

To make Medium work, we log user data. By using Medium, you agree to our [Privacy Policy](#), including cookie policy.

TUTORIAL

This is How I Convert PDF to Markdown

Better than any online tool out there



Suman Sourabh  · [Follow](#)

5 min read · Sep 7, 2024



481



5



To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

Yc

I have used multiple online tools to convert PDF documents to Markdown format, but none of them came close to Marker.

Along with basic Markdown conversion, it formats tables, converts most equations to latex, extracts and stores images.

Here's how I use Marker to extract PDF content and convert them into valid Markdown.

Environment Used

Windows 11

Prerequisites

As per Marker's GitHub repository, it requires the installation of:

- Python
- PyTorch

Installation

You'll need python 3.9+ and PyTorch. You may need to install the CPU version of torch first if you're not using a Mac or a GPU machine. See [here](#) for more details.

1. Install Python > 3.8

Go to Python Downloads page and download the latest version of Python.



Install the setup by following the instructions.

2. Install PyTorch

***Note:** For PyTorch to be installed correctly, you must have Python 3.8 or higher installed on your system.*

To install PyTorch, go to its [official website](#) and you will see something like the image below:

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

Your OS

Linux

Mac

Windows

Package

Conda

Pip

LibTorch

Source

Language

Python

C++ / Java

Compute Platform

CUDA 11.8

CUDA 12.1

CUDA 12.4

ROCm 6.1

CPU

Run this Command:

```
pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118
```

You can tweak those options to see which one would fit best for your system. Once you have your command, open PowerShell or Command Prompt and paste your command there.

Here's the command that I used to install PyTorch:

```
pip3 install torch torchvision torchaudio --index-url https://download.pytorch.o
```

PyTorch will begin installing on your system...

Looking in indexes: https://download.pytorch.org/whl/cu118

Collecting torch

Downloading https://download.pytorch.org/whl/torch-2.4.1%2Bcu118-cp312-cp312-win_amd64.whl (4.0 MB)

----- 4.0/4.0 MB 5.5 MB/s eta 0:00:00

Collecting torchaudio

Downloading https://download.pytorch.org/whl/torchaudio-2.4.1%2Bcu118-cp312-cp312-win_amd64.whl (4.0 MB)

----- 4.0/4.0 MB 5.5 MB/s eta 0:00:00

Collecting filelock (from torch)

Downloading https://download.pytorch.org/whl/filelock-3.13.1-py3-none-any.whl (11 kB)

Collecting typing-extensions>=4.8.0 (from torch)

Downloading https://download.pytorch.org/whl/typing_extensions-4.9.0-py3-none-any.whl (32 kB)

Collecting sympy (from torch)

Downloading https://download.pytorch.org/whl/sympy-1.12-py3-none-any.whl (5.7 MB)

----- 5.7/5.7 MB 5.9 MB/s eta 0:00:00

Collecting networkx (from torch)

Downloading https://download.pytorch.org/whl/networkx-3.2.1-py3-none-any.whl (1.6 MB)

----- 1.6/1.6 MB 1.8 MB/s eta 0:00:00

Collecting jinja2 (from torch)

Downloading https://download.pytorch.org/whl/jinja2-3.1.3-py3-none-any.whl (133 kB)

Collecting fsspec (from torch)

Downloading https://download.pytorch.org/whl/fsspec-2024.2.0-py3-none-any.whl (170 kB)

Collecting setuptools (from torch)

Downloading https://download.pytorch.org/whl/setuptools-70.0.0-py3-none-any.whl (863 kB)

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

It will take some time to download and install as the main file amounts to a size of 2.7 GB.

