Page 1:

Poppler On Windows  
  
Python, PDFs, and Window’s Subsytem for Linux  
  
qt Matthew Earl Miller - Follow  
(tas) Published in Towards Data Science  
  
4minread - Jan 9,2020  
  
0) Listen (4 Share  
  
— Windows 10  
  
Openinapp 7  
  
@@| Medium = Q search  
  
L2)  
  
Poppler On Windows  
  
Intro:  
  
(Sign up ) Sign in  
  
Portable Document Format (PDFs) are everywhere and importing a popular python-  
  
package like PDF2Image, PDFtoText, or PopplerQt5 is a common approach to  
  
dealing with them. Unfortunately, unless you are working with a Linux machine,  
  
many users are reporting that these packages are returning errors because they rely  
  
on Poppler.  
  
Never heard of Poppler?  
  
Poppler is a utility for rendering PDFs and it’s common to Linux systems, but not  
  
Windows. So, naturally, if you want to use Poppler and its associated packages, we  
  
need to bridge the gap.

Page 2:

Let’s visit google and see what our options are...  
  
A quick Google (StackOverflow) search reveals that there are many other people  
  
having this problem and they are still looking for solutions.  
  
The Problem:  
  
Poppler and Python’s PDF-libraries, which leverage Linux-utilities, don’t play well  
with Windows.  
  
When we look for solutions, many of them are outdated, ineffective, too difficult,  
etc...  
  
The Solution:  
  
Of the purposed solutions, one solution appears to work well.  
Windows Subsystem for Linux (WSL).  
  
Actually, because of how powerful Windows Subsystem for Linux is, it’s a great  
  
solution for other problems which require Linux tools on a Windows machine.  
  
So, what is WSL?  
  
Windows Subsystem for Linux is a compatibility layer for Linux binary executables  
natively on Windows 10. It recently entered version two (WSL 2) and introduced a  
real Linux kernel. To put it plainly, WSL makes it feel like you’re working on a real  
  
Linux machine (and you are).  
  
Installation and Usage Guide — WSL  
  
Reference

Page 3:

In this section, we will, in five short steps, install and setup WSL. Afterwards, we  
will install and setup Poppler in a few short steps.  
  
Step 1:  
  
Run Window’s Powershell as an administrator.  
  
All Apps Documents Email Web  
  
Best match  
  
I> | Windows PowerShell  
  
App ;  
Apps  
Windows PowerShell ISE ?  
EY Windows PowerShell (x86) >  
Step 2:  
  
Enable WSL by executing the ‘Enable-WindowsOptionalFeature’ command:  
  
E¥ Administrator: Windows PowerShell ~ a U  
  
Windows PowerShell  
Copyright (C) Microsoft Corporation. All rights reserved.  
  
Try the new cross-platform PowerShell https://aka.ms/pscore6  
  
IPS C:\WINDOWS\system32>  
  
Enabling WSL  
  
1 Enable-WindowsOptionalFeature -Online -FeatureName Microsoft-Windows -Subsystem-Linux  
  
Windows\_Enable\_WSL hosted with @ by GitHub view raw  
  
Step 3:  
  
Activate the changes by restarting your computer.

Page 4:

Note that, Microsoft says, “This reboot is required in order to ensure that WSL can  
  
initiate a trusted execution environment.”  
  
Step 4:  
Now, you're back from a restart, your system’s WSL is enabled, and you are ready to  
  
install a Linux distribution.  
  
Go to the Window’s Store and search for WSL.  
  
<< Home Gaming Entertainment Productivity Deals 2 Search 2  
  
Best sellingsgames  
  
Download now  
  
x  
  
Access to over 100 PG games This is your 365 See what's new in Movies & TV Asphalt 9  
  
Getting WSL from Windows Store  
  
Step 5 (final):  
Click Ubuntu and choose to install. Note, mine is already installed, so you have to do  
  
some imagining here.  
  
@ This product is installed. | staunch |  
ubuntu a  
  
Canonical Group Limited \* Developer tools > Utilities  
kee OS Share  
  
Ubuntu on Windows allows one to use Ubuntu Terminal and run Ubuntu command  
line utilities including bash, ssh, git, apt and many more.  
  
Mare  
  
/E| EVERYOME  
  
Installing Ubuntu  
  
Installation and Usage Guide — Poppler:  
  
Step 1:

Page 5:

Enter WSL through a terminal like this one in VS Code. Notice that, once you enter  
WSL, the terminal prompt will change. You are now operating within a Linux  
  
machine! Exciting!  
  
SEARCH PR 5 | 16 OUTPUT DEBUG ILE TERMINAL  
  
Windows PowerShell  
Copyright (C) Microsoft Corporation. All rights reserved.  
  
