bigpatent-preprocessing

May 13, 2023

[ ]:
```
import tensorflow as tf
print("Tensorflow version " + tf.__version__)

try:
  tpu = tf.distribute.cluster_resolver.TPUClusterResolver()  # TPU detection
  print('Running on TPU ', tpu.cluster_spec().as_dict()['worker'])
except ValueError:
  raise BaseException('ERROR: Not connected to a TPU runtime; please see the␣
 ↪previous cell in this notebook for instructions!')

tf.config.experimental_connect_to_cluster(tpu)
tf.tpu.experimental.initialize_tpu_system(tpu)
tpu_strategy = tf.distribute.TPUStrategy(tpu)
```

```
Tensorflow version 2.12.0
Running on TPU  ['10.24.134.98:8470']
```

0.1 Installing the Datasets

[ ]:
```
!pip install datasets
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting datasets
  Downloading datasets-2.12.0-py3-none-any.whl (474 kB)
                            474.6/474.6 kB
18.1 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.10/dist-packages (from datasets) (1.22.4)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.10/dist-
packages (from datasets) (9.0.0)
Collecting dill<0.3.7,>=0.3.0 (from datasets)
  Downloading dill-0.3.6-py3-none-any.whl (110 kB)
                            110.5/110.5 kB
16.3 MB/s eta 0:00:00
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-
packages (from datasets) (1.5.3)
Requirement already satisfied: requests>=2.19.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from datasets) (2.27.1)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (4.65.0)
Collecting xxhash (from datasets)
  Downloading
xxhash-3.2.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (212 kB)
                           212.5/212.5 kB
27.4 MB/s eta 0:00:00
Collecting multiprocess (from datasets)
  Downloading multiprocess-0.70.14-py310-none-any.whl (134 kB)
                           134.3/134.3 kB
17.5 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]>=2021.11.1 in
/usr/local/lib/python3.10/dist-packages (from datasets) (2023.4.0)
Collecting aiohttp (from datasets)
  Downloading
aiohttp-3.8.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.0
MB)
                            1.0/1.0 MB
25.4 MB/s eta 0:00:00
Collecting huggingface-hub<1.0.0,>=0.11.0 (from datasets)
  Downloading huggingface_hub-0.14.1-py3-none-any.whl (224 kB)
                           224.5/224.5 kB
28.7 MB/s eta 0:00:00
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from datasets) (23.1)
Collecting responses<0.19 (from datasets)
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (6.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets) (23.1.0)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (2.0.12)
Collecting multidict<7.0,>=4.5 (from aiohttp->datasets)
  Downloading
multidict-6.0.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (114
kB)
                           114.5/114.5 kB
15.7 MB/s eta 0:00:00
Collecting async-timeout<5.0,>=4.0.0a3 (from aiohttp->datasets)
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Collecting yarl<2.0,>=1.0 (from aiohttp->datasets)
  Downloading
yarl-1.9.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (268 kB)
                           268.8/268.8 kB
31.8 MB/s eta 0:00:00
Collecting frozenlist>=1.1.1 (from aiohttp->datasets)
```

```
  Downloading frozenlist-1.3.3-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_6
4.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (149 kB)
                           149.6/149.6 kB
19.9 MB/s eta 0:00:00
Collecting aiosignal>=1.1.2 (from aiohttp->datasets)
  Downloading aiosignal-1.3.1-py3-none-any.whl (7.6 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from huggingface-hub<1.0.0,>=0.11.0->datasets) (3.12.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0.0,>=0.11.0->datasets) (4.5.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0->datasets)
(2022.12.7)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests>=2.19.0->datasets) (3.4)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2022.7.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.1->pandas->datasets) (1.16.0)
Installing collected packages: xxhash, multidict, frozenlist, dill, async-
timeout, yarl, responses, multiprocess, huggingface-hub, aiosignal, aiohttp,
datasets
Successfully installed aiohttp-3.8.4 aiosignal-1.3.1 async-timeout-4.0.2
datasets-2.12.0 dill-0.3.6 frozenlist-1.3.3 huggingface-hub-0.14.1
multidict-6.0.4 multiprocess-0.70.14 responses-0.18.0 xxhash-3.2.0 yarl-1.9.2
```

0.2 Loading the Dataset

[ ]:
```
from datasets import load_dataset

dataset = load_dataset("big_patent", "d")
```

```
Downloading builder script:   0%|          | 0.00/5.50k [00:00<?, ?B/s]

Downloading metadata:   0%|          | 0.00/22.9k [00:00<?, ?B/s]

Downloading readme:   0%|          | 0.00/9.70k [00:00<?, ?B/s]

Downloading and preparing dataset big_patent/d to /root/.cache/huggingface/datas
ets/big_patent/d/2.1.2/bc8ec8bdf469c0da5fef04becd32bb3b0b34df0b0baa088ae1237628d
d7a9caa…

Downloading data files:   0%|          | 0/3 [00:00<?, ?it/s]
```

```
Downloading data:   0%|          | 0.00/9.13G [00:00<?, ?B/s]
Downloading data:   0%|          | 0.00/506M [00:00<?, ?B/s]
Downloading data:   0%|          | 0.00/508M [00:00<?, ?B/s]
Extracting data files:   0%|          | 0/3 [00:00<?, ?it/s]
Generating train split:   0%|          | 0/10164 [00:00<?, ? examples/s]
Generating validation split:   0%|          | 0/565 [00:00<?, ? examples/s]
Generating test split:   0%|          | 0/565 [00:00<?, ? examples/s]
Dataset big_patent downloaded and prepared to /root/.cache/huggingface/datasets/
big_patent/d/2.1.2/bc8ec8bdf469c0da5fef04becd32bb3b0b34df0b0baa088ae1237628dd7a9
caa. Subsequent calls will reuse this data.
  0%|          | 0/3 [00:00<?, ?it/s]
```

Choose the category 'd' - Textiles

0.3 DATASET DESCRIPTION

Train Data-10164, Test Data-565, Validation-565.

[ ]:
```
print(dataset)
```

```
DatasetDict({
    train: Dataset({
        features: ['description', 'abstract'],
        num_rows: 10164
    })
    validation: Dataset({
        features: ['description', 'abstract'],
        num_rows: 565
    })
    test: Dataset({
        features: ['description', 'abstract'],
        num_rows: 565
    })
})
```

[ ]:
```
train_data = dataset['train']
```

[ ]:
```
import pandas as pd
```

[ ]:
```
train_data = pd.DataFrame(train_data)
```

[ ]:
```
train_data= train_data.head(80)
```

[ ]:
```
train_data.shape
```

[ ]:
```
(80, 3)
```

0.4 Preprocessing steps:

Converting description column to lower case.

