PDF parsing involves extracting structured or unstructured data (such as text, metadata, tables) from PDF files. This is essential for automating workflows like document classification, indexing, and renaming.

Key Techniques:

- Text Extraction: Libraries like PyPDF2 and pdfminer.six are popular for extracting raw text from PDFs.(Source)
- Optical Character Recognition (OCR): Tools like Tesseract or Parsio can convert scanned or image-based PDFs into machine-readable text.(Source)
- Template-Based Parsing: Useful for documents with consistent layouts, such as invoices or legal forms.
- AI-Powered Parsing: Advanced tools like Nougat or Table Transformer (TATR) use machine learning to handle complex documents, including those with tables or irregular layouts. (Source)

Best Practices:

- Use OCR for scanned documents to ensure text is machine-readable.
- Choose parsing tools based on document complexity and structure.(Source)
- Implement error handling to manage parsing failures gracefully.(Source)