以下に、**Windows上で動作するPythonフルプロジェクト**を再掲します。必要なファイルや構成とともに、**PyInstallerで .exe に変換可能な構成**になっています。

**✅ プロジェクト構成（例）**

pdf\_ocr\_splitter/

├─ input/

│ └─ orders.pdf ← 処理対象PDF

├─ output/ ← 分割PDF・CSV出力先（自動作成）

├─ tessdata/

│ └─ jpn.traineddata ← OCR用日本語モデル

├─ pdf\_process.py ← メインスクリプト

├─ requirements.txt ← 必要パッケージ一覧

└─ README.md ← 利用手順（任意）

**✅ 1. requirements.txt**

pytesseract

pdf2image

PyMuPDF

psycopg2-binary

Pillow

psycopg2-binary は PostgreSQL用ドライバです。

**✅ 2. Pythonコード：pdf\_process.py**

import os, re, csv

import pytesseract

import fitz # PyMuPDF

import psycopg2

from pdf2image import convert\_from\_path

# === 設定 ===

PDF\_PATH = "input/orders.pdf"

OUTPUT\_DIR = "output"

CSV\_PATH = os.path.join(OUTPUT\_DIR, "data.csv")

POPPLER\_PATH = r"C:\tools\poppler\bin" # Popplerのインストール先（環境に応じて変更）

TESSERACT\_PATH = r"C:\Program Files\Tesseract-OCR\tesseract.exe"

TESS\_LANG = "jpn"

# PostgreSQL接続情報

PG\_CONN = {

"host": "localhost",

"port": 5432,

"database": "your\_db",

"user": "your\_user",

"password": "your\_password"

}

# === 初期処理 ===

os.makedirs(OUTPUT\_DIR, exist\_ok=True)

pytesseract.pytesseract.tesseract\_cmd = TESSERACT\_PATH

# === DB接続準備 ===

conn = psycopg2.connect(\*\*PG\_CONN)

cur = conn.cursor()

# === CSV出力準備 ===

with open(CSV\_PATH, "w", newline="", encoding="utf-8") as csv\_file:

writer = csv.writer(csv\_file)

writer.writerow(["denno", "text"])

# PDF画像化

images = convert\_from\_path(PDF\_PATH, dpi=300, poppler\_path=POPPLER\_PATH)

denno\_pages = {} # 伝票番号ごとのページ保持

for i, image in enumerate(images):

temp\_img = os.path.join(OUTPUT\_DIR, f"page\_{i+1}.png")

image.save(temp\_img)

text = pytesseract.image\_to\_string(temp\_img, lang=TESS\_LANG)

match = re.search(r"(DEN\d{8})", text)

if match:

denno = match.group(1)

denno\_pages.setdefault(denno, []).append(i)

writer.writerow([denno, text])

cur.execute("INSERT INTO orders (denno, content) VALUES (%s, %s)", (denno, text))

# 分割PDF保存

doc = fitz.open(PDF\_PATH)

for denno, pages in denno\_pages.items():

new\_pdf = fitz.open()

for page\_num in pages:

new\_pdf.insert\_pdf(doc, from\_page=page\_num, to\_page=page\_num)

new\_pdf.save(os.path.join(OUTPUT\_DIR, f"{denno}.pdf"))

# DB反映

conn.commit()

cur.close()

conn.close()

print("完了しました。")

**✅ 3. .exe への変換手順（PyInstaller）**

**インストール**

pip install pyinstaller

**実行コマンド**

pyinstaller --onefile pdf\_process.py

**出力**

dist/pdf\_process.exe ← このファイルを配布・使用可能

**✅ 補足：Tesseract 言語データ配置方法**

tessdata フォルダに以下のように配置：

pdf\_ocr\_splitter/

├─ tessdata/

│ └─ jpn.traineddata

コマンドラインなどでOCRの言語として "jpn" を指定することで、**日本語OCR**が有効になります。