Final Presentation

Language Translation using Pytesseract from Image

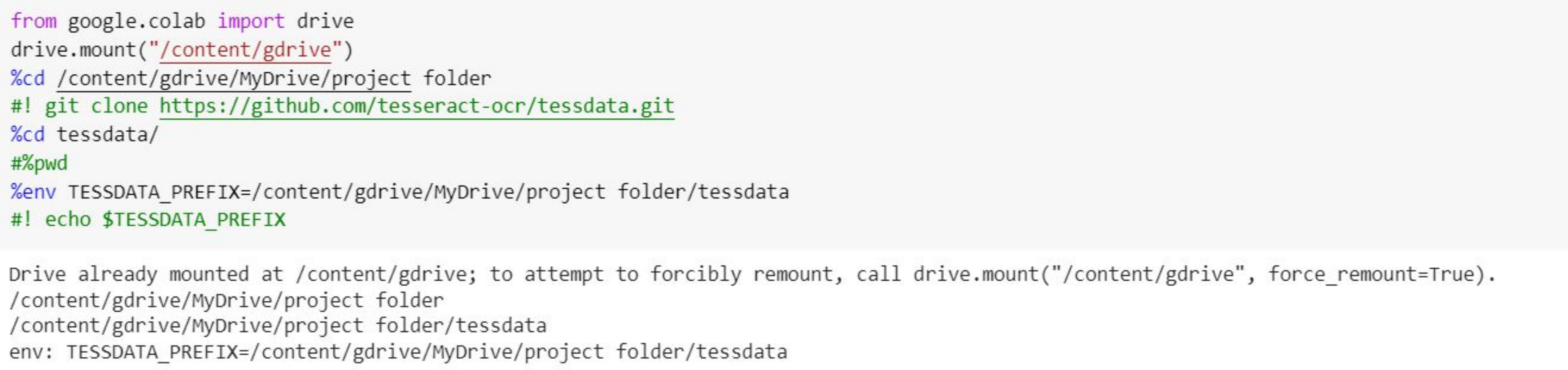
Kyle Stone, Vignesh Nokanaidu

Code goals

1. Accept an input image
2. Detect text in non-English languages
3. Translate the OCR’d text from the given input language into English
4. Display the results to our terminal

Setting up the environment and path

1. Download Tesseract’s language packs manually from GitHub and install them.
2. Set the `TESSDATA_PREFIX` environment variable to point to the directory containing the language packs.

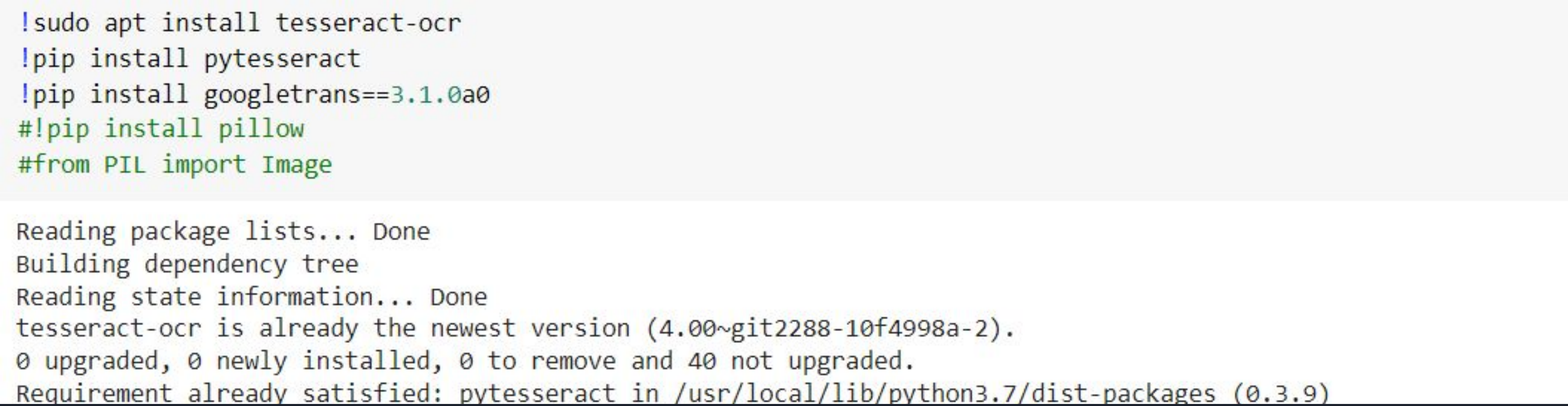


```
from google.colab import drive
drive.mount("/content/gdrive")
%cd /content/gdrive/MyDrive/project folder
#! git clone https://github.com/tesseract-ocr/tessdata.git
%cd tessdata/
#%pwd
%env TESSDATA_PREFIX=/content/gdrive/MyDrive/project folder/tessdata
#! echo $TESSDATA_PREFIX
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
/content/gdrive/MyDrive/project folder
/content/gdrive/MyDrive/project folder/tessdata
env: TESSDATA_PREFIX=/content/gdrive/MyDrive/project folder/tessdata
```

