AskPython

Convert PDF to TXT file using Python

By Siddhi Sawant / March 29, 2021

In this article, we're going to create an easy python script that will help us convert pdf to txt file. You have various applications that you can download and use for **pdf to txt** file conversion. There are a lot of online applications too available for this purpose but how cool would it be, if you could create your own pdf to txt file converter using a simple python script.

Let's get started!

Steps to Convert PDF to TXT in Python

Without any further ado, let's get started with the steps to convert pdf to txt.

Step 01 – Create a PDF file (or find an existing one)

- Open a new Word document.
- Type in some content of your choice in the word document.

- Now to File > Print > Save.
- Remember to save your pdf file in the same location where you save your python script file.
- Now your .pdf file is created and saved which you will later convert into a .txt file.

Step 02 – Install PyPDF2

- First, we will install an external module named **PyPDF2**.
- The PyPDF2 package is a **pure-python pdf library** that you can use for splitting, merging, cropping, and transforming pdfs. According to the PyPDF2 website, you can also use PyPDF2 to add data, viewing options, and passwords to the pdfs, too.
- For installing the PyPDF2 package, open your windows command prompt and use the pip command to install PyPDF2:

```
C:\Users\Admin>pip install PyPDF2
```

```
Collecting PyPDF2
  Downloading PyPDF2-1.26.0.tar.gz (77 kB)
     |████████████████████████████████| 77 kB 1.9 MB/s
Using legacy 'setup.py install' for PyPDF2, since package 'wheel' is not ins
Installing collected packages: PyPDF2
    Running setup.py install for PyPDF2 ... done
Successfully installed PyPDF2-1.26.0
```

This will successfully install your PyPDF2 package on your system. Once it's installed, you are good to go with your script.

Step 03 – Opening a new Python file for the script

- Open your python IDLE and press keys ctrl + N. This will open your text editor.

- You can use any other text editor of your prefered choice.
- Save the file as *your_pdf_file_name*.py.
- Save this .py file in the same location as your pdf file.

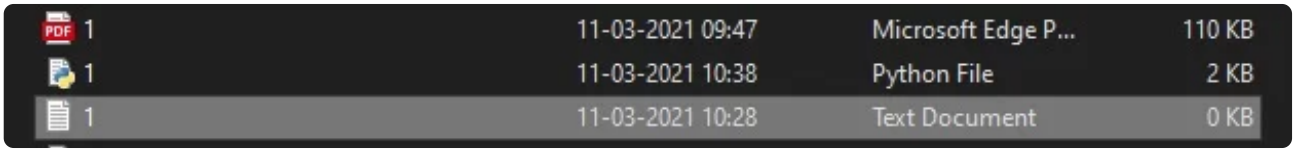
Let's get started with the Script Code

```
import PyPDF2

#create file object variable
#opening method will be rb
pdffileobj=open('1.pdf','rb')

#create reader variable that will read the pdffileobj
pdfreader=PyPDF2.PdfFileReader(pdffileobj)

#This will store the number of pages of this pdf file
x=pdfreader.numPages

#create a variable that will select the selected number of pages
pageobj=pdfreader.getPage(x+1)

#(x+1) because python indentation starts with 0.
#create text variable which will store all text datafrom pdf file
text=pageobj.extractText()

#save the extracted data from pdf to a txt file
#we will use file handling here
#dont forget to put r before you put the file path
#go to the file location copy the path by right clicking on the file
#click properties and copy the location path and paste it here.
#put "\\your_txtfilename"
file1=open(r"C:\Users\SIDDHI\AppData\Local\Programs\Python\Python38\\1.txt",
file1.writelines(text)
```

