Text Summarization of BBC News Using Pegasus NEWS model finetuning

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Exploratory Data Analysis

```
! pip install -q transformers[torch] datasets
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 547.8/547.8 kB 12.8 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 309.4/309.4 kB 30.9 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 40.8/40.8 MB 41.5 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 17.1 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.9/64.9 kB 10.3 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.1/194.1 kB 32.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 19.7 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.3/21.3 MB 68.6 MB/s eta 0:00:00
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the
cudf-cu12 24.4.1 requires pyarrow<15.0.0a0,>=14.0.1, but you have pyarrow 16.1.0 which is incompatible.
google-colab 1.0.0 requires requests==2.31.0, but you have requests 2.32.3 which is incompatible.
ibis-framework 8.0.0 requires pyarrow<16,>=2, but you have pyarrow 16.1.0 which is incompatible.
```

```
import os, sys, csv
os.environ['TOKENIZERS_PARALLELISM']='false'

import torch
import torch.nn

import pandas as pd
import numpy as np

from tqdm import tqdm

import matplotlib.pyplot as plt

from transformers import PegasusForConditionalGeneration, PegasusTokenizer, set_seed
from transformers import get_scheduler, DataCollatorForSeq2Seq, Seq2SeqTrainingArguments, Seq2SeqTrainer

from datasets import Dataset, DatasetDict, load_metric, load_dataset
```

```
model = 'google/pegasus-multi_news'
device="cuda" if torch.cuda.is_available() else "cpu"
vocab_size = 96103

max_input_length = 1024
max_target_length = 455

set_seed(42)
```

```
!unzip /content/drive/MyDrive/archive.zip -d /content/
```

```
  inflating: /content/bbc news summary/BBC News Summary/Summaries/tech/401.txt
```

Creating theh dataframe

```
data = []
topics = ['business', 'entertainment', 'politics', 'sport', 'tech']

articles_dir = '/content/BBC News Summary/News Articles/'
articles = pd.Series(name='Article', dtype=str)

summaries_dir = '/content/BBC News Summary/Summaries/'
summaries = pd.Series(name='Summary', dtype=str)

# changed from previous function, I think this is also one
# of the reasons why my model didn't work previously, my data was not preprocessed correctly.

for topic in topics:
    topic_len = 0
    full_path = os.path.join(articles_dir, topic)

    for path in os.listdir(full_path):
        if os.path.isfile(os.path.join(full_path, path)):
            topic_len += 1

    for x in range(1,topic_len):
        _data = pd.Series(dtype=str)
        with open((full_path + r'/' + str(x).zfill(3) + r'.txt'), 'r', encoding='ISO-8859-1') as f:
            file = f.read()
            file = file.replace('\n', ' ')
            fileseries = pd.Series(data=file)
            articles = pd.concat([fileseries, articles], ignore_index=True)

    articles = pd.DataFrame(articles)
    articles = articles.set_index(topic + articles.index.astype(str))

    s_dir_path = os.path.join(summaries_dir, topic)

    for x in range(1,topic_len):
        s_current_data = pd.Series(dtype=str)
        with open((s_dir_path + r'/' + str(x).zfill(3) + r'.txt'), 'r', encoding='ISO-8859-1') as f:
            file = f.read()
            file = file.replace('\n', ' ')
            fileseries = pd.Series(data=file)
            summaries = pd.concat([fileseries, summaries], ignore_index=True)

    summaries = pd.DataFrame(summaries)
    summaries = summaries.set_index(topic + summaries.index.astype(str))

articles = articles.rename(columns={0 : 'Articles'})
summaries = summaries.rename(columns={0 : 'Summaries'})

data = articles.join(summaries)
data.tail(25)

