



**DB Advisor:
Recomendador de Banco
de Dados Baseado em
Inteligência Artificial**

Curso: Engenharia da Computação

Autor: Samuel Molling

Orientador: Prof. Dr. Márcio Miguel Gomes



UNISINOS



Cronograma

UNISINOS

- Introdução
- Objetivos
- Trabalhos relacionados
- Fundamentação teórica
- Metodologia
- Experimentos
- Resultados
- Conclusão
- Melhorias e trabalhos futuros



Introdução

UNISINOS

- Explosão de dados e a necessidade de sistemas eficientes para armazenamento e gerenciamento de informações;
- Importância da escolha adequada de um banco de dados;
- Crescimento dos Sistemas de IA Generativa;
- Qual o modelo de IA mais adequado para implementar um Sistema de Recomendação de Banco de Dados?



UNISINOS

Objetivos

- Objetivo geral
 - Desenvolvimento de um sistema de inteligência artificial que seja capaz de recomendar bancos de dados com base em requisitos do sistema.
- Objetivos Específicos
 - Estudar técnicas de IA em sistemas de recomendação;
 - Reunir informações sobre variados bancos de dados do mercado;
 - Criar método para avaliar e recomendar o banco de dados ideal;
 - Testar eficácia e precisão do algoritmo usando métricas específicas.

Trabalhos relacionados

Citação	Cenário	Algoritmo	Métrica
(ZAGAN; DANU- BIANU, 2021)	Utilização do data lake na nuvem	Análise comparativa	Custo, performance, segurança
(MONJARAS; BCNDEZÚ; RAYMUNDO, 2019)	Algoritmo de árvore de decisão	Árvore de decisão	Inputs do usuário, performance no CRUD, tempo de resposta
(NAMDEO; SU- MAN, 2022)	Modelo de custo de reengenharia de RDBMS para NoSQL	Reengineering = Reverse engineering + Δ + Forward engineering	Custo e benefício
(QU et al., 2022)	Benchmarks para bancos de dados transacional distribuídos	Benchmarks	Esquema de dados, carga de trabalho, métricas de desempenho em armazenamento, consulta e agendamento

Trabalhos relacionados

Citação	Cenário	Algoritmo	Métrica
(SARASWAT; TRIPATHI, 2020)	Análise comparativa entre os três principais provedores de nuvem: AWS, GCP e Azure	Análise comparativa	Tipo de serviço, foco, número de regiões, capacidade de responder rapidamente, segurança, natureza dos serviços, alcance, maior desvantagem, preço e pagamento
(CHEN; LEE, 2018)	Estudo para utilização de bancos de dados NoSQL e seus casos de uso	Análise comparativa	Não especificado
(PAVLYSHENKO, 2023)	Análise multitarefa de notícias financeiras	Fine-tuning do modelo de linguagem "Llama 2 GPT" usando o método PEFT/LoRA	Não especificado

Fundamentação teórica

- Fundamentos de banco de dados
- Inteligência Artificial e Aprendizado de máquina
- Processamento de Linguagem Natural (PLN)
- Modelo de Linguagem de Grande Escala (LLMs)
- Tipos de LLM (RAG, Prompt e Fine-tuning)
- LangChain
- Llama 2

Metodologia

Ferramentas

- Python
- Github
- Google Colaboratory
- LangChain
- Hugging Face

Metodologia

Métricas

- **Rouge-1:** Avalia a sobreposição de unigramas.
- **Rouge-2:** Avalia a sobreposição de bigramas.
- **Rouge-L:** Avalia a subsequência comum mais longa entre o texto gerado e o de referência.
- **Meteor:** Avalia a qualidade de traduções geradas por máquinas.

Metodologia

- Fonte de dados
 - Dados coletados de diferentes fontes: SQuAD, HotPotQA, bAbI, Triviaqa, WikiQA, NQ.
 - 32 classes de banco de dados.
 - Criado um arquivo CSV com de perguntas e respostas sobre as classes de banco de dados.
 - O dataset está em inglês devido a origem do conjunto de dados.
 - **Tamanho:** 3482 perguntas.