[ ]:
```
lower_description =train_data['description'].str.lower()
```

[ ]:
```
lower_description
```

[ ]:
```
0     background of the invention \n     this invent…
1     cross-reference to related application \n     …
2     this is a division of application ser. no. 922…
3     field of the invention \n     the present inve…
4     [0001]     this application claims the benefit…
                           …
75    cross-reference to related applications \n    …
76    field of the invention \n     the present inve…
77    background of the invention \n     the present…
78    cross related applications \n     this applica…
79    cross reference to related application \n     …
Name: description, Length: 80, dtype: object
```

1 Importing the necessary libraries

[ ]:
```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import WordNetLemmatizer

# Download necessary NLTK resources
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

def preprocess_text(text):
    # Tokenize the text into sentences
    sentences = sent_tokenize(text)

    # Remove stop words and lemmatize the words in each sentence
    stop_words = set(stopwords.words('english'))
    lemmatizer = WordNetLemmatizer()
    clean_sentences = []
    for sentence in sentences:
        words = word_tokenize(sentence)
        words = [lemmatizer.lemmatize(w) for w in words if w not in stop_words]
```

```
        clean_sentences.append(' '.join(words))

    # Join the cleaned sentences into a single string
    cleaned_text = ' '.join(clean_sentences)

    return cleaned_text
```

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

Using the pre-process function to a data frame.

[ ]:
```
train_data['clean_description']=lower_description.apply(preprocess_text)
```

[ ]:
```
train_data['clean_description']
```

[ ]:
```
0     background invention invention relates method …
1     cross-reference related application applicatio…
2     division application ser . . 922,344 , filed o…
3     field invention present invention relates weft…
4     [ 0001 ] application claim benefit korean appl…
                           …
75    cross-reference related application [ 0001 ] a…
76    field invention present invention relates open…
77    background invention present invention relates…
78    cross related application application division…
79    cross reference related application applicatio…
Name: clean_description, Length: 80, dtype: object
```

Performing tokenizing, tagging on the dataframe.

[ ]:
```
#function to tokenize text into words
def tokenize_text(text):
    words = word_tokenize(text)
    return words

#function to tag words with part-of-speech (POS) tags
def tag_words(words):
    tagged_words = nltk.pos_tag(words)
    return tagged_words

#function to tag words with part-of-speech (POS) tags
def extract_entities(text):
    entities = nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(text)))
```

```
    return entities
```

[ ]:
```
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data…
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
```

[ ]:
```
True
```

[ ]:
```
train_data['tokens'] = train_data['clean_description'].apply(tokenize_text)
```

Converting dataframe to CSV.

[ ]:
```
train_data.to_csv('clean_big_data.csv', index=False)
```

[ ]:
```
import pandas as pd
```

[ ]:
```
clean_data = pd.read_csv('./clean_big_data.csv')
```

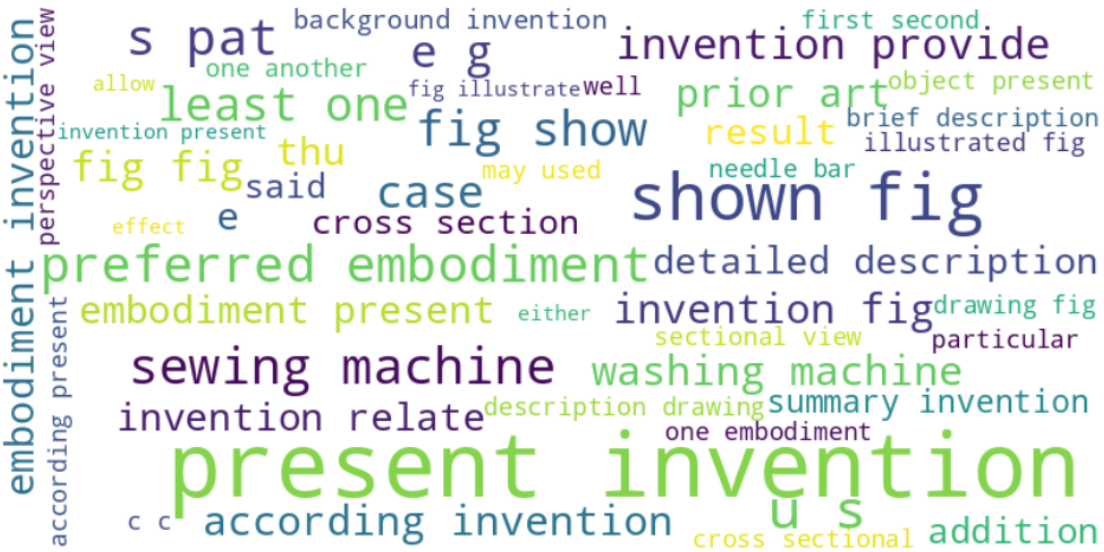
[ ]:
```
clean_data.head()
```

[ ]:
```
clean_text = ' '.join(clean_data['clean_description'])
```

Performing Cloud Operation on clean text.

