Universidade Federal da Paraíba

Eduardo Henrique Pessoa Alves
Marcos André Bezerra da Silva

Relatório para Projeto Final da disciplina de Processamento de Linguagem Natural

Nome do projeto: Burpee

João Pessoa
2023

Apresentação do Projeto

O objetivo deste projeto é desenvolver um chatbot de aprendizado com foco em Programação Orientada a Objetos (POO) pelo motivo de termos incluído um livro sobre POO. O problema abordado é a interação com o chatbot para obter informações e esclarecer dúvidas sobre vários conceitos.

Objetivos

Os principais objetivos deste projeto são:

- Desenvolver um chatbot interativo.
- Utilizar o chatbot como ferramenta de aprendizado.
- Integrar o chatbot com um sistema de processamento de dados.

LLMs possuem grande capacidade de escrever explicações, analogias, fazer comparações, explicar o mesmo conceito de maneiras diferentes. Porém, o conhecimento deles é limitado aos dados de treinamento e sua utilidade diminui bastante ao tentar utilizá-los em domínios mais específicos ou mais atuais do que à época a qual o modelo foi treinado. Fazer fine-tuning nesses modelos grandes é impraticável e requer muitos dados e poder computacional.

Uma alternativa viável é simular uma base de dados auxiliar utilizando banco de dados vetoriais. Isso permite utilizar as capacidades dos LLMs com qualquer conjunto de dados ou conhecimento novo. Podendo ser potencialmente utilizado para ajudar no ensino de conteúdos específicos.

Dados Utilizados e Pré-processamento dos Dados

O projeto pode utilizar qualquer livro em pdf arbitrário. Para a demonstração, foram utilizados dados do livro de Programação Orientada a Objetos (POO). O pré-processamento dos dados envolveu a extração do texto do livro, utilizando uma biblioteca de processamento de PDF. Em seguida, o texto foi dividido em documentos menores para possibilitar o processamento. Cada pedaço possui 1000 caracteres, foi vetorizado e armazenado num banco de dados vetorial local.

O modelo de embeddings utilizado para vetorizar os documentos foi o INSTRUCTOR. Ele é um modelo para embeddings de propósito geral. Foi treinado com os dados do EDI (Multitask Embeddings Data with Instructions), uma coleção de 330 datasets. Os dados são conjuntos de perguntas e um conjunto rotulado de informações que são úteis ou não para responder à pergunta.

O modelo INSTRUCTOR utiliza os modelos GTR (Giant T5-like Retrieval) como seus codificadores principais para gerar embeddings. São modelos baseados no T5, fine-tuned para os conjuntos de dados de compreensão de instruções.

Para criar os embeddings usando os modelos GTR, o modelo INSTRUCTOR concatena o texto de entrada e a instrução da tarefa juntos. O texto de entrada concatenado é então passado pelo modelo GTR. As representações da camada *hidden* da entrada são utilizadas para criar os embeddings.

