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

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.2)
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) (0.31.4)
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) (0.31.4)
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) (6.4.4)
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) (0.31.4)
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 22.8 MB/s eta 0:00:00
Downloading fsspec-2025.3.0-py3-none-any.whl (193 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 193.6/193.6 kB 12.4 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
```

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

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

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("train.csv", index=False)
val.to_csv("val.csv", index=False)
test.to_csv("test.csv", index=False)
train = "train.csv"
```

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

```python
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:   5777/0 [00:00<00:00, 104006.04 examples/s]

Generating val split:   1238/0 [00:00<00:00, 57927.98 examples/s]

Generating test split:   1238/0 [00:00<00:00, 55672.22 examples/s]
{'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
    })
})
```

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

checkpoint = "google/byt5-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(
tokenizer_config.json: 100%   2.59k/2.59k [00:00<00:00, 275kB/s]

config.json: 100%   698/698 [00:00<00:00, 72.1kB/s]

special_tokens_map.json: 100%   2.50k/2.50k [00:00<00:00, 256kB/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download. For better perfo
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to r
pytorch_model.bin: 100%   1.20G/1.20G [00:03<00:00, 303MB/s]
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download. For better perfo
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to r
model.safetensors: 100%   1.20G/1.20G [00:06<00:00, 355MB/s]

generation_config.json: 100%   147/147 [00:00<00:00, 16.6kB/s]
```

```
from transformers import DataCollatorForSeq2Seq
def preprocess(examples):
    model_inputs = tokenizer(examples["eng_sentence"],text_target=examples["sentence"], max_length=512, truncation=True)
    return model_inputs
```

Double-click (or enter) to edit

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

```
{'input_ids': [[227, 187, 152, 227, 187, 166, 227, 187, 181, 227, 187, 157, 227, 188, 134, 227, 187, 151, 227, 187, 154, 227, 187, 184,
227, 188, 139, 35, 118, 124, 113, 102, 35, 102, 114, 103, 104, 35, 118, 113, 108, 115, 115, 104, 119, 35, 108, 119, 118, 62, 109, 110,
42, 121, 57, 123, 103, 108, 76, 89, 111, 104, 106, 108, 73, 51, 93, 124, 114, 227, 188, 133, 117, 111, 60, 89, 49, 114, 35, 102, 107,
100, 119, 35, 106, 117, 114, 120, 115, 35, 49, 118, 107, 106, 117, 113, 109, 121, 114, 106, 48, 107, 110, 56, 58, 42, 35, 111, 114,
106, 108, 102, 35, 122, 105, 107, 112, 124, 114, 112, 120, 1], [227, 187, 135, 227, 187, 186, 227, 187, 156, 227, 187, 156, 227, 187,
184, 227, 188, 140, 227, 187, 139, 227, 187, 179, 227, 188, 135, 227, 187, 158, 109, 114, 108, 113, 35, 113, 108, 106, 107, 119, 35,
102, 124, 102, 111, 104, 35, 119, 114, 120, 117, 49, 114, 60, 124, 62, 106, 47, 113, 121, 42, 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, 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, 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, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1]], 'labels': [[227, 187, 152, 227, 187, 166, 227, 187, 181, 227, 187, 157, 227, 188, 134, 227, 187, 151, 227, 187, 154, 227, 187,
184, 227, 188, 139, 35, 118, 124, 113, 102, 35, 102, 114, 103, 104, 35, 118, 113, 108, 115, 115, 104, 119, 35, 227, 187, 166, 227, 187,
179, 227, 187, 174, 227, 187, 170, 227, 188, 139, 227, 187, 181, 227, 187, 138, 227, 187, 176, 227, 187, 187, 227, 187, 158, 227, 187,
132, 227, 187, 166, 227, 187, 150, 227, 188, 143, 227, 187, 173, 227, 187, 182, 227, 188, 131, 227, 187, 166, 227, 188, 138, 227, 187,
139, 227, 187, 140, 227, 187, 180, 227, 187, 156, 227, 188, 133, 227, 187, 161, 227, 187, 173, 227, 187, 152, 227, 188, 143, 227, 188,
134, 227, 187, 156, 35, 102, 107, 100, 119, 35, 106, 117, 114, 120, 115, 35, 227, 188, 134, 227, 187, 174, 227, 188, 140, 227, 188,
131, 227, 187, 161, 227, 187, 186, 227, 188, 139, 227, 187, 176, 227, 187, 156, 227, 188, 131, 227, 187, 133, 227, 188, 140, 227, 187,
181, 227, 187, 153, 227, 187, 185, 227, 187, 138, 35, 111, 114, 106, 108, 102, 35, 227, 188, 135, 227, 187, 151, 227, 188, 140, 227,
187, 154, 227, 187, 180, 227, 187, 156, 227, 187, 154, 227, 187, 184, 1], [227, 187, 135, 227, 187, 186, 227, 187, 156, 227, 187, 156,
227, 187, 184, 227, 188, 140, 227, 187, 139, 227, 187, 179, 227, 188, 135, 227, 187, 158, 109, 114, 108, 113, 35, 113, 108, 106, 107,
119, 35, 102, 124, 102, 111, 104, 35, 119, 114, 120, 117, 227, 188, 134, 227, 187, 156, 227, 187, 152, 227, 187, 180, 227, 187, 170,
227, 188, 131, 227, 187, 164, 227, 187, 186, 227, 187, 176, 227, 187, 138, 1]]}
```

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

