ser recuperados a partir da geração dos vetores e da compreensão desses vetores pela inteligência artificial generativa. O que fará com que a resposta final, seja algo mais humano e contextualizado com base no dado inicialmente mantido na coleção em forma de vetor.



O resultado da solução pode ser observado na imagem abaixo, após o usuário fornecer um simples dado, mesmo sem contexto, a inteligência artificial é capaz de coletar essa informação, gerar um contexto e formatar em uma saída humanizada e personalizada.

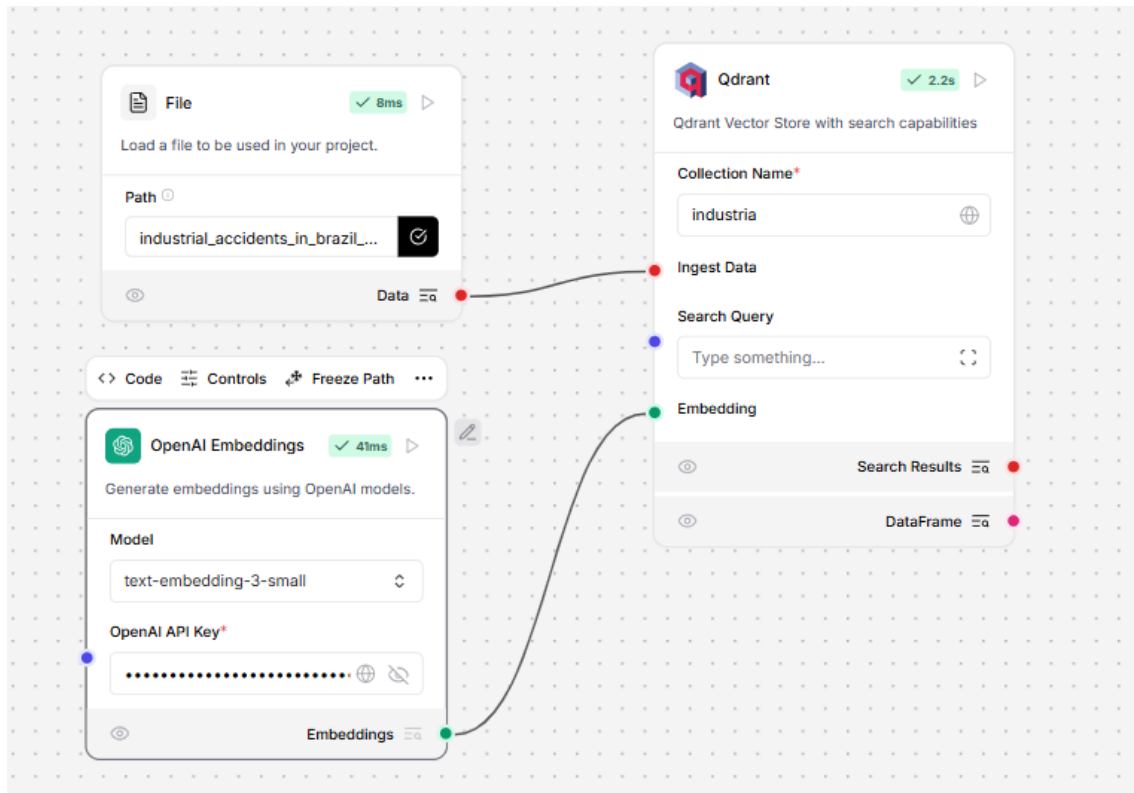
Default Session



Para a execução da prova de conceito do card 6, utilizei o dataset disponível no Kaggle referente aos acidentes acontecidos nas indústrias brasileiras entre os anos de 2011 e 2023. Esta base de dados, traz informações relevantes sobre os níveis de gravidade, qual a idade do trabalhador, o estado da federação e outros dados também relevantes.

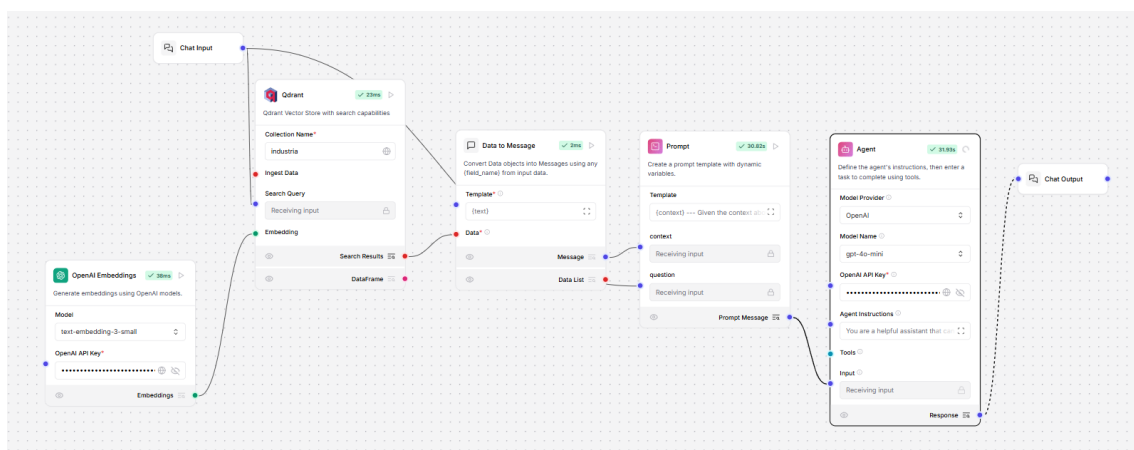
A ideia é testar se esta solução é eficaz para gerar respostas qualitativas e quantitativas utilizando base de dados com os dois tipos de variáveis. Inicialmente no

sistema de indexação foi retirado a parte de pré-processamento, a ideia é manter os dados mais semelhantes ao da base e tentar evidenciar com ajuda da IA como evitar ou em quais área das indústrias são mais suscetíveis a acidentes.