Tesseract with Non English Language

3. Installing the necessary softwares.



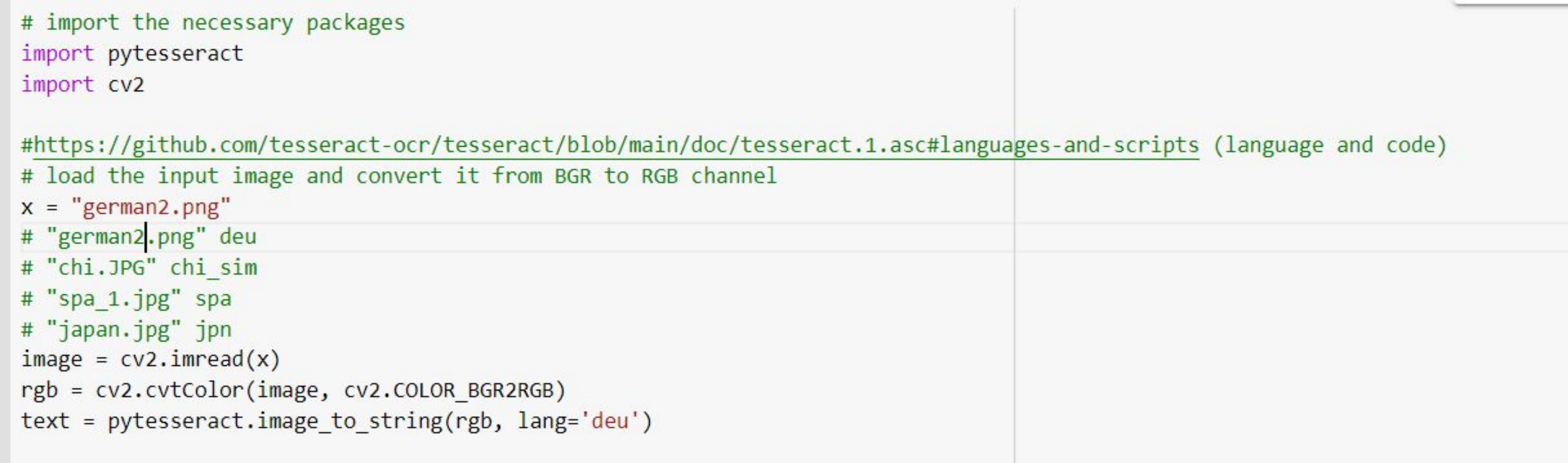
```
!sudo apt install tesseract-ocr
!pip install pytesseract
!pip install googletrans==3.1.0a0
#!pip install pillow
#from PIL import Image
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
tesseract-ocr is already the newest version (4.00~git2288-10f4998a-2).
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
Requirement already satisfied: pytesseract in /usr/local/lib/python3.7/dist-packages (0.3.9)
```

Tesseract with Non English Language continued...

> `cv2.imread` loads the image using the opencv module

> while the `cv2.cvrColor` swaps the color channels from Blue-Green-Red (BGR) to Red-Green-Blue (RGB) so the image is compatible with Tesseract, which takes an input image with an RGB color channel ordering.

