Python for Pdf

Umer Farooq Follow
Jan 23 · 3 min read

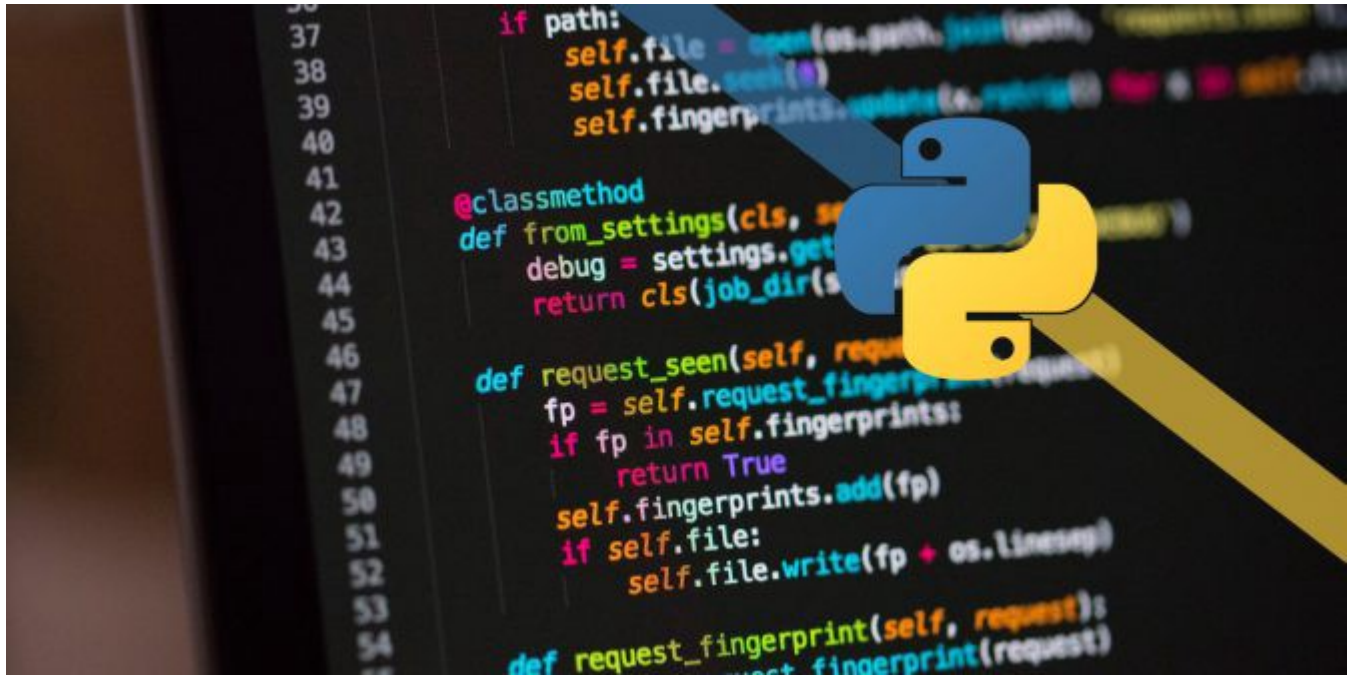

Table of content

- Introduction
- Why Python for PDF processing
- Common Python Libraries
- Extracting Text from pdf
- Reading the Table data from pdf
- Export Pdf into Excel
- Further Reading

Introduction

Being a high-level, interpreted language with a relatively easy syntax, Python is perfect even for those who don't have prior programming experience. Popular Python libraries are well integrated and provide the solution to handle unstructured data sources like Pdf and could be used to make it more sensible and useful

PDF is one of the most important and widely used digital media. used to present and exchange documents. PDFs contain useful information, links and buttons, form fields, audio, video, and business logic.

Why Python for PDF processing

PDF processing comes under text analytics. Most of the Text Analytics Library or frameworks are designed in Python only. This gives leverage on text analytics. Once you extract the useful information from PDF you can easily use that data into any Machine Learning or Natural Language Processing Model.

Common Python Libraries

Here is the list of some Python Libraries could be used to handle PDF files

1. **PDFMiner** is a tool for extracting information from PDF documents. Unlike other PDF-related tools, it focuses entirely on getting and analyzing text data.
2. **PyPDF2** is a pure-python PDF library capable of splitting, merging together, cropping, and transforming the pages of PDF files. It can also add custom data, viewing options, and passwords to PDF files. It can retrieve text and metadata from PDFs as well as merge entire files together.
3. **Tabula-py** is a simple Python wrapper of tabula-java, which can read the table of PDF. You can read tables from PDF and convert into pandas'

DataFrame. tabula-py also enables you to convert a PDF file into CSV/TSV/JSON file.

4. **Slate** is wrapper Implementation of PDFMiner
5. **PDFQuery** is a light wrapper around pdfminer, lxml and pyquery. It's designed to reliably extract data from sets of PDFs with as little code as possible.
6. **xpdf** Python wrapper for xpdf (currently just the "pdftotext" utility)

Extracting Text from pdf

First, we need to Install the

```
!pip install PyPDF2
```

Following is the code to extract simple Text from pdf using PyPDF2

```
# modules for

import PyPDF2

# pdf file object
# you can find find the pdf file with complete code in below
```

```
pdfFileObj = open('example.pdf', 'rb')

# pdf reader object

pdfReader = PyPDF2.PdfFileReader(pdfFileObj)

# number of pages in pdf

print(pdfReader.numPages)

# a page object

pageObj = pdfReader.getPage(0)

# extracting text from page.
# this will print the text you can also save that into String

print(pageObj.extractText())
```

You can read more Details from here

Reading the Table data from pdf

In order to work with the Table data in Pdf, we can use Tabula-py

```
pip install tabula-py
```

Following is the code to extract simple Text from pdf using PyPDF2

```
import tabula


# readinf the PDF file that contain Table Data
# you can find find the pdf file with complete code in below
# read_pdf will save the pdf table into Pandas Dataframe


df = tabula.read_pdf("offense.pdf")


# in order to print first 5 lines of Table


df.head()
```

If you Pdf file contain Multiple Table

```
df = tabula.read_pdf(“offense.pdf”,multiple_tables=True)
```

you can extract Information from the specific part of any specific page of PDF

```
tabula.read_pdf("offense.pdf", area=(126,149,212,462), pages=1)
```

If you want the output into JSON Format

```
tabula.read_pdf("offense.pdf", output_format="json")
```

Export Pdf into Excel

you can us Below code to convert the PDF Data into Excel or CSV

```
tabula.convert_into("offense.pdf", "offense_testing.xlsx",
output_format="xlsx")
```

Further Readings

you can find the complete code and Pdf files in This Github Link

1. This question on StackOverflow also has a lot of useful link in its Answer How to extract table as text from the PDF using Python?
2. Working with PDF files in Python using PyPDF2
3. Working with PDF and Word Documents

Python Pdf Artificial Intelligence Text Mining Data Science