UNIVERSIDADE FEDERAL DO AMAZONAS

INSTITUTO DE COMPUTAÇÃO

ENGENHARIA DE SOFTWARE

AMANDA HERICA DOS SANTOS FREIRE - 22051138

PAULA QUADROS DE MENDONÇA - 22052743

Trabalho Final de PLN - Relatório de Pré-Processamento

MANAUS - AM

2024

INTRODUÇÃO

O objetivo deste relatório é descrever detalhadamente as etapas de download, extração e pré-processamento dos textos das legislações acadêmicas. O processo incluiu a transformação de documentos PDF em arquivos de texto, seguida pela estruturação dos dados em um formato JSON para facilitar as etapas subsequentes de processamento e análise.

FERRAMENTAS UTILIZADAS

Python: Linguagem de programação principal.

Bibliotecas Python:

- PyMuPDF (fitz): Para extração de texto de PDFs.
- os e shutil: Para manipulação de arquivos e diretórios.
- json: Para criação e manipulação de arquivos JSON.

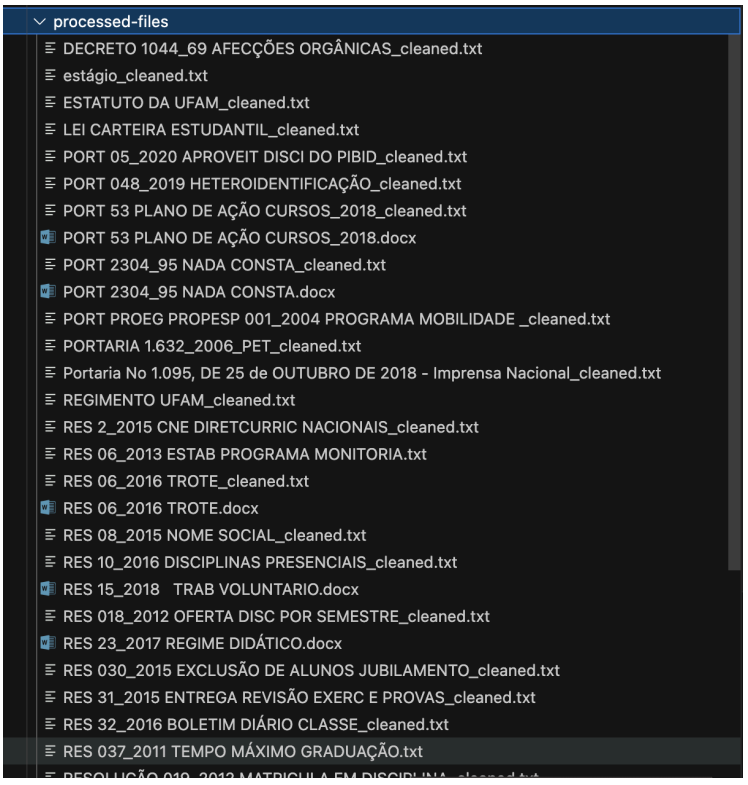
Google Colab: Ambiente de execução para processamento de dados e geração de perguntas e respostas.

Google Drive: Para armazenamento e acesso aos arquivos durante o processamento.

ETAPAS DE PRÉ PROCESSAMENTO

Parte 1: Download e Extração de Textos

1. Criação de Ambiente Python:
    - Configurei um ambiente virtual Python para garantir a instalação de todas as dependências necessárias.
2. Transformação de PDFs em Texto:
    - Utilizei a biblioteca PyMuPDF para converter documentos PDF em arquivos de texto (.txt).
    - Criei um script Python que lê cada PDF na pasta de entrada, extrai o texto e salva os resultados em uma pasta chamada processed-files.
3. Estruturação dos Dados em JSON:
    - Desenvolvi um segundo script Python para consolidar os textos extraídos em um único arquivo JSON.
    - O script percorre todos os arquivos .txt na pasta processed-files, lê o conteúdo e os organiza em um formato JSON estruturado, resultando no arquivo docs.json.



Parte 2: Geração de Perguntas e Respostas

1. Ambiente de Processamento:
    - Inicialmente, tentei gerar perguntas e respostas no ambiente Python local, mas encontrei problemas de compatibilidade com OpenSSL.
    - Como alternativa, utilizei um notebook do Google Colab para realizar o processamento.
2. Configuração do Google Colab:
    - Conectei o notebook do Colab ao Google Drive para acessar o arquivo docs.json.
    - Dividi o conteúdo do arquivo JSON em várias seções menores para facilitar o processamento.
3. Tokenização e Geração de Perguntas e Respostas:
    - Utilizei o modelo GPT-2 e a biblioteca Transformers da Hugging Face para tokenizar o texto e gerar perguntas e respostas.
    - Para evitar exceder os limites de tokens do GPT-2, limitei o tamanho dos tokens a 248 por seção.
    - O script gera perguntas e respostas para cada seção do texto e armazena os resultados em um arquivo JSON chamado dados_sinteticos.json.

Desafios Enfrentados

- Compatibilidade com OpenSSL: Enfrentei problemas com a versão do OpenSSL no ambiente local, o que levou à migração do processamento para o Google Colab.
- Limite de Tokens do GPT-2: Tive que dividir o conteúdo em seções menores para garantir que o modelo GPT-2 pudesse processar os textos corretamente.
- Tempo de Processamento: O processo de geração de perguntas e respostas é demorado, exigindo otimizações e divisões de conteúdo para manter a eficiência.

CONCLUSÃO

O pré-processamento dos dados foi essencial para preparar as legislações em um formato adequado para análise e geração de perguntas e respostas. Apesar dos desafios encontrados, o uso combinado de ferramentas Python e o ambiente do Google Colab permitiram concluir o processo com sucesso. A base de dados resultante está estruturada e pronta para etapas posteriores de treinamento de modelos e análise.