

How to convert webpage into PDF by using Python

Asked 8 years, 1 month ago Modified 1 year, 3 months ago Viewed 256k times



118



59



I was finding solution to print webpage into local file PDF, using Python. one of the good solution is to use Qt, found here, <https://bharatikunal.wordpress.com/2010/01/>.

It didn't work at the beginning as I had problem with the installation of PyQt4 because it gave error messages such as 'ImportError: No module named PyQt4.QtCore', and 'ImportError: No module named PyQt4.QtCore'.

It was because PyQt4's not installed properly. I used to have the libraries located at C:\Python27\Lib however it's not for PyQt4.

In fact, it simply needs to download from <http://www.riverbankcomputing.com/software/pyqt/download> (mind the correct Python version you are using), and install it to C:\Python27 (my case). That's it.

Now the scripts runs fine so I want to share it. for more options in using Qprinter, please refer to <http://qt-project.org/doc/qt-4.8/qprinter.html#Orientation-enum>.

[python](#) [html](#) [pdf](#) [qprinter](#)

ShareFollow

edited Jun 1, 2020 at 20:35
 0m3r
11.7k 15 30 65

asked Apr 29, 2014 at 8:10
 Mark K
7,735 13 50 98

1 Note that you can post a Q&A simultaneously if you're self-answering, and the usual quality rules still apply to both parts. – [jonrsharpe](#) Jan 29, 2018 at 22:24

9 Answers

Sorted by:

Highest score (default)



173



You also can use [pdflkit](#):

Usage

```
import pdfkit
pdfkit.from_url('http://google.com', 'out.pdf')
```

Install

MacOS: `brew install Caskroom/cask/wkhtmltopdf`

Debian/Ubuntu: `apt-get install wkhtmltopdf`

Windows: `choco install wkhtmltopdf`

See [official documentation](https://github.com/JazzCore/python-pdfkit/wiki/Installing-wkhtmltopdf) for MacOS/Ubuntu/other OS: <https://github.com/JazzCore/python-pdfkit/wiki/Installing-wkhtmltopdf>

By clicking "Accept all cookies", you agree Stack Exchange can store cookies on your device and disclose information in accordance with our [Cookie Policy](#).

ShareFollow

Accept all cookies Customize settings

edited Jul 31, 2020 at 10:10
 user13618029

answered May 20, 2014 at 13:24
 NorthCat
9,007 16 45 49

- This is awesome, way easier than messing around with reportlab or using a print drive to convert. Thanks so much. – Dowlers May 27, 2015 at 18:56
- PDFKit requires a running X Server (or "virtual" X Server). :(See here: [github.com/JazzCore/python-pdfkit/wiki/...](#) – Tim Ludwinski Sep 5, 2017 at 19:57
- It seems like windows does not support pdfkit. Is that true? – Kane Chew Nov 15, 2017 at 16:47
- Perfect !! Even download the embeded images, don't bother use that ! You'll have to `apt-get install wkhtmltopdf` – Tinmarino Jan 28, 2018 at 21:16
- pdfkit depends on non-python package wkhtmltopdf, which in turn requires a running X server. So while nice in some environments, this is not an answer that works generally in python. – Rasmus Kaj Feb 19, 2018 at 16:33




WeasyPrint

```
pip install weasyprint # No longer supports Python 2.x.

python
>>> import weasyprint
>>> pdf = weasyprint.HTML('http://www.google.com').write_pdf()
>>> len(pdf)
92059
>>> open('google.pdf', 'wb').write(pdf)
```

