```
! pip install transformers
! pip install datasets
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.51.3)
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) (0.31.2)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
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->transform
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0-
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.2)
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) (2025.4.26)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (2.14.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec>=2021.11.1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.11.1->datasets) (20
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.14.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.31.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.6.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.20.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0.0,>=0.14.0->datasets) (3.18
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0.0,>=0.14.
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->datasets) (3.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->datasets) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->datasets) (2025.4.2
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17
```

```
pip install -U datasets
```

```
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (2.14.4)
Collecting datasets
  Downloading datasets-3.6.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from datasets) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-packages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Collecting fsspec<=2025.3.0,>=2023.1.0 (from fsspec[http]<=2025.3.0,>=2023.1.0->datasets)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: huggingface-hub>=0.24.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.31.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (6.0.2)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<=2025.3.0,>=20
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->da
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets) (2025.4
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fs
```

```
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[ht
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http]
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[h
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[ht
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[h
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->datasets) (1.
Downloading datasets-3.6.0-py3-none-any.whl (491 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 491.5/491.5 kB 26.8 MB/s eta 0:00:00
Downloading fsspec-2025.3.0-py3-none-any.whl (193 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 193.6/193.6 kB 14.9 MB/s eta 0:00:00
Installing collected packages: fsspec, datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
  Attempting uninstall: datasets
    Found existing installation: datasets 2.14.4
    Uninstalling datasets-2.14.4:
      Successfully uninstalled datasets-2.14.4
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the sourc
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2025.3.0 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and platform_machine == "x86_64", but you have nvi
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you ha
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system == "Linux" and platform_machine == "x86_64", but you have nvid
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system == "Linux" and platform_machine == "x86_64", but you have nvid
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system == "Linux" and platform_machine == "x86_64", but you have n
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system == "Linux" and platform_machine == "x86_64", but you have n
```

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, DataCollatorForSeq2Seq, Seq2SeqTrainingArguments, Seq2SeqTrainer
import torch
import numpy as np
from torch.utils.data import Dataset
```

```
from sklearn.model_selection import train_test_split
import pandas as pd

df = pd.read_csv("drive/MyDrive/ai/test2.csv")
train, temp = train_test_split(df, test_size=0.3, random_state=42)
val, test = train_test_split(temp, test_size=0.5, random_state=42)
train.to_csv("train.csv", index=False)
val.to_csv("val.csv", index=False)
test.to_csv("test.csv", index=False)
train = "train.csv"
```

```
train = pd.read_csv("train.csv")
```

```
from datasets import load_dataset
dataset = {
    "train":      "train.csv",
    "val":        "val.csv",
    "test":       "test.csv"
}
raw_datasets = load_dataset("csv", data_files=dataset)
# ตรวจสอบตัวอย่างแรก
print(raw_datasets["train"][0])
raw_datasets
```

```
Generating train split:        3572/0 [00:00<00:00, 94688.88 examples/s]

Generating val split:        766/0 [00:00<00:00, 36609.35 examples/s]

Generating test split:        766/0 [00:00<00:00, 39572.32 examples/s]
{'sentence': 'ฉันintegrate real user monitoringด้วยRUM tool', 'eng_sentence': 'ฉันintegrate real user monitoringfh;pRUM tool'}
DatasetDict({
    train: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 3572
    })
    val: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 766
    })
    test: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 766
    })
})
```

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

checkpoint = "google/mt5-small"
tokenizer  = AutoTokenizer.from_pretrained(checkpoint)
model      = AutoModelForSeq2SeqLM.from_pretrained(checkpoint)
```

```
/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(
You are using the default legacy behaviour of the <class 'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is expected, and si
/usr/local/lib/python3.11/dist-packages/transformers/convert_slow_tokenizer.py:564: UserWarning: The sentencepiece tokenizer that you ar
  warnings.warn(
```

```
from transformers import DataCollatorForSeq2Seq
prefix = "fix typo to natural Thai sentence: "
def preprocess_function(examples):
    inputs = [prefix + doc for doc in examples["eng_sentence"]]
    model_inputs = tokenizer(inputs,text_target=examples["sentence"], max_length=64, truncation=True)
    return model_inputs
```

