UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUIZ GUSTAVO ALVES NERY

AVALIAÇÃO DE FERRAMENTAS DE EXTRAÇÃO DE TEXTO EM DOCUMENTOS PDF

CAMPINA GRANDE - PB

2024

LUIZ GUSTAVO ALVES NERY

AVALIAÇÃO DE FERRAMENTAS DE EXTRAÇÃO DE TEXTO EM DOCUMENTOS PDF

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Cláudio de Souza Baptista

CAMPINA GRANDE - PB

2024

LUIZ GUSTAVO ALVES NERY

AVALIAÇÃO DE FERRAMENTAS DE EXTRAÇÃO DE TEXTO EM DOCUMENTOS PDF

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Cláudio de Souza Baptista

Orientador – UASC/CEEI/UFCG

Joseana de Macêdo Fechine

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em:

17/05/2024

CAMPINA GRANDE - PB

RESUMO

Este estudo aborda a importância da extração precisa de informações em documentos PDF, destacando os desafios enfrentados devido à falta de uniformidade na estrutura e layout desses documentos. A extração de texto em documentos PDF, especialmente em contextos como Diários Oficiais, é crucial para automatizar processos e otimizar a análise de informações relevantes. A métrica ROUGE é utilizada para avaliar a qualidade da extração de texto pelas ferramentas e a importância de extrair todas as informações do texto original preservando a ordem de leitura. Diante da ineficiência e alto custo associado à extração manual de texto de documentos em formato PDF, este estudo visa proporcionar percepções significativas que auxiliam na escolha da ferramenta mais adequada, considerando os diferentes cenários de aplicação na extração de texto. A avaliação das ferramentas escolhidas, juntamente com a mensuração dos resultados através de métricas pertinentes à avaliação dos textos extraídos, aprimora a eficácia e a eficiência na análise dessas ferramentas.

Evaluation of Text Extraction Tools in PDF Documents

ABSTRACT

This study addresses the importance of precise information extraction in PDF documents, highlighting the challenges faced due to the lack of uniformity in the structure and layout of these documents. Text extraction in PDF documents, especially in contexts such as Official Gazettes, is crucial for automating processes and optimizing the analysis of relevant information. The ROUGE metric is used to evaluate the quality of text extraction by the tools and the importance of extracting all the information from the original text while preserving the reading order. Given the inefficiency and high cost associated with manual text extraction from PDF format documents, this study aims to provide significant insights that assist in choosing the most suitable tool, considering the different application scenarios in text extraction. The evaluation of the chosen tools, along with the measurement of the results through metrics relevant to the evaluation of the extracted texts, enhances the effectiveness and efficiency in the analysis of these tools.

Avaliação de ferramentas de Extração de texto em Documentos PDF

Luiz Gustavo Alves Nery
Unidade Acadêmica de Sistemas e Computação - UASC
Universidade Federal de Campina Grande - UFCG
Campina Grande, Paraíba, Brasil
luiz.nery@ccc.ufcg.edu.br

Cláudio de Souza Baptista
Unidade Acadêmica de Sistemas e Computação - UASC
Universidade Federal de Campina Grande
Campina Grande
Campina Grande, Paraíba, Brasil
baptista@computacao.ufcg.edu.br

PALAVRAS-CHAVE
PDF, Extração de Texto, Ferramenta, Eficiência, Métrica, Layout.

1. INTRODUÇÃO

No contexto da crescente digitalização da informação, os documentos digitais assumiram um papel central, tornando-se o formato padrão em diversas esferas. Esses documentos constituem uma vasta fonte de informações que podem ser processadas em larga escala. As possibilidades incluem a utilização desses documentos tanto em processos judiciais quanto administrativos. Ademais, o armazenamento eletrônico desses documentos facilita o compartilhamento dos mesmos, otimizando a comunicação e a colaboração entre as equipes, promovendo um ambiente de trabalho integrado. Além disso, podem ser utilizados também desde a indexação até a extração de informações, com base em seu conteúdo. No entanto, extrair o conteúdo desses documentos de forma precisa e fiel ao original é desafiador [11]. O formato predominante para documentos digitais é o PDF, amplamente utilizado devido à sua excelente legibilidade, embora não mantenha informações sobre o relacionamento entre elementos textuais.

O Google sozinho indexa atualmente mais de 3 bilhões de documentos em PDF, mais do que qualquer outro formato de documento, exceto o HTML [4]. O PDF continua sendo um dos formatos eletrônicos mais populares. Entretanto, o PDF é um formato baseado em layout, especificando as posições e fontes dos caracteres individuais dos quais o texto é composto. Muitas aplicações requerem, em vez disso, informações sobre os blocos de construção semânticos do texto. Isso inclui a compreensão das palavras utilizadas, bem como a estrutura do texto, que é dividida em parágrafos e seções. Além disso, é crucial entender os papeis semânticos desempenhados por diferentes partes do texto. Isso pode envolver a determinação de se um determinado trecho de texto pertence ao corpo principal do texto, a uma nota de rodapé ou a uma legenda. Essas informações semânticas geralmente não são fornecidas como parte do PDF [5]. Isso torna a recuperação do texto original uma tarefa desafiadora. Documentos como formulários, recibos, notas fiscais, processos jurídicos e publicações em diários oficiais frequentemente apresentam um alto volume de texto e variações significativas no layout. Essas diferenças estruturais, juntamente de elementos como tabelas, figuras e fórmulas, agregam ainda mais complexidade ao processo de extração.

Portanto, diante de uma ampla variedade de soluções disponíveis para o processamento de PDFs, é crucial determinar qual delas é a mais adequada para a extração do conteúdo textual. Este estudo visa realizar uma análise detalhada das ferramentas selecionadas para extração de texto de documentos PDFs e avaliar os resultados por meio de métricas relacionadas à extração de texto e à reconstrução estruturada do conteúdo original. O objetivo final é oferecer insights valiosos para a escolha da melhor solução. A maioria das ferramentas não especifica para quais problemas elas são realmente úteis, por isso a importância desse estudo em avaliar as diversas ferramentas. Todas as ferramentas realizam a identificação de palavras e consideram a ordem das palavras (não seriam muito úteis se não o fizessem). Apenas as ferramentas mais sofisticadas oferecem limites de parágrafos e funções semânticas [6].

