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

## Fase 1 - Construção do Dataset

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

# Resumo Projeto

## Informações nos documentos do IFSP

- O IFSP é especializado na oferta de Educação Profissional e Tecnológica.
- Regulamentação através de Estatutos, Instruções Normativas, Portarias e Editais.
- Alguns desses documentos podem ser publicados pelos atuais 37 campi.
- Documentos são armazenados em formato PDF
- Inexistência de um sistema em software que permita recuperar informações desses documentos.
- **Objetivo:**
  - Implementar um sistema RAG que permita extrair informações confiáveis de um ou mais documentos
  - Usuário não precisa ter conhecimento da estrutura organizacional do Instituto e seus diversos tipos de documentos.

# Índice

- Documentos públicos do IFSP
- Descrição dos documentos
- Documentos coletados
- Definição de perguntas e respostas
- Análise dos documentos
- Definição das bibliotecas necessárias para extração do conteúdo
- Extração de conteúdo e remoção de informações irrelevantes

# Documentos públicos do IFSP

- IFSP
  - Reitoria
  - Araraquara
  - Boituva
  - Capivari
- PDFs e Docx
- Repositórios com arquiteturas diferentes
  - Por exemplo, alguns documentos estão no Drive do Google e outros no Drive da instituição
- Não existe um padrão na organização dos arquivos
  - Alguns campi organizam seus arquivos em pastas por Ano e Mês, mas outros colocam todos os arquivos em uma única pasta

# Descrição dos documentos

- **Estatuto:**
  - Regulamento que estabelece a estrutura geral do IFSP, sua natureza e finalidades
- **Projeto Político Pedagógico (PPP):**
  - Documento que detalha objetivos, diretrizes e ações da comunidade escolar mapeando suas necessidades, propósitos e expectativas.
- **Portarias:**
  - Documento que contém instruções acerca da aplicação de leis ou regulamentos, recomendações de caráter geral, normas de execução de serviço, nomeações, demissões, punições, ou qualquer outra determinação da sua competência (UFSC, 2024)

# Documentos coletados

| Origem | Tipo de documento | Formato | # Documentos |
|---|---|---|---|
| Reitoria | Estatuto | PDF | 1 |
| Araraquara | 1) Projeto Político Pedagógico<br>2) Portarias | PDF | 1<br>1656 |
| Boituva | 1) Projeto Político Pedagógico<br>2) Portarias | PDF | 1<br>1180 |
| Capivari | 1) Projeto Político Pedagógico<br>2) Portarias | PDF | 1<br>1737 |
| | | **Total** | **4577** |

# Definição de perguntas e respostas

**Perguntas com respostas que requerem consulta a mais de um documento**

**Pergunta**: Em quais comissões o servidor Ramon Abílio participou?

**Resposta**: Ramon Abílio participou das comissões: Comissão de Avaliação de Atividades Docentes, Comissão de Inventário, Comitê de Pesquisa, Inovação e Pós-graduação do Câmpus (COMPESQ)

**Fonte**:

PORTARIA Nº 60/2023 - DRG/BTV/IFSP DE 29 DE SETEMBRO DE 2023

PORTARIA BTV.0135/2019, DE 20 DE NOVEMBRO DE 2019

PORTARIA Nº 85/2022 - DRG/BTV/IFSP DE 23 DE NOVEMBRO DE 2022

# Definição de perguntas e respostas

**Perguntas com respostas encontradas em um único documento**

**Pergunta**: O IFSP tem somente ensino médio?

**Resposta**: O IFSP oferece educação profissional técnica de nível médio, prioritariamente na forma de cursos integrados, e também cursos de nível superior e pós-graduação.

