How to Use Tesseract on Windows with Python

https://medium.com/@ahmetxgenc/how-to-use-tesseract-on-windows-fe9d2a9ba5c6

Tesseract is an optical character recognition software which developed by Google. Its an open source OCR tool. There are many versions of tesseract but we will use the 4.0 version.

In version 4, Tesseract has implemented a Long Short Term Memory (LSTM) based recognition engine. LSTM is a kind of Recurrent Neural Network (RNN). The LSTM-based recognition works much more effectively than the old (CNN-based) recognition processes.

Thanks to tesseract, we will be able to save the contents of our images as text files.

Installation

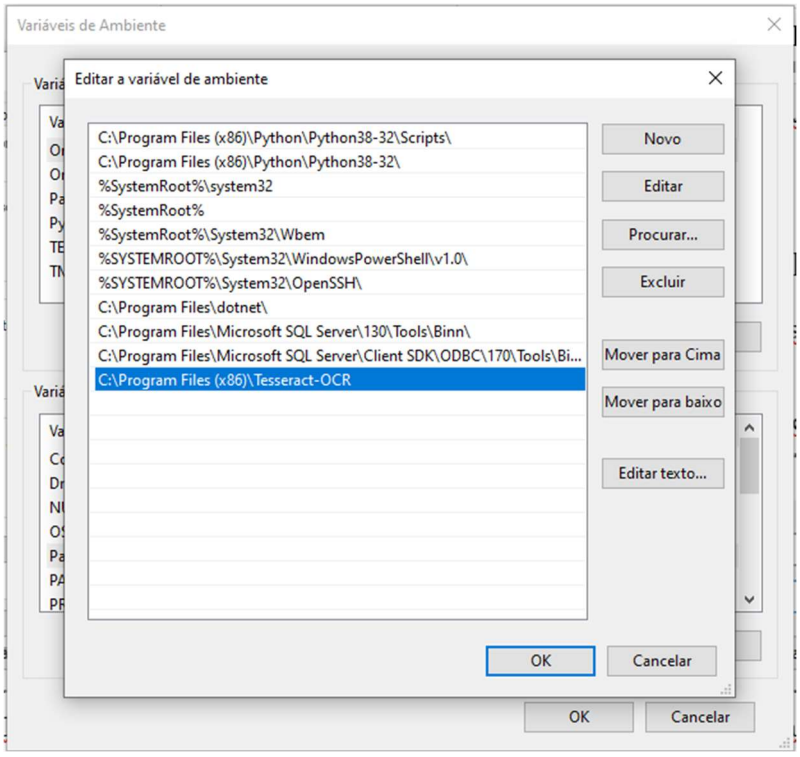
The installation is depends on your operating system. Now we're going to go through the ***windows***. First, let's download and install tesseract thorugh this link (http://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-setup-4.00.00dev.exe). (It downloads an exe file.) We setup the exe file easily.

After that, we should add an PATH to windows system variables. Actually it's an easy step. Firstly we find and copy the root folder of the tesseract installation. It will shold be like that :

