

```
1 PDF EXTRACTOR - README
2 =====
3
4 Version: 1.0.0
5 Date: 2025-09-03
6
7 Overview
8 -----
9 Pdf Extractor is a Windows desktop tool that extracts *text* and *tables* from PDF
10 files.
11 It uses:
12 - PdfPig (for text)
13 - Tabula (tabula.jar via Java) for table detection
14 - A lightweight heuristic for simple key-value blocks
15 It can also export a structured JSON with page size, reading order ("flow"), table
16 bounding boxes, and rows.
17
18 What's Included
19 -----
20 This ZIP contains:
21 - `CheckExtractor.exe` - the app
22 - `Tools\tabula.jar` - the table extractor engine
23 - `jre\` - a portable Java Runtime (if bundled in this package)
24 - `README.txt` - this file
25 - (Optional) `LICENSE.txt`
26
27 System Requirements
28 -----
29 - Windows 10 or 11 (x64)
30 - ~200-250 MB free space
31 - No admin rights required when using the ZIP package
32 - If *no* `jre\` folder is included, a system Java Runtime (JRE 17 or newer) must be
33 installed and on PATH
34
35 How to Run (ZIP package)
36 -----
37 1) Extract the ZIP to a folder where you have write permissions (e.g.,
38 `C:\Apps\CheckExtractor\` or Desktop).
39 2) Open the folder and run `CheckExtractor.exe`.
40 3) If Windows SmartScreen warns about an "Unknown publisher", click More info → Run
41 anyway** (see Code Signing below).
42
43 Using the App
44 -----
45 - Browse: choose a PDF file
46 - Extract: shows the document text and any detected tables (rendered below).
47 Table markers are added in-text.
48 - JSON: outputs a JSON object with pages → flow (text + table refs), tables (id +
49 bbox + rows), and textWithMarkers
50 - Clear: resets the UI
51
52 Known Limitations
53 -----
54 - Scanned PDFs (image-only) require OCR (not included). If needed, we can add
55 Tesseract or a cloud OCR.
56 - Some unusual table layouts may not be detected by Tabula; a heuristic fallback
57 converts "Label Value" blocks into a 2-column table.
58 - Single-file publish is disabled intentionally to ensure `tabula.jar` and the local
59 `jre\` folder are available.
60
61 Privacy & Security
62 -----
63 - All processing happens locally on your machine. No data leaves your device.
64 - PDF files are only read; no modifications are written to your files.
65
66 Troubleshooting
67 -----
68 - App doesn't start / crashes: ensure you extracted the ZIP and are not running
69 from inside it.
70 - Tables don't appear / JSON shows `tables`: []`: the file may not contain
71 detectable tables. Try a different PDF. If the PDF has clean tables but still fails,
72 contact support with a sample.
73 - "Java not found" / "tabula.jar not found": ensure the ZIP contains
```

`jre\bin\java.exe` or install Java 17+; the `Tools\tabula.jar` must exist next to the exe.

- **SmartScreen warning:** the app is not code-signed yet. See Code Signing below.

Code Signing

Unsigned executables can trigger Windows SmartScreen. We plan to add a code-signed build to reduce warnings. In the meantime, use **More info → Run anyway** to proceed.

Support

Website: <https://softwaretools.online>

Contact: <https://softwaretools.online/contact>

Inquiry Form: <https://tally.so/r/wdOojy>