

★ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



# A Step-by-Step Guide to Parsing PDFs using the pdfplumber Library In Python



Azhar Sayyad · [Follow](#)

2 min read · Jan 16, 2023



91



3



A Step-by-Step Guide to Parsing PDFs using the pdfplumber Library.



In this Tutorial, we will be looking the process of using the pdfplumber library in Python to parse PDFs. pdfplumber is a powerful library that allows

for easy extraction of text and data from PDFs, making it a valuable tool for data analysis and automation tasks.

**Step 1: Install the pdfplumber Library** To begin, you will need to install the pdfplumber library. This can be done using pip by running the command:

```
pip install pdfplumber
```

**Step 2: Import the library** Once the library is installed, you can import it into your Python script by using the following command:

```
import pdfplumber
```

**Step 3: Open the PDF** To open a PDF, you will need to create a pdfplumber.PDF object by passing the path to the PDF file to the open() function. For example:

```
with pdfplumber.open("path/to/pdf") as pdf:
```

**Step 4: Extract Text** pdfplumber provides several methods for extracting text from a PDF. The simplest method is the `extract_text()` method, which

returns a string containing all the text in the PDF. For example:

```
text = pdf.extract_text()
print(text)
```

Step 5: Extract Data pdfplumber also provides several methods for extracting data from a PDF. One such method is the `extract_table()` method, which returns a list of lists containing the data from tables in the PDF. For example:

```
tables = pdf.extract_table()
print(tables)
```

Step 6: Extract Images pdfplumber also allows you to extract images from a PDF. This can be done using the `get_image()` method, which returns an object containing the image data and meta-data. For example:

```
images = pdf.get_image()
print(images)
```

Complete Code:

```
import pdfplumber

# Open the PDF
with pdfplumber.open("path/to/pdf") as pdf:
```

```
# Extract the text
text = pdf.extract_text()
print(text)

# Extract the data
tables = pdf.extract_table()
for table in tables:
    print(table)

# Extract the images
images = pdf.get_images()
for image in images:
    print(image["page_number"])
    with open(f"image_{image['page_number']}.jpg", "wb") as f:
        f.write(image["data"])
```

In this guide, we have covered the basics of using the pdfplumber library to parse PDFs in Python. With pdfplumber, you can easily extract text, data, and images from PDFs, making it a valuable tool for data analysis and automation tasks you can use regular expression (RegExp ) to find particular text or string from extracted data.

You can now use this information to parse your own PDFs, and extract the information you need from it.

More Reference you Can watch This Video :

## [19] Convert a multi-page PDF file into csv / excel with Python



Python

Pdf

Csv

Pdfplumber

Extraction



**Written by Azhar Sayyad**

18 Followers · 8 Following

Follow

Just a software dev trying to make sense of the code chaos. Here to share the 'wisdom' I've picked up between debugging marathons.

## Responses (3)



Pranavdg

What are your thoughts?



Xc

Jul 1, 2024



There might be a problem when the pdf file is more than one page. We need to add " for page in pdf: " so that 'extract\_text' etc. can be successfully used,



8



1 reply

[Reply](#)



LordZhiHao

Aug 29, 2024



Great tutorial! Short and concise, and exactly what I needed for my project. Thanks and cheers!



2

[Reply](#)



Stephen Pace

Jul 11, 2024



Great example! For your next one, can you do a demo of where you convert a PDF that is text (but text in an image), use `get_image()` to get it, and then do an (OCR?) `image_to_text` to output the text?



1




1 reply

[Reply](#)

## More from Azhar Sayyad

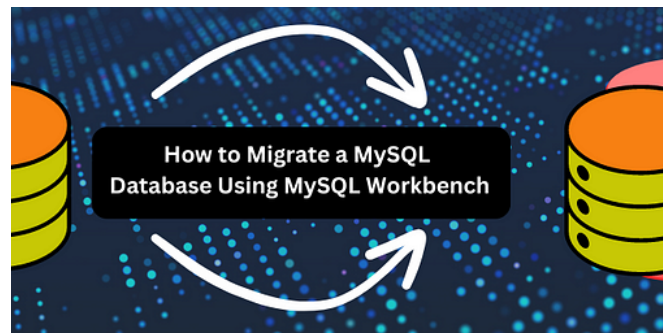



 Azhar Sayyad

## Extracting Data from Scanned PDFs Using OCR with Tesseract in...

Quick Backstory: A Task That Turned into an Adventure

★ Oct 22, 2024 🖱️ 2




 Azhar Sayyad

## How to Migrate a MySQL Database Using MySQL Workbench

Migrating a MySQL database can be crucial for various reasons, such as upgrading...

