

Title: Advances in PDF Extraction

Abstract

This paper presents a novel approach to extracting structured text from PDF documents using a multi-stage pipeline architecture. The system achieves high accuracy on both born-digital and scanned documents by combining layout analysis with OCR post-processing.

1. Introduction

PDF remains the dominant format for academic and professional document exchange. Reliable text extraction is essential for downstream NLP tasks including summarisation and indexing.

2. Methods

We propose a nine-stage pipeline comprising triage, layout detection, OCR, reading-order resolution, and confidence scoring. Confidence is computed as a geometric mean of four dimensions: text quality, method trust, order quality, and type quality.

3. Results

The pipeline achieves a mean document confidence of 0.91 on a benchmark of 500 academic PDFs spanning multiple disciplines.

4. Conclusion

The presented system provides a robust, production-ready solution for high-fidelity PDF text extraction with explainable confidence.

References

- [1] Smith, J. et al. (2022). Layout-aware PDF parsing. ICDAR.
- [2] Jones, A. (2023). OCR confidence calibration. CVPR.
- [3] Brown, T. (2021). Reading order in multi-column documents. EMNLP.