Team 78 Neo
PDF Data Masking Application:
Documentation + Setup

About:

This web page can mask both text and images inside pdfs to identify and mask the following data:

- Names: in English, Chinese, Korean, Arabic and Malay
- Phone numbers: in standard, domestic and international formats
- Emails
- Hospitals and clinic names

Tools used:

PyMuPDF - extract text from pdf

Pytesseract OCR - Extract text in any language from image and find coordinates of image text to be masked

PIL - Apply image filters for OCR

OpenCv - opening and masking text pixels in image

re - identify email formats, phone no. formats and hospital/clinic mentions

XLM Roberta NER Model - Identifying names in any language from text

PyMuPDF (fitz)

- **What it is**: PyMuPDF is a Python binding for MuPDF, a lightweight PDF and XPS viewer.
- **How it's used**: In this project, PyMuPDF is used to extract text and images from PDF documents. It allows you to load a PDF, read its content page by page, and manipulate it, such as adding redaction annotations to mask sensitive information.

Pytesseract OCR

- **What it is**: Pytesseract is a Python wrapper for Google's Tesseract-OCR Engine, which is an optical character recognition tool.
- **How it's used**: Pytesseract OCR is used to extract text from images within the PDF. It supports multiple languages, making it suitable for processing text in various languages.

Pillow (PIL)

- **What it is**: Pillow is a Python Imaging Library that adds image processing capabilities to your Python interpreter.
- **How it's used**: In this project, Pillow is used to open images extracted from PDFs and apply filters to enhance the images for better OCR results. It also converts images to formats suitable for further processing.

OpenCV

- **What it is**: OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision.
- **How it's used**: OpenCV is used to manipulate image pixels, such as masking text areas in images. It converts images between different color spaces and performs operations like drawing rectangles over text to mask it.

re (Regular Expressions)

- **What it is**: The `re` module provides support for regular expressions in Python.
- **How it's used**: Regular expressions are used to identify patterns in text, such as email addresses, phone numbers, and mentions of hospitals or clinics. These patterns are compiled and used to search for and mask sensitive information in both the text and images of the PDF.

XLM-RoBERTa NER Model

- **What it is**: XLM-RoBERTa is a multilingual variant of the RoBERTa model, a transformer-based model for natural language understanding tasks. The NER (Named Entity Recognition) model identifies entities like names, organizations, and locations in text.
- **How it's used**: The XLM-RoBERTa NER model is employed to identify personal names in various languages from the text extracted from the PDF. The identified names are then masked to protect sensitive information.

Integration in the Project

- **PyMuPDF** is used to load the PDF and extract text and images from each page.
- **Pytesseract OCR** extracts text from images within the PDF, enabling you to find and mask sensitive information present in images.
- **Pillow** processes images before OCR to improve text extraction quality and also helps in saving manipulated images.
- **OpenCV** handles the masking of text within images by drawing black rectangles over sensitive text.
- **re** is used to compile patterns for recognizing and masking emails, phone numbers, and mentions of "Clinic" or "Hospital" along with the preceding word.
- **XLM-RoBERTa NER Model** detects personal names in various languages from the text, which are then added to the list of patterns to be masked.

Setup (Windows):

Step 1:
Install required libraries by running pip command in terminal:

```
pip install streamlit pymupdf pytesseract pillow opencv-python-headless numpy transformers
```

Step 2:
install Tesseract-OCR:

Download the Tesseract-OCR installer from this link:
https://github.com/UB-Mannheim/tesseract/wiki

Run the installer and follow the installation instructions to download Tesseract-OCR under the path C:\Program Files\Tesseract-OCR

Step 3:
Add the Tesseract-OCR installation path (C:\Program Files\Tesseract-OCR) to your system PATH.

To add the Tesseract-OCR installation path to your system PATH on Windows:

1. Open the Start Menu and search for "Environment Variables."
2. Select "Edit the system environment variables."
3. In the System Properties window, click "Environment Variables..."
4. Under "System variables," find and select "Path," then click "Edit..."
5. In the Edit Environment Variable window, click "New" and enter C:\Program Files\Tesseract-OCR.
6. Click "OK" to close each window.
7. Open a new Command Prompt window and type echo %PATH% to verify the addition.

Step 4:
Download pytesseract trained model addons for the languages chinese, arabic, malaysian, and korean using the following links:
Arabic: https://github.com/tesseract-ocr/tessdata/blob/main/ara.traineddata
Chinese simplified: https://github.com/tesseract-ocr/tessdata/blob/main/chi_sim.traineddata
Chinese Traditional: https://github.com/tesseract-ocr/tessdata/blob/main/chi_tra.traineddata
Korean: https://github.com/tesseract-ocr/tessdata/blob/main/kor.traineddata
Malaysian: https://github.com/tesseract-ocr/tessdata/blob/main/msa.traineddata

Once downloaded, store all these files under C:\Program Files\Tesseract-OCR\tessdata

Step 5:
Extract zip file
Go to main.py
Change paths mentioned in the code if needed according to how they’re stored in computer.

Step 6:
Run main.py
A command similar to this will be returned as output:

```
streamlit run c:\PDF Masking\main.py
```

Paste this command in terminal with path enclosed in double quotes.

```
streamlit run “c:\Code\PDF Masking\main.py”
```

This command opens the web page made using streamlit framework.

Step 7:
Click on “Browse files” button to upload desired pdf.
Click on “Download Masked PDF” button to download the masked pdf.
To upload another file, click on x symbol next to last uploaded file name and click on “Browse files” button again to upload another pdf.