O problema central que este trabalho de conclusão de curso visa abordar reside na ineficiência e no alto custo associado à extração manual de documentos em formato PDF. Encontrar soluções de extração de texto que consigam obter adequadamente o texto-alvo, apesar de variações significativas de layout, bem como a falta de uniformidade na estrutura do documento. Essas características dificultam a tarefa de extração do conteúdo, dado que pode haver inconsistências no texto extraído, impactando assim soluções que se utilizam do texto extraído para outras tarefas.

De forma a ilustrar a problemática citada, foi considerado no contexto deste trabalho o domínio dos Diários Oficiais, mais especificamente o Diário Oficial do Estado de São Paulo. Tratam-se de documentos extensos e com conteúdo relevante à sociedade, agregando alterações em regulamentos, leis, contratos e outras informações críticas. Entidades privadas e públicas, impactadas pelo conteúdo publicado, possuem interesse no conteúdo publicado, bem como no seu processamento. Entretanto, a coleta manual desses dados torna-se uma tarefa difícil, consumindo tempo, recursos financeiros e humanos significativos. Essa ineficiência compromete a agilidade nas tomadas de decisão e a capacidade de estar em conformidade com as regulamentações em constante evolução. O número de documentos não estruturados ou semiestruturados produzidos em todos os tipos de organizações continua a aumentar rapidamente [7]. Isso evidencia ainda mais a importância da extração adequada e fidedigna do seu conteúdo textual.

As seções seguintes do trabalho estão estruturadas como segue.. A seção 2 é a Fundamentação Teórica do estudo que aborda a métrica Rouge utilizada. Na seção 3 são endereçados alguns Trabalhos Relacionados, retratando pesquisas e estudos que possuem correlação com a temática deste trabalho. Na seção 4 estão Materiais e Métodos, no qual são apresentados os fluxos da metodologia do estudo e da implementação. Na Seção 5 estão os Resultados que foram gerados pela aplicação da métrica e por fim a seção 6 apresenta a Conclusão deste trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção de fundamentação teórica, abordam-se os conceitos essenciais relacionados às métricas de avaliação Rouge, F1-score, Precisão e o Recover.

2.1 F1-Score

O F1-Score é uma métrica de avaliação que é utilizada dentro da métrica Rouge, ela é amplamente utilizada na área de processamento de linguagem natural mas também tem grande impacto nessa parte de geração de textos automáticos. O F1-Score combina duas métricas importantes. A precisão mede a proporção de verdadeiros positivos em relação ao total de positivos previstos pelo texto de referência. Ou seja, representa a capacidade de não classificar incorretamente os caracteres corretos como incorretos. Já a revocação mede a proporção de verdadeiro positivos em relação ao total de positivos reais presentes no documento.

No contexto da métrica Rouge, o F1-score é calculado considerando uma média harmônica entre a precisão e a revocação da Rouge, proporcionando uma medida agregada da fidelidade do texto que foi extraído pela ferramenta em relação ao texto de referência. Uma alta pontuação de F1 (próxima de 1) indica que o texto gerado é tanto completo quanto preciso em relação ao texto de referência, enquanto uma pontuação baixa (próxima de 0) sugere deficiências em uma ou ambas as áreas.

A Precisão e Revocação são calculadas a partir das seguintes fórmulas:

$$Precisão = \frac{Número\ de\ n-gramas\ extraídos\ corretamente}{Total\ de\ n-gramas\ extraídos\ pela\ ferramenta}$$

$$Revocação = \frac{Número\ de\ n-gramas\ extraídos\ corretamente}{Total\ de\ n-gramas\ do\ texto\ de\ referência}$$

O F1-score é calculado usando a seguinte fórmula:

$$F1\ Score = \frac{2\ x\ Precisão\ x\ Revocação}{Precisão + Revocação}$$

2.1.1 Métrica ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

A sumarização automática de texto é um campo rico de pesquisa. Por exemplo, workshops de avaliação de tarefas compartilhadas para sumarização foram realizados por mais de uma década na Conferência de Compreensão de Documentos (DUC) e, posteriormente, na Conferência de Análise de Texto (TAC) [1]. Um elemento importante dessas tarefas compartilhadas é a avaliação dos sistemas participantes. Inicialmente, a avaliação manual era realizada, onde juízes humanos eram encarregados de avaliar a qualidade dos resumos gerados automaticamente. No entanto, em um esforço para tornar a avaliação mais escalável, a medida automática Rouge 1 foi introduzida no DUC-2004 [1]. Rouge determina a qualidade de um resumo automático através da comparação de unidades sobrepostas, como n-gramas, sequências de palavras e pares de palavras com resumos escritos por humanos.

A métrica Rouge é composta por variantes, cada uma calculando a similaridade entre dois conjuntos de texto com base na presença de n-gramas compartilhados. Algumas das principais são ROUGE-N e ROUGE-L, cada uma enfatizando aspectos específicos da similaridade textual.

ROUGE-N mede a sobreposição de n-gramas entre o texto de referência e o texto gerado, avaliando a qualidade literal do texto gerado em relação ao texto de referência. O resultado é um valor entre 0 e 1, onde 1 indica a maior similaridade possível e 0 indica a menor similaridade. ROUGE-L é baseado na maior subsequência comum entre o texto de referência e o texto gerado, levando em consideração a similaridade entre as duas sequências de texto no nível de sentença e fluência dos textos extraídos.

A métrica ROUGE é uma métrica muito boa para avaliação de sistemas de processamento de linguagem natural, mas é muito valiosa para comparação de resumos e textos no geral que sejam gerados automaticamente, pois permite uma avaliação objetiva e automatizada, sendo fácil de calcular e de interpretar os resultados da qualidade do texto extraído e facilitando a comparação na análise. Contudo, uma das limitações inerentes a essa abordagem é a sua incapacidade de considerar a semântica das palavras, especificamente o significado individual de cada termo. Por exemplo, pode ocorrer de haver duas palavras, uma no texto gerado e outra no texto de referência, que possuem o mesmo significado, mas que não serão consideradas similares segundo a métrica ROUGE.