ShareFollow

edited Aug 12, 2020 at 8:48

 **Sunit Gautam**
4,115 ●2 ●16 ●27

answered Dec 23, 2015 at 15:04

 **JohnMudd**
13.4k ●2 ●25 ●24

- Can I provide file path instead of url? – Piyush S. Wanare Sep 29, 2017 at 6:57
- I think I will prefer this project as it's dependencies are python packages rather than a system package. As of Jan 2018 it seems to have more frequent updates and better documentation. – stvsmith Jan 4, 2018 at 23:13
- There are too many things to install. I stopped at libpango and went for the pdfkit. Nasty for system wide wkhtmltopdf but weasyprint also require some system wide installs. – visoft Jul 17, 2018 at 8:39
- this won't convert `javascripts` in the html file. for that you need to use `pdfkit` – suhailvs May 22, 2019 at 11:16
- I would believe the option should be `'wb'`, not `'w'`, because `pdf` is a `bytes` object. – Anatoly Scherbakov Aug 13, 2019 at 7:01



thanks to below posts, and I am able to add on the webpage link address to be printed and present time on the PDF generated, no matter how many pages it has.

Add text to Existing PDF using Python

https://github.com/disflux/django-mtr/blob/master/pdfgen/doc_overlay.py

To share the script as below:

```
import time
from pyPdf import PdfFileWriter, PdfFileReader
import StringIO
```

```

import sys
from reportlab.pdfgen import canvas
from reportlab.lib.pagesizes import letter
from xhtml2pdf import pisa
import sys
from PyQt4.QtCore import *
from PyQt4.QtGui import *
from PyQt4.QtWebKit import *

url = 'http://www.yahoo.com'
tem_pdf = "c:\\tem_pdf.pdf"
final_file = "c:\\younameit.pdf"

app = QApplication(sys.argv)
web = QWebView()
#Read the URL given
web.load(QUrl(url))
printer = QPrinter()
#setting format
printer.setPageSize(QPrinter.A4)
printer.setOrientation(QPrinter.Landscape)
printer.setOutputFormat(QPrinter.PdfFormat)
#export file as c:\\tem_pdf.pdf
printer.setOutputFileName(tem_pdf)

def convertIt():
    web.print_(printer)
    QApplication.exit()

QObject.connect(web, SIGNAL("loadFinished(bool)"), convertIt)

app.exec_()
sys.exit

# Below is to add on the weblink as text and present date&time on PDF generated

outputPDF = PdfFileWriter()

```

ShareFollow

edited May 23, 2017 at 12:18



Community Bot

1 • 1

answered Apr 30, 2014 at 7:31



Mark K

7,735 • 13 • 50 • 98

Thanks for sharing your code! Any advice for making this work for local pdf files? Or is it as easy as prepending "file:///" to the url? I'm not very familiar with these libraries... thanks – [user2426679](#) Oct 31, 2014 at 18:02

@user2426679, you mean convert online PDF into local PDF files? – [Mark K](#) Nov 25, 2014 at 1:48

thanks for your reply... sorry for my tardiness. I ended up using wkhtmltopdf since it was able to handle what I was throwing at it. But I was asking how to load a pdf that was local to my hdd. Cheers – [user2426679](#) Dec 28, 2014 at 23:15

@user2426679 sorry I still don't get you. maybe because I am a newbie to Python too. You meant read local PDF files in Python? – [Mark K](#) Jan 22, 2015 at 8:05

There were some issues with `html5lib`, which is used by xhtml2pdf. This solution fixed the problem: github.com/xhtml2pdf/xhtml2pdf/issues/318 – [Blairg23](#) Oct 14, 2016 at 21:22



14



here is the one working fine:

```

import sys
from PyQt4.QtCore import *
from PyQt4.QtGui import *

```

```

from PyQt4.QtWebKit import *
from PyQt4.QtCore import *

app = QApplication(sys.argv)
web = QWebView()
web.load(QUrl("http://www.yahoo.com"))
printer = QPrinter()
printer.setPageSize(QPrinter.A4)
printer.setOutputFormat(QPrinter.PdfFormat)
printer.setOutputFileName("file0K.pdf")

def convertIt():
    web.print_(printer)
    print("Pdf generated")
    QApplication.exit()

QObject.connect(web, SIGNAL("loadFinished(bool)"), convertIt)
sys.exit(app.exec_())

