PDF to Text OCR Converter

Overview

This Python script extracts text from PDF files by converting them into images and applying Optical Character Recognition (OCR) using Tesseract. The extracted text is saved into a `.txt` file for easy access and further processing.

Pipeline Explanation

1. **Input File Validation**:
   - Validates the existence of the PDF file before proceeding.

2. **Tesseract Setup**:
   - Configures Tesseract OCR executable path (necessary on Windows).
   - Verifies the presence of the Tesseract installation.

3. **PDF-to-Image Conversion**:
   - Converts each page of the PDF to an image using the `pdf2image` library.
   - Poppler binaries are required for this process on Windows.

4. **Text Extraction**:
   - Processes each image with Tesseract OCR using the `--psm 6` configuration (suitable for single-column text).
   - Extracted text is stored in memory.

5. **Output Handling**:

   - Compiles text from all pages and writes it to a `.txt` file in UTF-8 encoding.

   - Includes page breaks (`--- Page X ---`) in the output text.

6. **Error Handling**:

   - Captures and reports issues with missing files, Tesseract setup, or conversion errors.

How to Run the Code

Prerequisites

- Python 3.x

- Required Python libraries:

  - `pdf2image`

  - `pytesseract`

- Installed software:

  - [Tesseract OCR](https://github.com/tesseract-ocr/tesseract)

  - [Poppler for Windows](http://blog.alivate.com.au/poppler-windows/) (Windows users only)

Steps

1. Clone or download this script.

2. Install required Python packages:

   ```bash

   pip install pdf2image pytesseract

   ```

3. Install Tesseract OCR:

   - On Windows: Download and install from [Tesseract GitHub](https://github.com/tesseract-ocr/tesseract).
   - On Linux/Mac: Use package managers like `apt` or `brew`.

4. Configure Tesseract path in the script:

   ```python
   pytesseract.pytesseract.tesseract_cmd = r'C:\Path\To\Tesseract\tesseract.exe'
   ```

5. Run the script:

   ```bash
   python script_name.py
   ```

6. Provide the input PDF path, output text file path, and Poppler path (Windows only).

Example Usage:

```python
pdf_path = r"C:\Users\Example\Downloads\sample.pdf"
output_txt_path = r"C:\Users\Example\Desktop\output.txt"
poppler_path = r"C:\Path\To\Poppler\bin"

convert_pdf_to_text(pdf_path, output_txt_path, poppler_path=poppler_path)
```

Debugging and Logs

- Prints the number of pages processed.

- Displays the first 100 characters of extracted text per page for quick verification.

Future Enhancements

- Add support for batch processing of multiple PDF files.

- Implement evaluation metrics (e.g., CER, WER) for OCR accuracy.

- Integrate preprocessing for handwritten text recognition.

License