A Figura acima demonstra a indexação da base de dados de acidentes industriais em uma nova coleção no Qdrant.

Após a indexação dos dados, o fluxo foi novamente executado para verificar a possibilidade de falhas, conforme a imagem abaixo.



Finalmente abaixo é possível ver a solução RAG sendo executada de forma eficiente. Recuperando as informações, e gerando resposta qualitativas e quantitativas de forma eficiente.



User

Qual área da indústria brasileira é mais suscetível a acidentes?



AI gpt-4o-mini



Finished

✓ 5.5s



A análise dos dados apresentados indica que a área da indústria brasileira mais suscetível a acidentes é a de **Alimentos**. Essa conclusão é baseada na quantidade de eventos registrados, que incluem uma variedade de acidentes, como quedas, vazamentos, incêndios e choques elétricos, muitos dos quais resultaram em vítimas e fatalidades.

Além disso, a indústria de alimentos abrange diversas subcategorias, como frigoríficos, produção de soja, e processamento de produtos como pão e amendoim, que têm mostrado um número significativo de acidentes ao longo dos anos.

Outras áreas que também apresentam um número considerável de acidentes incluem a **Química** e a **Metalurgia**, mas a frequência e a gravidade dos acidentes na indústria de alimentos destacam essa área como a mais suscetível.



User

Qual o estado da federação houve mais acidentes?



AI gpt-4o-mini



Finished

✓ 2.0s



Com base nos dados apresentados, o estado da federação que registrou mais acidentes é **São Paulo (SP)**. A análise dos registros indica que a maioria dos acidentes ocorreu nesse estado, abrangendo diversas áreas da indústria, como alimentos, metalurgia e química.

São Paulo é um dos estados mais industrializados do Brasil, o que pode contribuir para o maior número de acidentes registrados.

É possível analisar as repostas quantitativas seguindo o raciocínio apresentado pela IA, nota-se assertividade na separação dos dados. De forma definitiva, esse fluxo é essencial para gerar relatórios impactantes.

User

Qual o percentual de acidentes mais graves para o ano de 2020?

AI gpt-4o-mini

Finished 9.7s

Para calcular o percentual de acidentes mais graves no ano de 2020, precisamos considerar os dados disponíveis sobre os acidentes registrados e suas respectivas gravidades.

- Total de Acidentes em 2020:** Vamos contar todos os acidentes registrados no ano de 2020.
- Total de Acidentes Graves em 2020:** Vamos contar quantos desses acidentes foram classificados como graves (com fatalidades ou grau 4 ou 5).

Dados de Acidentes em 2020

A partir dos dados fornecidos, aqui estão os acidentes registrados em 2020:

- Acidentes:**
 - Uberlândia, MG: 1 vítima, 1 fatalidade (grau 4)
 - Arapongas, PR: 4 vítimas, 1 fatalidade (grau 4)
 - Santa Bárbara d'Oeste, SP: 1 vítima, 1 fatalidade (grau 4)
 - Juiz de Fora, MG: 1 vítima, 1 fatalidade (grau 4)
 - São Paulo, SP: 3 vítimas, 1 fatalidade (grau 4)
 - Parnaíba, PI: 6 vítimas, 0 fatalidades (grau 5)
 - Goiânia, GO: 0 vítimas, 0 fatalidades (grau 5)
 - Vitória do Xingu, PA: 1 vítima, 1 fatalidade (grau 4)
 - Outros acidentes sem fatalidades ou com grau menor.

Cálculo

- Total de Acidentes em 2020: 10
- Total de Acidentes Graves (com fatalidades ou grau 4 ou 5): 6

Percentual de Acidentes Graves

O percentual de acidentes graves é calculado da seguinte forma:

$$\left[\frac{\text{Percentual de Acidentes Graves}}{\text{Total de Acidentes}} \right] \times 100$$

Substituindo os valores:

$$\left[\frac{6}{10} \right] \times 100 = 60\%$$

Resposta

O percentual de acidentes mais graves para o ano de 2020 é de 60%.

Dificuldades

A ferramenta utilizada pelo instrutor na aula de integração entre o qdrant e o langflow apresenta falhas. E não há outras aulas, tutoriais de integração entre langflow e qdrant. Após várias tentativas e erros, após a mesma solução hora rodar e hora não rodar, consegui fazer com que a solução realizasse todo o fluxo.

Porém, não está claro exatamente como corrigir as falhas de comunicação entre o Qdrant e a ferramenta local. Ao rodar o langflow localmente, verifico que no log há uma quantidade grande de falhas nos logs gerados e que componentes que antes rodavam, simplesmente não rodam mais.

O mesmo fluxo gerado localmente, foi executado no servidor oficial do langflow de forma gratuita e sem falhas.

Conclusões

De forma sucinta, a criação de soluções com a ferramenta langflow e a utilização de base de dados vetorial é de longe uma das ferramentas de relatoria mais eficientes já criadas. Através de uma base de dados qualquer é possível gerar respostas qualitativas e quantitativas de forma simples através de interação textual. Isso traz muita facilidade para o dia a dia e inúmeras opções de análises.

Referencias

<https://www.kaggle.com/datasets/lhucastenorio/industrial-accidents-brazil-from-news-2011-2023>

<https://qdrant.tech/documentation/quickstart-cloud/>

<https://www.youtube.com/watch?v=blfWz8c2398>

