Manual for Tesseract-OCR

Practical1: Use pytesseract library in Python for optical character recognition from (i) an image file (ii) a multi-page pdf file.

Tesseract-OCR in Jupyter Notebook

1. Down load the tesseract EXE and install. https://osdn.net/projects/sfnet_tesseract-ocr-alt/downloads/tesseract-ocr-setup-3.02.02.exe/
2. Check the path of tesseract-ocr
3. Open the jupyter notebook and terminal for it
4. Write pip install pytesseract

5. Copy the following code in the jupyter note book. Remember to update the path as per the need.

Code

```
import pytesseract as tess

try:

  import Image

except ImportError:

  from PIL import Image

#Changing the directory as the pytesseract library is only wrapper and the real engine has to be referenced seperately

tess.pytesseract.tesseract_cmd = r'C://Program Files//Tesseract-OCR//tesseract'

img = Image.open(r'E://OCR.jpg')

print(img)

text = tess.image_to_string(img)

print(text)
```

Tesseract-OCR in Google Colab

1. Install tesseract

```
!sudo apt install tesseract-ocr
```

```
!pip install pytesseract
```

2. Import the files

```
import pytesseract
import shutil
import os
import random
try:
  from PIL import Image
except ImportError:
  import Image
from google.colab.patches import cv2_imshow
import cv2
from matplotlib import pyplot as plt
```

3. Read the Image files

```
from google.colab import files
uploaded = files.upload()
```

After running these two lines of the code a browsing option populates, that allows you to select a file from your local derive. Select an Image file from the local drive.

4. Write this code for image to string conversion and run. Remember to

```
img = cv2.imread('OCRimg_2_noisy.png')
img1 = Image.open('OCRimg_2_noisy.png')
cv2_imshow(img)
plt.imshow(img1)
ocrinfo = pytesseract.image_to_string(img)
print(ocrinfo)
```

installing the tesseract-OCR engine and accessing from the python code is demonstrated above.

You are required to explore other functions in the library like bounding box, masking etc.

Read from a multipage PDF file