After a few minutes, PyTorch will be installed.

```
Installing collected packages: mpmath, typing-extensions, sympy, setuptools, pillow, numpy, networkx, MarkupSafe, fsspec, filelock, jinja2, torch, torchvision, torchaudio
Successfully installed MarkupSafe-2.1.5 filelock-3.13.1 fsspec-2024.2.0 jinja2-3.1.3 mpmath-1.3.0 networkx-3.2.1 numpy-1.26.3 pillow-10.2.0 setuptools-70.0.0 sympy-1.12 torch-2.4.1+cu118 torchaudio-2.4.1+cu118 torchvision-0.19.1+cu118 typing-extensions-4.9.0
```

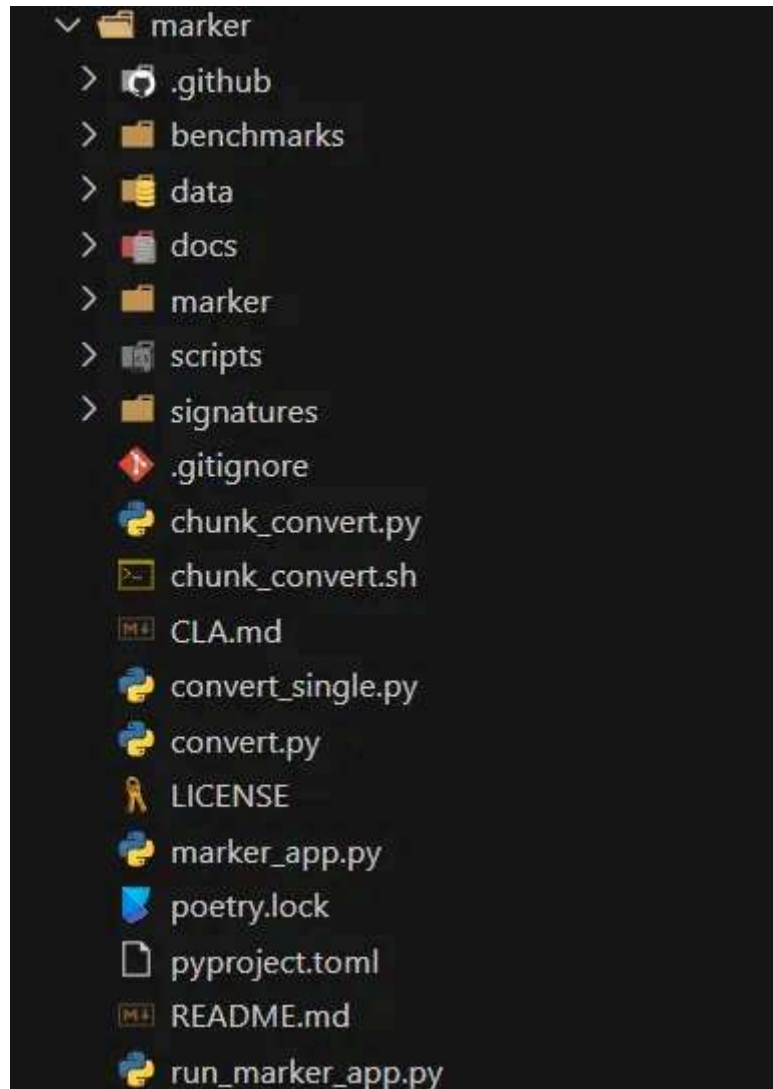
Now, the prerequisites are over. Next, you can go ahead and move to the actual Marker stuff.

Clone Marker

You can clone Marker project on your local system with the following command:

To make Medium work, we log user data. By using Medium, you agree to our [Privacy Policy](#), including cookie policy.

After cloning, the Marker GitHub repo would look something like this:



We have cloned the repository but still we cannot convert the PDFs into Markdown format as we haven't installed Marker.

