

Text Extraction Libraries for PDF (Python)

<i>Library</i>	<i>Best For</i>	<i>Notes</i>
PyPDF2	Basic text extraction	Simple but not good at preserving layout; cannot handle tables well.
PyMuPDF (fitz)	Text with layout + metadata	Extracts text by blocks, words, and coordinates — useful for table-like structures.
PDFMiner.six	Detailed text extraction	Fine-grained control over text positioning; steeper learning curve.
PDFPlumber	Text + structured extraction	Built on PDFMiner; provides easy access to words, lines, bounding boxes.
pdfrw	Low-level PDF manipulation	Good for editing/merging PDFs, but not ideal for extraction.
Slate	Wrapper over PDFMiner	Easier interface but largely outdated.
textract	Multiple file formats	Handles PDFs but relies on other tools; heavier dependency.