Sep 19, 2024 🖱️ 9

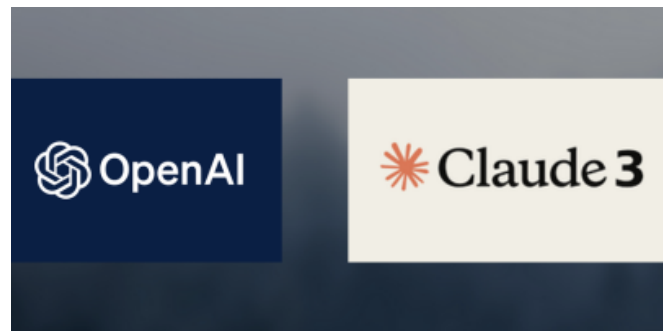



 Azhar Sayyad

## Top 5 Open-Source PDF Scraping Tools Using Python

PDF scraping is often tricky since PDFs aren't designed for easy data extraction. Luckily...

★ Oct 25, 2024



 Azhar Sayyad

## The Lazy Dev's Guide to AI: Work Smarter, Not Harder

Let's face it — we developers are inherently lazy (in a good way). We automate repetitive...

★ Dec 8, 2024



See all from Azhar Sayyad

Recommended from Medium




 Sudarshan Koirala

**Docling from IBM | Open Source Library To Make Documents AI...**

End to End Guide for Complete Beginners

★ Feb 15  51  



 Nutan

**How to Extract Images from a PDF in Python**

In this blog, we will extract images from a pdf file using pymupdf python module.

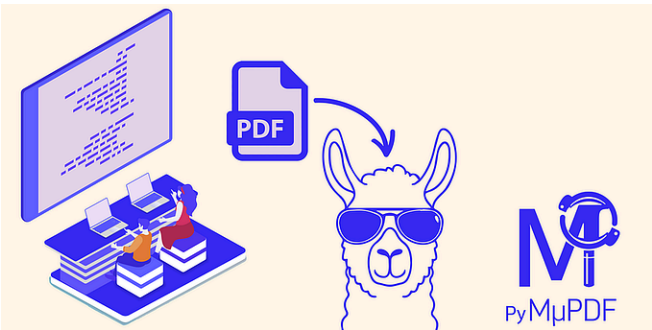
Nov 21, 2024  2  




 Jonathan Tan

**Extracting Structured Data from Word Documents to CSV Using...**

In educational content creation, research, and administrative processes, Word documents...



 DhanushKumar

**Using PyMuPDF4LLM: A Practical Guide for PDF Extraction in LLM &...**

Extracting and processing text from PDFs for machine learning, LLMs, or RAG setups can...



★ Oct 30, 2024



Oct 30, 2024

👏 113

💬 3



```
pdf_file = 'clcoding.pdf'
docx_file = 'clcoding.docx'
cv = Converter(pdf_file)
cv.convert(docx_file)
cv.close()
```

#source code --> [clcoding.com](https://clcoding.com)

```
[INFO] Start to convert clcoding.pdf
[INFO] [1/4] Opening document...
[INFO] [2/4] Analyzing document...
```

/c  
Pythoncoding  
clcoding



Python Coding

## PDF to docx using Python

The above Python code uses the pdf2docx library to convert a PDF file into a DOCX...

★ Nov 21, 2024

👏 15

💬 2



★ Mar 9

👏 695



Anoop Maurya

## Ollama-OCR Now Supports PDFs!



Stuck behind a paywall? Read for Free!

See more recommendations