PDF to DOCX Conversion with OCR  
  
This Python script converts PDF files to DOCX files using OCR (Optical Character Recognition)  
for text extraction.  
  
Table of Contents  
  
Overview  
Requirements  
Usage  
Configuration  
Dependencies  
Contributing  
License  
  
Overview  
  
The script utilizes the pdf2image library to convert PDF pages to images, pytesseract for  
OCR, and the docx library for creating Word documents. It iterates through PDF files in a  
specified directory, performs OCR on each page, and generates a corresponding DOCX file.  
Requirements  
  
e Python 3  
  
Install required Python packages:  
pip install pdf2image docx pytesseract  
  
e  
o Ensure Tesseract OCR is installed and in your system's PATH.  
  
Usage  
1. Place your PDF files in the raw data directory.  
  
Run the script:  
python pdf\_to\_docx\_with\_ocr.py  
  
2.

3. Converted DOCX files will be saved in the output directory.  
  
Configuration  
  
e Customize the input (raw data) and output (output) directories in the script.  
  
Dependencies  
e pdf2image: Converts PDF pages to images.  
  
e pytesseract: Python wrapper for Tesseract OCR engine.  
e docx: Python library for creating Word documents.  
  
Contributing  
  
Feel free to contribute by opening issues or submitting pull requests. Your feedback and  
contributions are welcome!  
  
License  
  
This project is licensed under the MIT License - see the LICENSE file for details.