Double-click (or enter) to edit

```
preprocess_function(raw_datasets['train'][:2])
```

```
{'input_ids': [[15480, 259, 139677, 288, 4926, 18448, 259, 98923, 267, 120194, 185061, 2784, 12394, 52342, 11926, 296, 325, 72615,
16080, 1], [15480, 259, 139677, 288, 4926, 18448, 259, 98923, 267, 259, 31850, 17752, 478, 75546, 260, 2622, 1582, 765, 55213, 325,
277, 260, 268, 188828, 1]], 'attention_mask': [[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], [1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]], 'labels': [[120194, 185061, 2784, 12394, 52342, 8084, 72615, 16080, 1], [259,
31850, 17752, 478, 75546, 46865, 40487, 2361, 188828, 1]]}
```

```
tokenized_datasets = raw_datasets.map(
    preprocess_function,
    batched=True,
    remove_columns=["sentence", "eng_sentence"]
)
tokenized_datasets
```

```
Map: 100%                                   3572/3572 [00:00<00:00, 12294.40 examples/s]

Map: 100%                                   766/766 [00:00<00:00, 9061.68 examples/s]

Map: 100%                                   766/766 [00:00<00:00, 8918.97 examples/s]

DatasetDict({
    train: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 3572
    })
    val: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 766
    })
    test: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 766
    })
})
```

```
!pip install --upgrade transformers
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.51.3)
Collecting transformers
  Downloading transformers-4.52.3-py3-none-any.whl.metadata (40 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.2/40.2 kB 3.5 MB/s eta 0:00:00
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) (0.31.2)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
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->transform
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0-
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.2)
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) (2025.4.26)
Downloading transformers-4.52.3-py3-none-any.whl (10.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.5/10.5 MB 129.3 MB/s eta 0:00:00
Installing collected packages: transformers
  Attempting uninstall: transformers
    Found existing installation: transformers 4.51.3
    Uninstalling transformers-4.51.3:
      Successfully uninstalled transformers-4.51.3
Successfully installed transformers-4.52.3
```

```
from transformers import TrainingArguments
from transformers import Seq2SeqTrainingArguments, Seq2SeqTrainer, DataCollatorForSeq2Seq
data_collator = DataCollatorForSeq2Seq(tokenizer, model=model)
training_args = Seq2SeqTrainingArguments(
    output_dir="./results",
    eval_strategy="epoch",
    logging_strategy="epoch",
    save_strategy="epoch",
    logging_steps=50,
    report_to="tensorboard",
    learning_rate= 5e-5,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=64,
    num_train_epochs=5,
    weight_decay=0.01,
    predict_with_generate=True,
)
```

```
from transformers import  Seq2SeqTrainer

trainer = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["val"],
    data_collator=data_collator,
```

```
    tokenizer=tokenizer,
)
```

```
<ipython-input-25-da6551846644>:3: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Seq2SeqTrainer.__i
  trainer = Seq2SeqTrainer(
```

[560/560 04:08, Epoch 5/5]

Epoch	Training Loss	Validation Loss
1	0.058300	0.590162
2	0.052000	0.597822
3	0.044400	0.610854
4	0.040100	0.620292
5	0.033500	0.623904

```
TrainOutput(global_step=560, training_loss=0.045652806758880615, metrics={'train_runtime': 248.9938, 'train_samples_per_second':
```

```
train_output = trainer.train()
```

[560/560 04:35, Epoch 5/5]

Epoch	Training Loss	Validation Loss
1	0.139100	0.514416
2	0.128300	0.500830
3	0.132000	0.489360
4	0.123300	0.498152
5	0.120700	0.496500