Steps to Install Marker

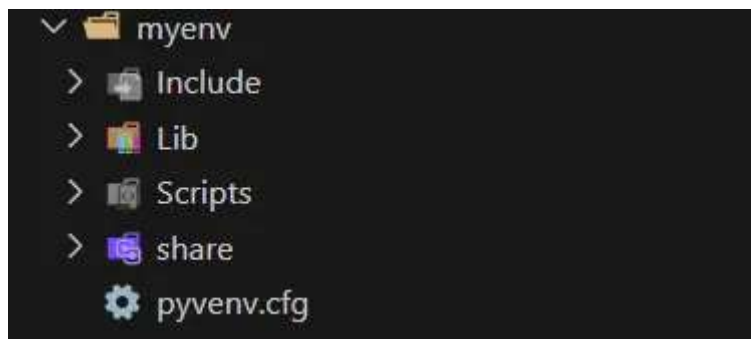
1. Create new environment

Outside of the newly cloned Marker GitHub repository, create a new

en To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

```
python -m venv myenv
```

This will create a **myenv** folder consisting of multiple files.



2. Activate the environment

```
myenv\Scripts\activate
```

This will activate the the newly created environment.

```
(myenv) PS D:\projects\marker-pdf>
```

3. Install “marker-pdf”

This command will actually install **marker-pdf** with the **pip** package manager.

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

Now we are ready to convert PDF documents to Markdown files!

4. Convert PDF format to Markdown

To convert a PDF to Markdown, we need two things:

- Input path of the PDF
- Output path

Because the command for the conversion is something like this:

```
marker_single "input_path" "output_path" - batch_multiplier 2 - max_pages 12
```

Hence, inside the cloned marker GitHub project folder, I will create two folders:

- **pdfs:** My input folder
- **output:** My output folder

And I will use a sample PDF for the Markdown conversion and paste it inside the **pdfs** folder.

Open in app ↗

Sign up

Sign in

Medium

🔍 Search

✍ Write



 **output**

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

Now, to convert the PDF “*Get_Started_With_Smallpdf.pdf*”, I will use the following command:

```
marker_single "D:/projects/marker-pdf/marker/pdfs/Get_Started_With_Smallpdf.pdf"
```

Here's what the other two arguments mean according to [Marker GitHub repo](#):

--batch_multiplier is how much to multiply default batch sizes by if you have extra VRAM. Higher numbers will take more VRAM, but process faster. Set to 2 by default. The default batch sizes will take ~3GB of VRAM.

--max_pages is the maximum number of pages to process. Omit this to convert the entire document.

Once the command is executed, Marker will initiate the conversion and save the Markdown to the **output** folder.

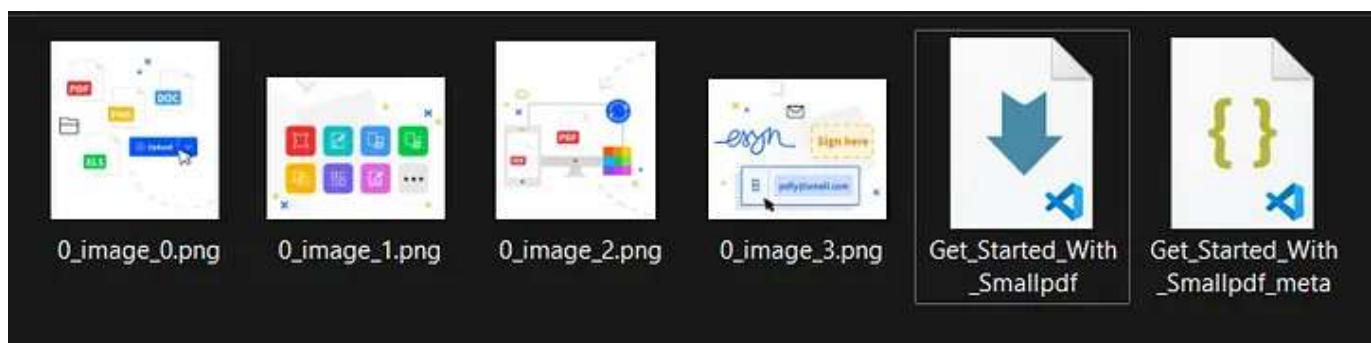
```
Loaded detection model vikp/surya_det3 on device cpu with dtype torch.float32
Loaded detection model vikp/surya_det3 on device cpu with dtype torch.float32
Loaded detection model vikp/surya_layout3 on device cpu with dtype torch.float32
Loaded reading order model vikp/surya_order on device cpu with dtype torch.float32
Loaded recognition model vikp/surya_rec2 on device cpu with dtype torch.float32
Loaded texify model to cpu with torch.float32 dtype
Detecting bboxes: 100% | 1/1 [00:02<00:00, 2.50s/it]
Detecting bboxes: 100% | 1/1 [00:01<00:00, 1.50s/it]
Finding reading order: 100% | 1/1 [00:03<00:00, 3.01s/it]
Saved markdown to the D:/projects/marker-pdf/marker/output/Get_Started_With_Smallpdf folder
```