[ ]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

[ ]:
```
wordcloud = WordCloud(width=800, height=400, max_words=50,␣
 ↪background_color='white').generate(clean_text)
# Display the word cloud
plt.figure(figsize=(12, 8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Noticed that words like present,invention,preferred,sewing machine,shown,fig,embodiment are the most repeteated words.

[ ]:
```
text_count = [len(sentence.split()) for sentence in clean_data.
↪clean_description]
```

Performing the average and maximum count operations on text and abstract columns

[ ]:
```
from numpy.ma.extras import average
print("Average word count of Description :",average(text_count))
print("Max word count in description :", max(text_count))
```

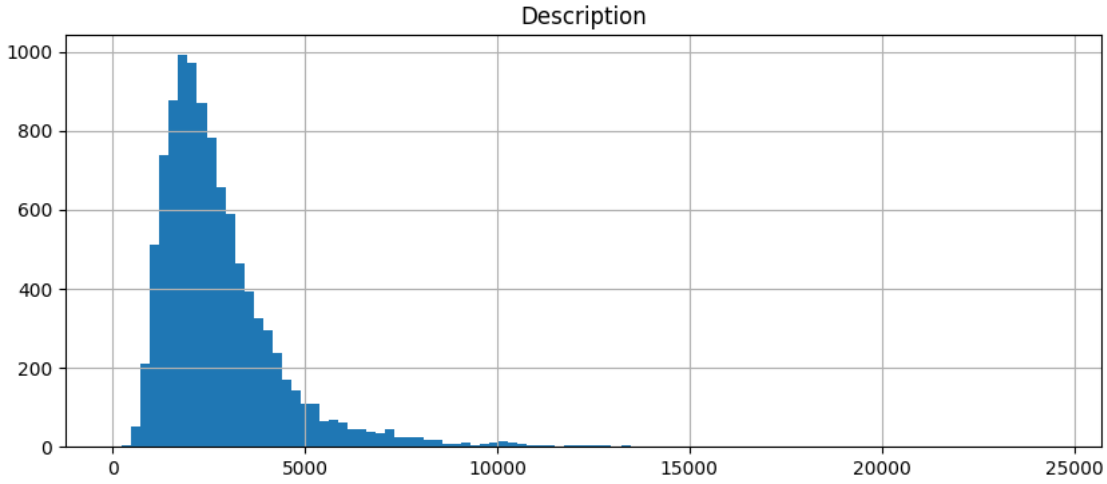
```
Average word count of Description : 2822.900137741047
Max word count in description : 24478
```

[ ]:
```
abstract_count = [len(sentence.split()) for sentence in clean_data.abstract]
```

[ ]:
```
print("Average word count of abstracts :",average(abstract_count))
print("Max word count in abstracts :", max(abstract_count))
```

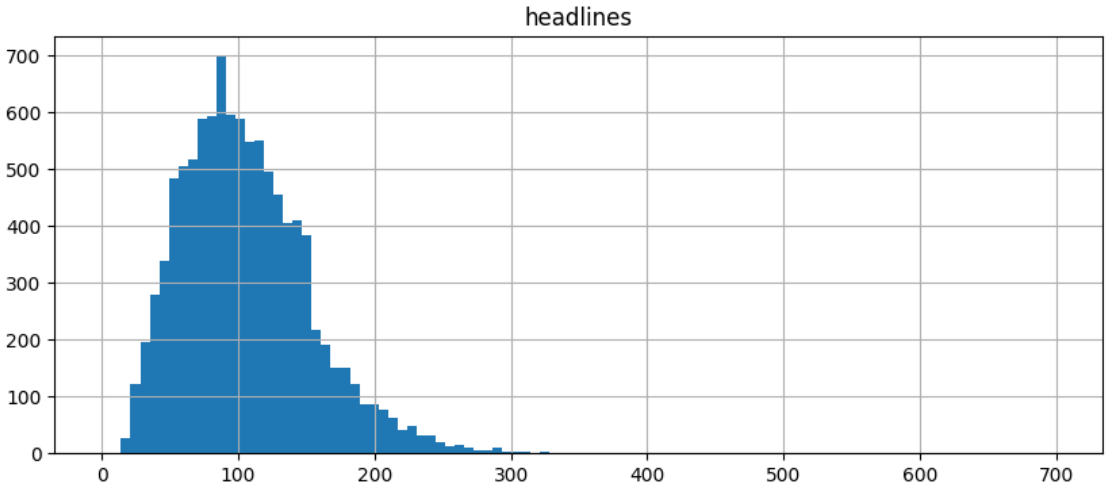
```
Average word count of abstracts : 105.05588351042897
Max word count in abstracts : 699
```

[ ]:
```
pd.DataFrame({'Description': text_count}).hist(bins=100, figsize=(10, 4),␣
↪range=[0, max(text_count)])
plt.show()
```



The patent descriptions are left-skewed.

[ ]:
```
pd.DataFrame({'headlines': abstract_count}).hist(bins=100, figsize=(10, 4),␣
  ↪range=[0, max(abstract_count)])
plt.show()
```



The patent abstract is normally distribute with few long abstracts.

[ ]:
```
clean_data.abstract= clean_data.abstract.apply(lambda x: f'_START_ {x} _END_')
```

[ ]:
```
start_token = 'sostok'
end_token = 'eostok'
clean_data.abstract = clean_data.abstract.apply(lambda x: f'{start_token} {x}␣
  ↪{end_token}')
```

[ ]:
```
# rare word analysis
def get_rare_word_percent(tokenizer, threshold):
    # threshold: if the word's occurrence is less than this then it's rare word

    count = 0
    total_count = 0
    frequency = 0
    total_frequency = 0

    for key, value in tokenizer.word_counts.items():
        total_count += 1
        total_frequency += value
        if value < threshold:
            count += 1
            frequency += value

    return {
        'percent': round((count / total_count) * 100, 2),
        'total_coverage': round(frequency / total_frequency * 100, 2),
        'count': count,
        'total_count': total_count
    }
```

[ ]:
```
df= clean_data[:10]
```

Printing the cleaned dataframe.

[ ]:
```
df
```

[ ]:
```
                                          description  \
0  BACKGROUND OF THE INVENTION \n     This invent…
1  CROSS-REFERENCE TO RELATED APPLICATION \n     …
2  This is a division of application Ser. No. 922…
3  FIELD OF THE INVENTION \n     The present inve…
4  [0001]     This application claims the benefit…
5  This application is a divisional of applicatio…
6  (This application claims the benefit of U.S. P…
7  TECHNICAL FIELD \n       [0001]    The present…
8  BACKGROUND OF THE INVENTION \n       [0001]   …
9  PRIOR APPLICATION \n     This application is a…

                                             abstract  \
0  sostok _START_ A method of forming fiber mixtu…
1  sostok _START_ The fibers of recycled paper ar…
2  sostok _START_ Non-woven, bias laid fabrics, w…
3  sostok _START_ Multiple-color air jet looms su…
4  sostok _START_ A method of performing a spinni…
```

```
5  sostok _START_ An air handler for collecting a…
6  sostok _START_ A device and method to implemen…
7  sostok _START_ The present invention introduce…
8  sostok _START_ Multiple groups of sensors are …
9  sostok _START_ The continuous digester is for …
```

```
                                clean_description
0  background invention invention relates method …
1  cross-reference related application applicatio…
2  division application ser . . 922,344 , filed o…
3  field invention present invention relates weft…
4  [ 0001 ] application claim benefit korean appl…
5  application divisional application ser . . 09/…
6  ( application claim benefit u.s. provisional a…
7  technical field [ 0001 ] present invention rel…
8  background invention [ 0001 ] present inventio…
9  prior application application u.s. national ph…
```

[ ]:
```
df = pd.read_csv('./clean_big_data.csv')
```

[ ]:
```
df= df[:10]
```

[ ]:
```
df= df['clean_description']
```

[ ]:
```
df
```

[ ]:
```
0    background invention invention relates method …
1    cross-reference related application applicatio…
2    division application ser . . 922,344 , filed o…
3    field invention present invention relates weft…
4    [ 0001 ] application claim benefit korean appl…
5    application divisional application ser . . 09/…
6    ( application claim benefit u.s. provisional a…
7    technical field [ 0001 ] present invention rel…
8    background invention [ 0001 ] present inventio…
9    prior application application u.s. national ph…
Name: clean_description, dtype: object
```

rrr-of-bigpatent-preprocessing

May 13, 2023

0.1 Installing Transformers

[ ]:
```
!pip install transformers
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting transformers
  Downloading transformers-4.29.0-py3-none-any.whl (7.1 MB)
                             7.1/7.1 MB
43.3 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers) (3.12.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.14.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from transformers) (1.22.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers) (6.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (2022.10.31)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers) (2.27.1)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1 (from transformers)
  Downloading
tokenizers-0.13.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(7.8 MB)
                             7.8/7.8 MB
92.8 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.10/dist-packages (from transformers) (4.65.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages
(from huggingface-hub<1.0,>=0.11.0->transformers) (2023.4.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.11.0->transformers) (4.5.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
```