```

ShareFollow

edited Oct 26, 2019 at 17:32



FractalSpace

5,183 ● 2 ● 39 ● 47

answered Apr 29, 2014 at 8:11



Mark K

7,735 ● 13 ● 50 ● 98

Interestingly, the web page links are generated as text rather than links in the generated PDF. – [amergin](#) Nov 24, 2014 at 17:16

Anyone know why this would be generating blank pdfs for me? – [boson](#) Oct 4, 2016 at 14:33



13



Per this answer: [How to convert webpage into PDF by using Python](#), the advice was to use [pdftkit](#). You also have to install [wkhtmltopdf](#).

If you have a local `.html` file, you then need to use this command:

```
pdftkit.from_file('test.html', 'out.pdf')
```

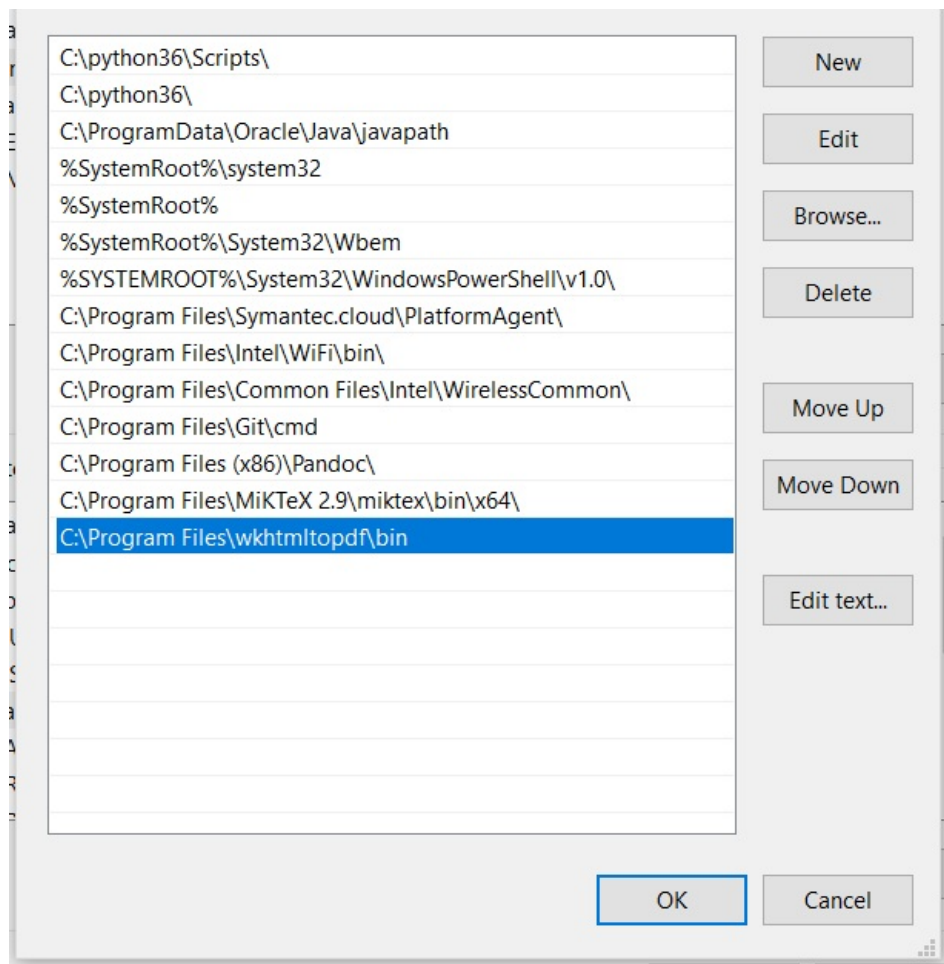
But this will throw an error if you haven't added the wkhtmltopdf executables to your system path. This was the part that tripped me up and I wanted to share.

