w.

– [Stop harming Monica](https://stackoverflow.com/users/2142055/stop-harming-monica)

[Apr 22, 2018 at 12:44](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python#comment86947755_49965818)

* @Goyo thanks. But I'm not able to run this commando in the terminal. It doesn't recognize lowriter as an executable command. Why is that?

– [Also](https://stackoverflow.com/users/9020421/also)

[Apr 22, 2018 at 13:17](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python#comment86948294_49965818)

* 3

How could I know? It is probably related to the way you installed libreoffice. But you'd better figure it out, you can't expect python to run a program when you are unable to run it yourself.

– [Stop harming Monica](https://stackoverflow.com/users/2142055/stop-harming-monica)

[Apr 22, 2018 at 14:20](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python#comment86949519_49965818)

* What is your operating system? Use [my answer from two weeks ago](https://stackoverflow.com/a/49763069/5100564) but modify the paths.

– [Jim K](https://stackoverflow.com/users/5100564/jim-k)

[Apr 23, 2018 at 18:59](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python#comment86992552_49965818)

* @JimK Thank you very much. I saw your post recently, and I used it as a help for my trouble, and it worked perfectly for converting the pdf to odg. (I checked your answer for that reason). However, I was looking for the conversion to .docx, that seems to be more difficult... In any case, my operating system is Windows 7.

– [Also](https://stackoverflow.com/users/9020421/also)

[Apr 24, 2018 at 8:55](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python#comment87010973_49965818)

[Show **2** more comments](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python)

5 Answers

Sorted by:



3

I am not aware of a way to convert a pdf file into a Word file using libreoffice.  
However, you can convert from a pdf to a html and then convert the html to a docx.  
Firstly, get the commands running on the command line. (The following is on Linux. So you may have to fill in path names to the soffice binary and use a full path for the input file on your OS)

soffice --convert-to html ./my\_pdf\_file.pdf

then

soffice --convert-to docx:'MS Word 2007 XML' ./my\_pdf\_file.html

You should end up with:

my\_pdf\_file.pdf  
my\_pdf\_file.html  
my\_pdf\_file.docx

Now wrap the commands in your subprocess code

[Share](https://stackoverflow.com/a/50218260/20377102)

[Edit](https://stackoverflow.com/posts/50218260/edit)

Follow

answered May 7, 2018 at 16:02

[[](https://stackoverflow.com/users/4637585/rolf-of-saxony)](https://stackoverflow.com/users/4637585/rolf-of-saxony)

[Rolf of Saxony](https://stackoverflow.com/users/4637585/rolf-of-saxony)

**20.4k**33 gold badges3636 silver badges5656 bronze badges

[Add a comment](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python)

3

I use this for multiple files

####

from pdf2docx import Converter

import os

# # # dir\_path for input reading and output files & a for loop # # #

path\_input = '/pdftodocx/input/'

path\_output = '/pdftodocx/output/'

for file in os.listdir(path\_input):

cv = Converter(path\_input+file)

cv.convert(path\_output+file+'.docx', start=0, end=None)

cv.close()

print(file)

[Share](https://stackoverflow.com/a/66108935/20377102)

[Edit](https://stackoverflow.com/posts/66108935/edit)

Follow

answered Feb 8, 2021 at 20:31

[[](https://stackoverflow.com/users/15171734/simon)](https://stackoverflow.com/users/15171734/simon)

[simon](https://stackoverflow.com/users/15171734/simon)

**31**11 bronze badge

[Add a comment](https://stackoverflow.com/questions/49965818/convert-pdf-to-docx-with-python)

1

My approach does not follow the same methodology of using subsystems. However this one does the job of reading through all the pages of a PDF document and moving them to a docx file. Note: It only works with text; images and other objects are usually ignored.

#Description: This python script will allow you to fetch text information from a pdf file

#import libraries

import PyPDF2

import os

import docx

mydoc = docx.Document() # document type

pdfFileObj = open('pdf/filename.pdf', 'rb') # pdffile loction

pdfReader = PyPDF2.PdfFileReader(pdfFileObj) # define pdf reader object

# Loop through all the pages

for pageNum in range(1, pdfReader.numPages):

pageObj = pdfReader.getPage(pageNum)

pdfContent = pageObj.extractText() #extracts the content from the page.

print(pdfContent) # print statement to test output in the terminal. codeline optional.

mydoc.add\_paragraph(pdfContent) # this adds the content to the word document

mydoc.save("p