```

	Articles	Summaries
tech2195	Yukos loses US bankruptcy battle A judge has ...	The court ruling is a blow to efforts to get d...
tech2196	US trade gap hits record in 2004 The gap betw...	The Commerce Department said the trade deficit...
tech2197	Mixed signals from French economy The French ...	Despite the apparent shortfall in annual econo...
tech2198	Sluggish economy hits German jobs The number ...	But officials said stagnant growth was still s...
tech2199	Rank 'set to sell off film unit' Leisure grou...	Leisure group Rank could unveil plans to demer...
tech2200	Call centre users 'lose patience' Customers t...	The drop in patience comes as the number of ca...
tech2201	India widens access to telecoms India has rai...	"We need at least $20bn (Â£10.6bn) in investme...
tech2202	India's rupee hits five-year high India's rup...	India's rupee has hit a five-year high after S...
tech2203	Parmalat boasts doubled profits Parmalat, the...	On Tuesday, the company's administrator, turna...
tech2204	China keeps tight rein on credit China's effo...	The breakneck pace of economic expansion has k...
tech2205	Air passengers win new EU rights Air passenge...	In addition, if a flight is cancelled or delay...
tech2206	Telegraph newspapers axe 90 jobs The Daily an...	"Journalists are the lifeblood of any newspape...
tech2207	Peugeot deal boosts Mitsubishi Struggling Jap...	Struggling Japanese car maker Mitsubishi Motor...
tech2208	Indonesians face fuel price rise Indonesia's ...	Indonesia's government has confirmed it is con...
tech2209	Ask Jeeves tips online ad revival Ask Jeeves ...	Ask Jeeves has become the third leading online...
tech2210	Court rejects $280bn tobacco case A US govern...	A US government claim accusing the country's b...
tech2211	Ethiopia's crop production up 24% Ethiopia pr...	In 2003, crop production totalled 11.49 millio...
tech2212	India calls for fair trade rules India, which...	At a conference on developing enterprise hoste...
tech2213	Jobs growth still slow in the US The US creat...	The job gains mean that President Bush can cel...
tech2214	Japan narrowly escapes recession Japan's econ...	On an annual basis, the data suggests annual g...
tech2215	Pernod takeover talk lifts Domecq Shares in U...	Pernod has reduced the debt it took on to fund...
tech2216	High fuel prices hit BA's profits British Air...	Rod Eddington, BA's chief executive, said the ...

tech2217	Yukos unit buyer faces loan claim The owners ...	Yukos' owner Menatep Group says it will ask Ro...
tech2218	Dollar gains on Greenspan speech The dollar h...	The dollar has hit its highest level against t...
tech2219	Ad sales boost Time Warner profit Quarterly p...	TimeWarner said fourth quarter sales rose 2% t...

```python
data.replace('\n', ' ', inplace=True, regex=True)
data.replace('  ', ' ', inplace=True, regex=True)
data.reset_index(drop=True, inplace=True)
data
```

	Articles	Summaries
0	US cyber security chief resigns The man making...	Amit Yoran was director of the National Cyber ...
1	Be careful how you code A new European directi...	This goes to the heart of the European project...
2	Spam e-mails tempt net shoppers Computer users...	A third of them read unsolicited junk e-mail a...
3	BT program to beat dialler scams BT is introdu...	BT is introducing two initiatives to help beat...
4	New consoles promise big problems Making games...	Mr Walsh suggested that new studios should mak...
...	...	...
2215	Pernod takeover talk lifts Domecq Shares in UK...	Pernod has reduced the debt it took on to fund...
2216	High fuel prices hit BA's profits British Airw...	Rod Eddington, BA's chief executive, said the ...
2217	Yukos unit buyer faces loan claim The owners o...	Yukos' owner Menatep Group says it will ask Ro...
2218	Dollar gains on Greenspan speech The dollar ha...	The dollar has hit its highest level against t...
2219	Ad sales boost Time Warner profit Quarterly pr...	TimeWarner said fourth quarter sales rose 2% t...