3. TRABALHOS RELACIONADOS

No trabalho de Ramalho [8] é abordado a aplicação de técnicas de aprendizado de máquina e processamento de linguagem natural na extração de informações de licitações a partir do Diário Oficial do Estado do Acre. É explorada a complexidade da extração de dados de documentos PDF e propõe soluções para automatizar esse processo. O trabalho investiga a relação entre tecnologias de processamento de linguagem natural e a eficiência na análise de licitações.

O trabalho de Cardoso [9] aborda a aplicação de técnicas de aprendizado de máquina e processamento de texto para extrair informações relevantes de diários oficiais. Além disso, ele explora o uso de ferramentas como o framework Scrapy para web crawling e web scraping, bem como bibliotecas para processamento de arquivos PDF. O fluxo de trabalho inclui etapas como reconhecimento óptico de caracteres (OCR) e refluxo de texto para lidar com o formato específico dos documentos. O estudo investiga a eficácia de diferentes classificadores em relação ao tamanho da janela de texto e à vetorização.

O estudo de Neves Junior [10], explora a extração de informações e mineração de dados no Diário Oficial de Pernambuco. É proposto um método para construir uma aplicação que utiliza algoritmos para indexar o conteúdo da base do Diário Oficial,

transformando as informações anteriormente disponíveis em texto para um formato estruturado. Essa abordagem visa aplicar técnicas de mineração de dados e tornar os dados mais acessíveis e significativos. O estudo de caso baseou-se em documentos publicados no Diário Oficial entre janeiro de 2007 e abril de 2017. Embora os resultados mostraram a viabilidade da indexação e estruturação desses dados, ainda é necessário aprimorar a padronização das informações.

O trabalho de Rastogi [11] aborda a extração de texto de documentos em formato PDF. A extração de texto é uma etapa fundamental para processar dados de PDFs, especialmente quando desejamos utilizar esses dados em outros contextos. O estudo explora várias bibliotecas disponíveis para a extração de texto de documentos PDF, considerando diferentes tecnologias e abordagens. Essas bibliotecas desempenham um papel crucial na recuperação precisa do conteúdo textual, superando desafios relacionados a fontes, layout de exibição e formatação tabular.

4. MATERIAIS E MÉTODOS

Nesta seção são descritas as quatro etapas essenciais, conforme mostrado na figura 1. A primeira etapa é a Definição de Tipos de Documentos, a segunda são os Critérios para Seleção das Ferramentas, a terceira são os Critérios para escolha da Métrica e a última é a Avaliação das Ferramentas a parte em que está toda a implementação, da extração do conteúdo até o armazenamento dos resultados. Além disso, todos os artefatos e trechos de código utilizados nessas etapas estão disponíveis no repositório GitHub[1], criado para armazenar e organizar a visualização de todos esses arquivos.

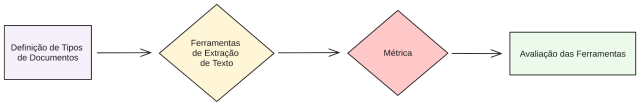

Figura 1: Imagem que representa o fluxograma da metodologia do trabalho

4.1 Definição de Tipos de Documentos

A primeira etapa da metodologia aplicada à problemática apresentada buscou justamente selecionar os principais tipos de documentos a serem utilizados. Dessa forma, avaliando as ferramentas de extração em diferentes layouts, sendo estes representativos de diferentes formatos de disposição das informações. Dentre algumas características consideradas para a seleção dos tipos de documentos, destacam-se: quantidade de colunas na organização do texto; presença de tabelas; tabelas de coluna dupla; somente texto; tabelas de 5 colunas.

A extensa quantidade de texto, as variações substanciais e a formatação complexa dos documentos frequentemente apresentam um desafio para a extração de informações, tornando este contexto propício para a experimentação das ferramentas em cenários diversos e mais sofisticados.

[1] https://github.com/GustavoNeery/dataset-diario-sao-paulo/tree/main

Foram selecionados cinco tipos de layouts para avaliar o desempenho das ferramentas na extração de texto. Os vários tipos de documentos analisados incluem tabelas, layout de uma coluna, documento apenas com texto, e algumas páginas que combinam texto e tabelas. Além disso, cada um desses modelos apresenta estruturas distintas.

O exemplo apresentado na Figura 2 é de uma página do Diário Oficial do Estado de São Paulo, selecionada para a realização deste estudo. Esse documento apresenta um layout de duas colunas, contendo texto que não segue um padrão de formatação específico. Além disso, inclui uma tabela de duas colunas inserida entre as demais informações. Neste cenário, é possível explorar a ferramenta e compará-la com as outras, tendo em vista que existe outro cenário do estudo em que o documento contém uma tabela com 5 colunas tornando a extração mais complexa pela quantidade maior de colunas na tabela.

Figura 2: Documento 3 - Página 17 do Diário Oficial do Estado de São Paulo de 07 de março de 2024

Além do formato descrito e apresentado na Figura 2, foram selecionados outros documentos com diferentes formatos, indentações, layouts e estruturas. Os tipos de documentos selecionados, bem como suas especificidades, estão detalhados na Tabela 1. Ao todo, foram utilizados cinco documentos, todos consistindo em páginas da mesma edição publicada no Diário Oficial do Estado de São Paulo no dia 7 de março de 2024.

	Página do Diário	**Layout do Documento**	**Estrutura da Tabela**
Documento 1	Página 1	Layout 4 colunas	Não possui

Documento 2	Página 4	Layout 2 colunas	Tabela com 5 colunas
Documento 3	Página 17	Layout 2 colunas	Tabela com layout distinto
Documento 4	Página 26	Layout 2 colunas	Tabela com layout distinto
Documento 5	Página 60	Layout 2 colunas	Não possui

Tabela 1: Descrição dos formatos e diferentes tipos de documentos selecionados

4.2 Critérios Para Seleção das Ferramentas

A escolha das ferramentas de extração de texto de documentos PDF é um aspecto crítico do trabalho, pois influencia diretamente na qualidade e precisão da análise. Será abordada a justificativa para a seleção de cada uma das ferramentas escolhidas de forma que seja garantido resultados confiáveis e significativos para a pesquisa, as quais atendam a critérios objetivos e importantes. A seguir, estão descritos cada critério utilizado:

1. A ferramenta deve ser gratuita;
2. A execução da ferramenta deve disponibilizar a possibilidade de realizar localmente;
3. O repositório do GitHub (quando disponível), deve possuir mais de 50 estrelas;
4. A entrada deve ser em formato .PDF e a saída em .TXT;
5. Para garantir a funcionalidade contínua da ferramenta à medida que as tecnologias evoluem, é imprescindível que o repositório do GitHub tenha recebido atualizações regulares ao longo dos últimos anos;
6. A ferramenta deve possuir uma boa documentação de fácil acesso;
7. Disponível em sistemas Linux/Unix.