```
Map: 100%                                5777/5777 [00:01<00:00, 3537.78 examples/s]

Map: 100%                                1238/1238 [00:00<00:00, 3567.18 examples/s]

Map: 100%                                1238/1238 [00:00<00:00, 3187.47 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
    })
})
```

```
!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 3.0 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 120.7 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=3e-4,
    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"],         # ชุด train ที่เรา map แล้ว
    eval_dataset=tokenized_datasets["val"],     # ชุด validation
    data_collator=data_collator,
    tokenizer=tokenizer,                               # บอก Trainer ยังต้องใช้ tokenizer ด้วย
)

# เริ่มเทรน
trainer.train()
```

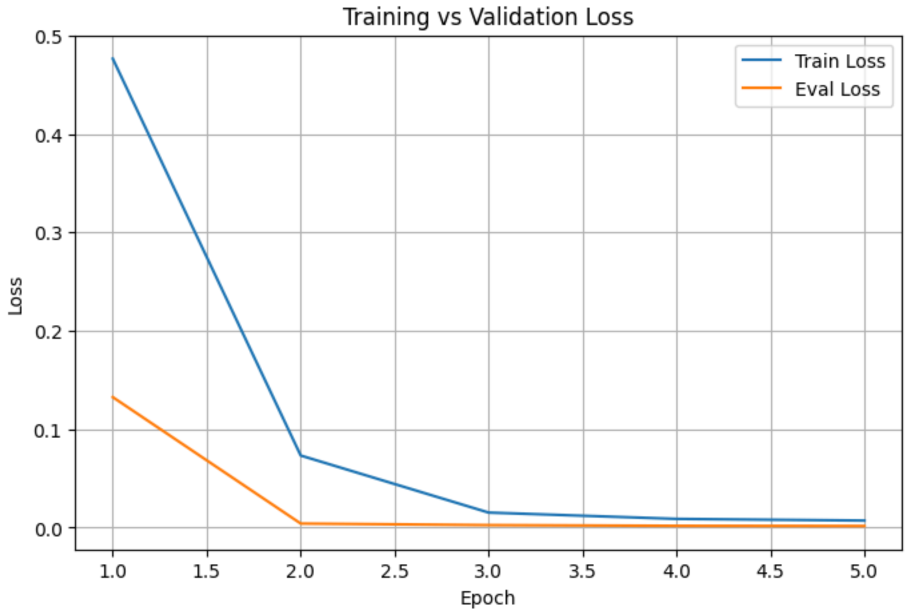
train เพิ่ม

```
from transformers import TrainingArguments
from transformers import Seq2SeqTrainingArguments, Seq2SeqTrainer, DataCollatorForSeq2Seq
data_collator = DataCollatorForSeq2Seq(tokenizer, model=model)
training_args = Seq2SeqTrainingArguments(
    output_dir="./resultss",
    eval_strategy="epoch",
    logging_strategy="epoch",
    save_strategy="epoch",
    logging_steps=50,
    report_to="tensorboard",
    learning_rate=3e-4,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=64,
    num_train_epochs=10,
    weight_decay=0.01,
    predict_with_generate=True,
)
```

