

Free RPA PDF Toolkit

1. PDF Extraction Tools

Extract text, tables, images, and metadata from PDFs.

- PyPDF2 / PyPDF4
 - Description: Python libraries for reading, splitting, merging, and extracting text/metadata from PDFs.
 - Use Case: Extract text from invoices or reports.
 - Implementation:
 - python
 -
 - Copy
 -
 - Download
 - from PyPDF2 import PdfReader
 - reader = PdfReader("document.pdf")
 - text = reader.pages[0].extract_text()print(text)
- PDFMiner / pdfplumber
 - Description: Advanced text extraction (including complex layouts) and table detection.
 - Use Case: Scrape data from multi-column PDFs.
 - Example:
 - python
 -
 - Copy
 -
 - Download
 - import pdfplumber
 - with pdfplumber.open("document.pdf") as pdf:
 - page = pdf.pages[0]
 - table = page.extract_table()
- Tabula-py
 - Description: Extract tables from PDFs into pandas DataFrames.
 - Use Case: Convert PDF tables to Excel/CSV.
 - Implementation:
 - python
 -
 - Copy
 -
 - Download
 - import tabula
 - tables = tabula.read_pdf("document.pdf", pages="all")
 - tables[0].to_csv("table.csv")

2. PDF Conversion Tools

Convert PDFs to other formats (Word, Excel, images, etc.).

- pdf2docx
 - Description: Convert PDF to editable Word documents.
 - Use Case: Automate document format conversion.
- LibreOffice (CLI)
 - Description: Use LibreOffice's command-line interface to convert PDFs to Word, Excel, or HTML.
 - Command:
 - bash
 -
 - Copy
 -
 - Download
 - libreoffice --headless --convert-to docx "document.pdf"
- Poppler Utilities (pdftoppm, pdftotext)
 - Description: Convert PDFs to images (PNG/JPEG) or text files.
 - Use Case: Extract pages as images for OCR processing.

3. PDF Manipulation Tools

Merge, split, encrypt, or watermark PDFs.

- PyPDF2
 - Merge PDFs:
 - python
 -
 - Copy
 -
 - Download
 - from PyPDF2 import PdfMerger
 - merger = PdfMerger()
 - merger.append("file1.pdf")
 - merger.write("merged.pdf")
 - Encrypt PDFs: Add passwords to PDF files.
- QPDF (Command Line)
 - Description: Split, compress, or linearize PDFs.
 - Example: Split a PDF into single pages:
 - bash
 -
 - Copy
 -
 - Download
 - qpdf --split-pages=1 input.pdf output-page-%d.pdf



4. PDF Automation with RPA Tools

Integrate PDF tasks into RPA workflows.

- Robot Framework + RPA.PDF
 - Description: Open-source RPA framework with a dedicated PDF library.
 - Use Case: Validate PDF content in an automated workflow.
 - Example:
 - robotframework
 - Copy
 - Download
 - *** Tasks ***
 - Extract PDF Text\${text} = Get Text From Pdf invoice.pdf
 - Log \${text}
- UiPath Community Edition (Free Tier)
 - Use "PDF Activities": Built-in activities for PDF extraction, merging, and OCR.
 - Integration: Combine with Python scripts for advanced tasks.
- Automation Anywhere Community Edition
 - PDF Command: Extract text, tables, or metadata using pre-built commands.

5. OCR for Scanned PDFs

Extract text from scanned/image-based PDFs.

- Tesseract OCR + PyTesseract
 - Description: Open-source OCR engine. Use with pdf2image to convert PDF pages to images.
 - Implementation:
 - python
 - Copy
 - Download
 - from pdf2image import convert_from_path
 - import pytesseract
 - images = convert_from_path("scanned.pdf")
 - text = pytesseract.image_to_string(images[0])
- OCRmyPDF
 - Description: Add OCR layers to scanned PDFs.
 - Command:
 - bash
 - Copy
 - Download
 - ocrmypdf --deskew --rotate-pages scanned.pdf searchable.pdf

6. PDF Form Handling

Fill or extract data from interactive PDF forms.

- pdfforms
 - Description: Python library to fill PDF forms programmatically.
 - Use Case: Automate tax form submissions.
- Pdftk (Free Version)
 - Description: Fill and flatten PDF forms via CLI.

7. PDF Reporting

Generate PDF reports dynamically.

- ReportLab (Open-Source Edition)
 - Description: Create PDFs from scratch using Python.
 - Example:
 - python
 - Copy
 - Download
 - from reportlab.lib.pagesizes import letter
 - from reportlab.pdfgen import canvas
 - c = canvas.Canvas("report.pdf", pagesize=letter)
 - c.drawString(100, 700, "Hello, Automated World!")
 - c.save()
- WeasyPrint
 - Description: Convert HTML/CSS to PDF (great for web-based reports).

8. Advanced Tools

- DiffPDF
 - Compare two PDFs for differences (text or visual).
- Apache PDFBox
 - Java library for PDF extraction, splitting, and encryption (works with Jython in RPA).

Sample Workflow (Invoice Processing)

- Extract Text: Use pdfplumber to pull invoice numbers and totals.
- OCR Scanned PDFs: Run OCRmyPDF if the invoice is scanned.
- Convert to Excel: Use tabula-py to extract tables.
- Generate Report: Use ReportLab to create a summary PDF.
- Merge & Email: Combine files with PyPDF2 and send via SMTP.

Limitations of Free Tools

- Scalability issues for large PDFs.
- Limited support for complex layouts.
- OCR accuracy varies with scan quality.

For enterprise needs, consider paid tools like Adobe Acrobat SDK, ABBYY FineReader, or iText.

