```
pip install opencv-python pytesseract nltk indic-nlp-library transformers
```

```
Requirement already satisfied: opencv-python in /usr/local/lib/python3.11/dist-packages (4.11.0.86)
Collecting pytesseract
  Downloading pytesseract-0.3.13-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Collecting indic-nlp-library
  Downloading indic_nlp_library-0.92-py3-none-any.whl.metadata (5.7 kB)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.50.3)
Requirement already satisfied: numpy>=1.21.2 in /usr/local/lib/python3.11/dist-packages (from opencv-python) (2.0.2)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11/dist-packages (from pytesseract) (24.2)
Requirement already satisfied: Pillow>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from pytesseract) (11.1.0)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.1.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
Collecting sphinx-argparse (from indic-nlp-library)
  Downloading sphinx_argparse-0.5.2-py3-none-any.whl.metadata (3.7 kB)
Collecting sphinx-rtd-theme (from indic-nlp-library)
  Downloading sphinx_rtd_theme-3.0.2-py2.py3-none-any.whl.metadata (4.4 kB)
Collecting morfessor (from indic-nlp-library)
  Downloading Morfessor-2.0.6-py3-none-any.whl.metadata (628 bytes)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from indic-nlp-library) (2.2.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.30.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
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: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.26.0->transf
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.26
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->indic-nlp-library) (2.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->indic-nlp-library) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->indic-nlp-library) (2025.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4
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) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.31
Requirement already satisfied: sphinx>=5.1.0 in /usr/local/lib/python3.11/dist-packages (from sphinx-argparse->indic-nlp-library) (8.
Requirement already satisfied: docutils>=0.19 in /usr/local/lib/python3.11/dist-packages (from sphinx-argparse->indic-nlp-library) (0
Collecting sphinxcontrib-jquery<5,>=4 (from sphinx-rtd-theme->indic-nlp-library)
  Downloading sphinxcontrib_jquery-4.1-py2.py3-none-any.whl.metadata (2.6 kB)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->indic-nlp-li
Requirement already satisfied: sphinxcontrib-applehelp>=1.0.7 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-
Requirement already satisfied: sphinxcontrib-devhelp>=1.0.6 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-ar
Requirement already satisfied: sphinxcontrib-htmlhelp>=2.0.6 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-a
Requirement already satisfied: sphinxcontrib-jsmath>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-arg
Requirement already satisfied: sphinxcontrib-qthelp>=1.0.6 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-arg
Requirement already satisfied: sphinxcontrib-serializinghtml>=1.1.9 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->s
Requirement already satisfied: Jinja2>=3.1 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->indic-nlp
Requirement already satisfied: Pygments>=2.17 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->indic-
Requirement already satisfied: snowballstemmer>=2.2 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->
Requirement already satisfied: babel>=2.13 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->indic-nlp
Requirement already satisfied: alabaster>=0.7.14 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->ind
Requirement already satisfied: imagesize>=1.3 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argparse->indic-
Requirement already satisfied: roman-numerals-py>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from sphinx>=5.1.0->sphinx-argpar
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2>=3.1->sphinx>=5.1.0->sphinx-ar
Downloading pytesseract-0.3.13-py3-none-any.whl (14 kB)
Downloading indic_nlp_library-0.92-py3-none-any.whl (40 kB)
```

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

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  tesseract-ocr-hin
0 upgraded, 1 newly installed, 0 to remove and 30 not upgraded.
Need to get 913 kB of archives.
After this operation, 1,138 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-ocr-hin all 1:4.00~git30-7274cfa-1.1 [913 kB]
Fetched 913 kB in 2s (551 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package tesseract-ocr-hin.
```

```
(Reading database ... 126315 files and directories currently installed.)
Preparing to unpack .../tesseract-ocr-hin_1%3a4.00~git30-7274cfa-1.1_all.deb ...
Unpacking tesseract-ocr-hin (1:4.00~git30-7274cfa-1.1) ...
Setting up tesseract-ocr-hin (1:4.00~git30-7274cfa-1.1) ...
```

```
import os
os.environ['TESSDATA_PREFIX'] = '/usr/share/tesseract-ocr/4.00/tessdata'
```