Com base nos requisitos acima, foram selecionadas as seguintes ferramentas: Apache PDFBox, PDFToText, Slate3k, Mupdf e PDFMiner. A seguir, detalhamos cada uma das ferramentas selecionadas.

4.2.1 Apache PDFBox

O PDFBox é uma ferramenta Java que faz parte da biblioteca Apache e é conhecido por sua precisão na extração de texto de documentos PDF. Ele implementa algoritmos para lidar com diferentes tipos de layout e fontes, o que é crucial para garantir a extração precisa de texto de documento com muitos layouts divergentes. É uma biblioteca popular e bem estabelecida, amplamente adotada na comunidade de desenvolvimento. Isso significa que há uma vasta base de usuários e um forte suporte da comunidade, incluindo atualizações regulares e correções de bugs.

Essa biblioteca oferece diversas funcionalidades para manipulação de arquivos PDF, mas, para os propósitos deste estudo, o foco foi na funcionalidade de extração do conteúdo textual presente nos PDFs. Essa abordagem viabiliza a manipulação do texto contidos nos documentos PDF.

4.2.2 PDFToText

PDFToText é uma ferramenta para extrair texto de documentos PDF. Ela oferece uma maneira direta de converter PDFs em texto puro, sem a necessidade de bibliotecas ou dependências adicionais. Além disso, PDFToText está disponível como parte do pacote Poppler, que é uma biblioteca de software livre amplamente distribuída. Isso torna fácil e conveniente a instalação e uso da ferramenta em diferentes plataformas e sistemas operacionais.

A ferramenta permite realizar a extração de texto com opções diversas que podem ser inseridas como parâmetro no momento da execução do comando para extração, um exemplo é a extração com preservação de layout. No entanto, o foco principal do trabalho consiste em extrair o texto independentemente do layout. Além disso, a ferramenta possibilita extrair o texto de uma gama de páginas PDF passando algumas opções como parâmetro na execução.

4.2.3 Slate3k

O Slate3k é uma biblioteca Python que simplifica o processo de extração de texto de documentos PDF. Baseada na biblioteca PDFMiner, oferece uma abordagem para recuperar o conteúdo textual presente em arquivos. O principal aspecto dessa ferramenta é a facilidade de uso. Essa simplicidade torna o Slate3k uma escolha atraente para desenvolvedores que desejam extrair rapidamente o texto de documentos PDF, sem complicações adicionais. Essa dependência no PDFMiner garante que o Slate3k herde a confiabilidade e a precisão do PDFMiner. Além disso, essa ferramenta tem flexibilidade e escalabilidade, pois além da extração de texto, o Slate3k permite que os desenvolvedores explorem outras funcionalidades da biblioteca que ele herda. Se necessário, é possível aprofundar-se na API do PDFMiner para realizar operações mais complexas, como a extração de metadados, imagens e estruturas hierárquicas.

O Slate3k é uma ferramenta prática e eficaz para a extração de texto de PDFs, especialmente quando a simplicidade e a rapidez são essenciais. Sua integração com o PDFMiner oferece uma solução completa para manipulação de conteúdo textual em documentos PDF, contribuindo para projetos acadêmicos e aplicativos .

4.2.4 Mupdf

Mupdf é um leitor de arquivos PDF leve e prático, ideal para aqueles que desejam abrir manuais e outros textos criados nessa extensão, mas não possuem em seus computadores um programa específico para essa finalidade. Mupdf é conhecido por seu desempenho rápido e eficiente na manipulação de documentos PDF. Ele é otimizado para processamento rápido e de baixo consumo de recursos, tornando-o ideal para lidar com grandes volumes de documentos. Mupdf é uma biblioteca multiplataforma que pode ser integrada em diferentes ambientes de desenvolvimento e sistemas operacionais, proporcionando uma solução flexível e portátil. Ele foi projetado para ter um desempenho rápido e usar poucos recursos do sistema, tornando-o ideal para dispositivos com capacidades limitadas, como smartphones ou tablets.

Além de sua funcionalidade de visualização, o Mupdf oferece a capacidade de extrair texto e imagens de documentos em formato PDF. Essa característica é especialmente útil para a manipulação e processamento de conteúdo contido nesses arquivos. Quando há a necessidade de obter o texto de um PDF para fins de edição ou análise, o Mupdf apresenta-se como uma escolha sólida.

4.2.5 PDFMiner

O PDFMiner oferece recursos avançados de análise de documentos PDF, incluindo a capacidade de extrair não apenas texto, mas também informações estruturadas, como metadados e marcadores. Além disso, é capaz de lidar com uma ampla variedade de idiomas e fontes, incluindo fontes não padrão e scripts complexos, garantindo uma extração precisa e confiável de texto em documentos multilíngues. O PDFMiner é escrito em Python, o que facilita sua integração em projetos Python existentes.

A principal função do PDFMiner é a extração de texto. Ele obtém a localização exata do texto em cada documento, incluindo informações sobre fontes e formatação. Além disso, realiza uma análise automática de layout, compreendendo a estrutura do documento e a disposição do texto. Ele lida com criptografia básica (RC4 e AES) presente em alguns PDFs. O PDFMiner suporta diferentes tipos de fontes, incluindo Type1, TrueType, Type3 e CID. Funciona bem com idiomas que usam caracteres complexos, como chinês, japonês e coreano. Ademais, possui um parser PDF extensível, permitindo personalizações específicas. Sua capacidade de extrair e analisar informações, bem como sua flexibilidade, torna-o uma opção válida para aplicação neste estudo.

