```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import os
IMG_FOLDER = '/content/drive/MyDrive/NLP driven Invoice Management System/invoice images'

print('Found', len(os.listdir(IMG_FOLDER)), 'files')
```

```
Found 2000 files
```

```
#installing dependencies
!pip install pytesseract pillow transformers
#pytesseract for Optical Character Recognition(OCR)
#pillow to open image files
#transformers to use pre-trained transformer models
```

```
Requirement already satisfied: pytesseract in /usr/local/lib/python3.11/dist-packages (0.3.13)
Requirement already satisfied: pillow in /usr/local/lib/python3.11/dist-packages (11.2.1)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11/dist-packages (from pytesseract) (24.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.32.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->trans
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.3
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->t
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.4.2
```

```
!apt update
!apt install tesseract-ocr
!tesseract --version
```

```
Get:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,632 B]
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:3 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Get:4 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease [1,581 B]
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Hit:7 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:8 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Get:9 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Hit:10 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:11 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [9,017 kB]
Get:12 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  Packages [1,765 kB]
Get:13 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,740 kB]
Get:14 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [2,984 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,553 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,295 kB]
Fetched 21.8 MB in 6s (3,603 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
36 packages can be upgraded. Run 'apt list --upgradable' to see them.
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' d
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
tesseract-ocr is already the newest version (4.1.1-2.1build1).
0 upgraded, 0 newly installed, 0 to remove and 36 not upgraded.
tesseract 4.1.1
 leptonica-1.82.0
  libgif 5.1.9 : libjpeg 8d (libjpeg-turbo 2.1.1) : libpng 1.6.37 : libtiff 4.3.0 : zlib 1.2.11 : libwebp 1.2.2 : libopenjp2 2.4.0
 Found AVX2
 Found AVX
 Found FMA
 Found SSE
 Found libarchive 3.6.0 zlib/1.2.11 liblzma/5.2.5 bz2lib/1.0.8 liblz4/1.9.3 libzstd/1.4.8
```

```
#lets check whether we can read images and extract text using ocr
from PIL import Image
```

```
import pytesseract
import os

def ocr_image(image_path):
    img = Image.open(image_path)
    text = pytesseract.image_to_string(img)
    return text

#lets check one sample image
sample_image_path = os.path.join(IMG_FOLDER, os.listdir(IMG_FOLDER)[0])
print("Sample image path:", sample_image_path)

text = ocr_image(sample_image_path)
print("Extracted text:\n", text)
```

```
PO Number :01

Bill to:Hunter White

‘580 Johns Trafficway Apt. 099

Lake Elizabethfort, CA 70291 US

Tel:+(383)777-6857

Email:robbinsjoseph@example.net

Site:http://www.jackson-roberts.com/
‘ty © Description Unitprice | Amount
3.00 427900__| Nor campaign. 66.31 198598
5.00 492288 | Care require. 99.14 495.70
6.00 ‘377740 | Nor past. 7453 447.18

‘SUB_TOTAL : 1141.81 EUR
TAX:VAT (6.32%): 72.16 EUR
BALANCE DUE : 1183.48 EUR

Total in words: one thousand, one hundred and-
ighty-three point four eight

Note: All payments to be made in cash.
Contact us for queries on these quotations.

Address:233 Foster Gardens
Johnburgh, ID 41900 US

Bailey Group
Email:paul73@example.com
```

```
# Empty list to store extracted data
data = []

# Loop over all images
for i, fname in enumerate(os.listdir(IMG_FOLDER)):
    image_path = os.path.join(IMG_FOLDER, fname)
    text = ocr_image(image_path)
    data.append({'filename': fname, 'extracted_text': text})

    if i % 100 == 0:
        print(f"Processed {i} images...")

# Convert to DataFrame
```

```
import pandas as pd
df = pd.DataFrame(data)

# Save as csv
csv_path = '/content/drive/MyDrive/NLP driven Invoice Management System/Phase 1 Output/invoice_texts.csv'
os.makedirs(os.path.dirname(csv_path), exist_ok=True)
df.to_csv(csv_path, index=False)
print(f"CSV saved to {csv_path}")
```

```
Processed 0 images...
Processed 100 images...
Processed 200 images...
Processed 300 images...
Processed 400 images...
Processed 500 images...
Processed 600 images...
Processed 700 images...
Processed 800 images...
Processed 900 images...
Processed 1000 images...
Processed 1100 images...
Processed 1200 images...
Processed 1300 images...
Processed 1400 images...
Processed 1500 images...
Processed 1600 images...
Processed 1700 images...
Processed 1800 images...
Processed 1900 images...
CSV saved to /content/drive/MyDrive/NLP driven Invoice Management System/Phase 1 Output/invoice_texts.csv
```