The cool thing about Marker is that it extracts all the images associated with

th To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.



It even generates a metadata file in json format



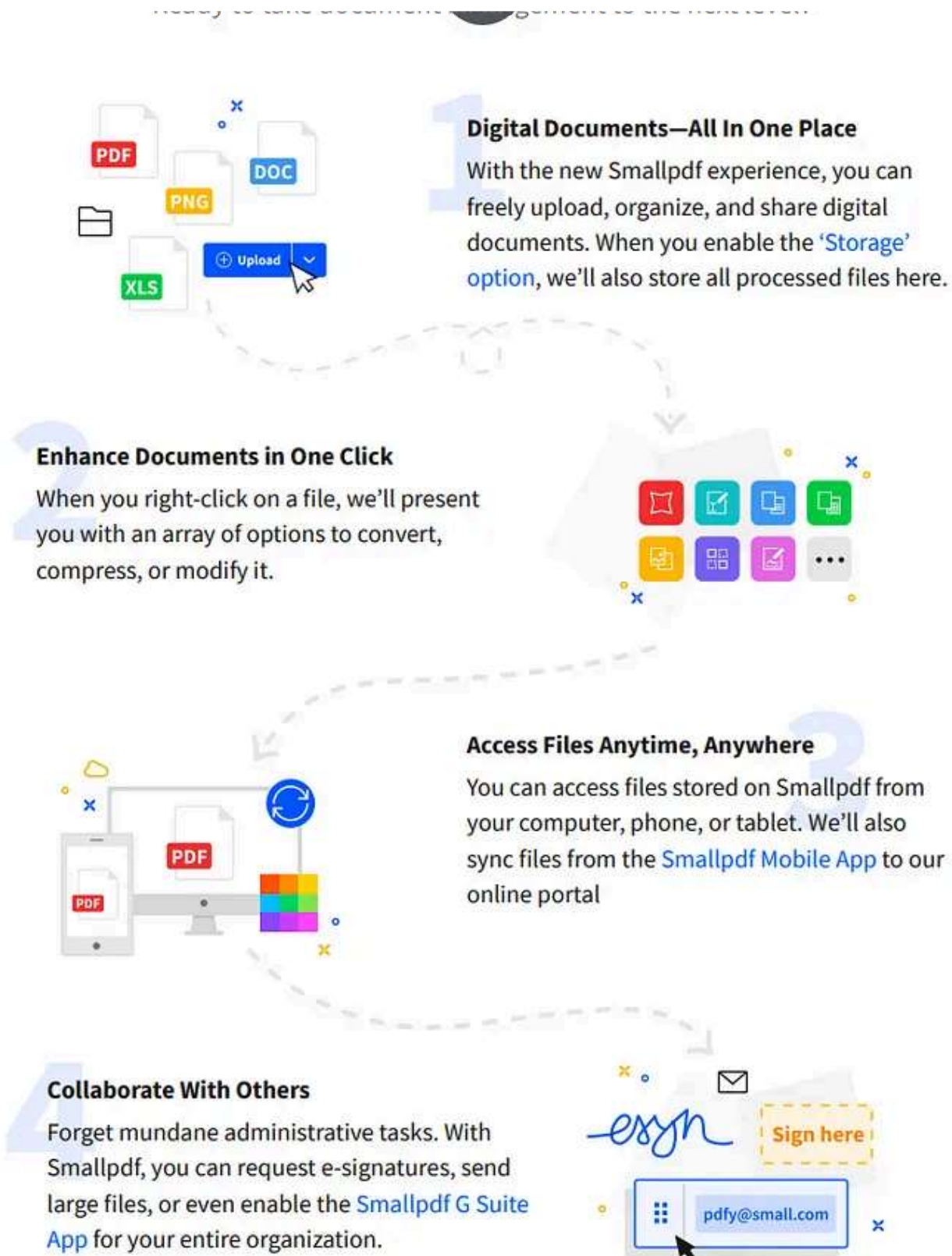
All the images are extracted in the .png format

