Extract Texts from PDF By Using Python Programing Language.

Hello everybody welcome to this python project

This python project is one of the real-world python use case projects.

So, in this python project, we create a text extractor program from a pdf file.

We can extract texts from pdf files.

So, this is one of the basic python projects you need to do as a python programmer.

Ok,

now we can start this python project by selecting and opening your favorite python IDE.

For me, I use PyCharm python IDE

And I also recommend using PyCharm for this project.

If you want to use another python IDE you can use any python ide you want to use.

Ok, now I start by opening PyCharm IDE and start Creating a new python file.

For that, I click and click new at the top, and the python file

After that, I give the file name

TEXT extractor

And when I hit enter, it creates a blank page.

And now I start writing this python project to create a program to extract texts from pdf file

For that

First, I import important python modules for this python project.

For this project we use

pyPdf2 Python module

This python module is used to extract texts from PDF files for that we need to use this python module.

For that, you need to install and import this python module.

If you are the first time using this

pyPDF2 python module you need to install it.

For that, you can use the PyCharm terminal section at the bottom of the PyCharm page

And when you click this terminal, it opens like this

And to install pypdf2

You write

Pip install pyPDF2

When you hit enter it starts to download and install automatically to your python project.

For me, I have already installed it

So, I import

By writing

Import pyPDF2

Like this

```
import PyPDF2
```

Ok, after we import this python module the next point is

Writing the program to open a PDF file.

For that

I write

PDF-File = open ()

And I pass my pdf file

For that, I already have a pdf file in the same folder of this python project.

And I write the pdf file name

“mypdf.pdf” pass as a string and

Open in reading binary mode for that

I write

“rb”

Like this

```
PDF_File = open("mypdf.pdf", "rb")
```

Ok, this code helps us to open the pdf file as read binary mode.

And after that, the next point is creating an object to read the pdf file by using the pyPDF2 python module.

For that

I call the object name PDF_reader

And

PDF_Reader = pypdf2.pdf file reader()

And I pass my opened pdf file

Which is I created above

So, I pass PDF_File inside this function

```
PDF_Reader = PyPDF2.PdfFileReader(PDF_File)
```

And.

This PDF_Reader object is used to read the open PDF file.

Ok,

Now I create a variable text and store by extracting the texts from the pdf files.

For that

I write

Text = PDF_Readr.getpage(0).extracttext()

like this

```
Text = PDF_Reader.getPage(0).extractText()
```

And to see the texts extracted from the pdf file

I write print()

Text

Like this

```
print(Text)
```

And when you run this program

It shows the texts which is extracted from the PDF Files.

So, you can extract any pdf file by using this python program.

you can also change the pdf file name and extract other texts from other PDFs.

Ok, this is all about this python project

I think you get some concepts from this video

Thanks for your time.

I will see you in the next python project.

Thanks again!!

!!! The Full Code Looks Like !!!

```
# extracting text from pdf file
import PyPDF2

PDF_File = open("mypdf.pdf", "rb")
PDF_Reader = PyPDF2.PdfFileReader(PDF_File)
Text = PDF_Reader.getPage(0).extractText()
print(Text)
```

By, Awoke Zemenu