```
!pip install pytesseract
!pip install jiwer
!pip install pdf2image
!apt install poppler-utils

!apt install tesseract-ocr # Install Tesseract OCR
!apt install libtesseract-dev
```

```
Requirement already satisfied: pytesseract in /usr/local/lib/python3.11/dis
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11
Requirement already satisfied: Pillow>=8.0.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: jiwer in /usr/local/lib/python3.11/dist-pack
Requirement already satisfied: click>=8.1.8 in /usr/local/lib/python3.11/di
Requirement already satisfied: rapidfuzz>=3.9.7 in /usr/local/lib/python3.1
Requirement already satisfied: pdf2image in /usr/local/lib/python3.11/dist-
Requirement already satisfied: pillow in /usr/local/lib/python3.11/dist-pac
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
poppler-utils is already the newest version (22.02.0-2ubuntu0.6).
0 upgraded, 0 newly installed, 0 to remove and 29 not upgraded.
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  tesseract-ocr-eng tesseract-ocr-osd
The following NEW packages will be installed:
  tesseract-ocr tesseract-ocr-eng tesseract-ocr-osd
0 upgraded, 3 newly installed, 0 to remove and 29 not upgraded.
Need to get 4,816 kB of archives.
After this operation, 15.6 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-ocr-e
Get:2 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-ocr-o
Get:3 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tesseract-ocr a
Fetched 4,816 kB in 1s (5,110 kB/s)
Selecting previously unselected package tesseract-ocr-eng.
(Reading database ... 125074 files and directories currently installed.)
Preparing to unpack .../tesseract-ocr-eng_1%3a4.00~git30-7274cfa-1.1_all.de
Unpacking tesseract-ocr-eng (1:4.00~git30-7274cfa-1.1) ...
Selecting previously unselected package tesseract-ocr-osd.
Preparing to unpack .../tesseract-ocr-osd_1%3a4.00~git30-7274cfa-1.1_all.de
Unpacking tesseract-ocr-osd (1:4.00~git30-7274cfa-1.1) ...
Selecting previously unselected package tesseract-ocr.
Preparing to unpack .../tesseract-ocr_4.1.1-2.1build1_amd64.deb ...
Unpacking tesseract-ocr (4.1.1-2.1build1) ...
Setting up tesseract-ocr-eng (1:4.00~git30-7274cfa-1.1) ...
Setting up tesseract-ocr-osd (1:4.00~git30-7274cfa-1.1) ...
Setting up tesseract-ocr (4.1.1-2.1build1) ...
Processing triggers for man-db (2.10.2-1) ...
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libarchive-dev libleptonica-dev
The following NEW packages will be installed:
  libarchive-dev libleptonica-dev libtesseract-dev
0 upgraded, 3 newly installed, 0 to remove and 29 not upgraded.
```

```
Need to get 3,743 kB of archives.
After this operation, 16.0 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libarchive-
Get:2 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libleptonica-de
Get:3 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libtesseract-de
Fetched 3,743 kB in 1s (4,013 kB/s)
Selecting previously unselected package libarchive-dev:amd64.
(Reading database ... 125121 files and directories currently installed.)
Preparing to unpack .../libarchive-dev_3.6.0-1ubuntu1.3_amd64.deb ...
```

```
import torch
import torch.nn as nn
import torch.optim as optim
import torchvision.transforms as transforms
from torch.utils.data import DataLoader, Dataset
import numpy as np
import cv2
import pytesseract
from transformers import TrOCRProcessor, VisionEncoderDecoderModel
import os # Import the os module
from pdf2image import convert_from_path

# Dataset Class
class OCRDataset(Dataset):
    def __init__(self, image_paths, labels, transform=None):
        self.image_paths = image_paths
        self.labels = labels
        self.transform = transform

    def __len__(self):
        return len(self.image_paths)

    def __getitem__(self, idx):
        image = cv2.imread(self.image_paths[idx], cv2.IMREAD_GRAYSCALE)
        image = cv2.resize(image, (128, 32))
        if self.transform:
            image = self.transform(image)
        label = self.labels[idx]