Awesome! We have converted the PDF into Markdown. But wait!! How's the output in Markdown look?

PDF Input

Here's the PDF that we supplied Marker as the input file

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.



Markdown Output

To make Medium work, we log user data. By using Medium, you agree to our [Privacy Policy](#), including cookie policy.

Welcome To Smallpdf

Ready to take document management to the next level?

!@_image_0.png

Digital Documents—All In One Place

With the new Smallpdf experience, you can

!@_image_1.png freely upload, organize, and share digital documents

Enhance Documents In One Click

When you right-click on a file, we'll present

!@_image_2.png you with an array of options to convert, compress, and share

Access Files Anytime, Anywhere

You can access files stored on Smallpdf from

!@_image_3.png

your computer, phone, or tablet. We'll also sync files from the Smallpdf Mobile app

Collaborate With Others

Forget mundane administrative tasks. With Smallpdf, you can request e-signatures and track document status

Pretty good, right?

Conclusion

In this tutorial, we used Marker to extract the content of a PDF and convert it into Markdown format.

Of course, the PDF had only one page, but Marker is capable of a lot many

pa To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

You can try and play with it yourself!

[Marker](#)[Machine Learning](#)[Pdf Converter](#)[Markdown](#)[Artificial Intelligence](#)

Written by Suman Sourabh

[Follow](#)

1.2K Followers · 4 Following

Frontend Developer | Freelance Technical Writer | sumansourabh.in | Open for Writing/Coding work: sumsourabh14@gmail.com

Responses (5)



Write a response

What are your thoughts?



Darvalab

Oct 11, 2024



Have you tried pymupdf4llm? I recommend it, the conversions are really very good

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

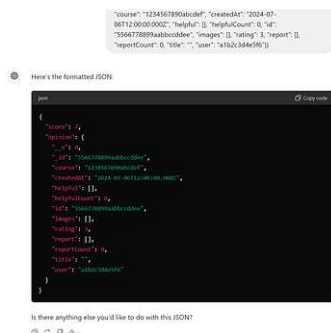



 67 3 replies [Reply.](#)



 70  1 reply [Reply.](#)

[See all responses](#)



Suman Sourabh 



In StartIt-Up by Suman Sourabh 

How I Use ChatGPT as a Frontend

De To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy,
Wi including cookie policy.
final result

Jul 6, 2024 🖱 2.1K 💬 63



 Suman Sourabh 

How I Created a Password Reset Flow in Next.js

This is THE BEST flow, trust me!

Jun 8, 2024 🖱 52 💬 4





How I Became a Full Stack Software

it)

Aug 5, 2023 🖱 1.92K 💬 38



 Suman Sourabh 

These Resources helped me become a Full Stack Developer

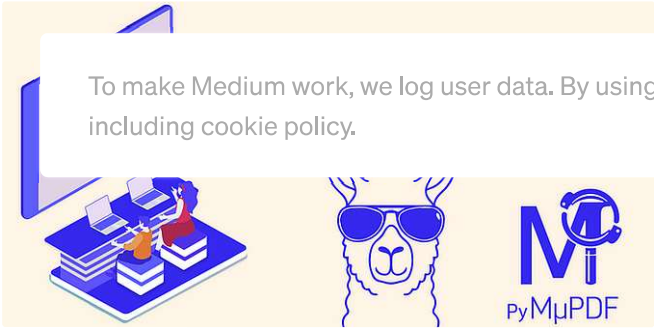
Includes 1 paid course and other FREE resources

