practical_1.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share S

Files

+ Code + Text

RAM Disk Editing

..
sample_data
MyPetDog.png

Disk 70.85 GB available

Practical 1

Name : Aayush Shah
Roll no. : 19BCE241
Course : Machine Learning(2CS501)
Practical : 1
Aim : Use pytesseract library in Python for optical character recognition from
(i) an image file (ii) a multi-page pdf file

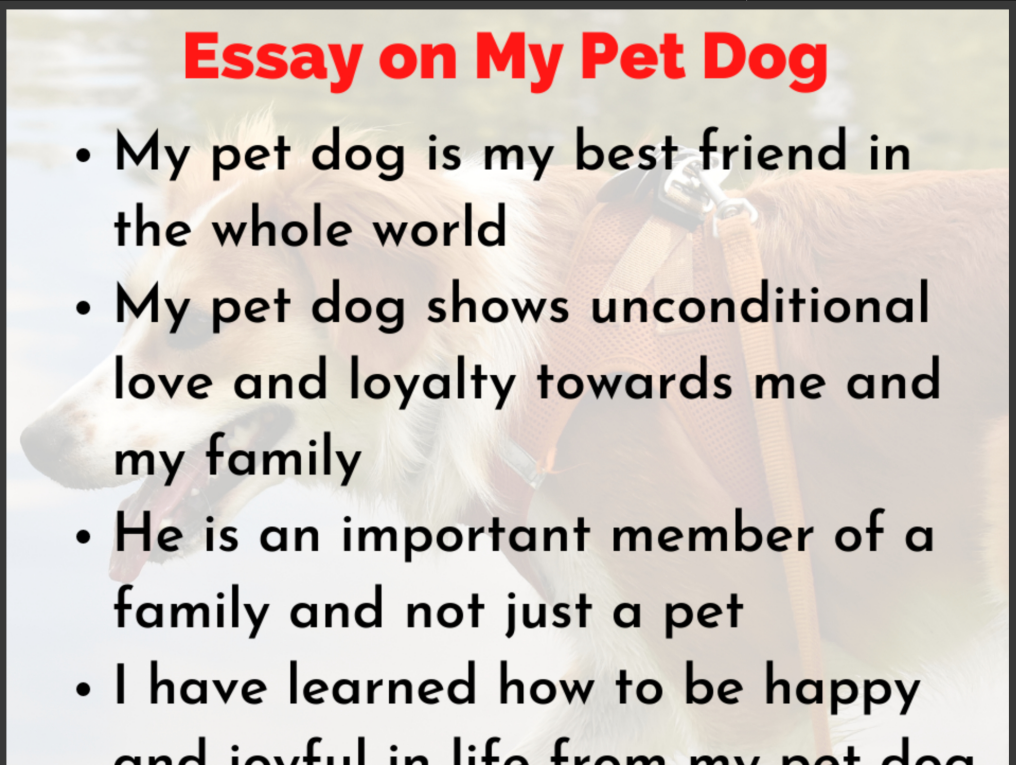
```
# Installation of required packages in google colab
!pip install pytesseract
!sudo apt install tesseract-ocr
```

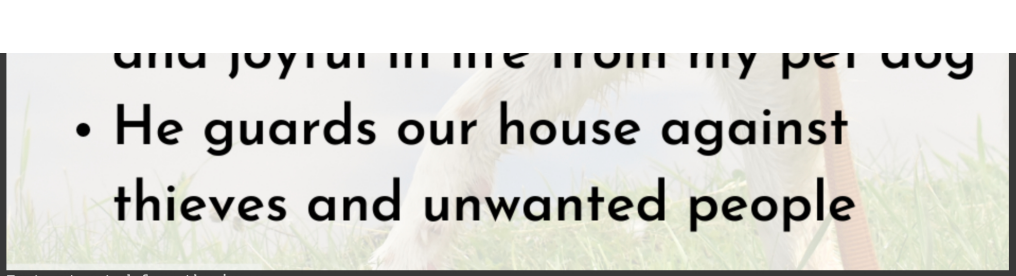
```
Requirement already satisfied: pytesseract in /usr/local/lib/python3.7/dist-packages (0.3.8)
Requirement already satisfied: Pillow in /usr/local/lib/python3.7/dist-packages (from pytesseract) (7.1.2)
Reading package lists... Done
Building dependency tree
Reading state information... Done
tesseract-ocr is already the newest version (4.00~git2288-10f4998a-2).
The following package was automatically installed and is no longer required:
  libnvidia-common-460
Use 'sudo apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
```

```
# Importing packages
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
```

```
# Uploaded png image named 'MyPetDog'
from google.colab import files
# uploaded = files.upload()
```

```
# Setting path and extracting text.
pytesseract.pytesseract.tesseract_cmd = r'/usr/bin/tesseract'
img = cv2.imread('MyPetDog.png')
img1 = Image.open('MyPetDog.png')
cv2_imshow(img)
plt.imshow(img1)
ocrinfo = pytesseract.image_to_string(img1)   #can be either img or img1; both works!
print('Text extracted from the image : ','\n',ocrinfo)
```






```
Text extracted from the image :
 Essay on My Pet Dog

« My pet dog is my best friend in
```

Explanation :

1. Imported pytesseract and tesseract-ocr packages which are used for optical character recognition engine for various operating systems. These will recognize and “read” the text embedded in images.
2. Then path for the tesseract library is set using 'pytesseract.pytesseract.tesseract_cmd'.
3. Using cv2's imread function, image is loaded from the path.
4. Then, using pytesseract's image_to_string function, the text from the image will be extracted and stored in the ocrinfo variable.
5. The extracted image and text are shown in the output using pytesseract, cv2, matplotlib and PIL packages.