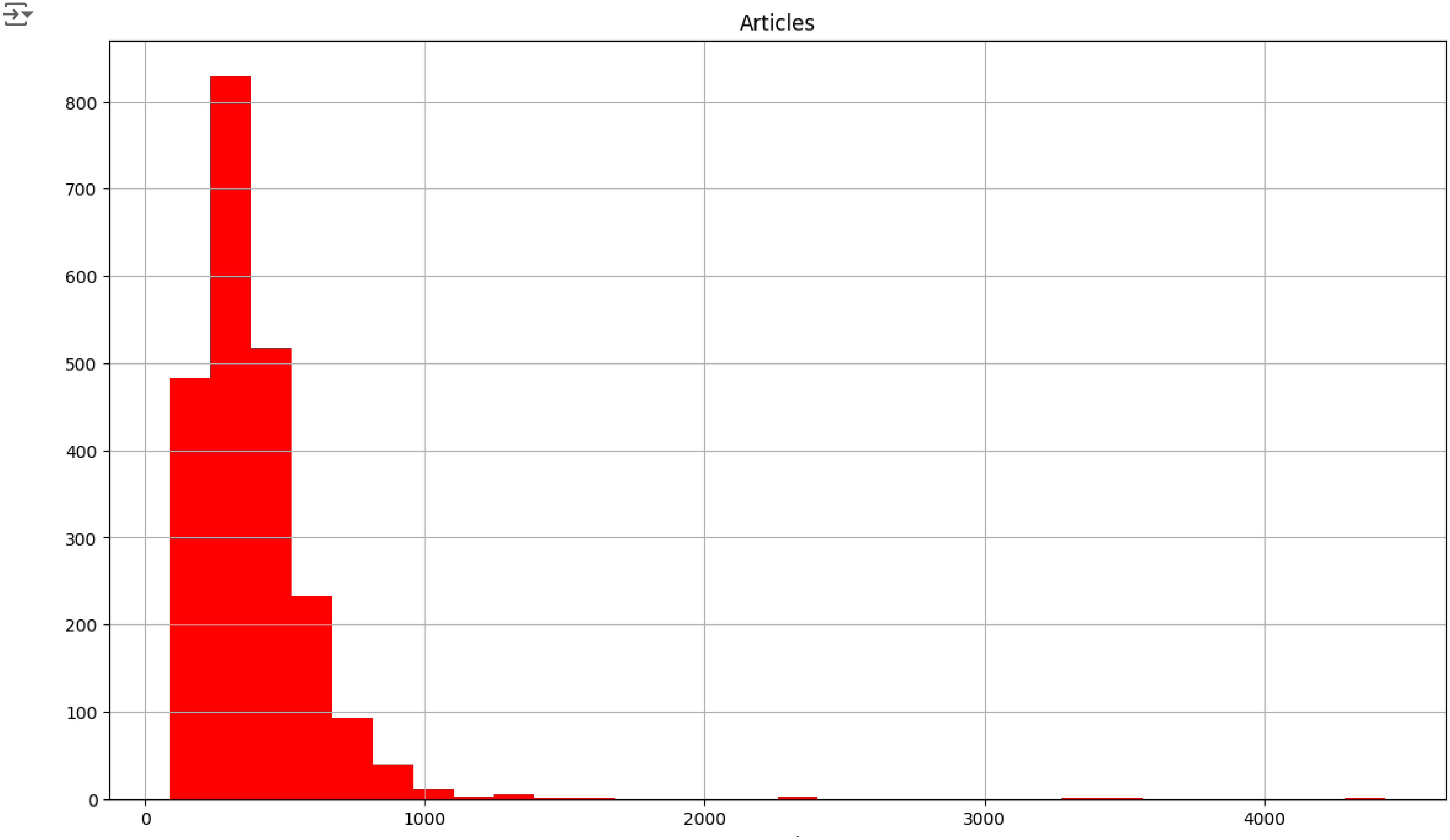
2220 rows × 2 columns

Next steps: Generate code with `data` View recommended plots

Length of Articles

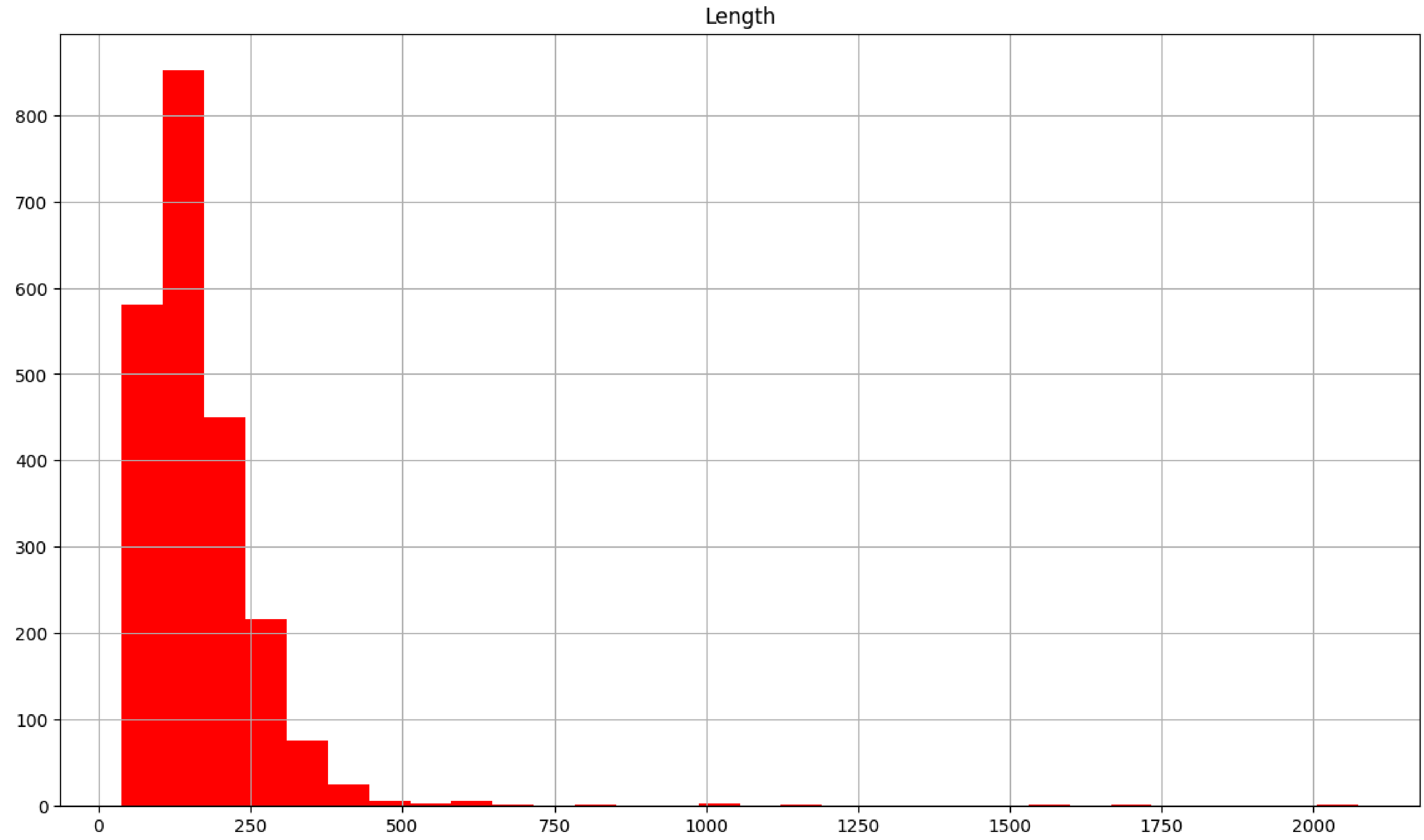
```
data['article_length'] = data['Articles'].str.split().apply(len)

data.hist("article_length", color="Red", figsize=(14,8), bins=30)
plt.suptitle('')
plt.xlabel('Length')
plt.ylabel('')
plt.title('Articles')
plt.show()
```


Articles
800
700
600
500
400
300
200
100
0
0
1000
2000
3000
4000

```
data['summary_length'] = data['Summaries'].str.split().apply(len)

data.hist("summary_length", color="Red", figsize=(14, 8), bins=30)
plt.suptitle('')
plt.xlabel('Summary')
plt.ylabel('')
plt.title('Length')
plt.show()
```



Trimmed Dataset Based on Article &/or Summary Length

```
data = data[data['article_length'] < 1024]
data = data[data['summary_length'] < 300]

data = data.drop(columns=['article_length', 'summary_length'])

data
```

	Articles	Summaries
0	US cyber security chief resigns The man making...	Amit Yoran was director of the National Cyber ...
2	Spam e-mails tempt net shoppers Computer users...	A third of them read unsolicited junk e-mail a...
3	BT program to beat dialler scams BT is introdu...	BT is introducing two initiatives to help beat...
7	Savvy searchers fail to spot ads Internet sear...	Almost 50% of those questioned said they would...
8	Broadband fuels online expression Fast web acc...	More than five million households in the UK ha...
...	...	...
2215	Pernod takeover talk lifts Domecq Shares in UK...	Pernod has reduced the debt it took on to fund...
2216	High fuel prices hit BA's profits British Airw...	Rod Eddington, BA's chief executive, said the ...
2217	Yukos unit buyer faces loan claim The owners o...	Yukos' owner Menatep Group says it will ask Ro...
2218	Dollar gains on Greenspan speech The dollar ha...	The dollar has hit its highest level against t...
2219	Ad sales boost Time Warner profit Quarterly pr...	TimeWarner said fourth quarter sales rose 2% t...

