10 PDF Parsing Practice Problems (Easy to Medium)

1. Extract all text from a PDF using PyPDF2.

2. Extract text only from Page 1 of your PDF.

3. Count the total number of pages in your PDF.

4. Detect whether your PDF is scanned (empty text extraction).

5. Extract headings (lines in uppercase or bold-like text).

6. Extract tables from your PDF using pdfplumber.

7. Extract metadata (Author, Title, Creation Date) from your PDF.

8. Save extracted text from your PDF into a .txt file.

9. Extract only email IDs from your PDF text using regex.

10. Create a combined extractor:

- Try PyPDF2

- If empty → try pdfplumber

- If still empty → print 'Scanned PDF detected'