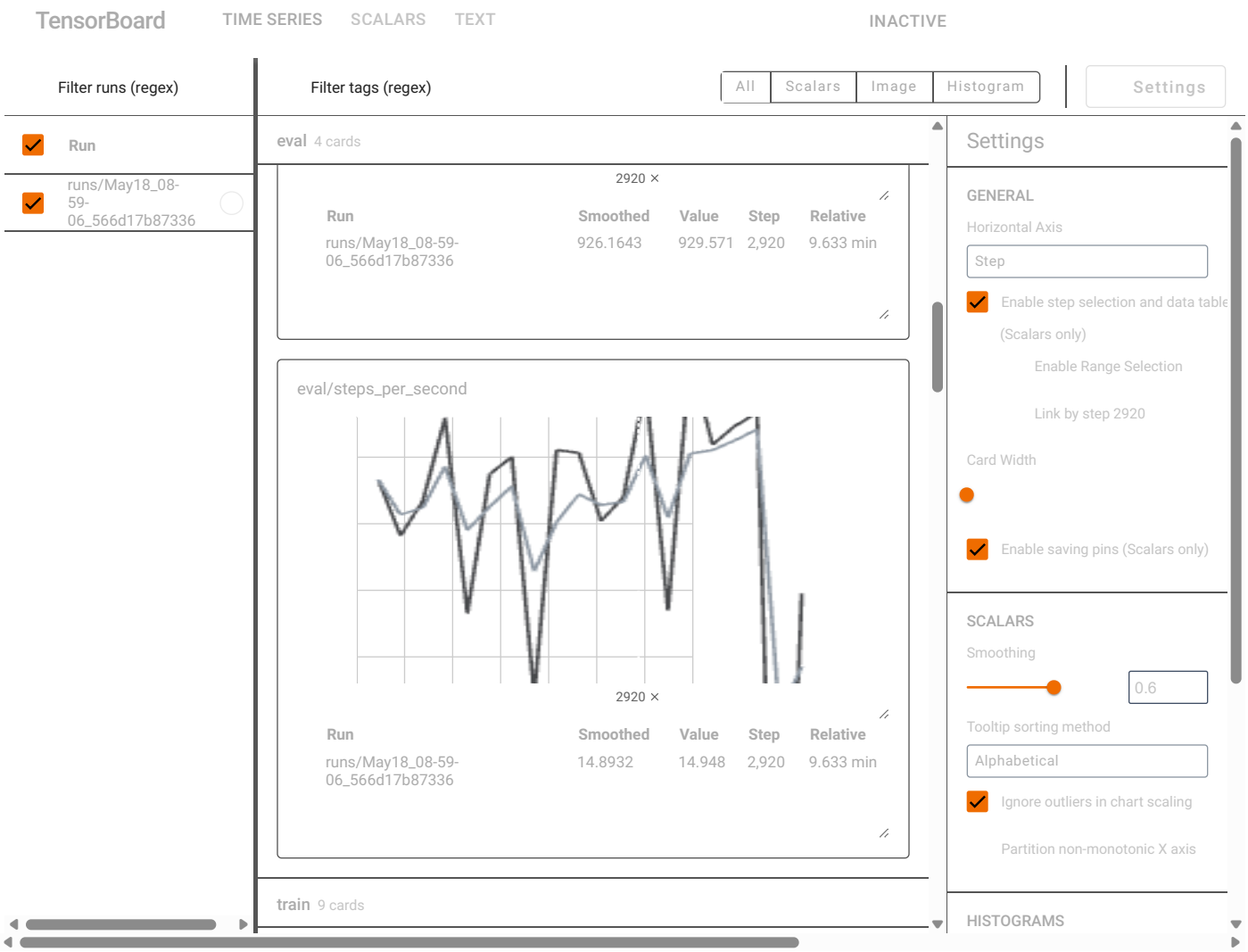
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-33-8b3fa3b1af09> in <cell line: 0>()
      2
      3 # —— ดึงค่า loss ออกมา ——
----> 4 logs = trainers.state.log_history
      5
      6 # Training loss จะถูกบันทึกตอนท้ายแต่ละ epoch

