Universidade de São Paulo – Instituto de Matemática e Estatística

Projeto de TCC (MAC0499) – 2025

Agente Conversacional para Transformação de Texto em SQL

Aluno: Eduardo Figueredo Pacheco
Orientador: Prof. Fabio Cozman

Resumo

Este projeto visa desenvolver um agente conversacional que transforme perguntas em linguagem natural sobre produtos em consultas SQL em um banco de dados. O objetivo principal é permitir que empresas interajam com seu banco de dados de forma simples e customizável, facilitando a extração de informações relevantes. Com o crescente interesse em soluções de inteligência artificial, este projeto explora a implementação de um agente que não apenas atende a consultas, mas também é adaptável às necessidades específicas de diferentes setores de mercado.

1 Introdução

A utilização de agentes conversacionais é uma tendência crescente no ambiente empresarial, proporcionando interações mais dinâmicas e eficientes. Este projeto busca aproveitar essa tendência ao desenvolver um agente que se comunica diretamente com bancos de dados por meio de consultas SQL. Essa comunicação permite que usuários façam perguntas em linguagem natural, e o agente as traduza para consultas SQL, facilitando a extração de informações relevantes e tornando o acesso aos dados mais intuitivo. A customização do agente é um aspecto central, permitindo que as empresas moldem o sistema de acordo com suas necessidades específicas.

1.1 Agentes Conversacionais

Os agentes conversacionais são sistemas que utilizam inteligência artificial para permitir interações em linguagem natural. Esses agentes têm se tornado ferramentas essenciais para empresas que desejam otimizar o atendimento e facilitar a análise de dados.

2 Ferramentas e Métodos

A linguagem principal para a realização deste projeto é Python, uma escolha estratégica por sua capacidade de integrar diferentes ferramentas de inteligência artificial e Modelos de Linguagem de Grande Escala (LLMs). Essa versatilidade permite que diversas tecnologias sejam acopladas conforme necessário. A seguir, serão listadas as prováveis e principais ferramentas que se pretendem utilizar para a realização do projeto. Entre as principais ferramentas estão o LangChain, o LlamaIndex e o MCP.

2.1 RAG (Retrieval-Augmented Generation)

O RAG é uma técnica que combina geração de texto e recuperação de informações, permitindo que os agentes forneçam respostas mais precisas ao consultar bases de dados externas. Essa abordagem garante que o agente tenha acesso a informações atualizadas e relevantes.

2.2 RAG + LangChain

A integração do RAG com LangChain proporciona uma estrutura robusta para o desenvolvimento de agentes conversacionais. LangChain facilita a

criação de fluxos de conversação, enquanto o RAG melhora a precisão das respostas ao buscar informações relevantes.

2.3 LlamaIndex

O LlamaIndex é uma ferramenta que permite a indexação eficiente de dados, facilitando a consulta e recuperação de informações contextuais em tempo real. Isso é particularmente útil para agentes que precisam lidar com grandes volumes de dados.

2.4 MCP (Model Context Protocol)

O MCP é um framework que permite a personalização de interações de agentes conversacionais, adaptando-se às necessidades específicas de diferentes setores. Isso proporciona flexibilidade e versatilidade na construção do agente.

2.5 LLMs (Modelos de Linguagem de Grande Escala)

Os Modelos de Linguagem de Grande Escala (LLMs) são componentes centrais para a geração de linguagem natural neste projeto. Há diversas abordagens disponíveis, que variam principalmente quanto à forma de acesso, nível de controle e desempenho oferecido.

Modelos como o **LLaMA 3**, de código aberto, oferecem a possibilidade de execução local, o que garante maior privacidade, customização e independência de serviços externos. Essa abordagem é especialmente interessante em contextos onde há preocupação com o uso e armazenamento dos dados. No entanto, ela exige infraestrutura computacional adequada e maior responsabilidade na configuração e manutenção do modelo.

Por outro lado, soluções como o **GPT-4 (OpenAI)** e o **Gemini (Google)** são acessadas como serviços prontos para uso. Essas ferramentas geralmente apresentam desempenho superior em tarefas complexas, maior estabilidade e funcionalidades avançadas, como suporte multimodal. Em contrapartida, envolvem custos por requisição e menor flexibilidade quanto à personalização ou controle sobre os dados.

A escolha do LLM mais adequado depende do contexto de aplicação. Modelos open source tendem a ser mais atrativos para pesquisa e prototipagem, enquanto os modelos oferecidos como serviço são vantajosos para aplicações em escala, com foco em robustez e facilidade de integração.

2.6 Outras Tecnologias

Além das ferramentas mencionadas, é importante considerar a possibilidade de incorporar outras tecnologias que possam facilitar ou melhorar o desempenho do projeto ao longo do desenvolvimento. Novas soluções de IA podem surgir e se adaptar às demandas do projeto.

3 Cronograma

- **Planejamento do projeto:** Esta atividade envolve a definição dos objetivos, escopo e metodologias que serão utilizadas ao longo do desenvolvimento do projeto.
- **Pesquisa Inicial:** Nesta fase, será realizada uma investigação bibliográfica para entender o estado da arte e as tecnologias pertinentes ao desenvolvimento do agente conversacional.
- **Construção do Agente:** Implementação e desenvolvimento do agente conversacional, estabelecendo sua arquitetura e funcionalidades para realizar consultas SQL.
- **Criação de Testes:** Desenvolvimento de diferentes cenários de testes para validar o funcionamento do agente e mensurar sua capacidade de responder perguntas com precisão, também visando obter dados para a seção de resultados do Trabalho de Conclusão de Curso.
- **Testes com Empresa:** O agente será testado em um ambiente real, interagindo com uma empresa parceira para avaliar sua eficácia e realizar ajustes necessários.
- **Relatório Final:** Elaboração de um documento final que compile todos os resultados obtidos, análises realizadas e conclusões do projeto.

Atividade	Abril	Maio	Jun.	Jul.	Ago.	Set.	Out.	Nov.
Planejamento do projeto	X							
Pesquisa Inicial	X	X						
Construção do Agente		X	X	X				
Criação de Testes					X	X		
Testes com Empresa						X	X	
Relatório Final						X	X	X

Table 1: Cronograma das Atividades

4 Referências

- LangChain Documentation. "An Overview of LangChain". *LangChain Documentation*. Disponível em: `https://python.langchain.com/docs/introduction/`. Acesso em: 20 out. 2023.
- LlamaIndex Documentation. "LlamaIndex Overview". *LlamaIndex Documentation*. Disponível em: `https://docs.llamaindex.ai/en/stable/DOCS_README/`. Acesso em: 20 out. 2023.
- Anthropic. "MCP Documentation". *Anthropic Documentation*. Disponível em: `https://docs.anthropic.com/en/docs/agents-and-tools/mcp`. Acesso em: 20 out. 2023.