Automated PDF text extraction: An adaptive solution *

Jacquelin HOUNSOU *ISE ENEAM*

This paper presents an optimized pipeline for extracting text from PDF documents, handling both native text and scanned images using Optical Character Recognition `OCR`. Our approach adapts to varying document formats and provides structured output. We compare our method to existing tools and discuss its advantages, limitations, and future improvements.

Keywords: Extraction, Unstructured Documents, Natural Language Processing - NLP

Introduction

Extracting information from PDF documents is a critical task in document processing. PDFs come in various formats, including those with embedded text and scanned images, making automated extraction challenging. Traditional methods struggle with inconsistencies in formatting and OCR quality. This work proposes an adaptive pipeline that ensures accurate and structured text extraction.

Related Work

Existing solutions such as 'pdfplumber', 'PyMuPDF', and 'pdf2text' perform well on text-based PDFs but fail with scanned images. Tesseract OCR improves recognition but requires preprocessing for better accuracy. Hybrid approaches combining text extraction and OCR have been proposed, but they often lack automation for handling different PDF types dynamically.

Proposed Solution

Our approach automates text extraction by:

- Detecting whether the PDF contains embedded text or scanned images.
- Using 'pdfplumber' for text-based PDFs and Tesseract OCR for scanned PDFs.
- Cleaning and structuring the extracted text for better usability.
- Processing any file placed in the designated input folder without manual intervention.

*Replication files are available on the author's Github account (https://github.com/MARIEL-J/extraction_info). **Current version**: March 30, 2025; **Corresponding author**: hounsoujacquelin@gmail.com.

Results and Evaluation

We tested our pipeline on diverse PDFs, including research papers, invoices, and scanned documents. Our method achieved a higher accuracy in text recovery compared to standard extraction tools. The automation significantly reduced manual effort in document processing.

Limitations and Future Work

While our solution improves text extraction, challenges remain with complex layouts, tables, and equations. Future enhancements will include NLP-based text structuring and integration with an interactive API for user-friendly document processing.

Conclusion

We presented an adaptive solution for PDF text extraction that outperforms traditional methods by integrating OCR dynamically. Our approach ensures better accuracy and efficiency, making it suitable for various document processing applications.