NameError: name 'trainers' is not defined
```

```
trainer.save_model("drive/MyDrive/ai/model3")
```

```
%load_ext tensorboard
%tensorboard --logdir ./results
```

```
The tensorboard extension is already loaded. To reload it, use:
  %reload_ext tensorboard
Reusing TensorBoard on port 6006 (pid 4818), started 0:11:47 ago. (Use '!kill 4818' to kill it.)
```




```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
import torch
import pandas as pd


def correct_typo(text: str) -> str:
    input_str = "fix typo to correct Thai sentence: " + text
    inputs = tokenizer(input_str, return_tensors="pt", truncation=True, padding="longest")
    inputs = {k: v.to(model.device) for k, v in inputs.items()}
    outputs = model.generate(**inputs, max_new_tokens=64)
    tttx = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return tttx


print(correct_typo("พรุ่งนี้มี meeting dy[]^d8hk8ole8yP vpjk]n,g9iup, presentation fh;pot8iy["))
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))
print(correct_typo("ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Water Sunscreen SPF50! PA!!!! xdxhv'zb;]he]7d57' 2 9jv!"))
print(correct_typo("บัญชีเริ่มต้นที่ใช้สำหรับเรียกใช้บริการ Windows Ffp,ulbmTbV-yho9je [yP=uouh0t.=h-hv,^]xit0e9y;8v,rb;g9viVgrnjvpnopyo9y;9ozjkog8inv-j
print(correct_typo("วันนี้ฉันต้องattend corporate meetinggrnjvpresent business plani;,57'vTb[kpfinancial forecastc]tmarketing strategy.shdy[stak
print(correct_typo("เล่นminecraftdyog5vtgrnvo"))
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))
print(correct_typo("train modelgliH0py'vt-vcodesojvp"))
```

```
Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncat
จาก: พรุ่งนี้มี meeting dy[]^d8hk8ole8yP vpjk]n,g9iup, presentation fh;pot8iy[ ได้เป็น: พรุ่งนี้มี meeting กับลูกค้ารายสัปดาห์ เพื่อเตรียม presentation ด้วย
จาก: เล่นrobloxdyog5vtgrnvo ได้เป็น: เล่นrobloxกันจนเพื่อน
ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Water Sunscreen SPF50+ PA++++ ปกป้องผิวล้ำลึกถึง 2 ต่อ!
บัญชีเริ่มต้นที่ใช้สำหรับเรียกใช้บริการ Windows ดาต้าของ ซอฟต์แวร์เพื่อใช้งานซอฟต์แวร์อัตโนมัติ
วันนี้ฉันต้องattend corporate meetingเพื่อpresent business planสำรวจfinancial forecastและmarketing strategyให้กับstakeholdersตั้งแต่การประชุม
เล่นminecraftกันจนเพื่อน
เล่นrobloxกันจนเพื่อน
```

train modelเสร็จกับรหัสcodeหน่อย

Double-click (or enter) to edit

```
!pip install python-Levenshtein nltk
```

```
Collecting python-Levenshtein
  Downloading python_levenshtein-0.27.1-py3-none-any.whl.metadata (3.7 kB)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Collecting Levenshtein==0.27.1 (from python-Levenshtein)
  Downloading levenshtein-0.27.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.6 kB)
Collecting rapidfuzz<4.0.0,>=3.9.0 (from Levenshtein==0.27.1->python-Levenshtein)
  Downloading rapidfuzz-3.13.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.2.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.5.0)
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)
Downloading python_levenshtein-0.27.1-py3-none-any.whl (9.4 kB)
Downloading levenshtein-0.27.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (161 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 161.7/161.7 kB 14.9 MB/s eta 0:00:00
Downloading rapidfuzz-3.13.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.1/3.1 MB 84.0 MB/s eta 0:00:00
Installing collected packages: rapidfuzz, Levenshtein, python-Levenshtein
Successfully installed Levenshtein-0.27.1 python-Levenshtein-0.27.1 rapidfuzz-3.13.0
```