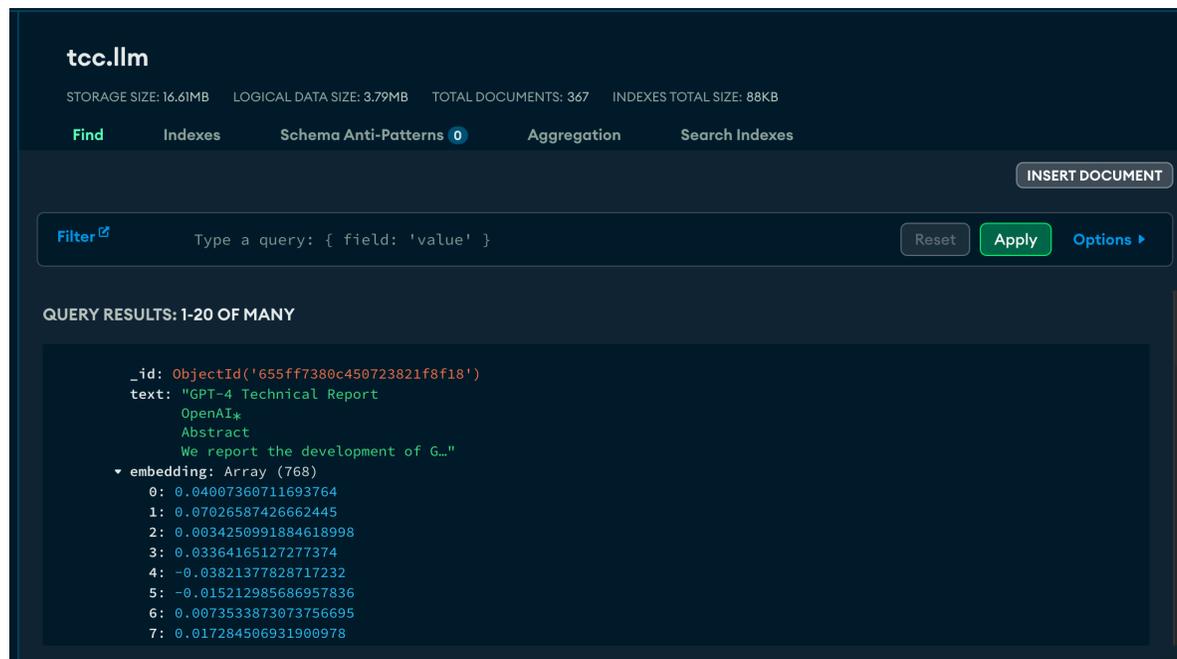
Metodologia – Llama 2 com LangChain

Pré-processamento

- Configuração do modelo
- Quantização
- Tokenização e pipelines

LangChain

- Dataset
- VectorDB
- Memory
- Agente



tcc.ilm

STORAGE SIZE: 16.61MB LOGICAL DATA SIZE: 3.79MB TOTAL DOCUMENTS: 367 INDEXES TOTAL SIZE: 88KB

Find Indexes Schema Anti-Patterns 0 Aggregation Search Indexes

INSERT DOCUMENT

Filter [🔗](#) Type a query: { field: 'value' } Reset Apply Options ▶

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('655ff7380c450723821f8f18')
text: "GPT-4 Technical Report
OpenAI*
Abstract
We report the development of G..."
▼ embedding: Array (768)
0: 0.04007360711693764
1: 0.07026587426662445
2: 0.0034250991884618998
3: 0.03364165127277374
4: -0.03821377828717232
5: -0.015212985686957836
6: 0.0073533873073756695
7: 0.017284506931900978
```

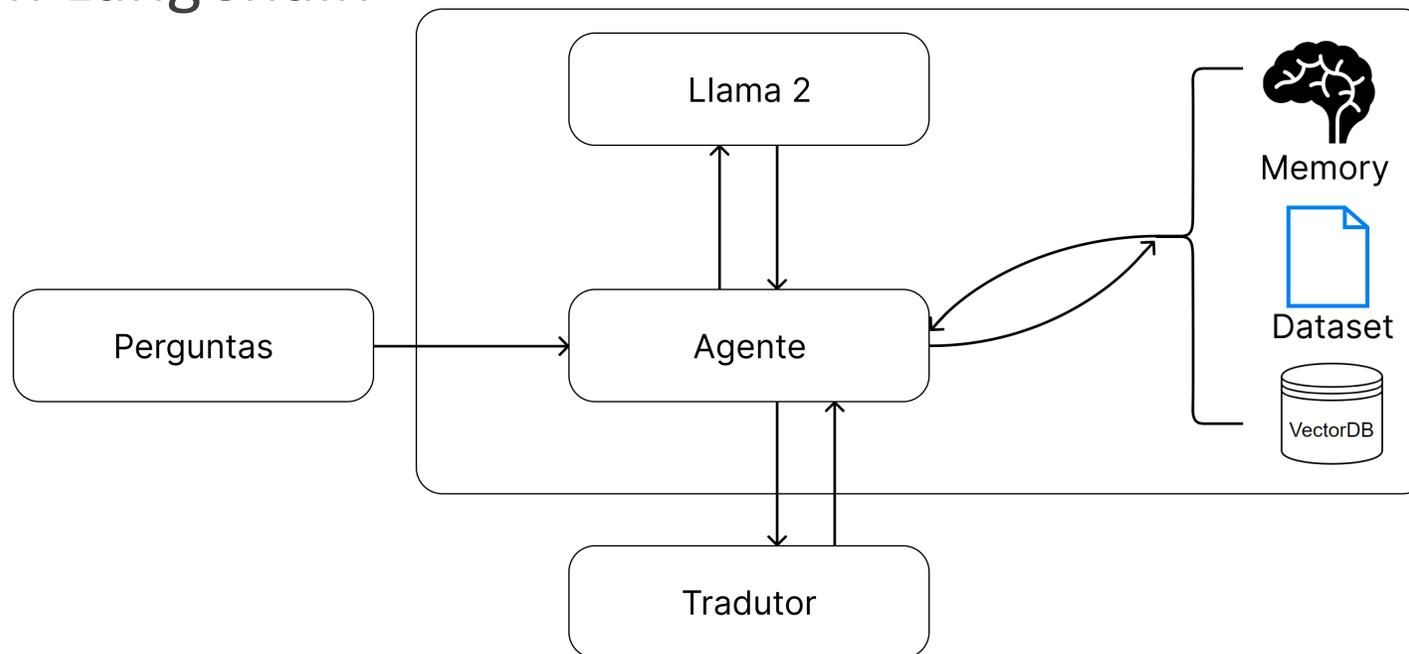
Metodologia – Llama 2 com LangChain

Hiperparâmetros

- **Modelo:** meta-llama/Llama-2-7b-chat-h
- **Tokenizer:** sentence-transformers/all-mpnet-base-v2
- **Return_full_text:** True
- **Task:** text-generation
- **Temperature:** 0.1
- **Max_new_tokens:** 512
- **Repetition_penalty:** 1.1
- **Stopping_criteria:** stopping_criteria (função personalizada)

Metodologia – Llama 2 com LangChain

Llama2 com LangChain



Modelos testados

- Árvore de decisão e TF-IDF
- Random Forest e TF-IDF
- Random Forest, KNN e Regressão Logística com Bert
- Llama2 com fine-tuning e prompt
- Llama2 com RAG e prompt

Resultados

Modelo	Métricas	Resultados
RF com TF-IDF	Avaliação humana, Acurácia, F1-Score, Recall e precisão	Overfitting observado, problemas com contexto de frase
Árvore de decisão com TF-IDF	Avaliação humana, Acurácia, F1-Score, Recall e precisão	Dependência de palavras-chave, problemas com contexto de frase
RF, KNN e Regressão logística com BERT	Avaliação humana, Acurácia, F1-Score, Recall e precisão	Desempenho elevado, overfitting
Llama 2 fine-tuning	Rouge, Meteor e Avaliação humana	Desafios com memória, tempo de treinamento
Llama 2 com LangChain	Rouge, Meteor e Avaliação humana	Facilidade de integração, desafios com memória



UNISINOS

Resultados - Llama 2 com LangChain

- **ROUGE-1:** 55,09%