4.3 Critérios para escolha da Métrica

A métrica Rouge (Recall-Oriented Understudy for Gisting Evaluation) é uma métrica amplamente utilizada na avaliação de sumarizações automáticas de texto e na comparação de texto gerado automaticamente com texto de referência [1][12]. Ela é especialmente útil para avaliar a qualidade de sistemas de resumo automático, tradução automática e geração de texto. Esse é um dos critérios que reforça a adequabilidade da métrica ao cenário deste estudo, uma vez que o texto obtido a partir das ferramentas de extração pode ser cruzado com uma anotação de referência, obtendo-se assim o valor da métrica para a referida ferramenta.

Além disso, o foco da Rouge é no Recall, isso significa que ela se concentra em capturar todas as informações do texto original no texto gerado, que no caso desta pesquisa é no texto extraído.

Foi realizada uma análise de trabalhos e estudos anteriores que utilizaram a mesma métrica para avaliar a qualidade de textos gerados [2]. Embora esses contextos fossem diferentes, como sumarização e geração de textos por IA, a métrica ROUGE avalia o texto de referência, o qual geralmente é anotado manualmente, e o texto hipótese, que é gerado, extraído ou sumarizado por alguma ferramenta, de forma muito similar ao cenário atual do estudo. Portanto, devido a esse fato, foi identificado que essa métrica se encaixava de forma ideal para avaliar o texto extraído pelas ferramentas.

Apesar da métrica ROUGE apresentar certas limitações, como a incapacidade de capturar todos os aspectos da qualidade do texto, ela é frequentemente empregada na avaliação de resumos e é mais adequada para o contexto desse estudo se comparada à métrica BLEU por exemplo, a qual foi analisada mas por ser muito simples e não possuir a capacidade de recuperar palavras e sequências no texto de referência não foi utilizada. Esta aplicação é análoga à avaliação do texto extraído pela ferramenta em questão. Ademais, a vasta quantidade de recursos e pesquisas disponíveis contribui significativamente para facilitar o uso efetivo dessa métrica ROUGE.

4.4 Avaliação das Ferramentas

A última etapa da metodologia aplicada neste estudo envolve a avaliação das ferramentas selecionadas, sendo esta conduzida utilizando-se dos tipos de documentos selecionados e computando a métrica obtida a partir da análise de trabalhos relacionados. A métrica selecionada, conforme especificado, é o ROUGE. De forma a detalhar os passos seguidos na avaliação das ferramentas, a Figura 2 possui um fluxograma resumindo as etapas do processo.

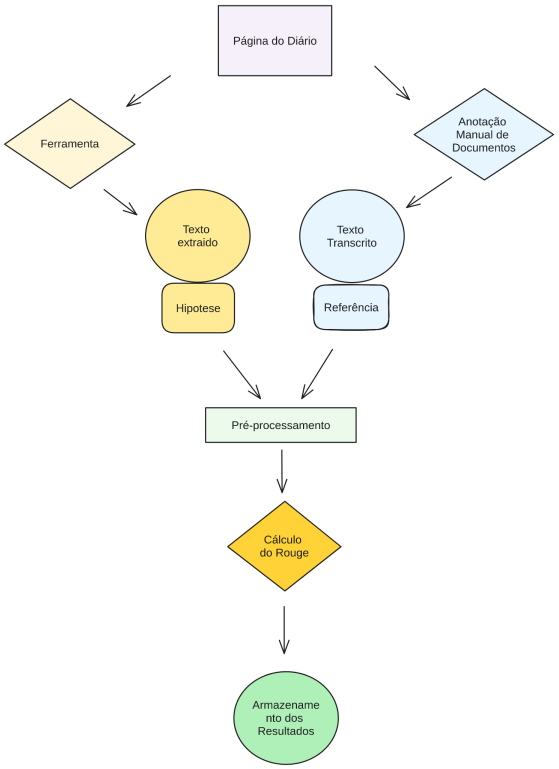

Figura 2: Fluxograma da parte de avaliação e implementação das ferramentas

A implementação da etapa de avaliação das soluções foi realizada utilizando-se do ambiente do Google Colab, estando o Notebook Python disponível no repositório GitHub citado anteriormente.

4.4.1 Anotação Manual de Documentos

Realizou-se a anotação manual das páginas selecionadas de documentos em formato PDF, criando-se um arquivo para cada documento, direcionando o conteúdo transcrito para esse novo arquivo. Posteriormente, esse arquivo serviu como referência na aplicação da métrica ROUGE. O foco principal recai sobre a extração precisa do conteúdo textual contido nestes documentos. No caso específico deste estudo, a atenção concentrou-se exclusivamente no texto, independentemente das complexidades de layout. Cada página do Diário Oficial do Estado de São Paulo pode conter uma combinação de elementos, como:

- Texto: informações gerais, parágrafos, títulos e subtítulos.
- Tabelas: dados organizados em colunas e linhas.
- Imagens: gráficos, ilustrações ou fotografias.

A extração manual permitiu isolar o texto relevante, respeitando a ordem de leitura do conteúdo dos documentos, ignorando outros componentes visuais. O exemplo presente na Figura 3 detalha um fragmento de anotação manual realizada a partir de um documento cujo layout contém tanto conteúdo textual simples como uma tabela.

Local: Aeroporto Moussa Nakhl Tobias - Estrada Municipal Murilo Vilaça Marangoni s/n – km 5,5 -Bauru, SP

Turma: 03 Data: 13/03/2024 Horário: 8h às 12h Local: Aeroporto Moussa Nakhl Tobias - Estrada Municipal Murilo Vilaça Marangoni s/n – km 5,5 -Bauru, SP Turma: 04 Data: 13/03/2024 Horário: 13h às 17h Local: Aeroporto Moussa Nakhl Tobias - Estrada Municipal Murilo Vilaça Marangoni s/n – km 5,5 -Bauru, SP

Turma: 04 Data: 13/03/2024 Horário: 13h às 17h Local: Aeroporto Moussa Nakhl Tobias - Estrada Municipal Murilo Vilaça Marangoni s/n – km 5,5 -Bauru, SP e não como constou.