On Windows, open your environment variables and add them to your `System variables` > `Path` like below. In my case, these `.exe` files were located here after I installed the wkhtmltopdf from an exe:

```
C:\Program Files\wkhtmltopdf\bin
```

Edit environment variable





ShareFollow

edited Mar 14, 2021 at 20:38

answered Jan 29, 2018 at 22:31



Jean-François Fabre ♦
132k ● 23 ● 123 ● 195



Jarad
15k ● 19 ● 84 ● 133

I was facing the same issue on Win10, this helped, thanks a ton. – kudo_shinichi Mar 6 at 12:18



11



Here is a simple solution using QT. I found this as part of an answer to a different question on StackOverFlow. I tested it on Windows.

```
from PyQt4.QtGui import QTextDocument, QPrinter, QApplication

import sys
app = QApplication(sys.argv)

doc = QTextDocument()
location = "c://apython/Jim/html/notes.html"
html = open(location).read()
doc.setHtml(html)

printer = QPrinter()
printer.setOutputFileName("foo.pdf")
printer.setOutputFormat(QPrinter.PdfFormat)
printer.setPageSize(QPrinter.A4);
printer.setPageMargins (15,15,15,15,QPrinter.Millimeter);

doc.print_(printer)
print "done!"
```

ShareFollow

edited Mar 12, 2015 at 13:31

answered Jan 20, 2015 at 20:38



Jim Paul
181 ● 1 ● 4

6



I tried @NorthCat answer using pdfkit.

It required wkhtmltopdf to be installed. The install can be downloaded from here. <https://wkhtmltopdf.org/downloads.html>

Install the executable file. Then write a line to indicate where wkhtmltopdf is, like below. (referenced from [Can't create pdf using python PDFKIT Error : " No wkhtmltopdf executable found:"](#))

```
import pdfkit

path_wkhtmltopdf = "C:\\Folder\\where\\wkhtmltopdf.exe"
config = pdfkit.configuration(wkhtmltopdf = path_wkhtmltopdf)

pdfkit.from_url("http://google.com", "out.pdf", configuration=config)
```

ShareFollow

answered Oct 18, 2019 at 2:09



Mark K

7,735 ● 13 ● 50 ● 98

where did it go after I clicked .deb and installed on software centre? – [mLstudent33](#) Nov 7, 2020 at 23:55

5



This solution worked for me using PyQt5 version 5.15.0

```
import sys
from PyQt5 import QtWidgets, QtWebEngineWidgets
from PyQt5.QtCore import QUrl
from PyQt5.QtGui import QPageLayout, QPageSize
from PyQt5.QtWidgets import QApplication

if __name__ == '__main__':
    app = QtWidgets.QApplication(sys.argv)
    loader = QtWebEngineWidgets.QWebEngineView()
    loader.setZoomFactor(1)
    layout = QPageLayout()
    layout.setPageSize(QPageSize(QPageSize.A4Extra))
    layout.setOrientation(QPageLayout.Portrait)
    loader.load(QUrl('https://stackoverflow.com/questions/23359083/how-to-convert-webpage-into-pdf-by-using-python'))
    loader.page().pdfPrintingFinished.connect(lambda *args: QApplication.exit())

    def emit_pdf(finished):
        loader.page().printToPdf("test.pdf", pageLayout=layout)

    loader.loadFinished.connect(emit_pdf)
    sys.exit(app.exec_())
```

ShareFollow

answered Aug 6, 2020 at 19:39



Y.kh

143 ● 2 ● 5

I tried this and get this error: Traceback (most recent call last): File "C:/Users/brentond/Documents/Python/PdfWebsite.py", line 2, in <module> from PyQt5 import QtWidgets, QtWebEngineWidgets ImportError: DLL load failed: The specified module could not be found. – [Dan](#) Jan 20, 2021 at 17:03

You have to install the PyQt5 package first: pip install PyQt5 – [Y.kh](#) Jan 21, 2021 at 16:17