```
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(2022.12.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers) (3.4)
Installing collected packages: tokenizers, transformers
Successfully installed tokenizers-0.13.3 transformers-4.29.0
```

Importing Tensorflow and Transformers

[ ]:
```
# Import required libraries
import tensorflow as tf
import  transformers
# Load tokenizer and model
```

Installing Sentencepiece

[ ]:
```
!pip install sentencepiece
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting sentencepiece
  Downloading
sentencepiece-0.1.99-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(1.3 MB)
                                    1.3/1.3 MB
16.2 MB/s eta 0:00:00
Installing collected packages: sentencepiece
Successfully installed sentencepiece-0.1.99
```

[ ]:
```
pip install --upgrade transformers
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-
packages (4.29.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers) (3.12.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.14.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-
packages (from transformers) (1.22.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers) (6.0)
```

```
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (2022.10.31)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers) (2.27.1)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.13.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-
packages (from transformers) (4.65.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages
(from huggingface-hub<1.0,>=0.11.0->transformers) (2023.4.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.11.0->transformers) (4.5.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(2022.12.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers) (3.4)
```

Initializing the Facebook/Bart-large-cnn Model

[ ]:
```
from transformers import BartTokenizer, BartForConditionalGeneration

model_name = 'facebook/bart-large-cnn'
model = BartForConditionalGeneration.from_pretrained(model_name)
tokenizer = BartTokenizer.from_pretrained(model_name)
```

```
Downloading (…)lve/main/config.json:   0%|          | 0.00/1.58k [00:00<?, ?B/s]
Downloading pytorch_model.bin:   0%|          | 0.00/1.63G [00:00<?, ?B/s]
Downloading (…)neration_config.json:   0%|          | 0.00/363 [00:00<?, ?B/s]
Downloading (…)olve/main/vocab.json:   0%|          | 0.00/899k [00:00<?, ?B/s]
Downloading (…)olve/main/merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
```

[ ]:
```
input_text = clean_data['clean_description'].tolist()[:10]
inputs = tokenizer(input_text, truncation=True, padding='longest',␣
↪max_length=1024, return_tensors='pt')
```

Printing the Summaries

[ ]:
```
for i in range(10):
    input_ids = inputs.input_ids[i][:512]  # Truncate to a maximum length of␣
↪512 tokens
```

```
    attention_mask = inputs.attention_mask[i][:512]  # Truncate to a maximum␣
↪length of 512 tokens
    input_shape = input_ids.shape  # Get the shape of input_ids
    input_ids = input_ids.unsqueeze(0)  # Add a batch dimension
    attention_mask = attention_mask.unsqueeze(0)  # Add a batch dimension
    outputs = model.generate(input_ids, attention_mask=attention_mask,␣
↪max_length=512, num_beams=4)
    summary = tokenizer.decode(outputs[0], skip_special_tokens=True)
    print(f"Summary for row {i+1}: {summary}\n")
```

Summary for row 1: German laid-open applications (offenlegungsschriften) nos. 1,685,596 (to which corresponds u.s. pat. no. 3,577,599) and 2,063,415 disclose a method. according to the method, such an amount of fiber material is removed each time from a plurality of fiber bales of one fiber type until a quantity corresponding to the intended proportion is reached.

Summary for row 2: This application is a continuation-in-part of the co-pending application of the same inventor, titled "method and apparatus for cleaning fibers. This invention relates generally to the cleaning of fibrous material, and is more particularly concerned with the removal of ink and other contaminants from fibers used in paper making and the like.

Summary for row 3: This is a division of application ser. no. 922,344, filed oct. 23, 1986 now u.s. pat.No. 4,877,470. The invention is directed to method and apparatus for forming bias laid, non-woven fabrics. The yarns in at least two of the layers of fabric are laid at an angle of from 30° to 150° to the long axis of the fabric.

Summary for row 4: The present invention relates to weft insertion in a multiple-color air jet loom. It provides a method and an apparatus for inserting different weft threads in and through warp sheds. The weft carrying force remains constant at all times for different wefts threads, and that the speed at which the weftthreads are taken through the warp sheds varies from one weft type to another.

Summary for row 5: This application claims the benefit of korean applications no. p2003-51511 filed on jul. 25, 2003. The invention relates to a method of performing a spinning operation for a washing machine. A microprocessor determines a load weight of wet clothes to measure spinning operation parameters, which helps balance the load in the tub.

Summary for row 6: Meltblowing and spunbond processes are commonly employed to manufacture nonwoven webs and laminates. Much of the air is heated and moving at very high velocities. Without properly collecting and disposing of the process air, the air would likely disturb personnel working around the manufacturing apparatus and other nearby equipment.

```
Summary for row 7: The invention is related to the field of converting a
standard manually controlled valve into an electronically controlled automatic
valve. It is also related to protecting real property against damage or excess
water usage when outdoor spigots (valves) are left on or hoses break. When used
this way, the invention will serve to conserve water.

Summary for row 8: Bast fabrics have gained more and more popularity with
people, as they are low electrostatic and have the antibacterial speciality and
great absorbency. However, bast fibre, especially jute fibres include great
amount of lignin, is more rigid and brittle than cotton fibres. The end breakage
rate may reach up to 300-400 times per hour for each machine averagely.

Summary for row 9: The invention relates generally to nip presses used to exert
pressing forces on moving webs for the formation of, for example, paper, textile
material, plastic foil and other related materials. The invention provides a
method and apparatus for measuring and removing rotational variability from the
nip pressure profile of the covered roll.