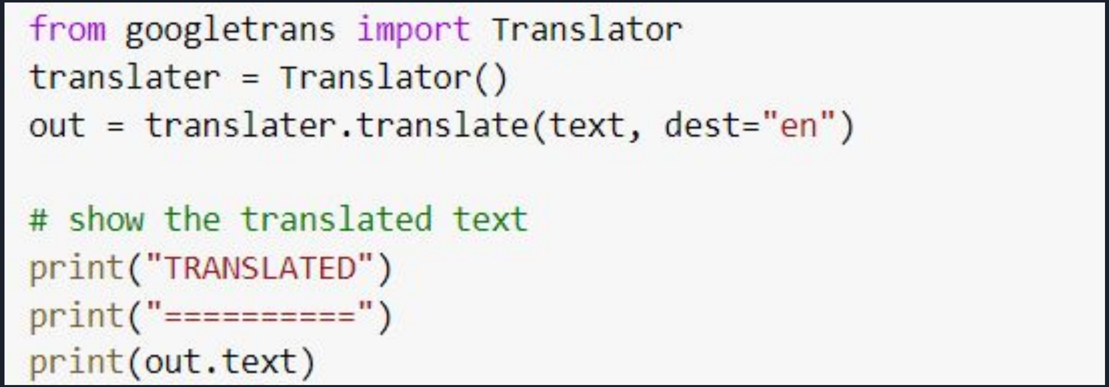


```
# import the necessary packages
import pytesseract
import cv2

#https://github.com/tesseract-ocr/tesseract/blob/main/doc/tesseract.1.asc#languages-and-scripts (language and code)
# load the input image and convert it from BGR to RGB channel
x = "german2.png"
# "german2.png" deu
# "chi.JPG" chi_sim
# "spa_1.jpg" spa
# "japan.jpg" jpn
image = cv2.imread(x)
rgb = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
text = pytesseract.image_to_string(rgb, lang='deu')
```

Initiating text translation

Now that we have Tesseract set up and have added support for a non-English language, we can move on to the text translation.

Next, we will wrap up this section by showing the OCR’d results from Tesseract in the native language

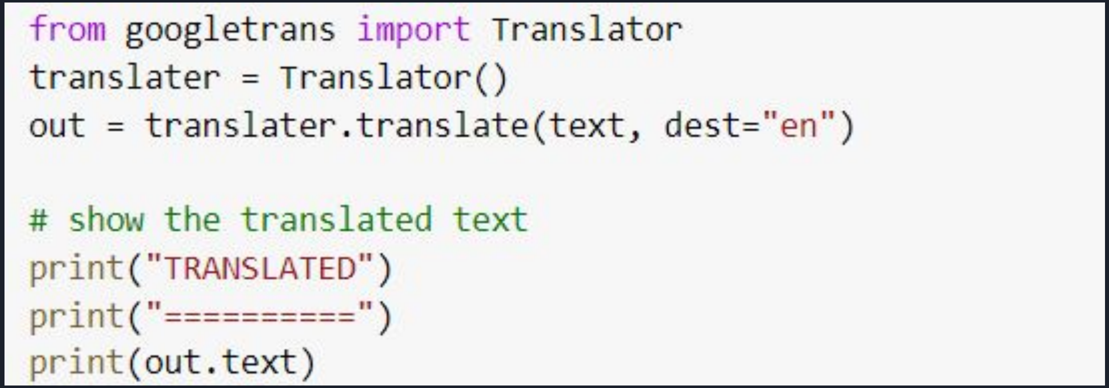


```python
from googletrans import Translator
translater = Translator()
out = translater.translate(text, dest="en")

# show the translated text
print("TRANSLATED")
print("==========")
print(out.text)
```

Initiating text translation

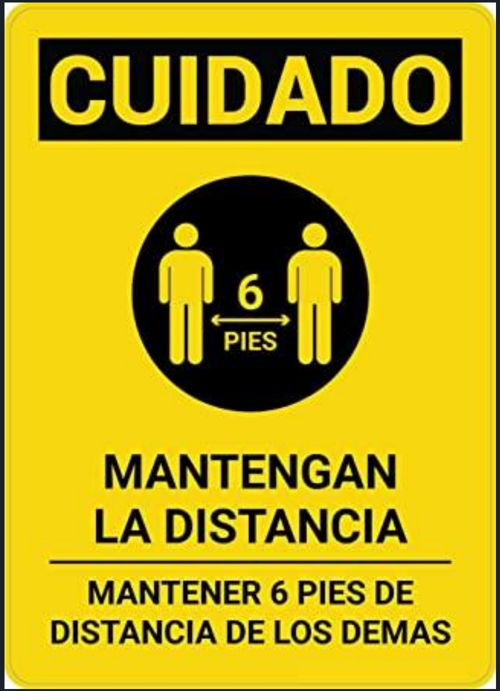
The extracted text is then converted to the required language (which in this case is English) using google translate software - we have installed before.