```
import pandas as pd
from Levenshtein import distance as levenshtein_distance
from nltk.translate.bleu_score import corpus_bleu
test_df = pd.read_csv("test.csv")
refs = test_df["sentence"].tolist()
preds = [correct_typo(t) for t in test_df["eng_sentence"]]

#Character-Level Accuracy
total_chars = sum(len(r) for r in refs)
correct_chars = sum(
    sum(1 for a,b in zip(r,p) if a==b)
    for r,p in zip(refs,preds)
)
char_acc = correct_chars / total_chars

#Average Levenshtein Distance
avg_edit = sum(
    levenshtein_distance(r, p)
    for r,p in zip(refs,preds)
) / len(refs)

#Exact Match Accuracy
exact_match = sum(1 for r,p in zip(refs,preds) if r==p) / len(refs)

#BLEU Score
list_of_refs = [[r.split()] for r in refs]
hypotheses  = [p.split()    for p in preds]
bleu_score  = corpus_bleu(list_of_refs, hypotheses)

print(f"Character-Level Accuracy : {char_acc:.4f}")
print(f"Average Edit Distance    : {avg_edit:.2f}")
print(f"Exact Match Accuracy     : {exact_match:.4f}")
print(f"BLEU Score               : {bleu_score:.4f}")
```

```
Character-Level Accuracy : 0.7538
Average Edit Distance    : 4.33
Exact Match Accuracy     : 0.5209
BLEU Score               : 0.6797
```

train เพิ่ม

New Section

```
df = pd.read_csv("drive/MyDrive/ai/merged.csv")
train, temp = train_test_split(df, test_size=0.3, random_state=42)
val, test = train_test_split(temp, test_size=0.5, random_state=42)
train.to_csv("train2.csv", index=False)
val.to_csv("val2.csv", index=False)
test.to_csv("test2.csv", index=False)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-c057c19be9b3> in <cell line: 0>()
----> 1 df = pd.read_csv("drive/MyDrive/ai/merged.csv")
      2 train, temp = train_test_split(df, test_size=0.3, random_state=42)
      3 val, test = train_test_split(temp, test_size=0.5, random_state=42)
      4 train.to_csv("train2.csv", index=False)
      5 val.to_csv("val2.csv", index=False)

NameError: name 'pd' is not defined
```

Next steps: Explain error

```
datasets = {
    "train":      "train2.csv",
    "val":        "val2.csv",
    "test":       "test2.csv"
}
raw_datasets2= load_dataset("csv", data_files=datasets)
# ตรวจสอบตัวอย่างแรก
print(raw_datasets2["train"][0])
raw_datasets2
```

```
{'sentence': 'ตราบใดที่ sync code snippet ระหว่างอุปกรณ์สำเร็จฉันโพสต์ใน chat group ให้เพื่อนเข้าถึง logic ได้ทันที', 'eng_sentence': "ตราบใดที่ sync cod
DatasetDict({
    train: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 5777
    })
    val: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 1238
    })
    test: Dataset({
        features: ['sentence', 'eng_sentence'],
        num_rows: 1238
    })
})
```

Double-click (or enter) to edit

```
from transformers import MT5ForConditionalGeneration, MT5Tokenizer

checkpoint_dir = "drive/MyDrive/ai/model3"
model = MT5ForConditionalGeneration.from_pretrained(checkpoint_dir)
tokenizer = AutoTokenizer.from_pretrained(checkpoint_dir)


from transformers import DataCollatorForSeq2Seq
prefix = "fix typo to natural Thai sentence: "
def preprocess_functions(examples):
    inputs = [prefix + doc for doc in examples["eng_sentence"]]
    model_inputs = tokenizer(inputs,text_target=examples["sentence"], max_length=64, truncation=True)
    return model_inputs


tokenized_datasets2 = raw_datasets2.map(
    preprocess_function,
    batched=True,
    remove_columns=["sentence", "eng_sentence"]
)
tokenized_datasets2
```