COORDENADORIA DE UNIDADES PRISIONAIS DA REGIÃO METROPOLITANA DE SÃO PAULO CENTRO DE DETENÇÃO PROVISÓRIA ASP VANDA RITA BRITO DO REGO DE OSASCO CENTRO ADMINISTRATIVO PUBLICAÇÕES DE PDS - FEVEREIRO / 2024

CENTRO ADMINISTRATIVO Comunicado:Relação de pagamentos efetuados em fevereiro 2024, em Cumprimento ao art. 2º da Lei Estadual 7.857/92: DATA DE PAGAMENTO ORDEM BANCÁRIA CREDOR PD VALOR RESPECTIVO 02/2/2024 35418 LINK CARD ADMINISTRADORA DE BENEFICIOS LT 2024PD00014 R$ 260,93 02/2/2024 35419 LINK CARD ADMINISTRADORA DE BENEFICIOS LT 2024PD00015 R$ 1.087,52 02/2/2024 35420 LINK CARD ADMINISTRADORA DE BENEFICIOS LT 2024PD00016 R$ 8.172,34 07/2/2024 38428 VMI SISTEMAS DE SEGURANCA LTDA 2024PD00013 R$ 7.520,80 16/2/2024 45878 LINK CARD ADMINISTRADORA DE BENEFICIOS LT 2024PD00019 R$ 239,35 16/2/2024 45879 LINK CARD ADMINISTRADORA DE BENEFICIOS LT 2024PD00020 R$ 11.607,46 16/2/2024 46686 TELEFONICA BRASIL S A 2024PD00023 R$ 599,04 27/2/2024 54527 MARCELO BALTAZAR TAVARES DE SOUZA 2024PD00037 R$ 49,50 27/2/2024 54528 ADRIANO RIDOLFI 2024PD00038 R$ 49,50 27/2/2024 54529 GILBERTO DE OLIVEIRA NEVES 2024PD00039 R$ 49,50 27/2/2024 54530 SIMONE APARECIDA DA ROCHA 2024PD00040 R$ 49,50 27/2/2024 54531 SERGIO LOURENçO DOS SANTOS 2024PD00041 R$ 49,50 27/2/2024 54532 ANDRE DA SILVA BAPTISTA 2024PD00042 R$ 49,50 27/2/2024 54533 SILVIO SPIROS VAYANOS 2024PD00043 R$ 114,56 27/2/2024 54534 ADELINO GONçALVES FERREIRA ALVES 2024PD00044 R$ 89,10 27/2/2024 54535 ADELINO GONçALVES FERREIRA ALVES 2024PD00045 R$ 49,50 27/2/2024 54536 ADELINO GONçALVES FERREIRA ALVES 2024PD00046 R$ 49,50

Figura 3: Anotação Manual do Documento 2 - página 4 do Diário Oficial do Estado de São Paulo do dia 07 de março de 2024

4.4.2 Extração do texto com as Ferramentas

Para as cinco ferramentas utilizadas foi criado um código nas linguagens de programação Java e Python, utilizando ambas bibliotecas para realizarem a extração do texto dos documentos PDF selecionados. Exceto para o PDFToText que é a única ferramenta que não era necessário implementar um código tendo em vista que ela faz parte do pacote Poppler. O processo de extração iniciou-se com uma análise da documentação dos repositórios selecionados do GitHub que iriam servir como base para as implementações para cada uma das ferramentas. Com base na documentação de cada repositório foi feita a implementação subsequente. Posteriormente, para cada implementação, foi realizada a extração para os cincos documentos individualmente e os resultados da extração de cada um foram sendo convertidos em arquivos do tipo .TXT armazenados em um diretório designado.

4.4.3 Pré-processamento dos arquivos

O pré-processamento foi realizado tanto do texto extraído como do texto transcrito das páginas do documento do Diário Oficial do Estado de São Paulo. Esse pré-processamento faz a remoção das quebras de linhas contidas no texto, inserindo no lugar um espaçamento. Após esse processamento a métrica ROUGE pode ser aplicada aos textos. Pois, após o processamento, o texto de referência e o texto de hipótese contêm o mesmo número de linhas. Essa correspondência elimina a possibilidade da métrica comparar n-gramas de linhas diferentes. Na figura 4, a imagem à esquerda apresenta o texto antes do processamento, enquanto que a imagem à direita mostra esse mesmo texto após o processamento dos dados.

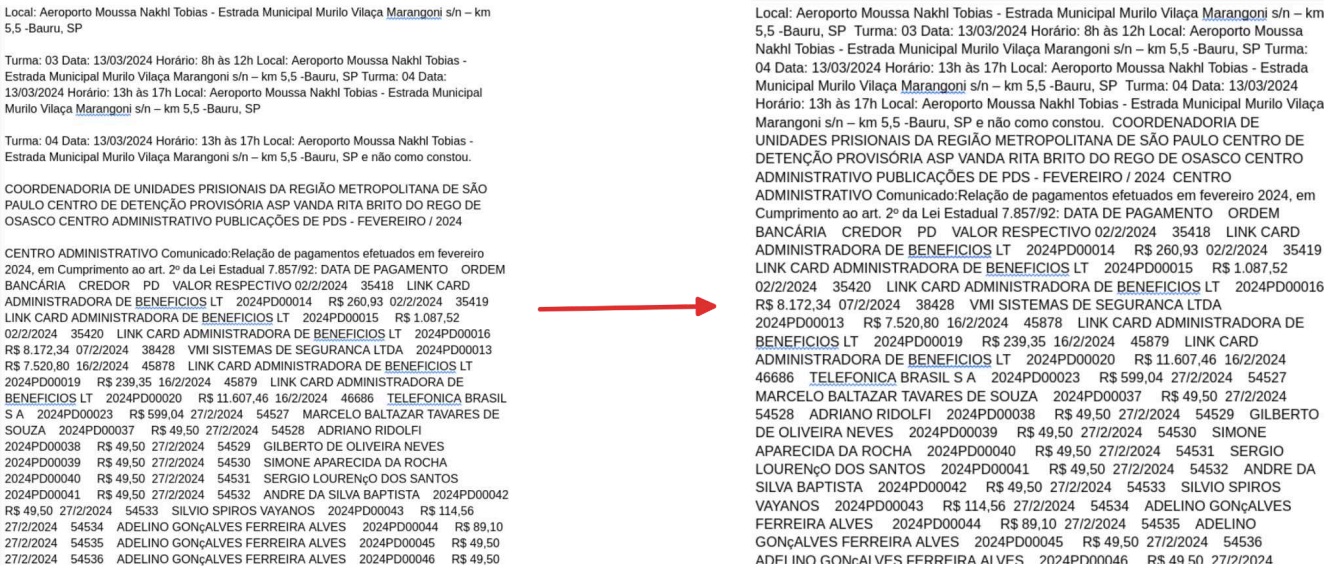

Figura 4: Exemplo de arquivos anotados manualmente pré e pós processamento do Documento 2 - Página 4 do Diário Oficial do Estado de São Paulo

É importante salientar que, apesar de alterar a disposição original do conteúdo extraído pelas ferramentas de processamento de arquivos PDF, não há prejuízo algum ao estudo realizado. A principal característica analisada é a ordem de leitura, ou seja, se a mesma é preservada no conteúdo obtido a partir das ferramentas de extração. Dado que não houve alteração na ordem dos elementos textuais, tendo sido removidas apenas as quebras de linhas, não há impacto ao computar a métrica de avaliação.