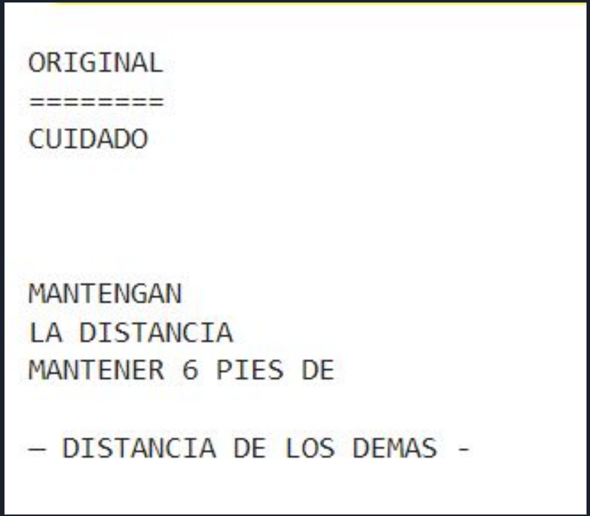
We wrap up by printing the results of the translated text. Now you have a complete workflow that includes OCR’ing the text in the native language and translated it into your desired language.



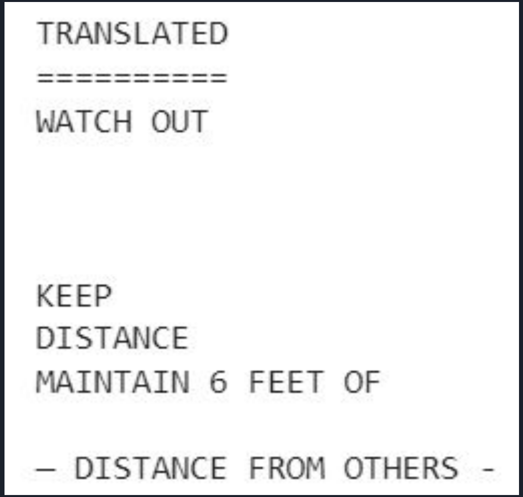
```
from googletrans import Translator
translater = Translator()
out = translater.translate(text, dest="en")

# show the translated text
print("TRANSLATED")
print("==========")
print(out.text)
```