        return image, label

# Model: CRNN
class CRNN(nn.Module):
    def __init__(self, num_classes):
        super(CRNN, self).__init__()
        self.cnn = nn.Sequential(
            nn.Conv2d(1, 64, kernel_size=3, stride=1, padding=1),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2, stride=2),
            nn.Conv2d(64, 128, kernel_size=3, stride=1, padding=1),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2, stride=2),
        )
        self.rnn = nn.LSTM(128, 256, bidirectional=True, batch_first=True)
        self.fc = nn.Linear(512, num_classes)
```

```
    def forward(self, x):
        x = self.cnn(x)
        x = x.squeeze(2).permute(0, 2, 1)
        x, _ = self.rnn(x)
        x = self.fc(x)
        return x

# Transformer OCR
class TransformerOCR:
    def __init__(self):
        self.processor = TrOCRProcessor.from_pretrained("microsoft/trocr-base-h
        self.model = VisionEncoderDecoderModel.from_pretrained("microsoft/trocr

    def recognize_text(self, image):
        inputs = self.processor(images=image, return_tensors="pt").pixel_values
        generated_ids = self.model.generate(inputs)
        text = self.processor.batch_decode(generated_ids, skip_special_tokens=T
        return text

# Evaluation Metrics
from jiwer import wer, cer

def evaluate(preds, labels):
    wer_score = wer(labels, preds)
    cer_score = cer(labels, preds)
    return wer_score, cer_score

# Training (Placeholder)
def train():
    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
    model = CRNN(num_classes=100).to(device)
    optimizer = optim.Adam(model.parameters(), lr=0.001)
    criterion = nn.CTCLoss()
    for epoch in range(10):
        # Training loop placeholder
        pass
    return model

# OCR Prediction with Tesseract
def tesseract_ocr(image_path):
    # Check if the image file exists
    if not os.path.exists(image_path):
        raise FileNotFoundError(f"Image file not found: {image_path}") # Raise

    image = cv2.imread(image_path)

    # Check if the image was loaded successfully
    if image is None:
        raise ValueError(f"Could not read image file: {image_path}") # Raise a

    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    text = pytesseract.image_to_string(gray)
    return text
```

```
# Function to convert PDF to images and perform OCR
def process_pdf(pdf_path):
    # Convert PDF to images
    images = convert_from_path(pdf_path)

    all_text = ""  # Store all extracted text

    for i, image in enumerate(images):
        # Save the image temporarily
        image_path = f"temp_page_{i}.jpg"
        image.save(image_path, "JPEG")

        # Perform OCR on the image
        try:
            tesseract_text = tesseract_ocr(image_path)
            trocr = TransformerOCR()
            transformer_text = trocr.recognize_text(cv2.imread(image_path))

            print(f"Page {i + 1}:")
            print("Tesseract OCR Output:", tesseract_text)
            print("Transformer OCR Output:", transformer_text)
            all_text += tesseract_text + transformer_text  # Append to all_text

        except (FileNotFoundError, ValueError) as e:
            print(f"Error processing page {i + 1}: {e}")

        finally:
            # Remove the temporary image file
            os.remove(image_path)

    return all_text  # Return all extracted text


# Example Usage
if __name__ == "__main__":
    pdf_path = "/content/Buendia - Instruccion.pdf"  # Replace with your PDF fi
    extracted_text = process_pdf(pdf_path)
    print("\nAll Extracted Text:\n", extracted_text)
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access
  warnings.warn(
preprocessor_config.json: 100%  224/224 [00:00<00:00, 17.4kB/
s]
Using a slow image processor as `use_fast` is unset and a slow processor wa
tokenizer_config.json: 100%  1.12k/1.12k [00:00<00:00, 64.9kB/
s]
```

```
vocab.json: 100%  899k/899k [00:00<00:00, 9.31MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 19.3MB/s]