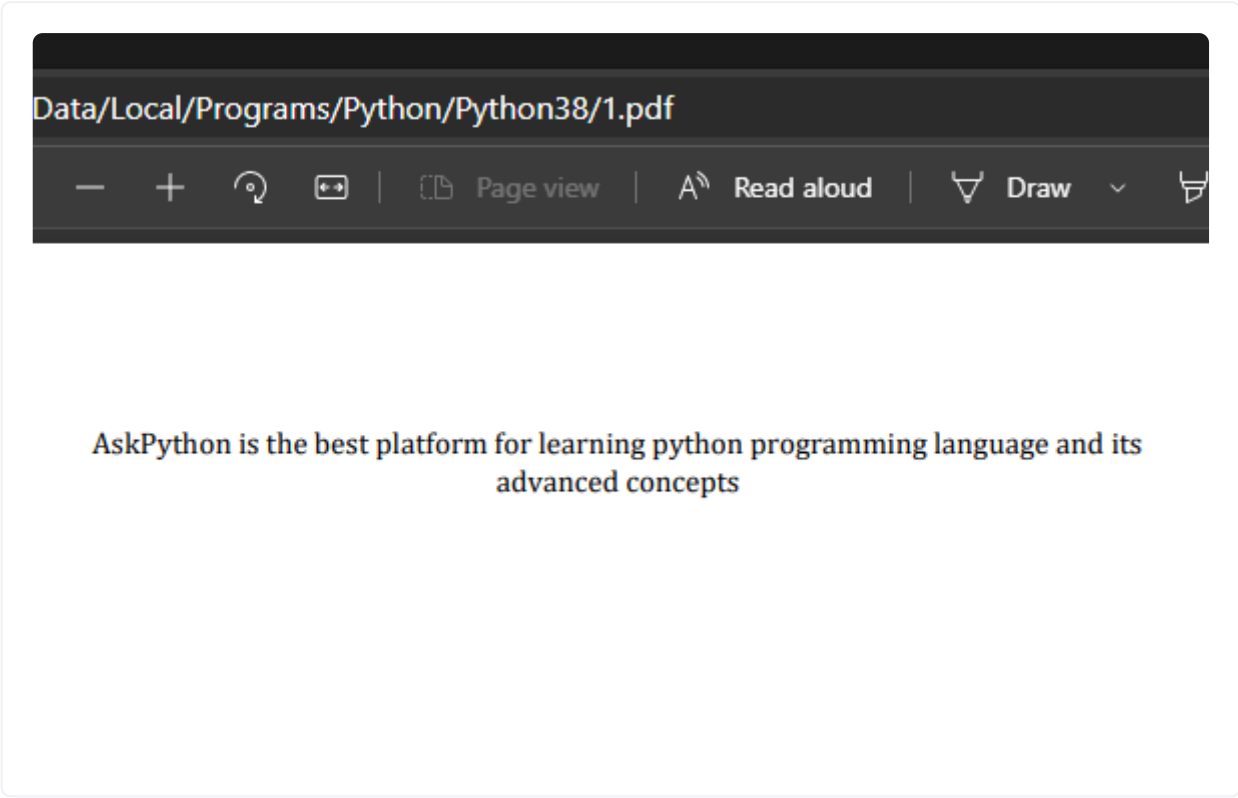
Here's a quick explanation of the code:

- We first create a Python file object and open the PDF file in “read binary (rb)” mode
- Then, we create the PdfFileReader object that will read the file opened from the previous step
- A variable is used to store the number of pages within the file
- The last part will write the identified lines from the PDF to a text file that you specify

Output:



PDF file Image :



Converted Txt file Image :

This was in brief about how to convert a pdf file to a txt file by writing your own python script. Try it out !

← Previous Post

Next Post →

Search ...

Recent Posts

Understanding Python 3's Decimal Formatting — :.0f vs. :.1f

Understanding and Avoiding Syntax Errors in Python Dictionaries

Copy Files in Python: Using shutil, os, and subprocess Modules

[SOLVED] 'Unexpected Keyword Argument' TypeError in Python

4 Handy Ways to Convert Bytes to Hex Strings in Python 3

Regular Expressions Difference Between [0-9] and [0-9.]

Regular Expression for a String With Certain Condition

Python Regex for Case-Insensitive Text Matching without re.compile()

How to Round 2 Decimals with Python?

[Fix] 'Can't Find a Default Python' Error: 2 Effective Solutions

Favorite Sites

GoLang Tutorials

VM-Help

Linux Tutorials

MySQL Tutorials

CodeForGeek

Mkyong

Copyright © 2023 AskPython · All Rights Reserved

AskPython is part of JournalDev IT Services Private Limited