PDF to Text

```
[ ]  1 !pip install pdf2image
```

```
Requirement already satisfied: pdf2image in /usr/local/lib/python3.7/dist-packages (1.16.0)
Requirement already satisfied: pillow in /usr/local/lib/python3.7/dist-packages (from pdf2image) (7.1.2)
```

```
[ ]  1 from PIL import Image
     2 import pytesseract
     3 import sys
     4 from pdf2image import convert_from_path
     5 import os
     6 import pathlib
```

```
[ ]  1 # !apt-get install poppler-utils
     2 # pathlib.Path('/generated_images/').mkdir(parents=True, exist_ok=True)
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
poppler-utils is already the newest version (0.62.0-2ubuntu2.12).
The following package was automatically installed and is no longer required:
  libnvidia-common-460
Use 'apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
```

```
[ ]  1 PDF_file = "DogsAsHumanCompanions.pdf"
     2 pages = convert_from_path(PDF_file, 500)
     3 image_counter = 1
     4 for page in pages:
     5   filename = "page_"+str(image_counter)+".jpg"
     6   page.save(filename, 'JPEG')
     7   image_counter = image_counter + 1
     8
     9 filelimit = image_counter-1
    10
    11 extracted_text = ''
    12
    13 for i in range(1, filelimit + 1):
    14   filename = "page_"+str(i)+".jpg"
    15   text = str(((pytesseract.image_to_string(Image.open(filename)))))
    16   text = text.replace('-\n', '')
    17   extracted_text += text
```

```
[ ]  1 print('Extracted text : \n',text)
```

```
Extracted text :
 their pets, the dog owners averaged 35.3 hours per week and the cat owners averaged
33.2 hours. For dog owners, 44% of this time was estimated as play, as compared with
36% for cat owners (J. Angus, personal communication).

Fig. 12.5. Attraction of young children to animals. Young toddlers respond to both
mechanical and live dogs, but a real dog elicits the stronger interest (Kidd & Kidd, 1987).
Photograph: Joan Borinstein.

Touch

A study of three- to four-year-old children’s interactions with dogs revealed that 67% of
these interactions involved body contact with the dog, such as putting a hand on the dog,
patting it or hitting it. In contrast, vocal and verbal behavior comprised only 9% of the
interactions (Millot & Filiatre, 1986). In a subsequent study touching was again the most
frequent behavior shown in the presence of a dog, accounting for 40% of all child-dog
interactions (Filiatre et al., 1988).

In an analysis of 1105 photographs of dogs or cats in a family setting submitted to a
national photographic contest, Katcher & Beck (1985) found that 97% of the pictures
illustrated people and animals touching each other, generally with the heads of the animal
and human close together. Over 92% showed a dyadic relationship, with one person and
```

```
one animal occupying the center of the photograph. Touching was also a primary mode
of interaction with a dog in a study of nursing home residents (Neer, Dorn & Grayson,
1987). Of the nine different types of interaction recorded involving the dog, grooming
and touching were the two most commonly employed by residents.

The value of dogs for different types of people

Albert & Bulcroft's (1987, 1988) Rhode Island study found that households with children
at home tended to have more pets than either widows or families with an 'empty nest', or
with an infant. However, feelings of attachment to the pet were lowest in families where
children were at home. Although pet ownership was highest among households
containing large families, attachment to pets was highest among people living alone and
among couples who did not have children living at home. The authors noted that the
single, divorced and widowed individuals and childless couples who were most attached
to their pets also expressed more anthropomorphic attitudes to their pets, particularly in
relation to dogs. In a longitudinal study of older people (a population that experiences
increasing losses), Lago, Connell & Knight (1985) found that persons who stayed at
home and spent more time with the animal also became more attached and formed a
stronger relationship with it.

An 'invisible cord' often seems to connect a dog to its owner (Serpell, 1986a). Almost
invariably, dogs are more attentive to their owners than their owners are to them. In a
study of ten families' interactions with their dogs, the associations between the dog and
the adult family members were found to differ between families with and without
children (Smith, 1983). In childless families the people and the dog interacted more
```

Explanation :

1. Imported package named pdf2Image which used to convert PDF to PIL Image object.
2. Generated file name based on the number of pages in pdf. Then generated images in for loop with the file name generated above.
3. Extracted the text from the image in the same way as we have previously done. Created empty string variable named extracted_text and appended every image's extracted text to it.
4. Displayed the extracted text as output.

Limitations :

1. By doing trial and error on this package, I came to know that it's accuracy is not as good and not as fast as the currently google's mobile app google lens is providing.
2. It is also not capable of recognizing handwritten text as apple's ipad does.
3. Also if the file contain the text written in two or more columns, then it might fail to recognize this pattern and can result in absurd text string.

Conclusion :

From this practical, I learned about the optical character recognition tool for python named pytesseract. Pytesseract is able to read and extract the text from files types like images, pdfs etc. with the use of this library, one can easily implement great ocr features in their apps and projects without significant effort.