```
!pip install datasets transformers rouge-score nltk
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: datasets in /usr/local/lib/python3.7/dist-packages (2.6.1)
Requirement already satisfied: transformers in /usr/local/lib/python3.7/dist-packages (4.24.0)
Requirement already satisfied: rouge-score in /usr/local/lib/python3.7/dist-packages (0.1.2)
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (3.7)
Requirement already satisfied: fsspec[http]>=2021.11.1 in /usr/local/lib/python3.7/dist-packages (from datas
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from datasets) (1.21.6
Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (6.0
Requirement already satisfied: multiprocess in /usr/local/lib/python3.7/dist-packages (from datasets) (0.70.
Requirement already satisfied: responses<0.19 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.1
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (2
Requirement already satisfied: xxhash in /usr/local/lib/python3.7/dist-packages (from datasets) (3.1.0)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.2.0 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from datasets) (21.3)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (from datasets) (6.0)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.7/dist-packages (from datasets) (4.64.
Requirement already satisfied: dill<0.3.6 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.3.5.1
Requirement already satisfied: aiohttp in /usr/local/lib/python3.7/dist-packages (from datasets) (3.8.3)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from datasets)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from datasets) (1.3.5)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.7/dist-packages (from aiohttp->
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datas
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in /usr/local/lib/python3.7/dist-packages (from a
Requirement already satisfied: asynctest==0.13.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->da
Requirement already satisfied: typing-extensions>=3.7.4 in /usr/local/lib/python3.7/dist-packages (from aioh
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.7/dist-packages (from aiohttp->da
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datase
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.7/dist-packages (from aiohttp->dat
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from pack
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests>=2
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests>=
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transformer
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.7/dist-packages (from rouge-score) (1.1
```

```
Requirement already satisfied: absl-py in /usr/local/lib/python3.7/dist-packages (from rouge-score) (1.3.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk) (1.2.0)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk) (7.1.2)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata-
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas
```

```
from huggingface_hub import notebook_login
notebook_login()
```

```
Login successful
Your token has been saved to /root/.huggingface/token
```

```
!apt install git-lfs
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
git-lfs is already the newest version (2.3.4-1).
The following package was automatically installed and is no longer required:
  libnvidia-common-460
Use 'apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 4 not upgraded.
```

```
from datasets import load_dataset, load_metric
from transformers import AutoTokenizer
from transformers import DataCollatorForSeq2Seq
from transformers import AutoModelForSeq2SeqLM, Seq2SeqTrainingArguments, Seq2SeqTrainer
from transformers import AutoModelForSeq2SeqLM, DataCollatorForSeq2Seq, Seq2SeqTrainingArguments, Seq2SeqTrainer
from transformers import create_optimizer, AdamWeightDecay
from transformers import AutoConfig
from transformers import T5Model
import nltk
import numpy as np
nltk.download('punkt')
```

```
metric = load_metric("rouge")
```

```
tokenizer = 0
prefix = "summarize: "
max_input_length = 1024
max_target_length = 128


#adds padding to input before traing the model on the dataset
def preprocess_function(examples):
    inputs = [prefix + doc for doc in examples["article"]]
    model_inputs = tokenizer(inputs, max_length=max_input_length, truncation=True)

    # Setup the tokenizer for targets
    with tokenizer.as_target_tokenizer():
        labels = tokenizer(examples["abstract"], max_length=max_target_length, truncation=True)

    model_inputs["labels"] = labels["input_ids"]
    return model_inputs



#tokenizer for the dataset
def my_tokenize(model_checkpoint, dataset, subset):
  global tokenizer

  sum = load_dataset(dataset, subset)

  tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
  tokenized_sum = sum.map(preprocess_function, batched=True)

  return (tokenizer, tokenized_sum)


#create new summerization model
def get_model(model_checkpoint, tokenizer):
  #make a model that is not pre-trained
  config = AutoConfig.from_pretrained(model_checkpoint)

  model = AutoModelForSeq2SeqLM.from_config(config)
```

```
    model.init_weights()

    data_collator = DataCollatorForSeq2Seq(tokenizer=tokenizer, model=model)
    return (model, data_collator)


def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    decoded_preds = tokenizer.batch_decode(predictions, skip_special_tokens=True)