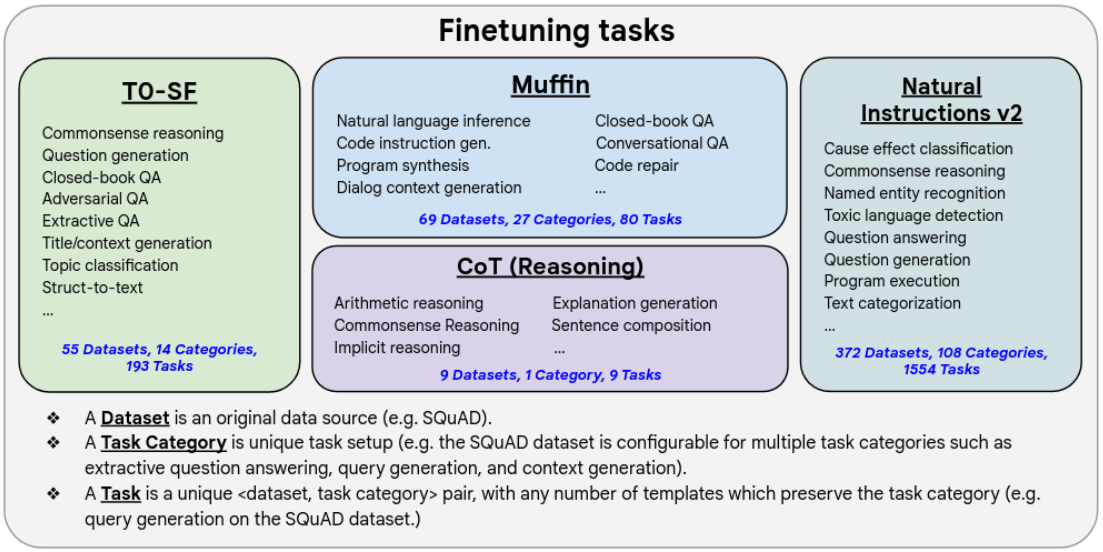
Rede Neural

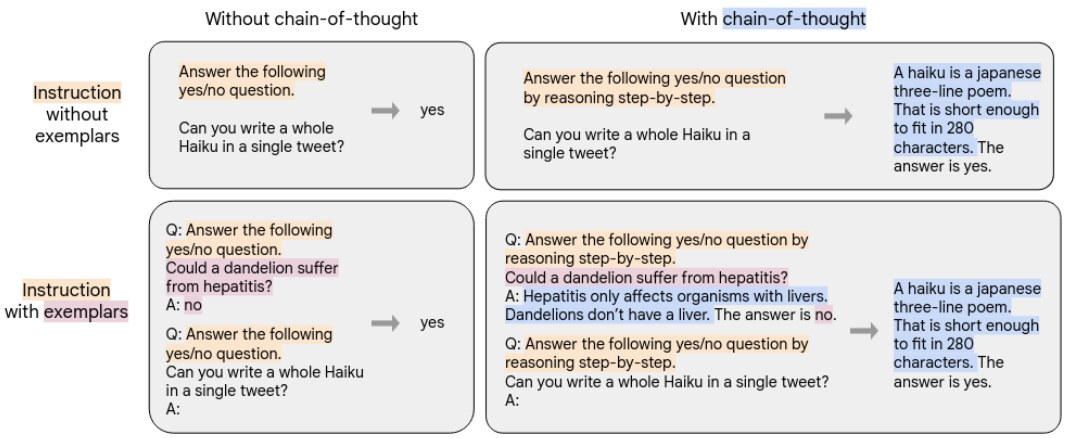
A arquitetura da rede neural utilizada no chatbot consiste em:

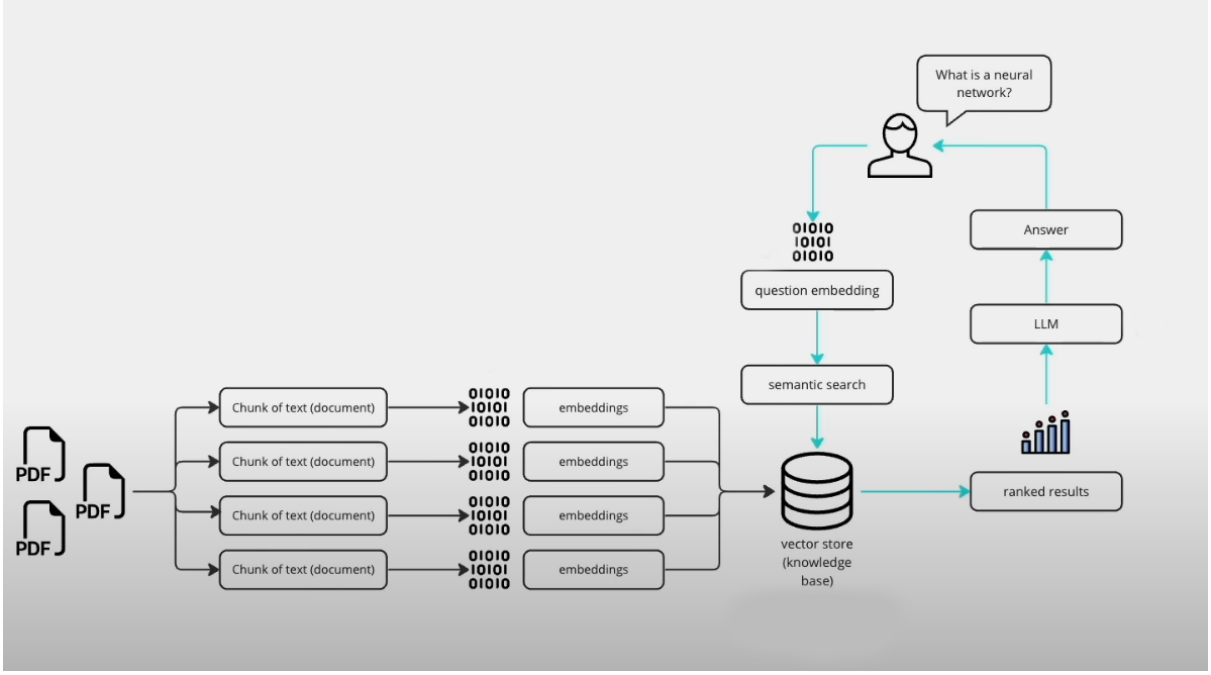
- Um modelo de linguagem pré-treinado (Hugging Face) para geração de respostas.
- Um mecanismo de recuperação de informações baseado em embeddings e busca em documentos.