Results_Spanish



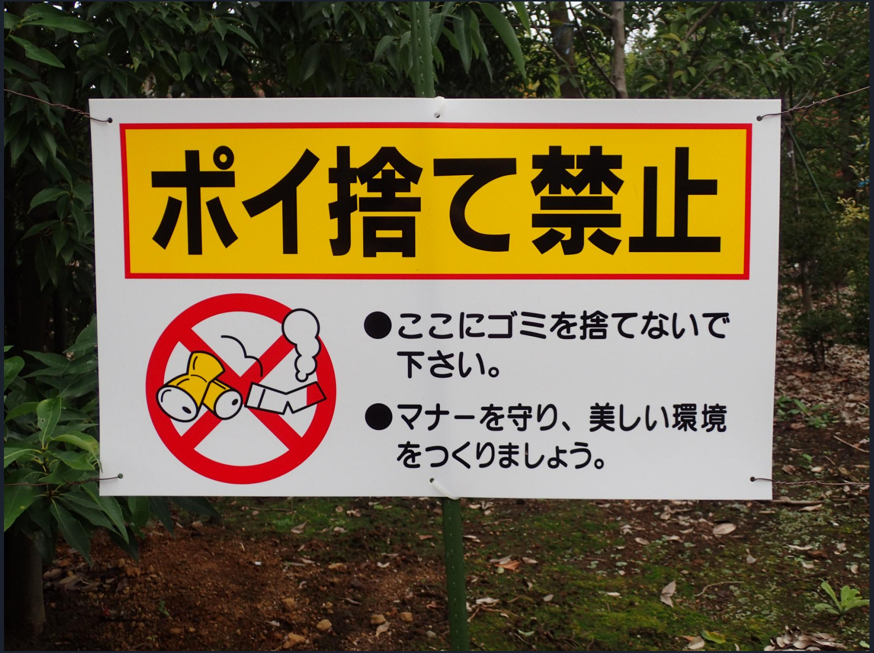


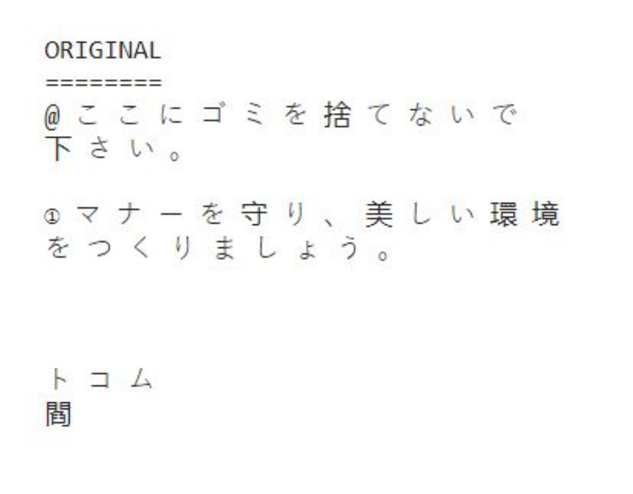
```
ORIGINAL
========
CUIDADO



MANTENGAN
LA DISTANCIA
MANTENER 6 PIES DE

— DISTANCIA DE LOS DEMAS -
```



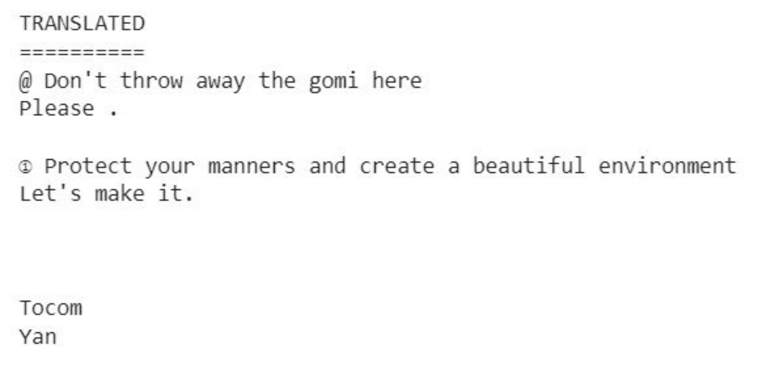
```
TRANSLATED
==========
WATCH OUT



KEEP
DISTANCE
MAINTAIN 6 FEET OF

— DISTANCE FROM OTHERS -
```

Results_Japanese


ポイ捨て禁止
●ここにゴミを捨てないで下さい。
●マナーを守り、美しい環境をつくりましょう。

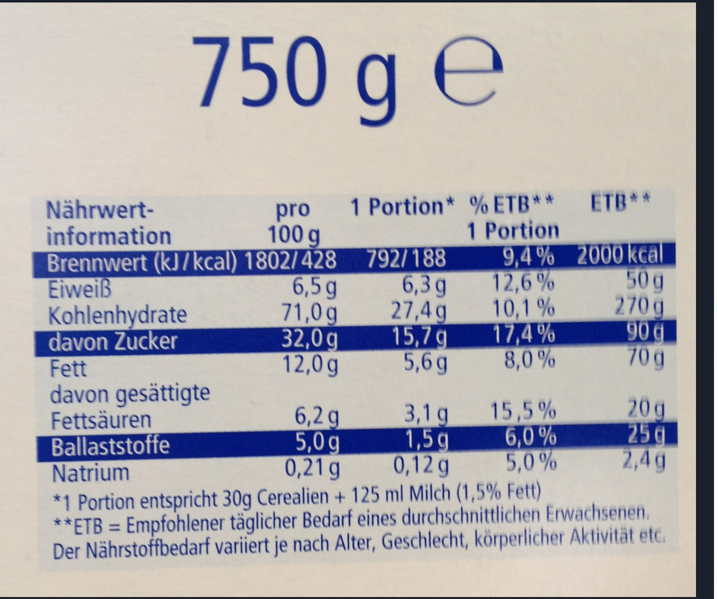


```
ORIGINAL
========
@ここにゴミを捨てないで
下さい。

①マナーを守り、美しい環境
をつくりましょう。



トコム
閻
```



```
TRANSLATED
==========
@ Don't throw away the gomi here
Please .

① Protect your manners and create a beautiful environment
Let's make it.



Tocom
Yan
```