Try the new cross-platform PowerShell https://aka.ms/pscore6  
  
PS C:\Users\Matthew\Desktop\Ken Batcher OCR> wsl  
  
Enter WSL  
  
Step 2:  
Conduct the following commands within the WSL-prompt. Note that, you can ignore  
some of the steps that deal with Tesseract-OCR and PyTesseract. These are for the  
  
demo-project which I share at the end of the article.

Page 6:

1 # Author: Matthew E. Miller  
  
2 # Date: 1/1/2020  
  
3 # Medium: https://medium.com/@matthew\_earl\_miller (where this is being published)  
4 # Github: https://github.com/matmil15  
  
5 # Linkedin: https://www. linkedin.com/in/matthew-miller-engineer/  
  
6 # StackOverflow: https://stackoverflow. com/users/11937169/matthew-e-miller?tab=profile  
7  
  
8 # Command 1: Enter Windows Subsystem for Linux  
  
9 PS C:\Users\Matthew\Desktop\Project> wsl  
  
10  
  
11 # Command 2: Cleanup  
  
12 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt-get clean  
  
13  
  
14. +# Command 3: Update  
  
15 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt-get update  
  
16  
  
17. # Command 4: Get Python 3 on your WSL  
  
18 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt install python3  
19  
2@ Command 5: Get Python PIP  
21 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt install python-pip  
22  
23 Command 6: Get poppler-utils  
24 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt install poppler-utils  
25  
26 Command 7: Get pdf2image (dependant on poppler and inspiration for article)  
27. ~—-user@device\_name: /mnt/c/Users/Matthew/Desktop/Project$ pip install pdf2image  
28  
29 Command 8: Get pathlib  
3@ user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ pip install pathlib  
31  
32 Command 9: Get pytesseract (if you're doing OCR)  
33 user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ pip install pytesseract  
34  
35 Command 10: Get tesseract-ocr (if you're doing OCR)  
  
36 ~=user@device\_name:/mnt/c/Users/Matthew/Desktop/Project$ sudo apt-get install tesseract-ocr  
Poppler\_WSL\_Commands hosted with @ by GitHub view raw  
  
Step 3 — Testing (final):  
  
Run a program with your newly acquired, ready-to-use, Poppler utilities.  
  
I’ve created this demo script, so you can use it if you don’t have your own. Although,  
  
you will need a PDF to mess with.

Page 7:

1 # Tesseract OCR  
2 import pytesseract  
3. from PIL import Image  
4 import sys  
5 from pdf2image import convert\_from\_path  
6 import os  
7 import io  
8  
9 # If you need to assign tesseract to path  
10 =# pytesseract.pytesseract.tesseract\_cmd = r'C:\Users\Matthew\AppData\Local\Tesseract-OCR\tesser  
11  
12 pdf\_path = 'pdfs/A Production Implementation of an Associative Arran Processor -STARAN - Rudolp  
13 output\_filename = "results.txt"  
14 pages = convert\_from\_path(pdf\_path)  
15 pg\_cntr = 1  
16  
17. ~=sub\_dir = str("images/" + pdf\_path.split('/')[-1].replace('.pdf','')[@:20] + "/")  
18 if not os.path.exists(sub\_dir):  
19 os.makedirs(sub\_dir)  
20  
21 for page in pages:  
22 if pg\_cntr <= 20:  
23 filename = "pg "+str(pg\_cntr)+'\_'+pdf\_path.split('/')[-1].replace('.pdf','.jpg')  
24 page.save(sub\_dir+filename)  
25 with io.open(output\_filename, ‘a+', encoding='utf8') as f:  
26 f .write(unicode("=========s=s==sssssssssssssssssesssssssssssssssssssssssa== PAGE " +  
27 f.write(unicode(pytesseract.image\_to\_string(sub\_dir+filename)+"\n"))  
28 f .write (unicode ("======s==sss=sssssssssssssssssssSsssssssssssssssssssssses Ssssssss5  
29 pg\_cntr = pg \_cntr +1  
CD >  
Demo Poppler On Windows OCR hosted with @ by GitHub view raw  
  
This code works by converting a PDF to JPG. Then, it conducts OCR and writes the  
OCR-results to an output-file.  
  
Conclusion:  
That’s it. You are certified Poppler-On-Windows.  
  
Enjoy the spoils of war! You have gained some seriously new and powerful skills.  
You are well on your way to becoming a more flexible developer (if you aren't  
already).  
  
Newly Acquired Skills:  
  
e Ability to successful manipulate PDFs with Python.

Page 8:

e Access to PDF2Image, PDFToText, or other Poppler-utils.  
  
e Windows Subsystem for Linux. \*\* A seriously powerful dev-tool \*\*  
  
Now What... What Can You Build?  
  
It’s so important to experiment with these new skills and solidify your  
  
understanding. True understanding comes with experience.  
  