O treinamento da rede neural envolveu o uso de técnicas de aprendizado supervisionado e transferência de aprendizado. O modelo de linguagem foi treinado em um grande conjunto de dados para gerar respostas coerentes e relevantes.

O Modelo de linguagem utilizado é o flan-t5-xxl da google que usa a arquitetura encoder-decoder com 11B de parâmetros. É uma versão fine-tuned do palm com diversos datasets (dos conjuntos Muffin, T0-SF, NIV2, and CoT) que contém perguntas e respostas sobre diversos domínios.







O texto do arquivo PDF é dividido em pedaços que são vetorizados e armazenados como documentos no banco de dados vetorial. São escolhidos os 3 documentos com maior similaridade vetorial. Utilizando a biblioteca langchain no python, a pergunta do usuário é enviada ao LLM e depois, para cada documento, a pergunta é feita novamente com os dados desse documento adicionados ao prompt, questionando o LLM se ele deve alterar a sua resposta com base no documento adicionado.

Exemplo de prompt:

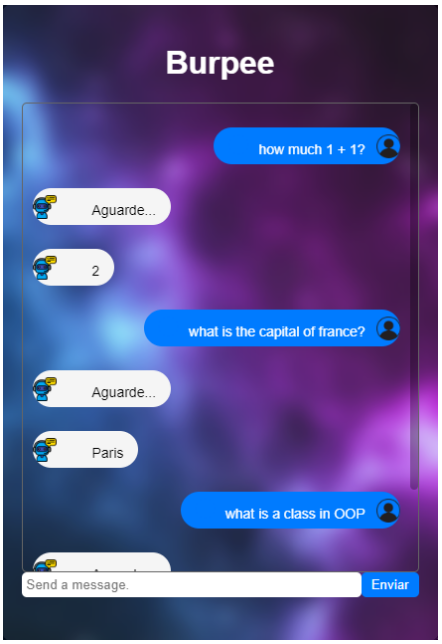
"We have the opportunity to refine the existing answer"
"(only if needed) with some more context below.\n"
"------------\n"
"{context_str}\n"
"------------\n"
"Given the new context, refine the original answer to better "
"answer the question. "
"If the context isn't useful, return the original answer."

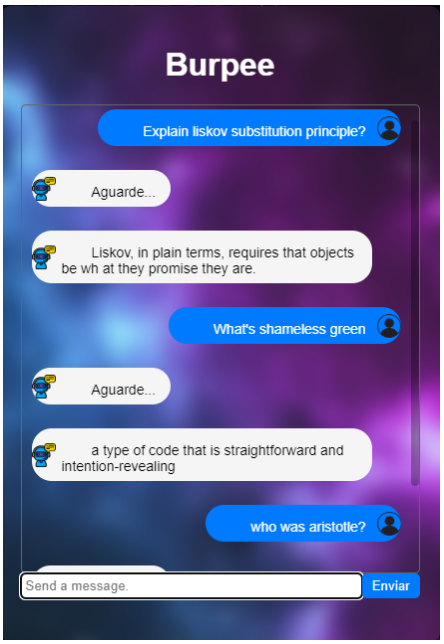
Resultados

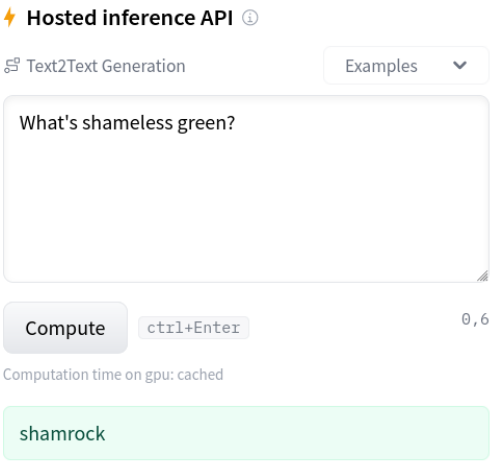
Os resultados obtidos com o projeto foram:

- Desenvolvimento de um chatbot funcional que interage de forma natural com os usuários.
- Integração do chatbot com um sistema de processamento de dados, permitindo o aprendizado baseado em conteúdo de um livro.
- Testes realizados com usuários demonstraram que o chatbot é capaz de fornecer respostas úteis e relevantes.

O projeto ainda está em andamento, e novas melhorias e funcionalidades estão sendo implementadas para aprimorar a experiência do usuário.







(Inferência do flan-t5-xxl no HF)

É possível observar que o chatbot mantém capacidades do modelo original, além de ser capaz de fornecer respostas sobre o domínio específico do livro, que não era possível no modelo original.