

ESTUDO COMPARATIVO DE FERRAMENTAS DE EXTRAÇÃO DE INFORMAÇÕES EM ARQUIVOS PDF

José Manuel António¹
Alfredo Ngola Macutunda²
Tales Paiva Nogueira³

RESUMO

1 - Instituto de Engenharias e Desenvolvimento Sustentável; Universidade da Integração Internacional da Lusofonia Afro-Brasileira.

Apoio Financeiro: Unilab.

Palavras-chave: EXTRAÇÃO DE DADOS, PDF, DIÁRIOS OFICIAIS, PROCESSAMENTO DE DOCUMENTOS.
Grande área do CNPq: Ciências Exatas e da Terra

Resumo

Introdução: documentos no formato PDF são amplamente utilizados em diversos contextos. Um desses usos ocorre em publicações oficiais de atos governamentais, como os Diários Oficiais da União, dos Estados e dos Municípios. Embora legível para seres humanos, é um formato de arquivo difícil de ser processado por máquinas devido à estrutura orientada à apresentação visual, dificultando a extração automatizada de informações. Nesse contexto, a extração de dados textuais, tabulares e imagéticos de PDFs tornou-se um desafio relevante para a transparência pública e o acesso à informação. Objetivo: o presente trabalho teve como objetivo fazer o levantamento sistemático de ferramentas e bibliotecas disponíveis para extração de dados em documentos PDF, avaliando capacidades, limitações e aplicabilidade no contexto de Diários Oficiais, propondo uma metodologia de seleção baseada em critérios objetivos. Metodologia: foram conduzidas buscas sistemáticas nas bases ACM, IEEE, Scopus utilizando a string de busca ((PDF) AND (extraction OR manipulation OR handling) AND (tool OR library OR lib)) com recorte temporal de 2017 a 2026. A triagem dos documentos seguiu etapas de remoção de duplicatas, exclusão de teses, dissertações, livros e artigos de revisão, além da aplicação de critérios de exclusão, que consideraram a especificidade da ferramenta para extração de texto, imagem e tabelas, disponibilidade pública, atualização e gratuidade ou período de teste. Complementarmente, foi realizada uma busca no repositório de projetos GitHub. Foram selecionadas ferramentas dedicadas exclusivamente à extração de dados de PDF, disponíveis publicamente, tendo a última atualização (commit) no mínimo desde 1º de janeiro de 2026 e ter pelo menos 1000 “estrelas” (indicador de popularidade de repositórios na plataforma). Resultados: com relação à revisão da bibliografia, os levantamentos resultaram em 532 documentos acadêmicos iniciais, dos quais, após triagem e aplicação dos critérios de exclusão, 56 artigos foram considerados relevantes, abrangendo 11 ferramentas específicas para extração de dados em PDF. Dentre as ferramentas identificadas, foram selecionadas PyMuPDF, pdfplumber, docling, markitdown, pdfminer.six, marker e paddleOCR. Com relação à aplicação prática sobre os Diários Oficiais, a extração de dados foi realizada sobre um conjunto de dados composto por Diários Oficiais do município de São Francisco do Conde (BA). Observou-se que a combinação de diferentes ferramentas é necessária para lidar com a heterogeneidade estrutural dos documentos, incluindo variações de layout, presença de tabelas e imagens escaneadas. Conclusões: Conclui-se que não há uma única ferramenta que atenda a todas as necessidades de extração de dados em PDF, sendo recomendada uma abordagem híbrida que combine diferentes bibliotecas conforme o tipo de conteúdo e a estrutura do documento. Quanto à contribuição para “Diversidade, Inclusão e Transformação Social”, o trabalho tem potencial para ampliar o acesso da sociedade às informações públicas contidas nos Diários Oficiais, promovendo maior transparência governamental e empoderamento do cidadão.

Palavras-chave: EXTRAÇÃO DE DADOS; PDF; DIÁRIOS OFICIAIS; PROCESSAMENTO DE DOCUMENTOS.

Unilab, Palmares, Discente, jmantonio2000@gmail.com¹

Unilab, Palmares, Discente, alfredomacutunda@outlook.com²

Unilab, Palmares, Docente, tales@unilab.edu.br³