Summary for row 10: This application is a u.s. national phase application that
is based on and claims priority from international application no.
pct/se2011/050075, filed 25 jan. 2011. The present invention concerns a
continuous digester that has a bottom scraper equipped with draining apparatus.
```

Length of the text in Clean Description column

[ ]:
```
sum = clean_data['clean_description'][0]
len(sum)
```

[ ]:
```
25793
```

Installing the Rogue metrics for evaluation

[ ]:
```
!pip install rouge
from rouge import Rouge

# Evaluation dataset with input data and reference summaries
eval_dataset = [
    {
        'input': 'summary',
        'reference_summary': clean_data['abstract'][:10]
    },
]
generated_summaries = []
for data in eval_dataset:
    input_ids = tokenizer.encode(data['input'], truncation=True,␣
 ↪max_length=512, padding='max_length', return_tensors='pt')
    outputs = model.generate(input_ids, max_length=512, num_beams=4)
```

```
    summary = tokenizer.decode(outputs[0], skip_special_tokens=True)
    generated_summaries.append(summary)

# Extract reference summaries from the evaluation dataset
reference_summaries = [data['reference_summary'] for data in eval_dataset]

# Convert reference summaries to strings
reference_summaries = [summary.tolist()[0] for summary in reference_summaries]

# Calculate ROUGE scores
rouge = Rouge()
scores = rouge.get_scores(generated_summaries, reference_summaries, avg=True)

print(f"ROUGE-1: {scores['rouge-1']['f']}")
print(f"ROUGE-2: {scores['rouge-2']['f']}")
print(f"ROUGE-L: {scores['rouge-l']['f']}")
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting rouge
  Downloading rouge-1.0.1-py3-none-any.whl (13 kB)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from rouge) (1.16.0)
Installing collected packages: rouge
Successfully installed rouge-1.0.1
ROUGE-1: 0.12244897480008349
ROUGE-2: 0.012903221740688066
ROUGE-L: 0.12244897480008349
```

Observed Rogue scores are o.12,0.012,0.12

bert

May 13, 2023

[2]:
```
from tensorflow.keras.preprocessing.sequence import pad_sequences
```

[3]:
```
import numpy as np
import pandas as pd
from keras.preprocessing.text import Tokenizer
from keras.models import Model
from keras.layers import Input, LSTM, Dense, Embedding
```

[8]:
```
!pip install rouge
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting rouge
  Downloading rouge-1.0.1-py3-none-any.whl (13 kB)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from rouge) (1.16.0)
Installing collected packages: rouge
Successfully installed rouge-1.0.1
```

[24]:
```
!pip install tensorflow==2.8
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting tensorflow==2.8
  Downloading tensorflow-2.8.0-cp310-cp310-manylinux2010_x86_64.whl (497.6 MB)
                           497.6/497.6
MB 3.1 MB/s eta 0:00:00
Requirement already satisfied: absl-py>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (1.4.0)
Requirement already satisfied: astunparse>=1.6.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (1.6.3)
Requirement already satisfied: flatbuffers>=1.12 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (23.3.3)
Requirement already satisfied: gast>=0.2.1 in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (0.4.0)
Requirement already satisfied: google-pasta>=0.1.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (0.2.0)
```

```
Requirement already satisfied: h5py>=2.9.0 in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (3.8.0)
Collecting keras-preprocessing>=1.1.1 (from tensorflow==2.8)
  Downloading Keras_Preprocessing-1.1.2-py2.py3-none-any.whl (42 kB)
                             42.6/42.6 kB
4.4 MB/s eta 0:00:00
Requirement already satisfied: libclang>=9.0.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (16.0.0)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (1.22.4)
Requirement already satisfied: opt-einsum>=2.3.2 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (3.3.0)
Requirement already satisfied: protobuf>=3.9.2 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (3.20.3)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (67.7.2)
Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (1.16.0)
Requirement already satisfied: termcolor>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (2.3.0)
Requirement already satisfied: typing-extensions>=3.6.6 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (4.5.0)
Requirement already satisfied: wrapt>=1.11.0 in /usr/local/lib/python3.10/dist-
packages (from tensorflow==2.8) (1.14.1)
Collecting tensorboard<2.9,>=2.8 (from tensorflow==2.8)
  Downloading tensorboard-2.8.0-py3-none-any.whl (5.8 MB)
                             5.8/5.8 MB
64.0 MB/s eta 0:00:00
Collecting tf-estimator-nightly==2.8.0.dev2021122109 (from
tensorflow==2.8)
  Downloading tf_estimator_nightly-2.8.0.dev2021122109-py2.py3-none-any.whl (462
kB)
                            462.5/462.5 kB
39.8 MB/s eta 0:00:00
Collecting keras<2.9,>=2.8.0rc0 (from tensorflow==2.8)
  Downloading keras-2.8.0-py2.py3-none-any.whl (1.4 MB)
                             1.4/1.4 MB
75.1 MB/s eta 0:00:00
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (0.32.0)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in
/usr/local/lib/python3.10/dist-packages (from tensorflow==2.8) (1.54.0)
Requirement already satisfied: wheel<1.0,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from
astunparse>=1.6.0->tensorflow==2.8) (0.40.0)
Requirement already satisfied: google-auth<3,>=1.6.3 in
/usr/local/lib/python3.10/dist-packages (from
tensorboard<2.9,>=2.8->tensorflow==2.8) (2.17.3)
```

```
Collecting google-auth-oauthlib<0.5,>=0.4.1 (from
tensorboard<2.9,>=2.8->tensorflow==2.8)
  Downloading google_auth_oauthlib-0.4.6-py2.py3-none-any.whl (18 kB)
Requirement already satisfied: markdown>=2.6.8 in
/usr/local/lib/python3.10/dist-packages (from
tensorboard<2.9,>=2.8->tensorflow==2.8) (3.4.3)
Requirement already satisfied: requests<3,>=2.21.0 in
/usr/local/lib/python3.10/dist-packages (from
tensorboard<2.9,>=2.8->tensorflow==2.8) (2.27.1)
Collecting tensorboard-data-server<0.7.0,>=0.6.0 (from
tensorboard<2.9,>=2.8->tensorflow==2.8)
  Downloading tensorboard_data_server-0.6.1-py3-none-manylinux2010_x86_64.whl
(4.9 MB)
                              4.9/4.9 MB
103.4 MB/s eta 0:00:00
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in
/usr/local/lib/python3.10/dist-packages (from
tensorboard<2.9,>=2.8->tensorflow==2.8) (1.8.1)
Requirement already satisfied: werkzeug>=0.11.15 in
/usr/local/lib/python3.10/dist-packages (from
tensorboard<2.9,>=2.8->tensorflow==2.8) (2.3.0)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from google-
auth<3,>=1.6.3->tensorboard<2.9,>=2.8->tensorflow==2.8) (5.3.0)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/usr/local/lib/python3.10/dist-packages (from google-
auth<3,>=1.6.3->tensorboard<2.9,>=2.8->tensorflow==2.8) (0.3.0)
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.10/dist-
packages (from google-auth<3,>=1.6.3->tensorboard<2.9,>=2.8->tensorflow==2.8)
(4.9)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
/usr/local/lib/python3.10/dist-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard<2.9,>=2.8->tensorflow==2.8) (1.3.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.21.0->tensorboard<2.9,>=2.8->tensorflow==2.8) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.21.0->tensorboard<2.9,>=2.8->tensorflow==2.8) (2022.12.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from
requests<3,>=2.21.0->tensorboard<2.9,>=2.8->tensorflow==2.8) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests<3,>=2.21.0->tensorboard<2.9,>=2.8->tensorflow==2.8)
(3.4)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/usr/local/lib/python3.10/dist-packages (from
werkzeug>=0.11.15->tensorboard<2.9,>=2.8->tensorflow==2.8) (2.1.2)
```