```
from transformers import  Seq2SeqTrainer

trainers = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],         # ชุด train ที่เรา map แล้ว
    eval_dataset=tokenized_datasets["val"],     # ชุด validation
    data_collator=data_collator,
    tokenizer=tokenizer,                               # บอก Trainer ยังต้องใช้ tokenizer ด้วย
)

# เริ่มเทรน
trainers.train()
```

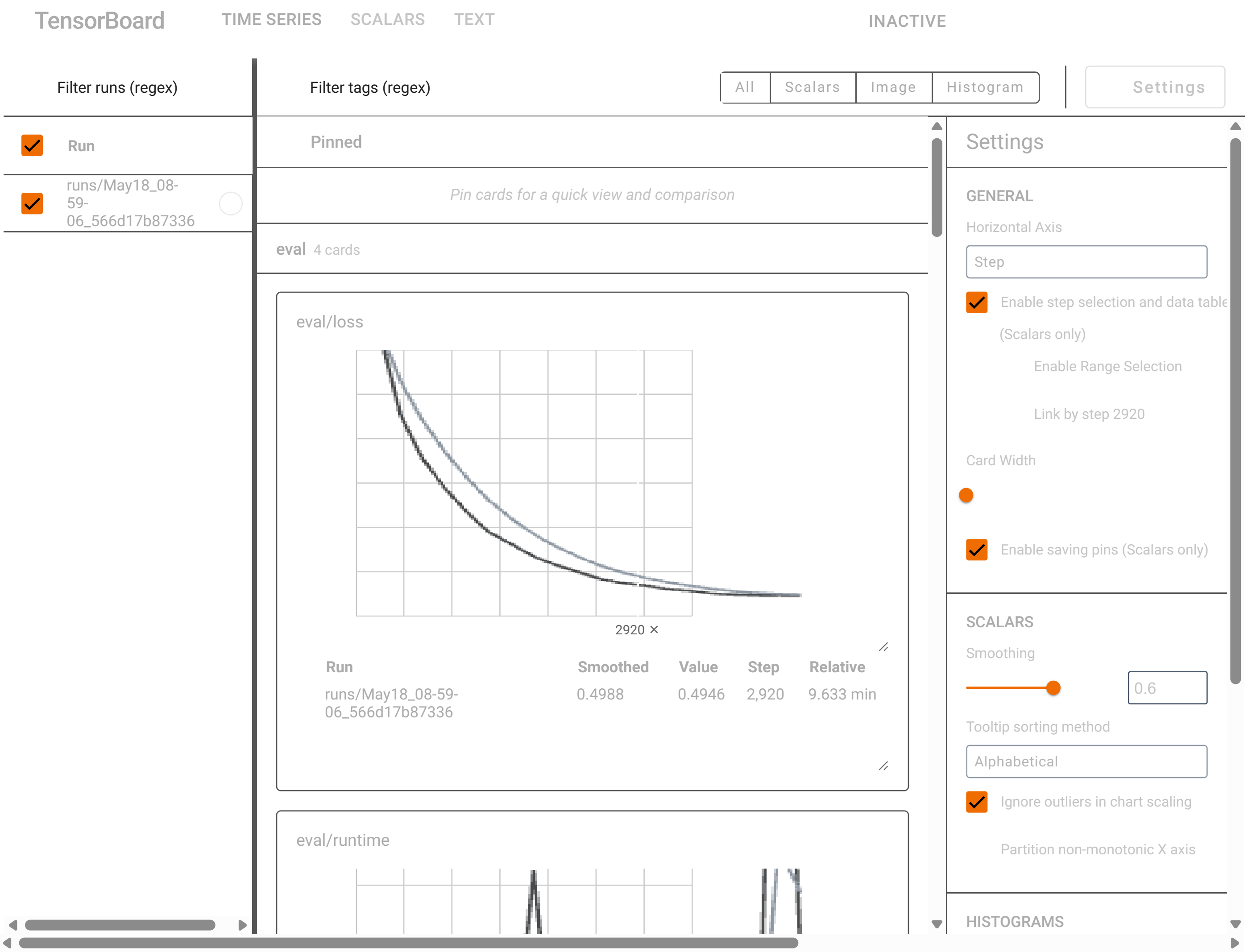
```
trainer.save_model("drive/MyDrive/ai/BYT5-SMALL")
```

