File Level Redaction & Privacy Tool

Documentation for @likithreddy10

Project Overview

This is a Flask-based web application designed to automatically redact sensitive information (PII) from PDF documents and blur faces in images. The tool is optimized for deployment on Render and utilizes PyMuPDF for document manipulation and OpenCV for image processing.

Current Status Functional (Searchable PDFs Only)

- **User Authentication:** Secure login and registration system.
- **PDF Redaction:** Automatic black-box redaction of PII using RegEx patterns on **selectable/searchable** text.
- **Image Processing:** Face detection and Gaussian blurring for .jpg and .png files.
- **Secure Storage:** Deployment-ready configuration with environment variables.

Test Cases & Redaction Logic

Category	Pattern Type	Example Format	Stat
Phone Numbers	10-digit Indian Mobile	9876543210	✅
Aadhaar Card Number	12-digit spaced	1234 5678 9012	✅
Aadhaar Card Number	12-digit continuous	123456789012	✅
Email Address	Standard format	user@example.com	✅
PAN Card	Indian Tax ID	ABCDE1234F	✅
Keywords	Case-insensitive	Confidential, OTP	✅
Faces	Human face detection	.jpg, .png	✅

Known Issues / Current Limitations

- **Scanned Documents (OCR required):** The tool **cannot** redact text inside a scanned PDF (e.g., a photo of an Aadhaar card saved as a PDF). It currently only "sees" text that can be highlighted or selected.
- **Handwritten Text:** It cannot detect or redact handwritten names or numbers.
- **Complex Layouts:** In extremely complex PDF layouts (multi-column), coordinates for black boxes may occasionally shift by a few pixels.

Environment Variables Required

To run this project securely on Render, the following variables must be set:

- SECRET_KEY: A unique string for session security.
- DATABASE_URL: Connection string for the database (defaults to SQLite if not provided).
- PYTHON_VERSION: Set to 3.10.12.

Future Work & Roadmap

1. **OCR Integration (Priority):** Implement Tesseract OCR or EasyOCR to convert scanned images/PDFs into machine-readable text.
2. **Custom Pattern Matching:** Allow users to upload a list of specific names or IDs they want to redact.
3. **Cloud Storage:** Move from local storage to AWS S3 or Cloudinary for permanent file persistence.
4. **AI-Based PII Detection:** Replace RegEx with an NLP model (like SpaCy or Microsoft Presidio).