```
Requirement already satisfied: pyasn1<0.6.0,>=0.4.6 in
/usr/local/lib/python3.10/dist-packages (from pyasn1-modules>=0.2.1->google-
auth<3,>=1.6.3->tensorboard<2.9,>=2.8->tensorflow==2.8) (0.5.0)
Requirement already satisfied: oauthlib>=3.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests-oauthlib>=0.7.0->google-
auth-oauthlib<0.5,>=0.4.1->tensorboard<2.9,>=2.8->tensorflow==2.8) (3.2.2)
Installing collected packages: tf-estimator-nightly, keras, tensorboard-data-
server, keras-preprocessing, google-auth-oauthlib, tensorboard, tensorflow
  Attempting uninstall: keras
    Found existing installation: keras 2.12.0
    Uninstalling keras-2.12.0:
      Successfully uninstalled keras-2.12.0
  Attempting uninstall: tensorboard-data-server
    Found existing installation: tensorboard-data-server 0.7.0
    Uninstalling tensorboard-data-server-0.7.0:
      Successfully uninstalled tensorboard-data-server-0.7.0
  Attempting uninstall: google-auth-oauthlib
    Found existing installation: google-auth-oauthlib 1.0.0
    Uninstalling google-auth-oauthlib-1.0.0:
      Successfully uninstalled google-auth-oauthlib-1.0.0
  Attempting uninstall: tensorboard
    Found existing installation: tensorboard 2.12.2
    Uninstalling tensorboard-2.12.2:
      Successfully uninstalled tensorboard-2.12.2
  Attempting uninstall: tensorflow
    Found existing installation: tensorflow 2.12.0
    Uninstalling tensorflow-2.12.0:
      Successfully uninstalled tensorflow-2.12.0
Successfully installed google-auth-oauthlib-0.4.6 keras-2.8.0 keras-
preprocessing-1.1.2 tensorboard-2.8.0 tensorboard-data-server-0.6.1
tensorflow-2.8.0 tf-estimator-nightly-2.8.0.dev2021122109
```

[52]:
```
df= pd.read_csv('./output (1).csv')
```

[25]:
```
X= df['clean_description']
y=df['abstract']
```

[41]:
```
max_length = 512
df['truncated_text'] = df['clean_description'].str.slice(0, max_length)
```

[39]:
```
from transformers import BertTokenizer, BertLMHeadModel

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertLMHeadModel.from_pretrained('bert-base-uncased')

def bert_summarize(text):
```

```
    inputs = tokenizer.encode_plus(text, add_special_tokens=True,␣
↪return_tensors='pt')
    input_ids = inputs['input_ids'].to(model.device)
    attention_mask = inputs['attention_mask'].to(model.device)

    output = model.generate(input_ids=input_ids, attention_mask=attention_mask,␣
↪max_new_tokens=100)
    summary = tokenizer.decode(output[0], skip_special_tokens=True)

    return summary
```

```
If you want to use `BertLMHeadModel` as a standalone, add `is_decoder=True.`
Some weights of the model checkpoint at bert-base-uncased were not used when
initializing BertLMHeadModel: ['cls.seq_relationship.bias',
'cls.seq_relationship.weight']
- This IS expected if you are initializing BertLMHeadModel from the checkpoint
of a model trained on another task or with another architecture (e.g.
initializing a BertForSequenceClassification model from a BertForPreTraining
model).
- This IS NOT expected if you are initializing BertLMHeadModel from the
checkpoint of a model that you expect to be exactly identical (initializing a
BertForSequenceClassification model from a BertForSequenceClassification model).
```

[42]:
```
df['summary'] = df['truncated_text'].apply(bert_summarize)
```

[47]:
```
from rouge import Rouge

def calculate_rouge(bert_summary, gt_summary):
    rouge = Rouge()
    scores = rouge.get_scores(bert_summary, gt_summary, avg= True)
    return scores[0]
```

[51]:
```
from rouge import Rouge

# Initialize Rouge
rouge = Rouge()

# Create empty lists to store the rouge scores
rouge1_scores = []
rouge2_scores = []
rougel_scores = []

# Iterate over the DataFrame rows and calculate the rouge scores
for index, row in df.iterrows():
    reference_summary = row['clean_description']
    generated_summary = row['summary']
    scores = rouge.get_scores(generated_summary, reference_summary)
```

```
    rouge1_scores.append(scores[0]['rouge-1']['f'])
    rouge2_scores.append(scores[0]['rouge-2']['f'])
    rougel_scores.append(scores[0]['rouge-l']['f'])

# Calculate the average rouge scores
avg_rouge1_score = sum(rouge1_scores) / len(rouge1_scores)
avg_rouge2_score = sum(rouge2_scores) / len(rouge2_scores)
avg_rougel_score = sum(rougel_scores) / len(rougel_scores)

print(f'Average Rouge-1 Score: {avg_rouge1_score:.4f}')
```

```
Average Rouge-1 Score: 0.1301
```

[ ]:

gpt2-updated

May 13, 2023

[1]:
```
!pip install transformers
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting transformers
  Downloading transformers-4.29.1-py3-none-any.whl (7.1 MB)
                             7.1/7.1 MB
32.2 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers) (3.12.0)
Collecting huggingface-hub<1.0,>=0.14.1 (from transformers)
  Downloading huggingface_hub-0.14.1-py3-none-any.whl (224 kB)
                            224.5/224.5 kB
23.7 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (1.22.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
packages (from transformers) (6.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (2022.10.31)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers) (2.27.1)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1 (from transformers)
  Downloading
tokenizers-0.13.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(7.8 MB)
                             7.8/7.8 MB
100.7 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.10/dist-packages (from transformers) (4.65.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages
(from huggingface-hub<1.0,>=0.14.1->transformers) (2023.4.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-
hub<1.0,>=0.14.1->transformers) (4.5.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
```

