State-Archives-Transcription

An app for OCR-based transcription of scanned, handwritten archives.

Table of Contents

Features

- PDF -> images -> text via Microsoft TrOCR
- Simple web UI for uploading/downloading transcriptions
- Local database of transcribed files

Prerequisites

- **Docker Desktop**
- **Python** 3.8+
- **Git**
- **pip** (install guide)
- **virtualenv** (optional, recommended)

Getting Started

Clone

```
git clone https://github.com/ksermon/State-Archives-Transcription.git
cd State-Archives-Transcription
```

Docker

From the project root run:

```
docker compose up --build
```

Open your browser at http://localhost:5000.

Note: The first build will download dependencies and model files, it will take much longer the first time.

Quick Start (Subsequent Runs)

From the project root again, since your image is already built, run:

```
docker compose up
```

Usage

1. Upload a PDF or image
2. Wait for OCR processing (page will reload automatically when complete)
3. Copy the resulting plain-text ***[export text and JSON coming soon]***

Project Structure

```
├── app/
│   ├── main/           # Flask blueprints (routes, errors, utils)
│   ├── models/         # SQLAlchemy ORM
│   └── utils/          # OCR engine, preprocessing
├── app/models/         # local TrOCR weights
└── entrypoint.sh       # container startup logic (DB init, etc.)
```

Contributing

(for Keeley, by Keeley, I will remember this)

1. Add or find an issue
2. Create a branch (`git checkout -b feat/xyz)
3. Commit your changes (git commit -m "Add xyz")
4. Push and open a PR

License