special_tokens_map.json: 100%  772/772 [00:00<00:00, 53.4kB/s]
config.json: 100%  4.17k/4.17k [00:00<00:00, 298kB/s]
model.safetensors: 100%  1.33G/1.33G [00:12<00:00, 112MB/s]
Config of the encoder: <class 'transformers.models.vit.modeling_vit.ViTMode
  "attention_probs_dropout_prob": 0.0,
  "encoder_stride": 16,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.0,
  "hidden_size": 768,
  "image_size": 384,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "model_type": "vit",
  "num_attention_heads": 12,
  "num_channels": 3,
  "num_hidden_layers": 12,
  "patch_size": 16,
  "qkv_bias": false,
  "transformers_version": "4.48.3"
}

Config of the decoder: <class 'transformers.models.trocr.modeling_trocr.TrO
  "activation_dropout": 0.0,
  "activation_function": "gelu",
  "add_cross_attention": true,
  "attention_dropout": 0.0,
  "bos_token_id": 0,
  "classifier_dropout": 0.0,
  "cross_attention_hidden_size": 768,
  "d_model": 1024,
  "decoder_attention_heads": 16,
  "decoder_ffn_dim": 4096,
  "decoder_layerdrop": 0.0,
  "decoder_layers": 12,
  "decoder_start_token_id": 2,
  "dropout": 0.1,
  "eos_token_id": 2,
  "init_std": 0.02,
  "is_decoder": true,
  "layernorm_embedding": true,
  "max_position_embeddings": 512,
  "model_type": "trocr",
  "pad_token_id": 1,
```

```
  "pad_token_id": 1,
  "scale_embedding": false,
  "transformers_version": "4.48.3",
  "use_cache": false,
  "use_learned_position_embeddings": true,
  "vocab_size": 50265
}
```


```
Some weights of VisionEncoderDecoderModel were not initialized from the mod
You should probably TRAIN this model on a down-stream task to be able to us
```


```
generation_config.json: 100%  190/190 [00:00<00:00, 16.9kB/s]
```


```
Page 1:
Tesseract OCR Output: Lowe

KN
INFINITAM ‘AMABLE
' NINO TESUS.

| SNCS Vos , Dulcifsimo Nifio
GANS | Jesus , queno folo OS Ex1/ai.33:
(GemYe)|| dignafteis de llamaros 38.

| Do@or de los Nifios, ©*4“* *
fino tambien de afsif- *”

tir como Nifio entre los Do&ores,

fe confagra humilde efta pequefa
Iaftruccion de los Nifios. Es afsi,

que ella tambien fe dirige a laju-

ventud ; pero aefta, como recuer-

do delo que aprendio, alos Ni

ios , como precifa explicacion de

lo que deben eftudiar. Por efte fo-
lagitulo.es muy vueftra 5 sy por

fer para Nifios, que confiais a la
educacion de vueftra Compafia,

lo es mucho mas. En Vos, ( Divi-

no Exemplar de todas las virta-

des ) tienen abreviado el mas {e-

q 2 gure
```

```
Transformer OCR Output: 0 0
Config of the encoder: <class 'transformers.models.vit.modeling_vit.ViTMode
  "attention_probs_dropout_prob": 0.0,
  "encoder_stride": 16,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.0,
  "hidden_size": 768,
  "image_size": 384,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "model_type": "vit",
  "num_attention_heads": 12,
  "num_channels": 3,
  "num_hidden_layers": 12,
  "patch_size": 16,
  "qkv_bias": false,
  "transformers_version": "4.48.3"
}

Config of the decoder: <class 'transformers.models.trocr.modeling_trocr.TrO
  "activation_dropout": 0.0,
  "activation_function": "gelu",
  "add_cross_attention": true,
  "attention_dropout": 0.0,
  "bos_token_id": 0,
  "classifier_dropout": 0.0,
  "cross_attention_hidden_size": 768,
  "d_model": 1024,
  "decoder_attention_heads": 16,
  "decoder_ffn_dim": 4096,
  "decoder_layerdrop": 0.0,
  "decoder_layers": 12,
  "decoder_start_token_id": 2,
  "dropout": 0.1,
  "eos_token_id": 2,
  "init_std": 0.02,
  "is_decoder": true,
  "layernorm_embedding": true,
  "max_position_embeddings": 512,
  "model_type": "trocr",
  "pad_token_id": 1,
  "scale_embedding": false,
  "transformers_version": "4.48.3",
  "use_cache": false,
  "use_learned_position_embeddings": true,
  "vocab_size": 50265
}

Some weights of VisionEncoderDecoderModel were not initialized from the mod
You should probably TRAIN this model on a down-stream task to be able to us
Page 2:
Tesseract OCR Output: i nstruccion de christiana y politica cortesania con

Luc. ibid.

P fal.114.6.
```