**Fonte**: Estatuto, Capítulo II, art. 6, página 4

Poderemos utilizar como referência a página de Perguntas Frequentes do próprio IFSP

# Análise dos documentos

- **Estatuto**
  - Em formato PDF com estrutura padronizada em Capítulos, Artigos e Parágrafos.
- **Projetos Político Pedagógico (PPP)**
  - Em formato PDF com estrutura geral parecida e organizada contendo capa, membros da comissão, folha de aprovação, listas de tabelas e figuras, sumário e capítulos com suas subseções.
  - Documentos longos, com mais de 140 páginas em média.
- **Portarias**
  - Formatos diversos, desde sua formatação (textos, tabelas) até conteúdo (objetivo da portaria). No geral, possuem de uma a duas páginas.
  - Existem portarias escaneadas (imagens) e outras digitais. Elas possuem um título, as mais antigas tem um “preâmbulo”, e depois vem o objetivo da portaria.

# Bibliotecas para extração do conteúdo

- Como vamos utilizar o LlamaIndex, testamos o recurso nativo do framework e pesquisamos ferramentas que pudessem ser integradas a ele.
- LlamaParse
  - API do LlamaIndex para parsear arquivos utilizando os frameworks do LlamaIndex.
  - Suporta diferentes tipos de arquivos: PDF, PPTX, DOCX…
  - Extração tabelas (*state-of-the-art*)
  - Exporta o resultado em Markdown ou JSON
    - O JSON traz, por exemplo, o texto completo da página e também uma lista de itens como: número da página, títulos, figuras, textos e tabelas
  - Tem limite de mil páginas por dia na versão gratuita (2K no nosso caso)

# Exemplo de Conteúdo

Markdown

Tabela


```
{
    "page": 25,
    "text": "    2.2. Estrutura Física\n\n\n          O campus do IFSP - Araraquara tem uma área constru
    "md": "# 2.2. Estrutura Física\n\nO campus do IFSP - Araraquara tem uma área construída de 8,6 mil me
    "images": [],
    "items": [
        {
            "type": "heading",
            "lvl": 1,
            "value": "2.2. Estrutura Física",
            "md": "# 2.2. Estrutura Física"
        },
        {
            "type": "text",
            "value": "O campus do IFSP - Araraquara tem uma área construída de 8,6 mil metros quadrados e
            "md": "O campus do IFSP - Araraquara tem uma área construída de 8,6 mil metros quadrados em u
        },
        {
            "type": "table",
            "rows": [
                [
                    "Local",
                    "Quantidade Atual",
                    "Quantidade prevista até 2023",
                    "Área (m²)"
```

# Extração de Conteúdo

- Extração do conteúdo do Estatuto, dos 3 PPPs e das portarias de Boituva
  - ~4h para extrair o conteúdo dos 1184 arquivos
  - Utilizamos duas chaves da API
  - 2K de páginas/dia
- Repositório no GitHub (https://github.com/rsabilio/ia024-projeto-rag)
  - PDF e JSON
  - Códigos

# Próximos passos

- Finalização do parser na base completa
- Refinamento: seleção dos itens de interesse dos arquivos JSON, como títulos, números de página, textos e tabelas para montagem do dataset, que será utilizado na fase seguinte.
- Implementação do Sistema de Indexação
- Implementação dos Mecanismos de *Augmentation* e *Generation*
- 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
- IFSP - Instituto Federal de São Paulo, Campus Boituva. **Direção** - Portarias https://btv.ifsp.edu.br/index.php/direcao-geral . Acesso em: 11 jun 2024
- IFSP - Instituto Federal de São Paulo, Campus Capivari. https://cpv.ifsp.edu.br/index.php/documentos-institucionais?id=156. Acesso em: 11 jun 2024
- IFSP - Instituto Federal de São Paulo, Campus Araraquara. **Transparência** - Documentos Institucionais. 2024. Disponível em:https://www.arq.ifsp.edu.br/documentos-institucionais. Acesso em : 11 jun 2024
- UFSC - Universidade Federal de Santa Catarina. Conceitos. 2024. Disponível em: https://legislacao.ufsc.br/conceitos/. Acesso em: 11 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 - Construção do Dataset
- 20 de junho - entrega III
- 27 de junho - entrega final