```
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(2022.12.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers) (3.4)
Installing collected packages: tokenizers, huggingface-hub, transformers
Successfully installed huggingface-hub-0.14.1 tokenizers-0.13.3
transformers-4.29.1
```

[17]:
```python
import pandas as pd
import torch
from transformers import GPT2Tokenizer, GPT2LMHeadModel

df= pd.read_csv("./output (1).csv")

# Load pre-trained GPT2 model and tokenizer
tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
model = GPT2LMHeadModel.from_pretrained('gpt2')
```

[11]:
```python
def gpt_summarize(text):
    # Tokenize the text
    inputs = tokenizer.encode_plus(text, return_tensors='pt', max_length=255,␣
 ↪truncation=True)
    input_ids = inputs['input_ids'].to(device)
    attention_mask = inputs['attention_mask'].to(device)

    # Generate summary
    summary_ids = model.generate(input_ids, attention_mask=attention_mask,␣
 ↪max_length=256, num_beams=5, early_stopping=True)

    # Decode the summary
    summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
    return summary
```

[12]:
```python
# Apply the gpt_summarize function to the 'description' column
df['summary'] = df['clean_description'].apply(gpt_summarize)
```

```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```

```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```

[14]:
```
!pip install rouge
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting rouge
  Downloading rouge-1.0.1-py3-none-any.whl (13 kB)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from rouge) (1.16.0)
Installing collected packages: rouge
Successfully installed rouge-1.0.1
```

[16]:
```
from rouge import Rouge

# Initialize Rouge
rouge = Rouge()

# Create empty lists to store the rouge scores
rouge1_scores = []
rouge2_scores = []
rougel_scores = []

# Iterate over the DataFrame rows and calculate the rouge scores
for index, row in df.iterrows():
    reference_summary = row['clean_description']
    generated_summary = row['summary']
    scores = rouge.get_scores(generated_summary, reference_summary)
    rouge1_scores.append(scores[0]['rouge-1']['f'])
    rouge2_scores.append(scores[0]['rouge-2']['f'])
    rougel_scores.append(scores[0]['rouge-l']['f'])

# Calculate the average rouge scores
avg_rouge1_score = sum(rouge1_scores) / len(rouge1_scores)
avg_rouge2_score = sum(rouge2_scores) / len(rouge2_scores)
avg_rougel_score = sum(rougel_scores) / len(rougel_scores)

print(f'Average Rouge-1 Score: {avg_rouge1_score:.4f}')
print(f'Average Rouge-2 Score: {avg_rouge2_score:.4f}')
print(f'Average Rouge-L Score: {avg_rougel_score:.4f}')
```

```
Average Rouge-1 Score: 0.2420
Average Rouge-2 Score: 0.1441
Average Rouge-L Score: 0.2420
```

```
import pandas as pd
import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
from string import punctuation
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
df=pd.read_csv('./output (1).csv')

df.head()
```

```
                                          description  \
0  BACKGROUND OF THE INVENTION \n     This invent...
1  CROSS-REFERENCE TO RELATED APPLICATION \n     ...
2  This is a division of application Ser. No. 922...
3  FIELD OF THE INVENTION \n     The present inve...
4  [0001]     This application claims the benefit...

                                             abstract  \
0  A method of forming fiber mixtures from differ...
1  The fibers of recycled paper are cleaned by ag...
2  Non-woven, bias laid fabrics, where the variou...
3  Multiple-color air jet looms successively inse...
4  A method of performing a spinning operation of...

                                  clean_description
0  background of the invention \n     this invent...
1  cross-reference to related application \n     ...
2  this is a division of application ser. no. 922...
3  field of the invention \n     the present inve...
4  [0001]     this application claims the benefit...
```

```
# Calculate the TF-IDF scores for the cleaned text
tfidf_vectorizer = TfidfVectorizer(use_idf=True)
tfidf_matrix = tfidf_vectorizer.fit_transform(df["clean_description"])

# Calculate cosine similarity between sentences
sentence_similarity_matrix = cosine_similarity(tfidf_matrix)

import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.

True
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from nltk.tokenize import sent_tokenize

# Define the TextRank algorithm for extractive summarization
```

```
def textrank_summarize(text, num_sentences):
    # Tokenize sentences
    sentences = sent_tokenize(text)

    # Calculate the TF-IDF scores for the sentences
    tfidf_vectorizer = TfidfVectorizer(stop_words='english')
    sentence_tfidf = tfidf_vectorizer.fit_transform(sentences)

    # Calculate the sentence similarity matrix
    sentence_similarity_matrix = cosine_similarity(sentence_tfidf)

    # Calculate the sentence scores using the TextRank algorithm
    sentence_scores = []
    for i, sentence in enumerate(sentences):
        score = 0
        for j, other_sentence in enumerate(sentences):
            if i != j:
                score += sentence_similarity_matrix[i][j]
        sentence_scores.append(score)

    # Sort the sentences by score and return the top num_sentences
sentences
    top_sentences_indices = sorted(range(len(sentence_scores)),
key=lambda i: sentence_scores[i], reverse=True)[:num_sentences]
    top_sentences = [sentences[i] for i in top_sentences_indices]

    # Join the top sentences back into a single string and return the
summary
    summary = " ".join(top_sentences)
    return summary

# Generate summaries for each row in the data frame
df["summary"] = df["clean_description"].apply(lambda x:
textrank_summarize(x, num_sentences=3))

# Print the results
print(df[["clean_description", "summary"]])
```

```
                                  clean_description  \
0  background of the invention \n     this invent...
1  cross-reference to related application \n     ...
2  this is a division of application ser. no. 922...
3  field of the invention \n     the present inve...
4  [0001]     this application claims the benefit...
5  this application is a divisional of applicatio...
6  (this application claims the benefit of u.s. p...
7  technical field \n       [0001]    the present...
8  background of the invention \n       [0001]   ...
9  prior application \n     this application is a...
```

```
                                                summary
0  for the fiber removal during the first pass, t...
1  as a result, it will be understood that the wa...
2  thus, for example, if the yarn carrying means ...
3  summary of the invention \n     it is an objec...
4  the method further includes the steps of measu...
5  more importantly and as will be discussed in g...
6  figure three, one preferred embodiment of the ...
7  example 15 \n       [0055]    an experiment is...
8  again, each sensor of the first set has corres...
9  2  shows a first embodiment of the cone divert...
```

```
df.head()
```

```
                                           description  \
0  BACKGROUND OF THE INVENTION \n     This invent...
1  CROSS-REFERENCE TO RELATED APPLICATION \n     ...
2  This is a division of application Ser. No. 922...
3  FIELD OF THE INVENTION \n     The present inve...
4  [0001]     This application claims the benefit...

