RAG aplicado a documentos institucionais do Instituto Federal de São Paulo (IFSP)

Ramon Abilio - r224253@dac.unicamp.br
Vagner Inácio - v208939@dac.unicamp.br

Índice

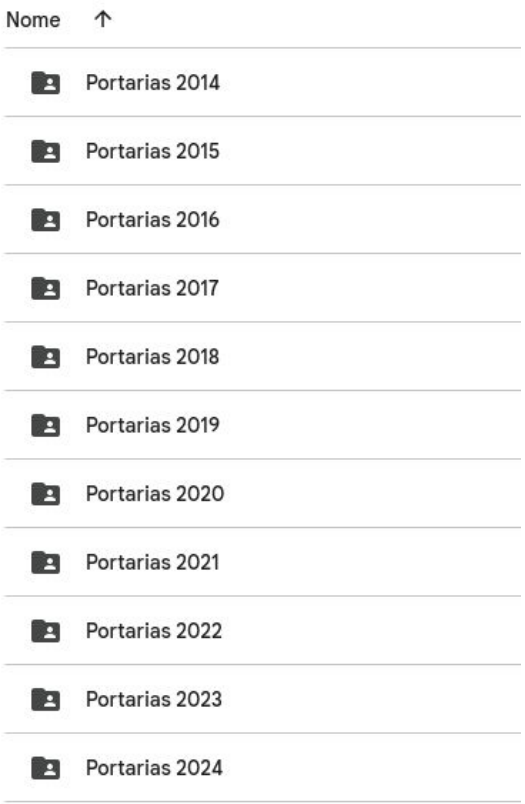
Descrição do Projeto

IFSP e seus documentos

- O IFSP é especializado na oferta de Educação Profissional e Tecnológica.
- Detém autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar.
- Regulamentação e divulgação através de instrumentos como Regimentos, Estatutos, Instruções Normativas, Portarias e Editais.
- Alguns desses documentos podem ser publicados, inclusive, pelos atuais 37 campi.
- Esses documentos são armazenados em formato PDF em um drive institucional ou no Google Drive e links são disponibilizados no site da Reitoria ou nos sites dos campi.

Acesso à documentos institucionais por meio do site da Reitoria

Acesso à Portarias editadas pela Direção Geral de um dos campus no Google Drive



Problema

- Documentos dispersos em, pelo menos, 38 locais.
- Documentos relacionados por numeração de portarias.
- PDF como formato padrão.
- Estrutura dos documentos e geração dos PDF com mudanças ao longo do tempo.
- Documentos escaneados e com modificação na estrutura do conteúdo de texto corrido para tabelas.
- Inexistência de um sistema em software que permita recuperar informações desses documentos.

Objetivo

O objetivo deste projeto é implementar um sistema Retrieval-Augmented Generation (RAG) que permita a um usuário extrair informações confiáveis de um ou mais documentos sem que ele tenha que conhecer a estrutura organizacional do Instituto e seus diversos tipos de documentos.

Metodologia

Fase 1 - Composição do dataset

- Coleta de documentos públicos do IFSP
- Definição de perguntas e respostas
- Análise dos documentos e definição das bibliotecas necessárias para extração do conteúdo
- Extração de conteúdo e remoção de informações irrelevantes (p.e. cabeçalhos, números de página)

Fase 2 - Indexação do conteúdo dos documentos

- Definição sobre como indexar os conteúdos, visto que além da resposta, é necessário indicar em qual(is) documento(s) a resposta foi baseada e podem existir documentos que revogam ou alteram documentos anteriores
- Implementação da indexação e recuperação (componente Retrieval do RAG) utilizando o Llamaindex (LIU, 2022)
- Avaliação do sistema utilizando métricas do framework RAGAS (RAGAS, 2023)

Fase 3 - Desenvolvimento e Implementação dos mecanismos de *Augmentation* e *Generation*

- Implementação dos mecanismos de *Augmentation* e *Generation* do RAG utilizando o Llamaindex, e o Large Language Model (LLM) Llama3 70b por meio da plataforma Groq (GROQ, 2024)
- Avaliação do sistema utilizando métricas do framework RAGAS

Datasets

O dataset será composto por documentos públicos do IFSP de um ou mais dos seguintes tipos:

- Estatutos
- Regimentos
- Instruções Normativas
- Portarias
- Resoluções

Métricas

Pretende-se utilizar métricas de avaliação disponíveis no framework RAGAS, como (RAGAS, 2023):

- Faithfulness
- Answer relevance
- Context recall
- Context precision
- Context relevancy
- Context entity recall

Resultados

Resultados **esperados** para a primeira entrega:

- Coleta de documentos públicos do IFSP.
- Definição de perguntas e respostas.
- Análise e extração de conteúdo dos documentos, removendo informações irrelevantes.

Resultados **preliminares** das entregas intermediárias

- Sistema de Indexação Implementado:
  - Definição e implementação da indexação utilizando Llamaindex.
  - Avaliação inicial e ajustes baseados em precisão e recall.
- Implementação de Mecanismos de *Augmentation* e *Generation*:
  - Implementação dos mecanismos com Llamaindex e Llama3 70b através plataforma Groq.
  - Avaliação inicial com métricas do framework RAGAS e ajustes necessários.

Resultados **finais** se for entrega final

- Sistema RAG Completo e Funcional.
- Documentação

Referências

- IFSP - Instituto Federal de São Paulo. **Transparência** - Documentos Institucionais. 2024. Disponível em: https://ifsp.edu.br/transparencia-documentos. Acesso em: 3 jun 2024
- GROQ. **Why Groq**. 2024. Disponível em: https://wow.groq.com/why-groq/. Acesso em: 3 jun 2024
- RAGAS. **Metrics**. 2023. Disponível em: https://docs.ragas.io/en/stable/concepts/metrics/index.html. Acesso em: 3 jun 2024
- LIU, Jerry. **LlamaIndex**. 2022. Disponível em: https://github. com/jerryjliu/llama_index. Acesso em: 4 jun 2024

Cronograma

Lista de atividades a serem feitas antes de cada entrega:

- 06 de junho - entrega I - Plano de Trabalho
- 13 de junho - entrega II
- 20 de junho - entrega III
- 27 de junho - entrega final