**Text Extraction and Export Tool**

This project is a web application built with Flask that allows users to upload PDF or image files, extract text from them using OCR (Optical Character Recognition), and export the extracted text as a PDF or Word document.

Features

File Upload: Supports uploading PDF and image files (PNG, JPG, JPEG).

PDF to Image Conversion: If a PDF is uploaded, each page is converted to an image.

OCR Text Extraction: Uses Tesseract OCR to extract text from each image.

Error Detection: Detects common issues in the extracted text, such as unrecognized symbols and empty pages, and displays these errors to the user.

Text Cleaning: Cleans up extra line breaks in the extracted text to improve readability.

Export Options: Allows users to export the cleaned text to PDF or Word format.

Setup Instructions

Install Python: Make sure Python is installed on your system.

Install Required Packages: Use the following command to install the required libraries:

bash

requirements:

Flask

pytesseract

PyMuPDF

fpdf

python-docx

re

Configure Tesseract: Ensure Tesseract OCR is installed on your system. You may need to set the path for Tesseract in your code or system settings if not already configured.

Run the Application:

bash

Copy code

python app.py

Access the Application: Open your web browser and go to http://127.0.0.1:5000 to access the tool.

File Structure

app.py: The main Flask application file, containing routes and logic for file upload, text extraction, error handling, and export.

templates/index.html: The front-end interface where users can upload files, view extracted text, and export data.

static/style.css: Optional CSS for styling the front-end.

uploads/: Folder for storing uploaded files temporarily.

exports/: Folder where the exported PDF and Word files are saved.

How to Use

Upload File: Select a PDF or image file to upload.

Extract Text: The extracted text will appear in the text area along with any errors found during extraction.

Export Text: Choose to export the extracted text to PDF or Word format by clicking the corresponding button.

Error Handling

If the OCR process encounters any issues, such as unrecognized symbols or no extracted text, the errors will be displayed on the page to help identify any problems.