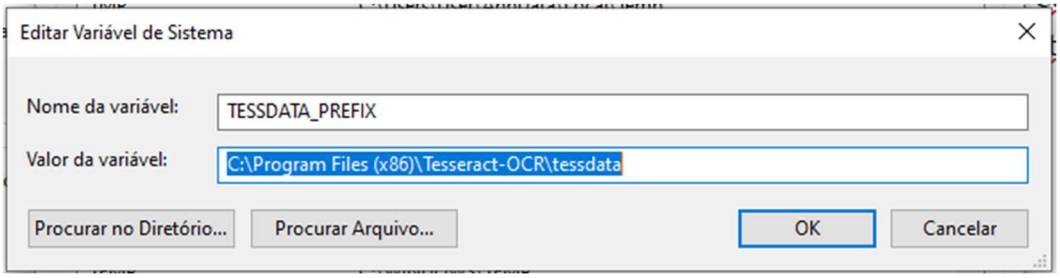
```
C:\Program Files\Tesseract-OCR
```

And then in the search bar of the windows Advanced System Settings

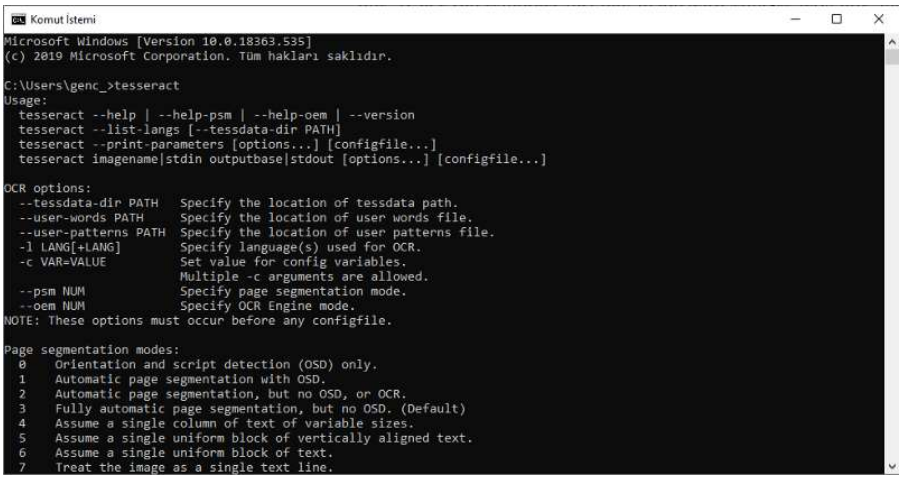
Advanced system settings > Advanced > Environment variables > PATH > New



We paste the source path which copied and we save this configurations. After this step the computer must be rebooted to apply configurations.
Make sure the TESSDATA_PREFIX environment variable is set to your "tessdata" directory



The tesseract installation completed. You can confirm the installation from the command line. When we run **tesseract** command on the command line, it should give us information about the program.



Now we can move on to the python part. To use tesseract on python, we should download ***pytesseract*** library. This library can be downloaded via pip to the environment you are using.

```
pip install pytesseract
```

Now the tesseract is ready to use!!

Coding

It's realy simple to use tesseract. The hard part is the optimizing the settings.
Because if you want to make a successful ocr, you need to be careful in image processing step and ocr settings.
Let's apply OCR to the receipt.

Importing The Libraries

```
import pytesseract
from PIL import Image
import cv2
import numpy as np
```

Setting DPI Value of Image

Dots per inch (DPI, or dpi) is a measure of video or image scanner dot density. DPI value is an important thing to run OCR. Because if DPI value is lower then 300, it may reduce the success of OCR.

