

Medical Prescription Text Extraction from PDFs

Prashanth Patil

LinkedIn: https://www.linkedin.com/in/prashanthpatil879/

Table of Contents

[Project Overview: 2](#_Toc178276428)

[Objective: 2](#_Toc178276429)

[Tools & Technologies Used: 2](#_Toc178276430)

[Project Workflow: 3](#_Toc178276431)

[Implementation Details: 4](#_Toc178276432)

[**Step 1: Converting PDFs to Images** 4](#_Toc178276433)

[**Step 2: Processing the Images** 4](#_Toc178276434)

[**Step 3: Extracting the Images** 5](#_Toc178276435)

[**Step 4: Regular Expression for Extracting Key Information** 5](#_Toc178276436)

[**Step 5: Saving Extracted Data** 6](#_Toc178276437)

[Challenges & Solutions: 7](#_Toc178276438)

[Conclusion: 7](#_Toc178276439)

# **Project Overview:**

This project is focused on automating the process of extracting key information from medical prescriptions stored in PDF format. The PDFs are first converted to images, followed by text extraction using Optical Character Recognition (OCR) and regular expressions to parse and structure the required details.

# **Objective:**

To extract key data such as doctor name, patient name, prescription date, address, medications, and refill details from PDF prescriptions by converting them into images, applying preprocessing steps, and using text extraction and regular expressions.

# **Tools & Technologies Used:**

* **Python Libraries**:
* pdf2image: To convert PDF files into images.
* Pillow: For image processing and saving the output.
* OpenCV: For image preprocessing and thresholding.
* PyTesseract: For OCR (Optical Character Recognition) to extract text from images.
* re: For parsing text using regular expressions.
* **Tesseract-OCR**: OCR engine for extracting text from images.
* **Poppler**: A PDF rendering library to assist with the conversion of PDF to images.
* **Operating System**: macOS

# **Project Workflow:**

* **Converting PDF files to Images**: Using pdf2image to convert each page of a PDF into an image.
* **Image Preprocessing**: Preprocessing the images using OpenCV to improve OCR accuracy.
* **Extracting Text from Images**: Extracting text from preprocessed images using PyTesseract.
* **Applying Regular Expressions**: Using regular expressions to identify and extract key data from the text.
* **Saving Extracted Data**: Saving the extracted data into separate .txt files for each processed image.

A screenshot of a cell phone

Description automatically generated

# **Implementation Details:**

## **Step 1: Converting PDFs to Images**

We use the pdf2image library to convert each page of the PDF into an image, which can then be processed further. The *convert\_from\_path* function is used to load the PDF, and the images are stored in a list. Each page from the PDF will be stored in the pages list as an image in array format.

A screenshot of a computer

Description automatically generated

## **Step 2: Processing the Images**

Before extracting text using OCR, we preprocess the images to enhance the text clarity. This includes converting the image to grayscale and resizing it to make the text more legible.

A screenshot of a computer program

Description automatically generated

## **Step 3: Extracting the Images**

After processing, the image is passed through PyTesseract to extract text. We also specify the Tesseract command path.

A computer screen shot of a code

Description automatically generated

**IMAGE:**

**TEXT:**

Dr John Smith, M.D

2 Non-Important Street,

New York, Phone (000)-121-2222

Name: Marta Sharapova Date: 2/11/2022

Address: 9 tennis court, new Russia, DC

K

Prednisone 20 mg

Lialda 2.4 gram

Directions:

Prednisone, Taper 5 mg every 3 days,

Finish in 2.5 weeks -

Lialda - take 2 pill everyday for 1 month

Refill: 2 times

A prescription paper with a shadow of a person's hand

Description automatically generated

## **Step 4: Regular Expression for Extracting Key Information**

We use regular expressions to extract structured data (like doctor’s name, patient name, etc.) from the extracted text.

A screenshot of a computer program

Description automatically generated

## **Step 5: Saving Extracted Data**

Finally, the extracted data is saved into a .txt file for each processed PDF page.

A screenshot of a computer program

Description automatically generated

**Final Text:**

Doctor: John Smith

Patient Name: Marta Sharapova Date

Date: 2/11/2022

Address: 9 tennis court, new Russia, DC

Prednisone 20 mg

Lialda 2

Medications:

- Prednisone: 20 mg

- Lialda: 2.4 gram

- Taper: 5 mg

Refill: 2 times

# **Challenges & Solutions:**

* **Challenge**: Low-quality images causing OCR to misread text.
  + **Solution**: Applied image preprocessing techniques like grayscale conversion and adaptive thresholding to enhance text visibility.
* **Challenge**: Handling varying formats of prescription documents.
  + **Solution**: Using regular expressions to capture a broad range of common patterns for data extraction.

# **Conclusion:**

This project successfully demonstrates the end-to-end pipeline from converting PDF prescriptions to images, applying image processing, extracting textual information using OCR, and parsing the required data using regular expressions. This automated solution significantly reduces the manual effort of extracting data from prescription documents.