**受注伝票が100ページくらいのPDFをOCRで読み込んでCSV出力するとともにPostgreSQLにINSERTする処理の機能開発を行う。**

**処理の中で読み込んだPDFを、中に記載されている伝票番号をもとに分割して指定のフォルダに保存したい。**

**可能なやり方はある？可能な場合、その際に使用する開発言語は何になる？**

**使用する言語の優先順位としては、EXCEL-VBA、VB、C#、PHP、Python、Java。**

はい、**PDFをOCRで読み取り→伝票番号ごとに分割→CSV出力→PostgreSQLにINSERT→分割PDFを指定フォルダに保存**という一連の処理は、以下の技術で実現可能です。

**✅ 結論（要件に対応できる開発言語）**

| **優先順位** | **言語** | **実現性** | **備考** |
| --- | --- | --- | --- |
| 1 | VBA | △ | 外部ツール連携で実現可。処理負荷・柔軟性が低い。OCRやPDF分割はライブラリ任せ。 |
| 2 | VB | ○ | C#と同様、.NETライブラリ活用で十分対応可能。 |
| 3 | C# | ◎ | PdfSharp, Tesseract, iTextSharp など豊富な.NETライブラリあり。 |
| 4 | PHP | △ | OCRやPDF操作に向かない。サーバーサイド処理には適しているが非推奨。 |
| 5 | Python | ◎最適 | PyMuPDF, pdfplumber, pytesseract, psycopg2 で全処理がシンプルに可能。 |
| 6 | Java | ○ | Apache PDFBox, Tesseract, JDBC 等で対応可能。C#よりやや記述が複雑。 |

**✅ 推奨：Python + PostgreSQL**

Pythonが最も開発効率が高く、柔軟性・安定性ともに優れています。

**🔧 使用ライブラリと処理概要**

| **処理内容** | **使用ライブラリ** |
| --- | --- |
| PDF読込・分割 | PyMuPDF または pdfplumber |
| OCR（画像文字認識） | pytesseract + pdf2image |
| CSV出力 | Python標準の csv モジュール |
| DB INSERT | psycopg2（PostgreSQL用） |
| 伝票番号抽出 | OCR文字列から正規表現で取得 |

**フロー例（Python）**

1. PDF全体を pdf2image で画像化
2. 各ページを pytesseract でOCR実行
3. 正規表現で「伝票番号」抽出
4. 伝票番号ごとにページをグループ化し、PyMuPDF などで分割保存
5. OCRデータをCSV出力
6. PostgreSQLに psycopg2 でINSERT

**📌 Excel VBA や VB でやりたい場合は？**

**VBAの場合（かなり工夫が必要）**

* OCR → Microsoft OCR Engine または MODI (古いOfficeに限る)
* PDF分割 → Adobe Acrobat COM API or サードパーティコマンド（例：pdftk、gswin64c.exe など）
* CSV/DB操作 → 可能（VBA標準機能＋ADODB）

→ 現実的には「PythonやC#等で外部ツールを作成 → VBAから呼び出す」形が現実的。

**✅ 各言語の実現可能性まとめ**

| **言語** | **OCR** | **PDF分割** | **DB INSERT** | **コメント** |
| --- | --- | --- | --- | --- |
| VBA | △（MODI等） | △（外部ツール必須） | ○（ADO） | あくまで「呼び出し」役が現実的 |
| VB/C# | ○（Tesseract.NET） | ○（PDFSharp, iText） | ○（Npgsql） | 実装量は多いが実用的 |
| PHP | △（OCR弱い） | △（fpdf等は編集限定） | ◎ | OCRに外部処理必要 |
| Python | ◎ | ◎ | ◎ | 最もシンプルかつ柔軟 |
| Java | ○ | ○ | ○ | C#より若干冗長 |

**✅ 補足：伝票番号の抽出精度について**

OCR処理では誤認識の可能性があります。伝票番号は以下のように対策するのが望ましいです：

* 明確なパターン（例：DEN12345678）を正規表現で抽出
* ページ冒頭・特定位置のみチェック
* OCR結果をロギングして手動確認フェーズを設ける

**📎 開発環境テンプレ（Python例）**

pip install pytesseract pdf2image PyMuPDF psycopg2

import pytesseract, fitz, csv, psycopg2

from pdf2image import convert\_from\_path

**✅ まとめ**

| **機能要件** | **実現可否** | **備考** |
| --- | --- | --- |
| PDFから伝票番号抽出（OCR） | ◎ | pytesseractが最適 |
| 伝票番号ごとにPDF分割 | ◎ | PyMuPDFなどで可能 |
| OCR結果CSV出力 | ◎ | 標準ライブラリで可能 |
| PostgreSQLへのINSERT | ◎ | psycopg2などで簡単 |
| EXCEL-VBAによる全処理の実装 | △ | 外部ツール必須・現実的でない |

ご希望があれば、**PythonやC#でのテンプレコード**も用意できます。どの言語での実装を希望されますか？