My Poppler-On-Windows Project:  
  
I built an OCR application to help document the historical work of emeritus  
professor and famous computer scientist, Dr. Kenneth E. Batcher. It uses a PDF to  
image tool for JPEG-conversion. Then, it does OCR on the image and writes the  
results to an output file. Since this proof of concept works well enough, it’ll  
  
eventually be used on document-scans instead of PDFs.  
  
You can find the project here.  
  
A production implem|  
associative array pro  
  
by JACK A, RUDOLPH  
  
Goodyear Aerospace Corporation  
Akron, Ohio  
  
INTRODUCTION  
  
The associative or content-addressed i  
been an attractive concept to computer de  
since Slade and McMahon's 1957 paper!  
  
“catalog” memory. Associative memories q  
from the continuing problem presented. by}  
coordinate-addressed memory which requi  
  
OCR App — In Development  
  
Python3 Python Programming Programming Windows 10 Vscode

Page 9:

© eae  
  
Written by Matthew Earl Miller  
  
14 Followers - Writer for Towards Data Science  
  
BSCS Kent State 2019, MSCS Case Western Reserve 2021, Software Developer and Graduate Student.  
LinkedIn — https://www.linkedin.com/in/matthew-miller-engineer  
  
More from Matthew Earl Miller and Towards Data Science  
  
@ matthew Earl Miller  
  
React Google Lighthouse FAB  
  
Easily Showcase Your App’s Performance  
  
2minread - Jul 27,2021  
  
8s Q [Ly

Page 10:

3 Cristian Leo iN Towards Data Science  
  
The Math behind Adam Optimizer  
  
Why is Adam the most popular optimizer in Deep Learning? Let’s understand it by diving into  
its math, and recreating the algorithm.  
  
16 minread - Jan 30,2024  
  
Sy isk OQ 13 Cy  
  
a3 Siavash Yasini iN Towards Data Science  
  
Python’s Most Powerful Decorator

Page 11:

And 5 ways to use it in data science and machine learning  
  
- 1minread - Feb 2,2024  
  
S) 2k Q6 Cj  
  
Selection View Go Debug Terminal Help  
  
g Matthew Earl Miller in Towards Data Science  
  
Rapidly Document Authorship with VS Code Snippets  
  
Be confident that you’re getting credit for your work.  
  
2minread - Jan3, 2020  
  
4 Q [y  
  
See all from Matthew Earl Miller  
  
See all from Towards Data Science

Page 12:

Recommended from Medium  
  
) George Stavrakis iN Towards Data Science  
  
Extracting text from PDF files with Python: A comprehensive guide  
  
A complete process to extract textual information from tables, images, and plain text from a  
PDF file  
  
+ . 17minread - Sep 22,2023  
  
&) 12k QQ 23 Cj

Page 13:

@ Gaurav Garg  
  
How to Extract Text from PDFs and Images for LLMs Use  
  
Large language models like GPT-3 rely on vast amounts of text data for training. While there  
are many open datasets available, sometimes...  
  
4minread - Aug 22,2023  
  
S) 175 Q1  
  
Lists  
  
General Coding Knowledge  
  
20 stories - 924 saves  
  
Coding & Development  
  
11stories - 449 saves  
  
io) Stories to Help You Grow as a Software Developer  
“ae 19 stories - 817 saves  
  
aa) Our Favorite Productivity Advice  
S| a 9stories - 383 saves

Page 14:

é Alice Yang  
  
with Read or Extract Text from PDF with Python—A Comprehensive  
Guide  
  
PDF documents such as research papers, legal documents, contracts, or reports often contain  
important textual information. By extracting...  
  
5minread - Sep 5, 2023  
  
6 Qi Ci  
  
we Search tracted a  
  
fa) Alex Stock

Page 15:

Extract Images from PDF Documents in Python  
  
Extracting images from a PDF file can be a useful and practical task in various situations.  
Whether you need to repurpose images for a...  
  
3minread - Oct12, 2023  
  
S26 Q Ci  
  
ES voxe.si ==.  
  
Optical Character  
  
Recognition with  
PyTesseract  
  
Parse PDFs and filter by  
contents in FiftyOne  
  
© jacob Marks, Ph.D. in Voxel51  
  
Optical Character Recognition with PyTesseract  
Parse PDFs and filter by contents in FiftyOne  
  
10 minread - Sep 21, 2023  
  
&) 278 QQ [a

Page 16:

G& Jason Roell  
  
Ultimate Python Cheat Sheet: Practical Python For Everyday Tasks  
  
This Cheat Sheet was born out of necessity. Recently, | was tasked with diving into a new  
Python project after some time away from the...  
  
33 minread - Jan 30,2024  
  
&) 13k Q 22 Cy  
  
See more recommendations