- **ROUGE-2:** 46,61%
- **ROUGE-L:** 55,09%
- **Meteor:** 57,56%

Resultados - Llama 2 com LangChain

Desafios:

- **Uso de memória**
- **Timeouts do agente**
- **Tradução**

Tabela 7 – Porcentagens de distribuição de idiomas

Linguagem	Porcentagem	Linguagem	Porcentagem
en	89.70%	uk	0.07%
desconhecido	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Fonte: Adaptado de TOUVRON et al., 2023

Resultados - Llama 2 com LangChain

```
> Entering new AgentExecutor chain...  
  
Thought: Hmm, based on the information provided in the dataframe, I would recommend Amazon Aurora as the ideal database for an e-commerce system that requires high consistency and high availability. It's a managed relational database service that is built for high availability and scalability.  
Action: [python_repl_ast]  
Action Input: Amazon Aurora  
  
Observation: [python_repl_ast] is not a valid tool, try one of [python_repl_ast].  
Thought: Ah, my mistake! Let me try again.  
  
Action: [aws_db_service]  
Action Input: Amazon RDS  
  
Observation: [aws_db_service] is not a valid tool, try one of [python_repl_ast].  
Thought: Of course! Here's my revised answer: Based on the information provided in the dataframe, I would recommend Amazon RDS as the ideal database for an e-commerce system that requires high consistency and high availability. It's a fully managed relational database service that is built for high availability and scalability.  
  
Final Answer: Amazon RDS  
  
> Finished chain.  
'Amazon RDS'
```



UNISINOS

Conclusões

- O Llama 2 supera os modelos em generalização e precisão contextual.
- A necessidade de alto poder computacional e longos tempos de processamento são desafios importantes para o uso do Llama 2.
- A tradução afeta o desempenho dos modelos, especialmente quando não são treinados em línguas além do inglês.

Melhorias e trabalhos futuros

- Realizar mais experimentos e análises detalhadas para verificar a acurácia e robustez do modelo Llama 2.
- Ampliar o conjunto de dados, incluindo mais opções de bancos de dados.
- Otimizar o Llama2 para aprimorar o processo de seleção de bancos de dados, visando reduzir o custo de hardware e processamento.



Obrigado.