4.4.4 Cálculo da métrica ROUGE

A aplicação da métrica ROUGE consistiu na comparação do texto extraído com o texto anotado manualmente a partir das etapas anteriores. A saída gerada pela aplicação da métrica retorna os valores de precisão, revocação e o F1-Score. O cálculo da métrica foi realizado de forma a computar os resultados para cada combinação possível de ferramentas de extração e tipos de documento. Desta forma, todos os documentos foram processados por cada uma das ferramentas, tendo assim a sua saída comparada com o texto de referência anotado manualmente.

A métrica foi aplicada a um documento que passou por um processo de pós-processamento, extraído por uma das ferramentas. Tendo em vista que o F1-score é a média harmônica, a ideia é que, para cada resultado gerado, somente esse valor seja levado em consideração para armazenamento no DataFrame, apesar da métrica gerar os outros dois valores. Posteriormente, calculou-se a média para cada DataFrame gerado a partir de cada documento, com base nos F1-scores de cada ferramenta. Essas médias foram armazenadas em um novo DataFrame, que contém os valores de todos os F1-scores de todas as ferramentas, considerando todos os documentos analisados.

5. RESULTADOS

A partir da análise das ferramentas, sendo esta a última etapa da metodologia descrita e aplicada neste estudo, foram obtidas as métricas ROUGE para cada combinação de documento e ferramenta de extração. Os resultados são detalhados nesta seção. Para cada documento, foram computados o F1-Score do ROUGE-1, ROUGE-2 e ROUGE-L.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	0.9238	0.9486	**0.9526**	0.9212	0.8493
Rouge-2	0.9116	0.9128	**0.9422**	0.8798	0.7666
Rouge-L	0.9238	0.9436	**0.9526**	0.9186	0.8413

Tabela 2: Tabela que representa o Data Frame criado com os valores gerados pelo cálculo da média da métrica ROUGE com base em todos os documentos.

A ferramenta Mupdf apresentou o melhor resultado dentre as ferramentas analisadas, obtendo um valor médio superior às demais. Este valor superior representa uma maior consistência do conteúdo extraído, sendo este comparado com uma anotação manual de referência. Por outro lado, o Slate3k teve o pior desempenho em comparação com as outras ferramentas. Suas médias para cada página foram significativamente inferiores, principalmente no ROUGE-2 em que são comparados bi-gramas. A Figura 6 representa graficamente a média dos resultados em termos de F1-Score para cada uma das ferramentas avaliadas.

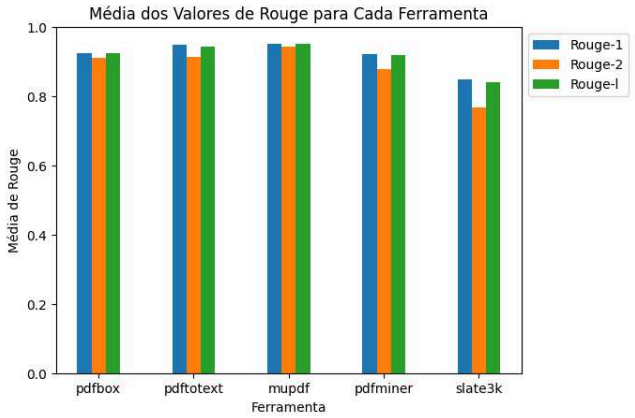

Figura 5: Gráfico com a média dos valores dos Rouge-1 Rouge-2 e Rouge-l para cada ferramenta

Apesar da superioridade dos resultados em termos médios da ferramenta MuPDF, ainda cabe uma análise detalhada por tipo de documento, de forma a identificar possíveis cenários onde ferramentas diferentes possuem uma melhor "afinidade" com tipos de documentos específicos.

5.1 Resultados específicos

Ao analisar a tabela 4, pode-se observar que, embora o PDFBox tenha obtido a terceira melhor média entre as ferramentas testadas, seus resultados no Documento 2, que contém uma tabela de 5 colunas, foram os piores. A métrica baseada nos dados extraídos indicou que, para esse cenário específico, o PDFBox se mostrou a ferramenta menos eficaz entre as cinco avaliadas. Surpreendentemente, o Slate3k, apesar de ter apresentado a pior média geral, obteve um desempenho superior ao do PDFBox no Documento 2.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	0.78	**0.971**	0.965	0.965	0.864
Rouge-2	0.776	0.936	**0.964**	0.926	0.792
Rouge-L	0.78	**0.971**	0.965	0.961	0.864

Tabela 3: Tabela que representa o Data Frame criado com os valores do F1-Score somente para o Documento 2.

Analisando os resultados apresentados na tabela 4, é evidente que tanto o PDFMiner quanto o Slate3k demonstram uma limitação semelhante ao serem testados em documentos que possuem uma estrutura com mais de três colunas, como é o caso do cenário para esse documento 1 em questão. Nesse contexto, os valores obtidos para ambas as ferramentas revelaram-se inferiores em relação às demais avaliadas. Essa similaridade nos resultados pode ser atribuída, em parte, ao fato de que ambas as ferramentas foram desenvolvidas com base na mesma biblioteca, o que sugere uma possível influência compartilhada de características ou funcionalidades no desempenho observado.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	**0.971**	0.932	0.967	0.81	0.81
Rouge-2	0.943	0.912	**0.948**	0.745	0.745
Rouge-L	**0.971**	0.932	0.967	0.81	0.81

Tabela 4: Tabela que representa o Data Frame criado com os valores do F1-Score somente para o Documento 1.