    # Replace -100 in the labels as we can't decode them.
    labels = np.where(labels != -100, labels, tokenizer.pad_token_id)
    decoded_labels = tokenizer.batch_decode(labels, skip_special_tokens=True)

    # Rouge expects a newline after each sentence
    decoded_preds = ["\n".join(nltk.sent_tokenize(pred.strip())) for pred in decoded_preds]
    decoded_labels = ["\n".join(nltk.sent_tokenize(label.strip())) for label in decoded_labels]

    result = metric.compute(predictions=decoded_preds, references=decoded_labels, use_stemmer=True)
    # Extract a few results
    result = {key: value.mid.fmeasure * 100 for key, value in result.items()}

    # Add mean generated length
    prediction_lens = [np.count_nonzero(pred != tokenizer.pad_token_id) for pred in predictions]
    result["gen_len"] = np.mean(prediction_lens)

    return {k: round(v, 4) for k, v in result.items()}


#set hyper paramaters
#change hyper paramters for better trained model
def get_my_hyper_params(model_checkpoint, my_epochs, floating_point):
  batch_size = 16
  model_name = model_checkpoint
  args = Seq2SeqTrainingArguments(
      f"{model_name}-science-papers",
      evaluation_strategy = "epoch",
      learning_rate=2e-5,
      per_device_train_batch_size=batch_size,
```

```
        per_device_eval_batch_size=batch_size,
        weight_decay=0.01,
        save_total_limit=20,
        num_train_epochs=my_epochs,
        predict_with_generate=True,
        fp16=floating_point,
        push_to_hub=True,
    )

    return args

#make the trainer
def get_trainer(model, tokenizer, tokenized_sum, data_collator, training_args):
    trainer = Seq2SeqTrainer(
      model=model,
      args=training_args,
      train_dataset=tokenized_sum["train"],
      eval_dataset=tokenized_sum["test"],
      tokenizer=tokenizer,
      data_collator=data_collator,
      compute_metrics=compute_metrics
      )
    return trainer
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
def my_train_model():
    model_name = "t5-small"
    dataset = "scientific_papers"
    subset = "arxiv"
    epochs = 5
    floating_point = True

    token_tuple = my_tokenize(model_name, dataset, subset)

    model_tuple = get_model(model_name, token_tuple[0])

    params = get_my_hyper_params(model_name, epochs, floating_point)
```

```
    trainer = get_trainer(model_tuple[0], token_tuple[0], token_tuple[1], model_tuple[1], params)
    trainer.train()
    return trainer

trainer = my_train_model()
```

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	4.473500	4.372655	9.960400	1.764100	8.621300	9.277900	19.000000
2	4.010400	3.938435	11.400100	2.147400	9.651600	10.660200	19.000000
3	3.823700	3.757975	11.180600	2.122900	9.388100	10.385300	19.000000
4	3.738200	3.673797	11.929800	2.322200	9.907700	11.045000	19.000000
5	3.699400	3.640550	12.356800	2.444900	10.237100	11.420900	19.000000

```
Saving model checkpoint to t5-small-science-papers/checkpoint-56000
Configuration saved in t5-small-science-papers/checkpoint-56000/config.json
Model weights saved in t5-small-science-papers/checkpoint-56000/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-56000/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-56000/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-46000] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-56500
Configuration saved in t5-small-science-papers/checkpoint-56500/config.json
Model weights saved in t5-small-science-papers/checkpoint-56500/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-56500/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-56500/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-46500] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-57000
Configuration saved in t5-small-science-papers/checkpoint-57000/config.json
Model weights saved in t5-small-science-papers/checkpoint-57000/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-57000/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-57000/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-47000] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-57500
Configuration saved in t5-small-science-papers/checkpoint-57500/config.json
Model weights saved in t5-small-science-papers/checkpoint-57500/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-57500/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-57500/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-47500] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-58000
Configuration saved in t5-small-science-papers/checkpoint-58000/config.json
Model weights saved in t5-small-science-papers/checkpoint-58000/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-58000/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-58000/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-48000] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-58500
Configuration saved in t5-small-science-papers/checkpoint-58500/config.json
```

```
Model weights saved in t5-small-science-papers/checkpoint-58500/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-58500/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-58500/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-48500] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-59000
Configuration saved in t5-small-science-papers/checkpoint-59000/config.json
Model weights saved in t5-small-science-papers/checkpoint-59000/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-59000/tokenizer_config.json
Special tokens file saved in t5-small-science-papers/checkpoint-59000/special_tokens_map.json
Deleting older checkpoint [t5-small-science-papers/checkpoint-49000] due to args.save_total_limit
Saving model checkpoint to t5-small-science-papers/checkpoint-59500
Configuration saved in t5-small-science-papers/checkpoint-59500/config.json
Model weights saved in t5-small-science-papers/checkpoint-59500/pytorch_model.bin
tokenizer config file saved in t5-small-science-papers/checkpoint-59500/tokenizer_config.json
```