```
Map: 100%                                   5777/5777 [00:00<00:00, 15271.25 examples/s]

Map: 100%                                   1238/1238 [00:00<00:00, 13750.18 examples/s]

Map: 100%                                   1238/1238 [00:00<00:00, 14254.94 examples/s]

DatasetDict({
    train: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 5777
    })
    val: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 1238
    })
    test: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 1238
    })
})
```

```
from transformers import DataCollatorForSeq2Seq
data_collator = DataCollatorForSeq2Seq(tokenizer, model=model)
```

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./results",
    eval_strategy="epoch",
    logging_strategy="epoch",
    save_strategy="epoch",
    logging_steps=50,
    report_to="tensorboard",
    learning_rate=3e-5,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=64,
    num_train_epochs=30,
    weight_decay=0.01,
    predict_with_generate=True,
)
```

```
from transformers import  Seq2SeqTrainer

trainers = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets2["train"],
    eval_dataset=tokenized_datasets2["val"],
    data_collator=data_collator,
    tokenizer=tokenizer,
)

trainers.train()
```

```
<ipython-input-21-d6892a95a06e>:3: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Seq2SeqTrainer._
  trainers = Seq2SeqTrainer(
```

[5430/5430 1:00:17, Epoch 30/30]

Epoch	Training Loss	Validation Loss
1	1.180900	0.756787
2	1.071700	0.704987
3	1.010200	0.675252
4	0.961100	0.649837
5	0.924300	0.621983
6	0.892600	0.605654
7	0.860900	0.586101
8	0.836700	0.574505
9	0.817200	0.558185
10	0.796900	0.550338
11	0.776200	0.535398
12	0.758100	0.522747
13	0.750600	0.517181
14	0.726000	0.511981
15	0.723200	0.502258
16	0.705400	0.496009
17	0.699300	0.489487
18	0.682100	0.485625
19	0.676200	0.480830
20	0.663900	0.478816
21	0.663300	0.475811
22	0.654500	0.472371
23	0.654100	0.470298
24	0.646000	0.466357
25	0.639900	0.464878
26	0.644500	0.462816
27	0.634600	0.462914
28	0.632100	0.460768
29	0.630700	0.460267
30	0.634800	0.459609

```
TrainOutput(global_step=5430, training_loss=0.7649353533159962, metrics={'train_runtime': 3617.6417, 'train_samples_per_second':
```

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
import torch
import pandas as pd


def correct_typo(text: str) -> str:
    input_str = "fix typo to natural Thai sentence: " + text
    inputs = tokenizer(input_str, return_tensors="pt", truncation=True, padding="longest")
    inputs = {k: v.to(model.device) for k, v in inputs.items()}
    outputs = model.generate(**inputs, max_new_tokens=64)
    tttx = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return tttx


print(correct_typo("พรุ่งนี้มี meeting dy[]^d8hk8ole8yP vpjk]n,g9iup, presentation fh;pot8iy["))
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))
print(correct_typo("ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Water Sunscreen SPF50! PA!!!! xdxhv'zb;]he]7d57' 2 9jv!"))
print(correct_typo("บัญชีเริ่มต้นที่ใช้สำหรับเรียกใช้บริการ Windows Ffp,ulbmTbV-yho9je [yP=uouh0t.=h-hv,^]xit0e9y;8v,rb;g9viVgrnjvpnopyo9y;9ozjkog8inv-jk
print(correct_typo("วันนี้ฉันต้องattend corporate meetinggrnjvpresent business plani;,57'vTb[kpfinancial forecastc]tmarketing strategy.shdy[stakeh
print(correct_typo("เล่นminecraftdyog5vtgrnvo"))
```