Observando os resultados da tabela 5, pode-se notar que, embora a ferramenta PDFBox tenha apresentado o desempenho mais fraco no cenário do Documento 2, ela se destacou como a melhor em todas as três métricas neste outro cenário. Isso ressalta a ideia de que cada ferramenta pode possuir áreas de especialização ou funcionalidades superiores em comparação com outras, dependendo do contexto em que são utilizadas.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	**0.956**	0.955	0.948	0.948	0.853
Rouge-2	**0.954**	0.885	0.944	0.921	0.726
Rouge-L	**0.956**	0.93	0.948	0.944	0.833

Tabela 5: Tabela que representa o Data Frame criado com os valores do F1-Score somente para o Documento 3.

Os valores apresentados na tabela 6 para este cenário de teste refletem uma semelhança, o que pode ser atribuído ao fato de que o documento 4 possui as mesmas características quando comparado com o documento 3. Torna-se evidente que, para o cenário de um documento com layout de duas colunas e tabela contendo menos de três colunas, o PDFBox demonstra um desempenho superior em todas as métricas em comparação com as outras ferramentas avaliadas. Porém, tanto o PDFToText como o Mupdf tiveram valores inferiores mas por poucos décimos, a maior diferença entre o Mupdf que teve a maior média no resultado final e o PDFBox nesse cenário foi de 0,018 décimos.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	**0.968**	0.956	0.95	0.95	0.876

Rouge-2	**0.965**	0.944	0.949	0.909	0.799
Rouge-L	**0.968**	0.956	0.95	0.945	0.858

Tabela 6: Tabela que representa o Data Frame criado com os valores do F1-Score somente para o Documento 4.

A análise dos resultados revela variações no desempenho das ferramentas de extração de texto de documentos PDF, conforme avaliado pelas métricas Rouge-1, Rouge-2 e Rouge-L. O PDFBox e o PDFMiner destacam-se, obtendo pontuações mais altas em todas as métricas. O Mupdf também apresenta resultados competitivos, embora inferiores em comparação com o PDFBox e o PDFMiner. Por outro lado, o Slate3k demonstra o pior desempenho em mais um cenário, especialmente evidenciado pela pontuação mais baixa na métrica Rouge-2.

	PDFBox	**PDFToText**	**Mupdf**	**PDFMiner**	**Slate3k**
Rouge-1	**0.944**	0.929	0.933	0.933	0.844
Rouge-2	**0.92**	0.887	0.906	0.898	0.771
Rouge-L	**0.944**	0.929	0.933	0.933	0.842

Tabela 7: Tabela que representa o Data Frame criado com os valores do F1-Score somente para o Documento 5.

6. CONCLUSÃO

Com base nos resultados apresentados, pode-se concluir que a aplicação da métrica ROUGE proporcionou o entendimento acerca do desempenho das diferentes ferramentas de extração de texto no contexto do DOE-SP. A ferramenta Mupdf destacou-se como a ferramenta melhor avaliada no cenário completo, considerando a média dos resultados de F1-Score em todos os documentos. Embora a ferramenta não tenha obtido o melhor desempenho na extração de todos os tipos de documento, demonstrou um desempenho melhor para o cenário completo.

A capacidade do Mupdf de lidar com diferentes layouts é o ponto principal de superioridade dessa ferramenta em relação às outras, mostrando que independentemente do layout da página do Documento PDF que ela fizer a extração, vai conseguir extrair sem cometer tantos erros. O Slate3k, por outro lado, apresentou o pior desempenho em comparação com as outras ferramentas. Suas médias para cada PDF foram significativamente inferiores, indicando limitações na capacidade de extração de acordo com o tipo de documento processado.

O estudo aplicado neste trabalho atingiu o seu objetivo, comparando de forma detalhada as ferramentas selecionadas para o objetivo proposto: extração de texto em documentos PDF preservando-se a ordem de leitura, extração de texto independente do layout do documento e identificação dos cenários em que as ferramentas se saem melhor. Além disso, com base nos resultados obtidos é possível ter uma avaliação e comparação das ferramentas utilizadas, destacando suas diferentes capacidades e limitações a depender do contexto de sua aplicação. Essa análise oferece uma base para a seleção e implementação de uma das ferramentas de extração de texto, permitindo que seja possível escolher a que garante mais qualidade e confiabilidade no texto extraído.

REFERÊNCIAS

[1] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74–81). Association for Computational Linguistics.

[2] Lin, C.-Y. (2004b). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop.

[3] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[4] Google Search Console. (2023). PDF Index Size. Retrieved from https://support.google.com/webmasters/answer/93710?hl=en.

[5] Adobe Systems Incorporated. (2008). The Portable Document Format (PDF) Reference Manual. San Jose, CA: Adobe Systems Incorporated.

[6] Islam, S. M., & Rahman, S. M. M. (2012). A survey of techniques for extracting text from PDF documents. Journal of Intelligent Information Systems, 39(2), 131-164.

[7] Gaudreault, J., & Viel, C. (2018). PDF Table Extraction for Semi-structured Documents. 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE.

[8] Ramalho, R. E. C. (2022). Utilizando técnicas de aprendizagem de máquina e NLP para extração de informações em licitações do Diário Oficial do Estado do Acre. Trabalho de Conclusão de Curso - Artigo, Curso de Bacharelado em Ciência da Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Paraíba, Brasil.

[9] Cardoso, A. A. de A. (2023). Extração de Dados com Aprendizagem de Máquina para Processamento de Informações em Diários Oficiais. Universidade de Brasília.

[10] Neves Junior, R. B., Melo, W. F. de M., Fagundes, R. A. de A., & Maciel, A. M. A. (2018). Extração de Informação e Mineração de Dados no Diário Oficial de Pernambuco. Revista de Engenharia e Pesquisa Aplicada, 3, Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.

[11] Rastogi, U. (2024). Study on Libraries for Text Extraction from PDF Document. International Research Journal of Engineering and Technology (IRJET), 5(6), 1-5.

[12] Ng, J.-P., & Abrecht, V. (2015). Better Summarization Evaluation with Word Embeddings for ROUGE. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1925–1930. DOI: 10.18653/v1/D15-1222.