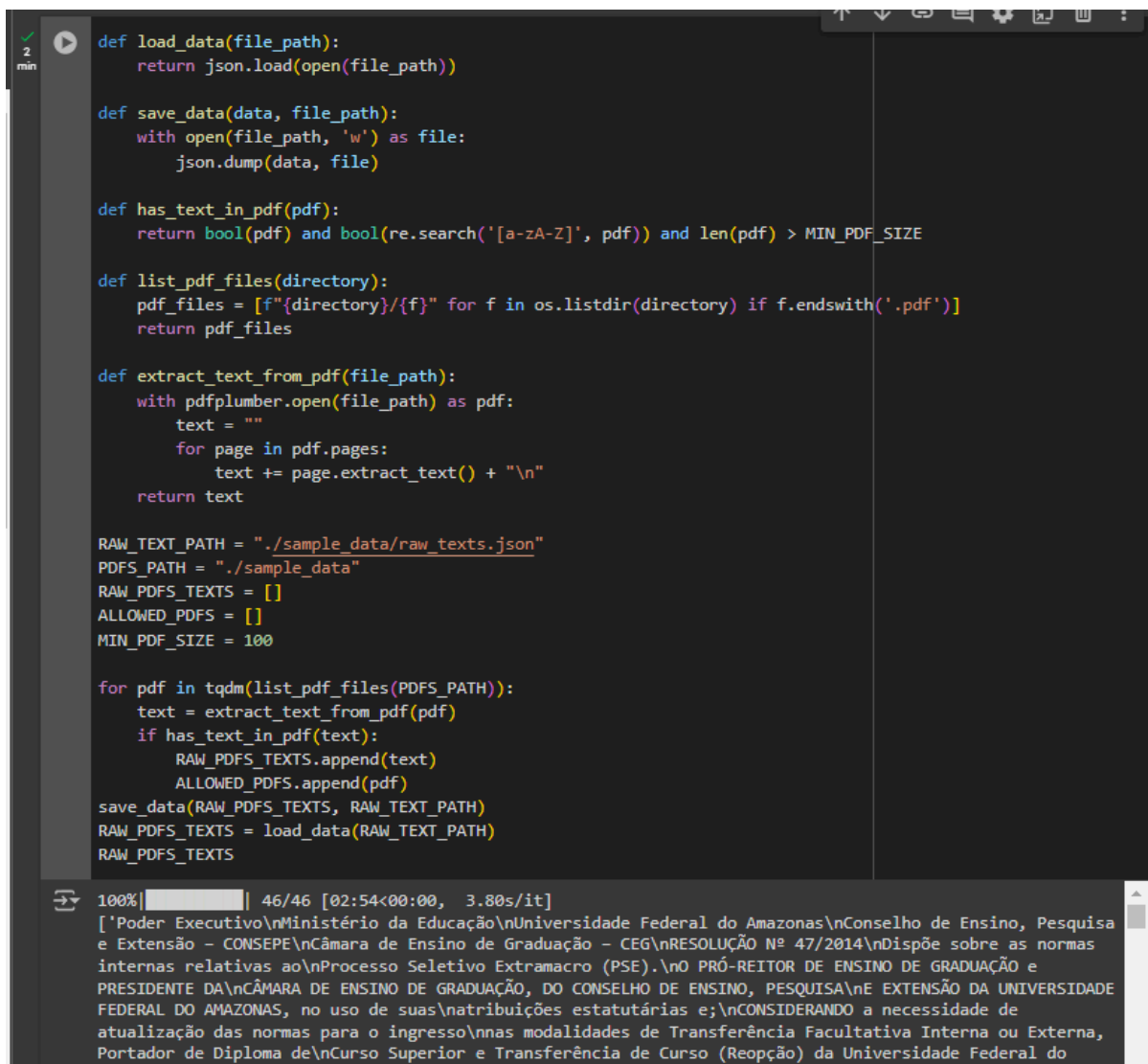


1. Relatório de Pré-processamento

1.1. Descrição detalhada das etapas de download, extração e pré-processamento dos textos das legislações.

- O download dos arquivos foi feito de forma manual, utilizando o navegador Google Chrome.
- A extração dos textos foi feita através da biblioteca “pdfplumber” conforme a figura abaixo:



```
def load_data(file_path):
    return json.load(open(file_path))

def save_data(data, file_path):
    with open(file_path, 'w') as file:
        json.dump(data, file)

def has_text_in_pdf(pdf):
    return bool(pdf) and bool(re.search('[a-zA-Z]', pdf)) and len(pdf) > MIN_PDF_SIZE

def list_pdf_files(directory):
    pdf_files = [f"{directory}/{f}" for f in os.listdir(directory) if f.endswith('.pdf')]
    return pdf_files

def extract_text_from_pdf(file_path):
    with pdfplumber.open(file_path) as pdf:
        text = ""
        for page in pdf.pages:
            text += page.extract_text() + "\n"
        return text

RAW_TEXT_PATH = "./sample_data/raw_texts.json"
PDFS_PATH = "./sample_data"
RAW_PDFS_TEXTS = []
ALLOWED_PDFS = []
MIN_PDF_SIZE = 100

for pdf in tqdm(list_pdf_files(PDFS_PATH)):
    text = extract_text_from_pdf(pdf)
    if has_text_in_pdf(text):
        RAW_PDFS_TEXTS.append(text)
        ALLOWED_PDFS.append(pdf)
save_data(RAW_PDFS_TEXTS, RAW_TEXT_PATH)
RAW_PDFS_TEXTS = load_data(RAW_TEXT_PATH)
RAW_PDFS_TEXTS
```

100% | 46/46 [02:54<00:00, 3.80s/it]

['Poder Executivo\nMinistério da Educação\nUniversidade Federal do Amazonas\nConselho de Ensino, Pesquisa e Extensão - CONSEPE\nCâmara de Ensino de Graduação - CEG\nRESOLUÇÃO Nº 47/2014\nDispõe sobre as normas internas relativas ao\nProcesso Seletivo Extramuro (PSE).\nO PRÓ-REITOR DE ENSINO DE GRADUAÇÃO e PRESIDENTE DA\nCÂMARA DE ENSINO DE GRADUAÇÃO, DO CONSELHO DE ENSINO, PESQUISA\nE EXTENSÃO DA UNIVERSIDADE FEDERAL DO AMAZONAS, no uso de suas\natribuições estatutárias e;\nCONSIDERANDO a necessidade de atualização das normas para o ingresso\nnas modalidades de Transferência Facultativa Interna ou Externa, Portador de Diploma de\nCurso Superior e Transferência de Curso (Reopção) da Universidade Federal do

- Para o pré-processamento dos textos, foi feita a remoção de caracteres e símbolos indesejados; a conversão dos textos em minúsculas; a remoção de espaços em branco extras; a remoção de stopwords (palavras comuns que podem não ser úteis para análise); e por fim, uma etapa de lematização de palavras. Conforme a figura abaixo:

```
def preprocess_text(text):
    lemmatizer = WordNetLemmatizer()
    text = text.lower()
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    text = re.sub(r'\s+', ' ', text).strip()
    words = text.split()
    stop_words = set(stopwords.words('portuguese'))
    words = [word for word in words if word not in stop_words]
    words = [lemmatizer.lemmatize(word) for word in words]
    cleaned_text = ' '.join(words)
    return cleaned_text

CLEANED_TEXT_PATH = "./sample_data/cleaned_texts.json"
CLEANED_PDFS_TEXTS = []

for text in tqdm(RAW_PDFS_TEXTS):
    CLEANED_PDFS_TEXTS.append(preprocess_text(text))
save_data(CLEANED_PDFS_TEXTS, CLEANED_TEXT_PATH)
CLEANED_PDFS_TEXTS = load_data(CLEANED_TEXT_PATH)
CLEANED_PDFS_TEXTS[-3:]

norma norteiam aproveitamento estudos vista otimizao processo administrao acadmica ufam paragrafo nico fin
desta resoluo considerase aproveitamento estudos processo aceita ufam estudos realizados cursos graduao
autorizados reconhecidos brasil ministrio educacao oriundos instituies estrangeiras educacao superior
mediante condies estabelecidas nesta resoluo art 2 aproveitamento estudos assenta aplicao trs critrios
bsicos i densidade identificao carga horria disciplina origem ufam ii qualidade identificao contedo
programtico disciplina origem ufam iii adequao identificao objetivos disciplina origem disciplina destino
universidade federal amazona conselho ensino pesquisa extenso art 3 dever solicitado aproveitamento
estudos via processual ae i disciplinas cursadas ufam identidade cdigo constantes quadro equivalncia
contido projeto pedaggico curso ii disciplinas cursadas outras instituies educacao superior inclusive
estrangeiras 1 poder aproveitada disciplina ufam base disciplina cursada y origem 2 forma inversa base
nica disciplina cursada y origem poder aproveitada disciplina ufam 3 ambos casos devero observados
```

1.2. Ferramentas utilizadas e desafios enfrentados durante o processo.

- Como mencionado anteriormente, a ferramenta utilizada para extração de textos foi a “pdfplumber” e o principal desafio foi extrair pdfs cujo os textos faziam parte de imagens, tornando inviável processar alguns desses arquivos.

1.3. Base de dados

- A base de dados está no arquivo “sample_data/cleaned_texts.json”