```
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))
print(correct_typo("ได้มาจากtwittersinvwfh0kdmujwso"))
```

```
จาก: พรุ่งนี้มี meeting dy[]^d8hk8ole8yP vpjk]n,g9iup, presentation fh;pot8iy[ ได้เป็น: พรุ่งนี้มี meeting กับลูกค้าทุกคน พร้อมเตรียม presentation ด้วยนะ
จาก: เล่นrobloxdyog5vtgrnvo ได้เป็น: เล่นrobloxกันเถอะ
ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Water Sunscreen SPF50+ PA++++ ปกป้องผิวล้ำลึกถึง 2 ต่อ!
บัญชีเริ่มต้นที่ใช้สำหรับเรียกใช้บริการ Windows ดาต้าของเว็บไซต์นี้จะใช้ข้อมูลประมวลผลผลอัตโนมัติเพื่อจัดการบัญชีผู้ใช้
วันนี้ฉันต้องattend corporate meetingเพื่อpresent business planสรุปผลการลงทุนfinancial forecastและmarketing strategyให้กับstakeholdersก่อนประชุม
เล่นminecraftกันเถอะ
เล่นrobloxกันเถอะ
ได้มาจากtwitterหรือยังจากที่โรงหนัง
```

```
# 2. นำเข้าโมดูล
import pandas as pd
from Levenshtein import distance as levenshtein_distance
from nltk.translate.bleu_score import corpus_bleu

# 3. เตรียมข้อมูลทดสอบ
#    สมมติคุณมี DataFrame test_df ที่แบ่งไว้แล้ว และฟังก์ชัน correct_typo()
test_df = pd.read_csv("test.csv")
refs = test_df["sentence"].tolist()
preds = [correct_typo(t) for t in test_df["eng_sentence"]]

# 4. Character-Level Accuracy
total_chars = sum(len(r) for r in refs)
correct_chars = sum(
    sum(1 for a,b in zip(r,p) if a==b)
    for r,p in zip(refs,preds)
)
char_acc = correct_chars / total_chars

# 5. Average Levenshtein Distance
avg_edit = sum(
    levenshtein_distance(r, p)
    for r,p in zip(refs,preds)
) / len(refs)

# 6. Exact Match Accuracy
exact_match = sum(1 for r,p in zip(refs,preds) if r==p) / len(refs)

# 7. BLEU Score
list_of_refs = [[r.split()] for r in refs]
hypotheses  = [p.split()    for p in preds]
bleu_score  = corpus_bleu(list_of_refs, hypotheses)

# 8. แสดงผล
print(f"Character-Level Accuracy : {char_acc:.4f}")
print(f"Average Edit Distance    : {avg_edit:.2f}")
print(f"Exact Match Accuracy     : {exact_match:.4f}")
print(f"BLEU Score               : {bleu_score:.4f}")
```

```
Character-Level Accuracy : 0.7926
Average Edit Distance    : 2.62
Exact Match Accuracy     : 0.6005
BLEU Score               : 0.7344
```

```
trainers.save_model("drive/MyDrive/ai/model")
```

```
import pandas as pd
from Levenshtein import distance as levenshtein_distance
from nltk.translate.bleu_score import corpus_bleu

test_df = pd.read_csv("test2.csv")
refs = test_df["sentence"].tolist()
preds = [correct_typo(t) for t in test_df["eng_sentence"]]

#Character-Level Accuracy
total_chars = sum(len(r) for r in refs)
correct_chars = sum(
    sum(1 for a,b in zip(r,p) if a==b)
    for r,p in zip(refs,preds)
)
char_acc = correct_chars / total_chars
```

```
#Average Levenshtein Distance
avg_edit = sum(
    levenshtein_distance(r, p)
    for r,p in zip(refs,preds)
```