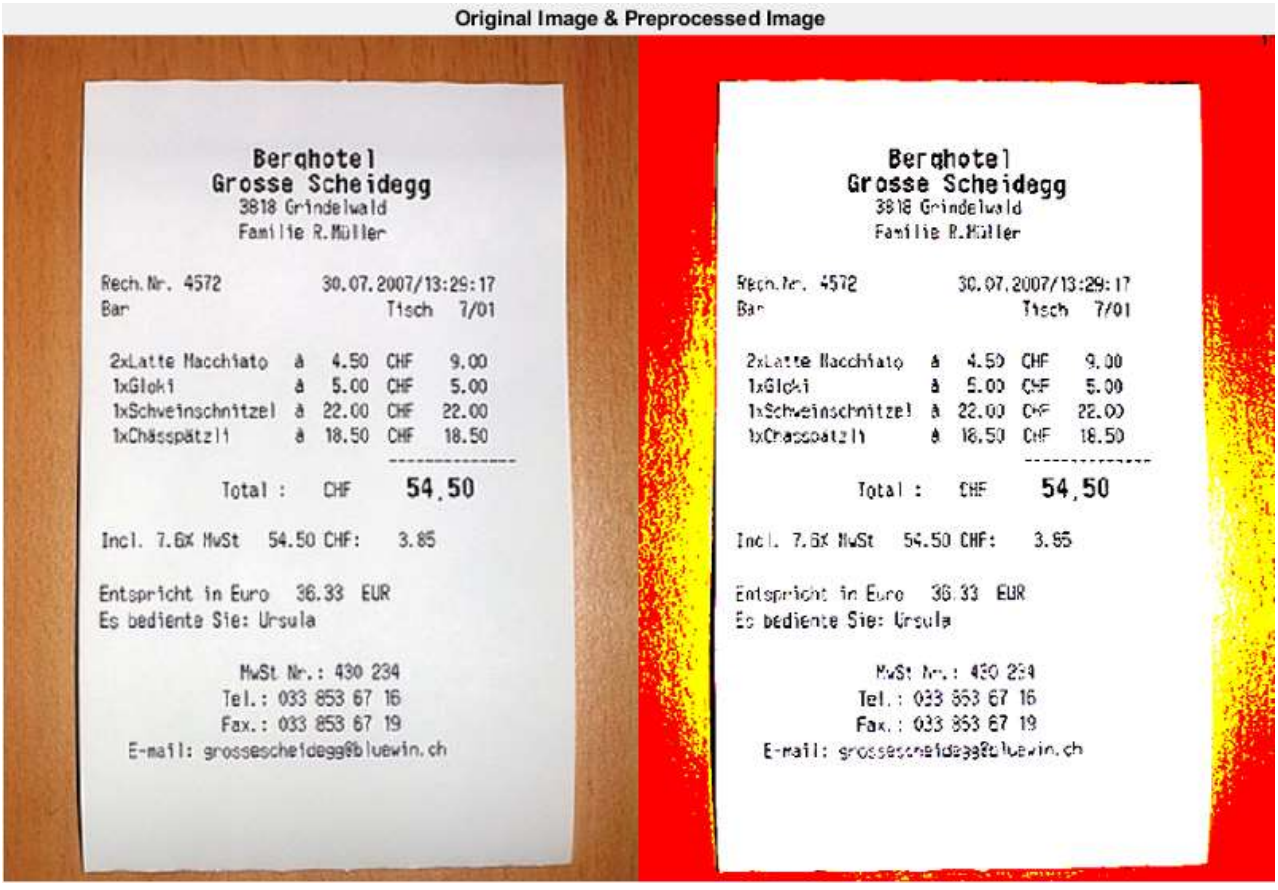
```
file_path= 'receipt.jpg'
im = Image.open(file_path)
im.save('ocr.png', dpi=(300, 300))
```

Applying Some Techniques to Make Image Cleaner

Firstly we scale our image with x2. If characters are small then we need to scale the image to recognize it. After that we apply a simple threshold technique. Its ***Binary Threshold.*** First you shold try with 127 value after that different variables can be tried. The thrashold change the pixel with black if the pixel value over the threshold value. If we make the image grayscale, it will give us a black and white image.

There is different threshold techniques. You can check the source website with this link (https://docs.opencv.org/3.4.0/d7/d4d/tutorial_py_thresholding.html).

```
image = cv2.imread('ocr.png')
image = cv2.resize(image, None, fx=2, fy=2, interpolation=cv2.INTER_CUBIC)
retval, threshold = cv2.threshold(image,127,255,cv2.THRESH_BINARY)
```



Running Tesseract

Now we can run tesseract. It has an ***image_to_string()*** function. It gives us a string as an output.

```
text = pytesseract.image_to_string(treshold)
```

Saving Output

We can save the output with the following code.

```
with open(“Output.txt”, “w”,5 ,”utf-8") as text_file:
text_file.write(text)
```

The output of the OCR is as follows:

```
Berghote 1
Grosse Scheidegg
3818 Grindelsald
Fam ie

Rech. tin, 4572 30. 07, 20077 13:29: 17
Ban Tach — 7/01

Pxlatte Macchiato — 4 4.50 CHF — 9,00
IxGlcki a 5.00 CF — 5.00
IxSchusinschnitze} A 22.00 OF 22.00
IxChasspatz li a 18,50

Total : _ {HF
Incl. 7.6% HuSt — 54.50 CHF: 3.85

Entsnricht in Euro — 36.33 EUR
Es bediente Sig: Ursula

BuSt Nn. : 430 25
Tel.: 033 853 67 16
Fax. : 033 853 67 19
E-mail: grossescn¢idegg@bluswin. ch
```

The result is very successful. If higher success is desired, different operations can be applied to the image.

The tesseract github page was referenced.