I do have it installed... But as far as I can see there is no PyQt5 method called QtwebEngineWidgets... At least not in 5.15.2 that I have installed in PyCharm. – [Dan](#) Jan 29, 2021 at 14:43

You also need to `pip install PyQtWebEngine` for this to work – [Dániel Kis-Nagy](#) Apr 29, 2021 at 15:52



If you use selenium and chromium, you do not need to manage cookies by you self, and you can generate pdf page from chromium's print as pdf. You can refer this project to realize it. <https://github.com/maxvst/python-selenium-chrome-html-to-pdf-converter>

modified base > https://github.com/maxvst/python-selenium-chrome-html-to-pdf-converter/blob/master/sample/html_to_pdf_converter.py

```
import sys
import json, base64

def send_devtools(driver, cmd, params={}):
    resource = "/session/%s/chromium/send_command_and_get_result" % driver.session_id
    url = driver.command_executor._url + resource
    body = json.dumps({'cmd': cmd, 'params': params})
    response = driver.command_executor._request('POST', url, body)
    return response.get('value')

def get_pdf_from_html(driver, url, print_options={}, output_file_path="example.pdf"):
    driver.get(url)

    calculated_print_options = {
        'landscape': False,
        'displayHeaderFooter': False,
        'printBackground': True,
        'preferCSSPageSize': True,
    }
    calculated_print_options.update(print_options)
    result = send_devtools(driver, "Page.printToPDF", calculated_print_options)
    data = base64.b64decode(result['data'])
    with open(output_file_path, "wb") as f:
        f.write(data)

# example
from selenium import webdriver
from selenium.webdriver.chrome.options import Options

url = "https://stackoverflow.com/questions/23359083/how-to-convert-webpage-into-pdf-by-using-python#"
webdriver_options = Options()
webdriver_options.add_argument("--no-sandbox")
webdriver_options.add_argument('--headless')
webdriver_options.add_argument('--disable-gpu')
driver = webdriver.Chrome(chromedriver, options=webdriver_options)
```

ShareFollow

answered Jul 26, 2020 at 13:31



Yuanmeng Xiao

158 • 1 • 7

Firstly i use weasyprint but it do not support cookies even you can write your own `default_url_fetcher` to handle cookies but later i occur issue when install it in Ubuntu16. Then i use wkhtmltopdf it suport cookie setting but it caused many OSERROR like -15 -11 when handle some page. – [Yuanmeng Xiao](#) Jul 26, 2020 at 13:35

Thank you for sharing Mr. @Yuanmeng Xiao. – [Mark K](#) Jul 27, 2020 at 1:16

Hi @YuanmengXiao I copied your code above and I get this error: Traceback (most recent call last): File "C:/Users/brentond/Documents/Python/PdfWebsite.py", line 39, in <module> driver = webdriver.Chrome(chromedriver, options=webdriver_options) NameError: name 'chromedriver' is not defined – [Dan](#) Jan 20, 2021 at 16:51

I then installed a module called chromedriver and imported it to the above code and now get this error Traceback (most recent call last): File "C:/Users/brentond/Documents/Python/PdfWebsite.py", line 33, in <module> import chromedriver File "C:\Program Files\ArcGIS\Pro\bin\Python\envs\arcgispro-py3\lib\site-packages\chromedriver_init_.py", line 16, in <module> raise RuntimeError("This package supports only Linux, MacOSX or Windows platforms") RuntimeError: This package supports only Linux, MacOSX or Windows platforms – [Dan](#) Jan 20, 2021 at 16:56

you should download chromedriver from chromedriver.chromium.org And you would better learn how to use selenium to driver chrome browser. – [Yuanmeng Xiao](#) Jan 21, 2021 at 21:59



Highly active question. Earn 10 reputation (not counting the [association bonus](#)) in order to answer this question. The reputation requirement helps protect this question from spam and non-answer activity.