```
import cv2
import pytesseract
from indicnlp.normalize.indic_normalize import IndicNormalizerFactory
from indicnlp.tokenize import indic_tokenize
from transformers import MarianMTModel, MarianTokenizer

# Optional: specify path to tesseract if not in PATH
# pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'

# Step 1: Preprocess Image
def preprocess_image(image_path):
    image = cv2.imread(image_path)
    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    denoised = cv2.fastNlMeansDenoising(gray, h=30)
    return denoised

# Step 2: OCR for Hindi
def extract_text_from_image(image):
    return pytesseract.image_to_string(image, lang='hin')

# Step 3: Normalize Hindi text
def normalize_text(text):
    factory = IndicNormalizerFactory()
    normalizer = factory.get_normalizer("hi")
    return normalizer.normalize(text)

# Step 4: Tokenization
def tokenize_text(text):
    return indic_tokenize.trivial_tokenize(text, lang='hi')

# Step 5: Translate Hindi to English
def translate_to_english(text):
    model_name = 'Helsinki-NLP/opus-mt-hi-en'
    tokenizer = MarianTokenizer.from_pretrained(model_name)
    model = MarianMTModel.from_pretrained(model_name)
    inputs = tokenizer(text, return_tensors="pt", padding=True, truncation=True)
    translated = model.generate(**inputs)
    return tokenizer.decode(translated[0], skip_special_tokens=True)

# Step 6: Display pipeline
def main(image_path):
    print("\n--- Hindi OCR and NLP Pipeline ---")
    image = preprocess_image(image_path)

    print("[1] Extracting Text...")
    raw_text = extract_text_from_image(image)
    print("Raw Text:\n", raw_text)

    print("\n[2] Normalizing...")
    norm_text = normalize_text(raw_text)
    print("Normalized:\n", norm_text)

    print("\n[3] Tokenizing...")
    tokens = tokenize_text(norm_text)
    print("Tokens:\n", tokens)

    print("\n[4] Translating to English...")
    translated = translate_to_english(norm_text)
    print("Translation:\n", translated)

# Change this to the path of your handwritten Hindi image
main("/content/images.jpg")
```

```
--- Hindi OCR and NLP Pipeline ---
[1] Extracting Text...
Raw Text:
 आजादी

कही है केर अगर,
जे उड़ने में कर मदद त।
सतत है काली आगर






जला कर रौशन कर तू
(विकार में उल्क- कर



बी नए कर सा रूटिवादी
सुलझा पर के आात तू.
रत, आदमी वा हो कोई बच्चा




[2] Normalizing...
Normalized:
 आजादी

कही है केर अगर,
जे उड़ने में कर मदद त।
सतत है काली आगर






जला कर रौशन कर तू
(विकार में उल्क- कर



बी नए कर सा रूटिवादी
सुलझा पर के आात तू.
रत, आदमी वा हो कोई बच्चा




[3] Tokenizing...
Tokens:
 ['आजादी\n\nकही', 'है', 'केर', 'अगर', ',', '\nजे', 'उड़ने', 'में', 'कर', 'मदद', 'त', '।', '\nसतत', 'है', 'काली', 'आगर\n\n', '\n\n', '\n',

[4] Translating to English...
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning: 
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secre
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%                                    42.0/42.0 [00:00<00:00, 1.55kB/s]
source.spm: 100%                                    1.06M/1.06M [00:00<00:00, 1.99MB/s]
target.spm: 100%                                    813k/813k [00:00<00:00, 1.51MB/s]
vocab.json: 100%                                    2.06M/2.06M [00:00<00:00, 2.89MB/s]
config.json: 100%                                    1.38k/1.38k [00:00<00:00, 101kB/s]
/usr/local/lib/python3.11/dist-packages/transformers/models/marian/tokenization_marian.py:175: UserWarning: Recommended: pip install sac
  warnings.warn("Recommended: pip install sacremoses.")
pytorch_model.bin: 100%                                    304M/304M [00:01<00:00, 240MB/s]
model.safetensors: 100%                                    304M/304M [00:01<00:00, 181MB/s]
generation_config.json: 100%                                    293/293 [00:00<00:00, 18.1kB/s]
Translation:
 Charer said that if he could help in his fly, he's constantly burning black fireer and lit up you (option-B) over the new root solver,
```