2085 rows × 2 columns

Next steps: Generate code with data View recommended plots

Convert df to datasets then split, I was doing wrong here before.

```
# Convert df
dataset = Dataset.from_pandas(data)

train_test_valid = dataset.train_test_split(test_size=0.3)
test_valid = train_test_valid['test'].train_test_split(test_size=0.4)

# Combine the train/test/valid into one datasetdict
dataset = DatasetDict({
    'train' : train_test_valid['train'],
    'test' : test_valid['test'],
    'valid' : test_valid['train']
})

print('Training Data Shape:', dataset['train'].shape)
print('Testing Data Shape:', dataset['test'].shape)
print('Validation Data Shape:', dataset['valid'].shape)
```

```
Training Data Shape: (1459, 3)
Testing Data Shape: (251, 3)
Validation Data Shape: (375, 3)
```

Get Model

```
tokenizer = PegasusTokenizer.from_pretrained(model, truncation=True, padding=True, batched=True)
pipe = None
model = PegasusForConditionalGeneration.from_pretrained(model)
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: 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
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%                                        88.0/88.0 [00:00<00:00, 7.34kB/s]

spiece.model: 100%                                          1.91M/1.91M [00:00<00:00, 8.01MB/s]

special_tokens_map.json: 100%                                        65.0/65.0 [00:00<00:00, 5.47kB/s]
/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download` is deprecated anc
  warnings.warn(
config.json: 100%                                          1.12k/1.12k [00:00<00:00, 96.4kB/s]

pytorch_model.bin: 100%                                        2.28G/2.28G [00:06<00:00, 379MB/s]
Some weights of PegasusForConditionalGeneration were not initialized from the model checkpoint at google/pegasus-multi_news and
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
generation_config.json: 100%                                        280/280 [00:00<00:00, 24.9kB/s]
```

```
# from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

# tokenizer = AutoTokenizer.from_pretrained("google/pegasus-multi_news")
# model = AutoModelForSeq2SeqLM.from_pretrained("google/pegasus-multi_news")
```

Tokenize and make batches

```
def convert(dataset_article):
    inputs = tokenizer(dataset_article['Articles'], max_length=max_input_length, truncation=True, padding=True)
    with tokenizer.as_target_tokenizer():
        targets = tokenizer(dataset_article['Summaries'], max_length=max_target_length, truncation=True, padding=True)
    return {"input_ids" : inputs["input_ids"],
            "attention_mask" : inputs["attention_mask"],
            "labels" : targets["input_ids"],}
```

```
converted_ds = dataset.map(convert, batched=True)
converted_ds.set_format(type="torch", columns=['input_ids', 'labels', 'attention_mask'])
```

```
Map: 100%                                    1459/1459 [00:04<00:00, 292.30 examples/s]
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:3946: UserWarning: `as_target_tokenizer` is depre
  warnings.warn(
Map: 100%                                    251/251 [00:00<00:00, 286.78 examples/s]
Map: 100%                                    375/375 [00:01<00:00, 308.56 examples/s]
```

Instantiate TrainingArguments

```
# training_args = Seq2SeqTrainingArguments(
#     output_dir="my_fine_tuned_t5_small_model",
#     evaluation_strategy="epoch",
#     learning_rate=2e-5,
#     per_device_train_batch_size=16,
#     per_device_eval_batch_size=16,
#     weight_decay=0.01,
#     save_total_limit=3,
#     num_train_epochs=4,
#     predict_with_generate=True,
#     fp16=True,
# )

# trainer = Seq2SeqTrainer(
#     model=model,
#     args=training_args,
#     train_dataset=encoded_ds['train'],
#     eval_dataset=encoded_ds["test"],
#     tokenizer=tokenizer,
#     data_collator=data_collator,
#     compute_metrics=compute_metrics,
# )
```

```
os.environ["PYTORCH_CUDA_ALLOC_CONF"] = "max_split_size_mb:1024"
```

```
batch_size = 1
num_of_epochs = 2
logging_steps = 2  #15
learning_rate=2e-5

model_name = 'pegasus-multi_news'
training_args = Seq2SeqTrainingArguments(output_dir=model_name,
                        num_train_epochs=num_of_epochs,
                        learning_rate=learning_rate,
                        per_device_train_batch_size=1,
                        per_device_eval_batch_size=1,
                        weight_decay=0.01,
                        save_strategy="epoch",
                        logging_strategy="epoch",
                        logging_first_step=True,
                        hub_strategy="checkpoint",
                        warmup_steps=50,
                        disable_tqdm=False,
                        logging_steps=logging_steps,
                        push_to_hub=True,
                        gradient_accumulation_steps=16,
                        log_level="error")