em 118.130,
», &i8.8.
Mattb.19.
14.
Marci, 10,
14.

Matt. 18.
2. PC.

Dios y con los hombres

guro diffeno ag edad : Ia Reli-
gion paracon Dios en la devora
afsiftécia 4 los Templos;la piedad
con los Padres en la obediencia
mas rendida; gy: modettia, y de~
feo de faber con los mayores,
guftando mas de olr, y pregun-
tar,que de definir,y refolver.Bien
que efto en vuettra infinita Sabi-.
duria fue foberana dignacion , y
en la natural ignorancia de los
Nifios es indifpenfable necefsi-
dad.

Ni tienen folamente en Vos
el diffefio , la luz ,y el exemplo,
fino tambien el amor, y protec-

cion. Vos, como fingularsMaef-

tro dé los Nifios , les dais enten-
dimiento , y comunicais la fabi-

. duria. Vos les prometeis el Reyno.
de los Cielos , y os indignais con

quien les aparta de Vos, y les

proponeis por norma del can-

dor , inocencia , y chriftiana hu-
3

‘mildad. Vueftro amor parece que

no pudo explicarfe mas tierno, y

liberal con los Nifos , pues no

contento de echarles vueftras dis
Me, -. divi-

divinas bendiciones., les unifteis

```
a vueſtro ſagrado pecho con ſ{ua-
vifsimos abrazos, Dichofa edad,
que os merecio tan regalados ca-
rifos ! | —
Y pues en la celeftial Jeru-
falen no ha mudado de condicion
vueftra Benignidad , profeguid,
© Nifio tierno, y Dios Eterno,
profeguid en bendecirles, y favo-
recerles. Sean tan fervorofamen-
te devotos de vueftra Admirable
Mapre , que fe porten como {us
hijos , y hermanos de leche con
Vos. Seran fabios, fi fueren caf-
tos ; que no entra vueftra Sabi-

- duria ,donde no ay mucha pure-

za de conciencia. Crezcan en
vueftro fanto temor, y amor, co-
como en los afios,y mucho mas.
Adelantenfe en la virtud , como
en Jas letras, y mucho mas ; haf-
ta que Ileguen , por vuettra imt-
tacion , a fer varones perfe&os,
y confumados , agradables 4
vueftros ojos, y provechofos 4
la Republica , que libra cafi to-
da fu, felizidad en la acercada
hl A crian-

Marci. to.
16°

Cant. 8.%«

Sap. 3. 4s

Ad Epheft
4. 13.

vi
About this book [Gi=si

1 Q @ AR

Transformer OCR Output: 25 9
Config of the encoder: <class 'transformers.models.vit.modeling_vit.ViTMode
  "attention_probs_dropout_prob": 0.0,
  "encoder_stride": 16,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.0,
  "hidden_size": 768,
  "image_size": 384,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "model_type": "vit",
```

```
  "num_attention_heads": 12,
  "num_channels": 3,
  "num_hidden_layers": 12,
  "patch_size": 16,
  "qkv_bias": false,
  "transformers_version": "4.48.3"
}

Config of the decoder: <class 'transformers.models.trocr.modeling_trocr.TrO
  "activation_dropout": 0.0,
  "activation_function": "gelu",
  "add_cross_attention": true,
  "attention_dropout": 0.0,
  "bos_token_id": 0,
  "classifier_dropout": 0.0,
  "cross_attention_hidden_size": 768,
  "d_model": 1024,
  "decoder_attention_heads": 16,
  "decoder_ffn_dim": 4096,
  "decoder_layerdrop": 0.0,
  "decoder_layers": 12,
  "decoder_start_token_id": 2,
  "dropout": 0.1,
  "eos_token_id": 2,
  "init_std": 0.02,
  "is_decoder": true,
  "layernorm_embedding": true,
  "max_position_embeddings": 512,
  "model_type": "trocr",
  "pad_token_id": 1,
  "scale_embedding": false,
  "transformers_version": "4.48.3",
  "use_cache": false,
  "use_learned_position_embeddings": true,
  "vocab_size": 50265
}

Some weights of VisionEncoderDecoderModel were not initialized from the mod
You should probably TRAIN this model on a down-stream task to be able to us
Page 3:
Tesseract OCR Output: erianza ‘dé la Nifiéz: ‘Afsi fea;
Divinifsimo Nifo , por vueftra
pracia, afsi fea, a yueftra mas
yor gloria. Amen.

CENSUS

KN
CENSURA DEL. R. P, ANTONIO CO,
* dornia de la Compania de Fefus , Maef-
tro que fué de Theologta , Examinador,
Synodal de los ObiPados de Gerona , Ura
gel,y Barcelona, Oc,

E orden del Iluftre Sefor Don Frans
cifco de Baftéro y de Vilana, Dr. en
ambos Drechos , Canonigo , y Saeriftan
Dignidad de la Santa Iglefia de Gerona, y;
```

```
Vicario General por el Illuſtriſsimo Señor,
D. Balthafar de Baftéro y Lledo , Obifpal
de Gerona,del Confejo de fu Mageftad,&c4
He vifto un Librico, cuyo tituloes: In/4
truccion de Chriftiana ,y Politica Corte/aa
nia, &c. Su Author D. Faufto Aguftin dé
Buendia , Colegial que fue en el Imperial
de Cordellas, &c. Y brevemente digo,
no folo que nada contiene contra lake, yj
buenas coftumbres , fino que muy atento,
el Author con entrambas , defcribe ,y en-
fefiatanculta, y difcreta la Virtud, cos
mo fanta la Policia, y Urbanidad. Los
Sefioritos , que fe criaren con eftos do-
cumentos , mereceran , quando hombres,
aver nacido Sefiores. Porque no folo fa-
bran fer Caballeros , fino samibien ® lo
wadas

= iT QO@OR

Transformer OCR Output: 0 1
Config of the encoder: <class 'transformers.models.vit.modeling_vit.ViTMode
  "attention_probs_dropout_prob": 0.0,
  "encoder_stride": 16,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.0,
  "hidden_size": 768,
  "image_size": 384,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "model_type": "vit",
  "num_attention_heads": 12,
  "num_channels": 3,
  "num_hidden_layers": 12,
  "patch_size": 16,
  "qkv_bias": false,
  "transformers_version": "4.48.3"
}

Config of the decoder: <class 'transformers.models.trocr.modeling_trocr.TrO
  "activation_dropout": 0.0,
  "activation_function": "gelu",
  "add_cross_attention": true,
  "attention_dropout": 0.0,
  "bos_token_id": 0,
  "classifier_dropout": 0.0,
  "cross_attention_hidden_size": 768,
  "d_model": 1024,
  "decoder_attention_heads": 16,
  "decoder_ffn_dim": 4096,
  "decoder_layerdrop": 0.0,
  "decoder_layers": 12,
  "decoder_start_token_id": 2,
  "dropout": 0.1,
  "eos_token_id": 2,
  "init_std": 0.02,
```

```
  "is_decoder": true,
  "layernorm_embedding": true,
  "max_position_embeddings": 512,
  "model_type": "trocr",
  "pad_token_id": 1,
  "scale_embedding": false,
  "transformers_version": "4.48.3",
  "use_cache": false,
  "use_learned_position_embeddings": true,
  "vocab_size": 50265
}
```


```
Some weights of VisionEncoderDecoderModel were not initialized from the mod
You should probably TRAIN this model on a down-stream task to be able to us
```