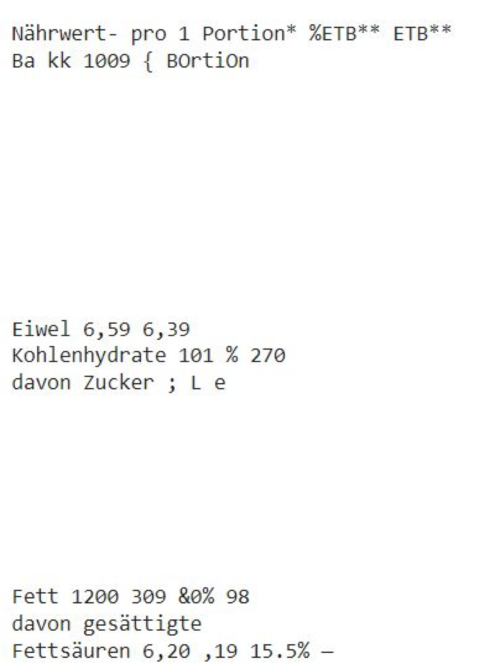
Results_German



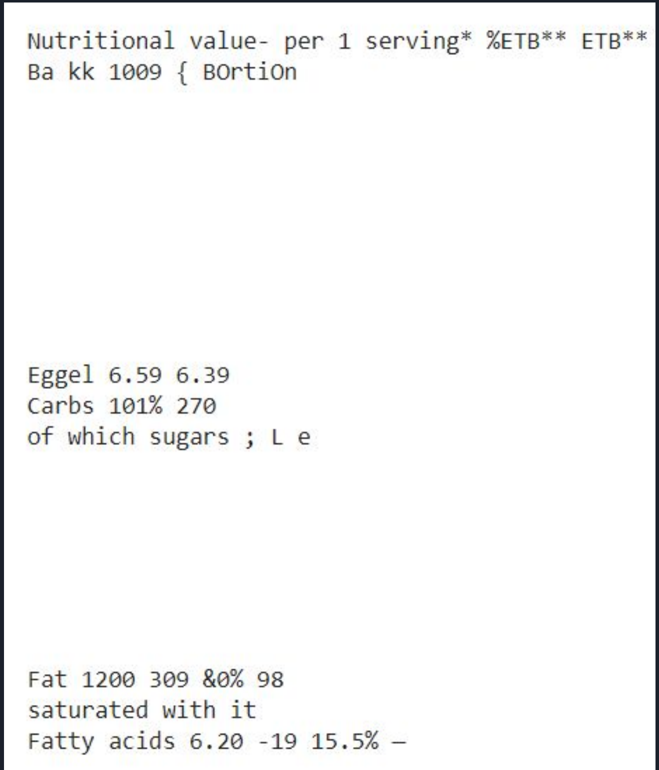
750 g ℮

Nährwert-information	pro 100 g	1 Portion*	% ETB** 1 Portion	ETB**
Brennwert (kJ/kcal)	1802/428	792/188	9,4 %	2000 kcal
Eiweiß	6,5 g	6,3 g	12,6 %	50 g
Kohlenhydrate	71,0 g	27,4 g	10,1 %	270 g
davon Zucker	32,0 g	15,7 g	17,4 %	90 g
Fett	12,0 g	5,6 g	8,0 %	70 g
davon gesättigte Fettsäuren	6,2 g	3,1 g	15,5 %	20 g
Ballaststoffe	5,0 g	1,5 g	6,0 %	25 g
Natrium	0,21 g	0,12 g	5,0 %	2,4 g

*1 Portion entspricht 30g Cerealien + 125 ml Milch (1,5% Fett)
**ETB = Empfohlener täglicher Bedarf eines durchschnittlichen Erwachsenen. Der Nährstoffbedarf variiert je nach Alter, Geschlecht, körperlicher Aktivität etc.