```

```
pip install rouge_score
```

```
Collecting rouge_score
  Downloading rouge_score-0.1.2.tar.gz (17 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: absl-py in /usr/local/lib/python3.10/dist-packages (from rouge_score) (1.4.0)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (from rouge_score) (3.8.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from rouge_score) (1.25.2)
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.10/dist-packages (from rouge_score) (1.16.0)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk->rouge_score) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk->rouge_score) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk->rouge_score) (2024.5.15)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk->rouge_score) (4.66.4)
```

```
Building wheels for collected packages: rouge_score
  Building wheel for rouge_score (setup.py) ... done
  Created wheel for rouge_score: filename=rouge_score-0.1.2-py3-none-any.whl size=24933 sha256=70379fd26b05d68adf92e884240d7aa26
  Stored in directory: /root/.cache/pip/wheels/5f/dd/89/461065a73be61a532ff8599a28e9beef17985c9e9c31e541b4
Successfully built rouge_score
Installing collected packages: rouge_score
Successfully installed rouge_score-0.1.2
```

```
rouge_metric = load_metric("rouge")
rouge_names = ["rouge1", "rouge2", "rougeL", "rougeLsum"]

def chunks(list_of_elements, batch_size):
    for i in range(0, len(list_of_elements), batch_size):
        yield list_of_elements[i : i + batch_size]

def compute_metrics(dataset, metric, model, tokenizer, column_text="Article", column_summary="Summary", batch_
    article_batches = list(chunks(dataset[column_text], batch_size))
    target_batches = list(chunks(dataset[column_summary], batch_size))

    for article_batch, target_batch in tqdm(
        zip(article_batches, target_batches), total=len(article_batches)):
        inputs = tokenizer(article_batch, max_length=max_input_length, truncation=True, padding="max_length",
        summaries = model.generate(input_ids=inputs["input_ids"].to(device),
                                   attention_mask=inputs["attention_mask"].to(device),
                                   length_penalty=0.8, num_beams=8, max_length=max_target_length)
        decoded_summaries = [tokenizer.decode(s, skip_special_tokens=True, clean_up_tokenization_spaces=True)
        decoded_summaries = [d.replace("<n>", " ") for d in decoded_summaries]
        metric.add_batch(predictions=decoded_summaries, references=target_batch)

    score = metric.compute()
    return score
```

```
<ipython-input-22-e5d9fb421053>:1: FutureWarning: load_metric is deprecated and will be removed in the next major version of data
  rouge_metric = load_metric("rouge")
```

Downloading builder script: 5.65k/? [00:00<00:00, 363kB/s]

```
The repository for rouge contains custom code which must be executed to correctly load the dataset. You can inspect the repositor
You can avoid this prompt in future by passing the argument `trust_remote_code=True`.

Do you wish to run the custom code? [y/N] y
```

```
seq2seq_data_collator = DataCollatorForSeq2Seq(tokenizer, model=model)
```

HGFACE

```
from huggingface_hub import login

login()
# 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
```

```
Token is valid (permission: fineGrained).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /root/.cache/huggingface/token
Login successful
```

```
import os

os.environ['HUGGINGFACE_HUB_TOKEN'] = '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'
```

```
from huggingface_hub import HfApi

api = HfApi()
user = api.whoami()
print(f"Logged in as: {user['name']}")

```

```
Start coding or generate with AI.
```

```
trainer = Seq2SeqTrainer(model=model,
                  args=training_args,
                  data_collator=seq2seq_data_collator,
                  train_dataset=converted_ds['train'],
                  eval_dataset=converted_ds['valid'],
                  tokenizer=tokenizer)
```

Train Model

```
import torch
import gc
torch.cuda.empty_cache()
```

```
torch.cuda.empty_cache()
gc.collect()
```

```
10960
```

```
import os
os.environ["PYTORCH_CUDA_ALLOC_CONF"] = "max_split_size_mb:1024"
```