Page 4:
Tesseract OCR Output: Caballero muy Santos. Por todo lo qual,
a mas de la licencia , que folicita ,’ meu
rece el:Cortefano Zelo del Author muchas
gracias de quantos fe intereffan en‘ tan
‘primorofa , como neceffaria educacion
de la primera edad de los''Nobles : de cu-
yo acierto fe deriva la primera utilidad,
y decoro 4 la Republica. Afsi lo ‘fien-
to, falvo, &c. En efte Colegio de San
Martin de Ja Compafia de Jefus de Gero-
na, a 15. de Julio de 1740.

Antonio Codornin, de Ia *
Compania de Fefus.

Die 15. Fulii, 1740.
« Imprimatur.

‘De Baftéro Vic, Gen. & Offic.

CENSUs

OK N |
CENSURA DEL R, P;: MARIANO ALBE-
rich de laCompaiiia de Fefus, Catbedra-
~tico antes de Theologia en el Colegio de
Barcelona, (oy Prefecto de fus Eftudios)
Retor que fue del Colegio de Cordellas , y
de Gerona , Calificador del Santo Oficio,

Examinador Synodal de muchos Obifpa-
dos, Orc. A a , :

M, P. S.

E orden de V. A. he vifto el Librito,

que conel titulo de Politica,y Chri/-

tiana Cortefania, Oc. defea facar a luz el
Sefor D. Faufto Aguftin de Buendia , pa-
ra darle a todos con el concierto,que pref-

```
cribe para las operaciones , que cada uno

debe pra@icar , defde el ofrecimiento de
las obras por la mafiana , hafta el examen
de la conciencia , ultima diligencia , que
debe cerrar nueftras taréas , antes de con-
ceder al cuerpo el feb defcanfo, para
tomarnos diariamente cuenta de todas las
```

Start coding or generate with AI.