```
logs = trainer.state.log_history
train_losses = [x['loss'] for x in logs if 'loss' in x and 'epoch' in x and 'eval_loss' not in x]
val_losses   = [x['eval_loss'] for x in logs if 'eval_loss' in x]
epochs = list(range(1, len(train_losses) + 1))
%matplotlib inline
import matplotlib.pyplot as plt

plt.figure(figsize=(8,5))
plt.plot(epochs, train_losses, label='Train Loss')
plt.plot(epochs, val_losses,   label='Eval Loss')
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.title('Training    Validation Loss')
plt.legend()
plt.grid(True)
plt.show()
```



```
%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.)
```



☑ Run

☑ runs/May18_08-59-06_566d17b87336

Run	Smoothed	Value	Step	Relative
runs/May18_08-59-06_566d17b87336	0.4988	0.4946	2,920	9.633 min

Settings

GENERAL

Horizontal Axis

Step

☑ Enable step selection and data table (Scalars only)

Enable Range Selection

Link by step 2920

Card Width

☑ Enable saving pins (Scalars only)

SCALARS

Smoothing

0.6

Tooltip sorting method

Alphabetical

☑ Ignore outliers in chart scaling

Partition non-monotonic X axis

HISTOGRAMS

Double-click (or enter) to edit

```
metrics = trainer.evaluate(tokenized_datasets["test"])
print(metrics)
```

```
[12/12 00:03]
{'eval_loss': 0.5668103098869324, 'eval_runtime': 3.8, 'eval_samples_per_second': 201.581, 'eval_steps_per_second': 3.158, 'epoch': 15.0
```

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

```
Requirement already satisfied: python-Levenshtein in /usr/local/lib/python3.11/dist-packages (0.27.1)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: Levenshtein==0.27.1 in /usr/local/lib/python3.11/dist-packages (from python-Levenshtein) (0.27.1)
Requirement already satisfied: rapidfuzz<4.0.0,>=3.9.0 in /usr/local/lib/python3.11/dist-packages (from Levenshtein==0.27.1->python-Leve
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)
```

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

# model_dir = "./drive/MyDrive/ai/test"
# tokenizer = AutoTokenizer.from_pretrained(model_dir)
# model     = AutoModelForSeq2SeqLM.from_pretrained(model_dir)
# device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
# model.to(device)
def correct_typo(text: str) -> str:
    inputs = tokenizer(text, 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=512)
    tttx = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return tttx
print(correct_typo("พรุ่งนี้มี meeting dy[]^d8hk8ole8yP vpjk]n,g9iup, presentation fh;pot8iy["))#พรุ่งนี้มี meeting กับลูกค้าคนสำคัญ อย่าลืมเตรียม presentati
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))#เล่นrobloxกันเถอะเพื่อน
print(correct_typo("ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Water Sunscreen SPF50! PA!!!! xdxhv'zb;]he]7d57' 2 9jv!"))#ใหม่! กันแดดอัจฉริยะ UV Adapt Hya Wa
print(correct_typo("เล่นminecraftdyog5vtgrnvo"))#เล่นminecraftกันเถอะเพื่อน
print(correct_typo("เล่นrobloxdyog5vtgrnvo"))#เล่นrobloxกันเถอะเพื่อน
print(correct_typo("train modelgliH0py'vt-vcodesojvp"))#train modelเสร็จยังอะขอcodeหน่อย


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 : 0.9928
Average Edit Distance    : 0.10
Exact Match Accuracy     : 0.9628
BLEU Score               : 0.9424
```

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

test_df = pd.read_csv("realworld.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}")
```