bit.ly/mli-linkedin

MACHINE LEARNING INDIA

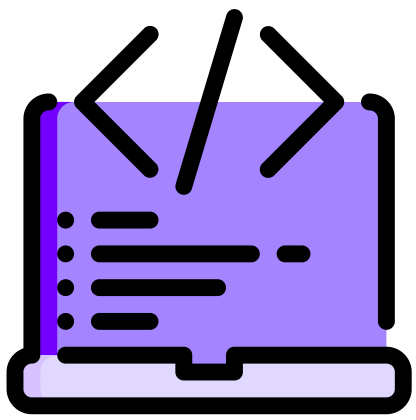
Easy-peazy

PDF TEXT EXTRACTION WITH PYTHON

In just 20 lines of code!

PDF

♥ Support!



It's super easy!

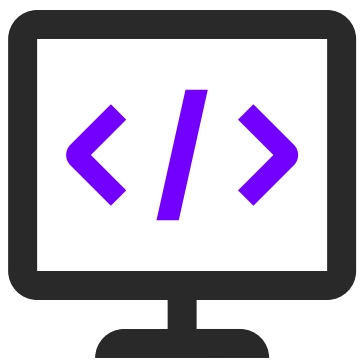
In this micro-tutorial, you will learn how to extract text from a given **PDF** in **Python**. We will be using the **PyPDF2 module** for extracting text from PDF files.

Note: The PDF can be a multipage PDF too.

Installing the module:

To install the **PyPDF2** module and some other related dependencies, we can use the **pip** command:

```
pip install PyPDF2
```



The details:

For extracting text from a PDF we will be using the **PdfFileReader** class which is used to initialize **PdfFileReader** object, taking a stream parameter, in which we will provide the file stream for the PDF file.

Link to the sample file:

http://www.africau.edu/images/default/sample.pdf

The code:

ReadingPdf.py - C:\Users\91884\Desktop\pgms.py\ReadingPdf.py (3.7.0)

File Edit Format Run Options Window Help

```
from PyPDF2 import PdfFileReader

#opening the pdf file in a read binary mode
file=open('C:/Users/91884/Desktop/sample pdf2.pdf','rb')

#instantiating the object
reader=PdfFileReader(file)

print("Printing the document info: ",(reader.getDocumentInfo()))
'''
Output:
        Printing the document info: {'/Creator': 'Rave (http://www.nevrona.com/rave)',
        '/Producer': 'Nevrona Designs', '/CreationDate': 'D:20060301072826'}
'''
print('****************************************')
print()
print("Number of Pages: ",reader.getNumPages()) # Number of Pages:  2

print("PDF File created by: " ,reader.getDocumentInfo().creator)
#Output : PDF File created by:  Rave (http://www.nevrona.com/rave)

print('****************************************')
pages=reader.getNumPages()
for i in range(0, pages):
        print("Page Number: ",i+1)
        print("- - - - - - - - - - - - - - - - - - - -")
        pageObj = reader.getPage(i)
        print(pageObj.extractText())
        print("- - - - - - - - - - - - - - - - - - - -")
# close the PDF file object
file.close()
```

The output:


ReadingPdf.py - C:\Users\91884\Desktop\pgms.py\ReadingPdf.py (3.7.0)
File Edit
Python 3.7.0 Shell
File Edit Shell Debug Options Window Help
Python 3.7.0 (v3.7.0:1bf9cc5093, Jun 27 2018, 04:59:51) [MSC v.1914 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\91884\Desktop\pgms.py\ReadingPdf.py ==========
Printing the document info: {'/Creator': 'Rave (http://www.nevrona.com/rave)', '/Producer': 'Nevrona Designs', '/CreationDate': 'D:20060301072826'}
****
Number of Pages: 2
PDF File created by: Rave (http://www.nevrona.com/rave)
****
Page Number: 1
- - - - - - - - - - - - - - - - - - - -
A Simple PDF File This is a small demonstration .pdf file - just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...
- - - - - - - - - - - - - - - - - - - -
Page Number: 2
- - - - - - - - - - - - - - - - - - - -
Simple PDF File 2 ...continued from page 1. Yet more text. And more text. And more text. And more text. And more text. And more text. Oh, how boring typing this stuff. But not as boring as watching paint dry. And more text. And more text. And more text. And more text. Boring. More, a little more text. The end, and just as well.
- - - - - - - - - - - - - - - - - - - -
>>>
@ML.INDIA

Other Applications of PyPDF2 Module:

- Rotating a PDF file page by any defined angle.
- Merging two or more PDF files at a defined page number.
- Appending two or more PDF files, one after another.
- Find all the meta information for any PDF, like creator, author, date of creation, etc.
- We can even create a new PDF file using the text coming from some text file.

Content curators:

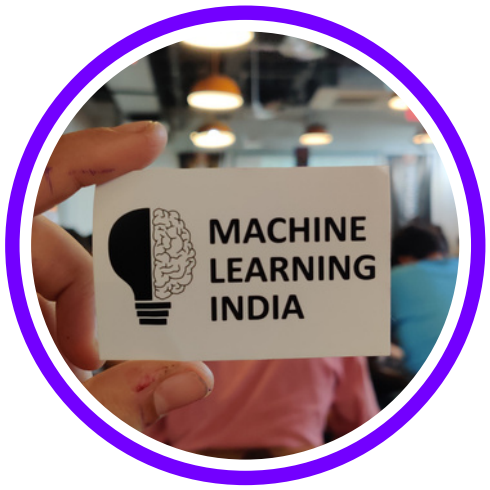
Bhavishya Pandit and Priyanka Kasture.

Important references:

- Extract Text from PDF in Python - PyPDF2 Module by **Abhishek Ahlawat** on **www.studytonight.com**.

Important note:

The links to these resources will be put up on our Telegram. Channel ID: **@machinelearning24x7**.



Wasn't that easy-peazy?

Let us know in the comments! If you like our content and find it **valuable**, do give us a **follow**! Your **love** and **support** inspires us to keep delivering the best we can! ❤️

Like. **Comment.** **Share.**