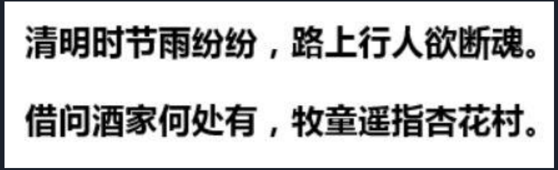


```
Nährwert- pro 1 Portion* %ETB** ETB**
Ba kk 1009 { BOrtiOn




Eiwel 6,59 6,39
Kohlenhydrate 101 % 270
davon Zucker ; L e



Fett 1200 309 &0% 98
davon gesättigte
Fettsäuren 6,20 ,19 15.5% —
```



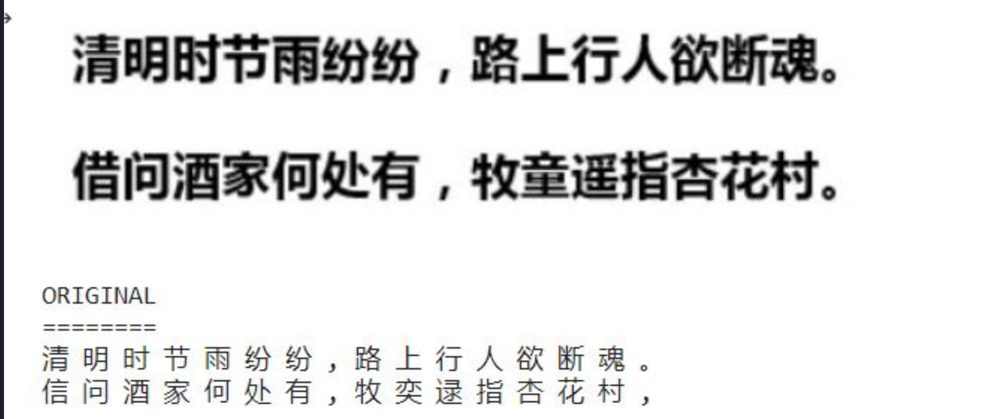
```
Nutritional value- per 1 serving* %ETB** ETB**
Ba kk 1009 { BOrtiOn




Eggel 6.59 6.39
Carbs 101% 270
of which sugars ; L e



Fat 1200 309 &0% 98
saturated with it
Fatty acids 6.20 -19 15.5% —
```

Results_Chinese



清明时节雨纷纷，路上行人欲断魂。

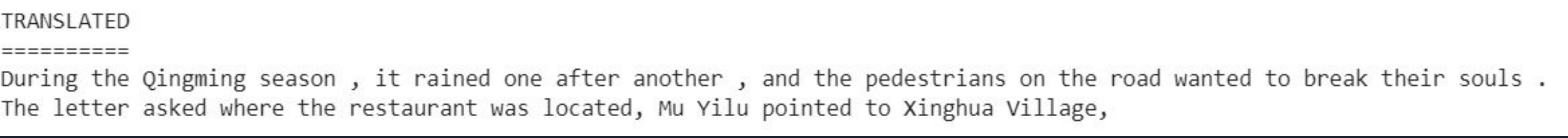
借问酒家何处有，牧童遥指杏花村。



清明时节雨纷纷，路上行人欲断魂。

借问酒家何处有，牧童遥指杏花村。

```
ORIGINAL
=========
清 明 时 节 雨 纷 纷 , 路 上 行 人 欲 断 魂 。
信 问 酒 家 何 处 有 , 牧 奕 逯 指 杏 花 村 ,
```



```
TRANSLATED
==========
During the Qingming season , it rained one after another , and the pedestrians on the road wanted to break their souls .
The letter asked where the restaurant was located, Mu Yilu pointed to Xinghua Village,
```

Summary

1. Manually download the Tesseract language packs
2. Set the `TESSDATA_PREFIX` environment variable to point the language packs
3. Verify that the language packs directory is correct
4. Extract the text using pytesseract command
5. Translate the text using google translate