Sep 19, 2023 🖱 473 💬 9



See all from Suman Sourabh

Recommended from Medium

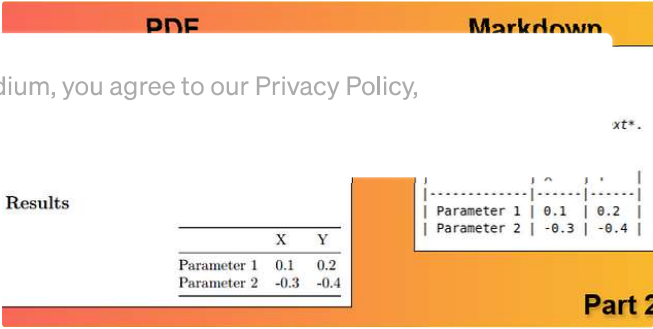


 Shravan Kumar

PyMuPDF4LLM is all You Need for Extracting Data from PDFs

This package converts the pages of a PDF to text in Markdown format using PyMuPDF....

Nov 1, 2024  333  7 



 In AI Advances by Dr. Leon Eversberg

Benchmarking PDF to Markdown Document Converters—Part 2

Testing 4 more Python Markdown converters on a benchmark PDF document for better...

 Feb 22  475  12 

Lists



Predictive Modeling w/ Python
20 stories · 1853 saves



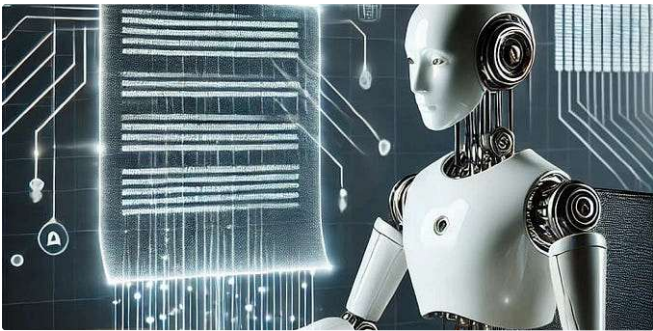
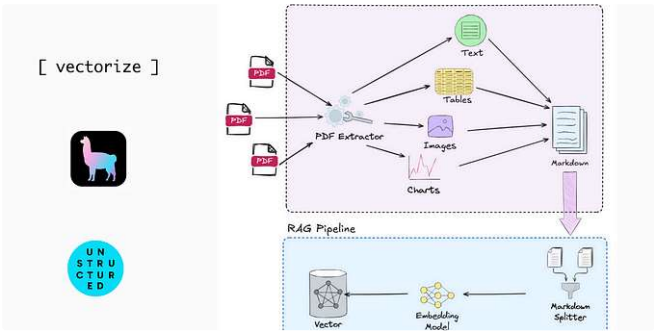
Natural Language Processing
1973 stories · 1617 saves



AI Regulation
6 stories · 705 saves



Practical Guides to Machine Learning
10 stories · 2221 saves





In Level Up Coding by Pavan Beladatti



Pankaj

WR/

To make Medium work, we log user data. By using Medium, you agree to our Privacy Policy, including cookie policy.

...

If you're building retrieval augmented generation (RAG) applications, you will...

Efficiently Convert PDFs to Structured Data for Large Language Models and Retrieval-...

Feb 19 1K 14



Oct 15, 2024 344 4



In Python in Plain English by Anoop Maurya



Kevin Meneses González

Why PyMuPDF4LLM is the Best Tool for Extracting Data from PDF...

How to 300x Your Productivity with These 13 AI Tools

Stuck behind a paywall? Read for Free!

Introduction — The story of how I stopped wasting time

Oct 18, 2024 2.4K 20



Feb 2 1.6K 53



See more recommendations