                                             abstract  \
0  A method of forming fiber mixtures from differ...
1  The fibers of recycled paper are cleaned by ag...
2  Non-woven, bias laid fabrics, where the variou...
3  Multiple-color air jet looms successively inse...
4  A method of performing a spinning operation of...

                                    clean_description  \
0  background of the invention \n     this invent...
1  cross-reference to related application \n     ...
2  this is a division of application ser. no. 922...
3  field of the invention \n     the present inve...
4  [0001]     this application claims the benefit...

                                                summary
0  for the fiber removal during the first pass, t...
1  as a result, it will be understood that the wa...
2  thus, for example, if the yarn carrying means ...
3  summary of the invention \n     it is an objec...
4  the method further includes the steps of measu...
```

```
!pip install rouge
```

```
Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Collecting rouge
  Downloading rouge-1.0.1-py3-none-any.whl (13 kB)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-
packages (from rouge) (1.16.0)
```

```
Installing collected packages: rouge
Successfully installed rouge-1.0.1
```

```
from rouge import Rouge
rouge = Rouge()
# Calculate ROUGE-1 scores for each row in the data frame
scores = df.apply(lambda row: rouge.get_scores(row["summary"],
row["abstract"])[0]["rouge-1"]["f"], axis=1)

# Print the average ROUGE-1 score
print("Average ROUGE-1 score: {:.2f}".format(scores.mean()))
```

```
Average ROUGE-1 score: 0.36
```

```
!pip install sumy

Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Collecting sumy
  Downloading sumy-0.11.0-py2.py3-none-any.whl (97 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 97.3/97.3 kB 5.5 MB/s eta
0:00:00
 sumy)
  Downloading docopt-0.6.2.tar.gz (25 kB)
  Preparing metadata (setup.py) ...  sumy)
  Downloading breadability-0.1.20.tar.gz (32 kB)
  Preparing metadata (setup.py) ... ent already satisfied:
requests>=2.7.0 in /usr/local/lib/python3.10/dist-packages (from sumy)
(2.27.1)
Collecting pycountry>=18.2.23 (from sumy)
  Downloading pycountry-22.3.5.tar.gz (10.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.1/10.1 MB 65.0 MB/s eta
0:00:00
ents to build wheel ... etadata (pyproject.toml) ... ent already
satisfied: nltk>=3.0.2 in /usr/local/lib/python3.10/dist-packages
(from sumy) (3.8.1)
Requirement already satisfied: chardet in
/usr/local/lib/python3.10/dist-packages (from breadability>=0.1.20-
>sumy) (4.0.0)
Requirement already satisfied: lxml>=2.0 in
/usr/local/lib/python3.10/dist-packages (from breadability>=0.1.20-
>sumy) (4.9.2)
Requirement already satisfied: click in
/usr/local/lib/python3.10/dist-packages (from nltk>=3.0.2->sumy)
(8.1.3)
Requirement already satisfied: joblib in
/usr/local/lib/python3.10/dist-packages (from nltk>=3.0.2->sumy)
(1.2.0)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.10/dist-packages (from nltk>=3.0.2->sumy)
(2022.10.31)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from nltk>=3.0.2->sumy) (4.65.0)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from pycountry>=18.2.23-
>sumy) (67.7.2)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.7.0->sumy)
(1.26.15)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.7.0->sumy)
(2022.12.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.7.0->sumy)
```

```
(2.0.12)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.7.0->sumy)
(3.4)
Building wheels for collected packages: breadability, docopt,
pycountry
  Building wheel for breadability (setup.py) ... e=breadability-
0.1.20-py2.py3-none-any.whl size=21696
sha256=9f69a6565dfd193a9391c04180fdecb929abd2a3918adc938d786ff6bfdeaa2
4
  Stored in directory:
/root/.cache/pip/wheels/64/22/90/b84fcc30e16598db20a0d41340616dbf9b1e8
2bbcc627b0b33
  Building wheel for docopt (setup.py) ... e=docopt-0.6.2-py2.py3-
none-any.whl size=13707
sha256=9cb8f266c4f81f80e6c0e579cf008e6cdd94c7110270caa86a9b66670938e49
9
  Stored in directory:
/root/.cache/pip/wheels/fc/ab/d4/5da2067ac95b36618c629a5f93f8094257005
06f72c9732fac
  Building wheel for pycountry (pyproject.toml) ... e=pycountry-
22.3.5-py2.py3-none-any.whl size=10681832
sha256=069bdbc196289092bce0cbab98484dad4c4dfd95e2999ed1f38a27661f9d214
4
  Stored in directory:
/root/.cache/pip/wheels/03/57/cc/290c5252ec97a6d78d36479a3c5e5ecc76318
afcb241ad9dbe
Successfully built breadability docopt pycountry
Installing collected packages: docopt, pycountry, breadability, sumy
Successfully installed breadability-0.1.20 docopt-0.6.2 pycountry-
22.3.5 sumy-0.11.0
```

```
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.lsa import LsaSummarizer
import pandas as pd

df=pd.read_csv('./output (1).csv')

import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.

True
```

```
# Define the number of sentences to include in the summary
num_sentences = 3
```

```python
# Initialize the LSA summarizer
summarizer = LsaSummarizer()

# Define a function to generate summaries for a given text
def lsa_summarize(text):
    # Parse the text into sentences
    parser = PlaintextParser.from_string(text, Tokenizer("english"))
    sentences = parser.document.sentences

    # Summarize the text using LSA
    summary = []
    for sentence in summarizer(parser.document, num_sentences):
        summary.append(str(sentence))

    # Join the summary sentences into a single string and return
    return " ".join(summary)

# Generate summaries for each row in the data frame
df["summary"] = df["clean_description"].apply(lsa_summarize)

# Print the results
print(df[["clean_description", "summary"]])
```

```
                                clean_description  \
0  background of the invention \n     this invent...
1  cross-reference to related application \n     ...
2  this is a division of application ser. no. 922...
3  field of the invention \n     the present inve...
4  [0001]     this application claims the benefit...
5  this application is a divisional of applicatio...
6  (this application claims the benefit of u.s. p...
7  technical field \n       [0001]    the present...
8  background of the invention \n       [0001]   ...
9  prior application \n     this application is a...

                                             summary
0  as a result, it is unavoidable that in the ind...
1  it is common to use sedimentation, centrifugin...
2  for many modern usages, particularly in areas ...
3  therefore, a certain amount of air under press...
4  however, it is very likely that some wet cloth...
5  moreover, the dampers must be readjusted each ...
6  the automatic valve can be controlled to turn ...
7  generally, bast fibres need to be humidified f...
8  [0002]    nipped rolls are used in a vast numb...
9  3  shows a second embodiment of the cone diver...
```

```
!pip install rouge
```

```
Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
```

```
Collecting rouge
  Downloading rouge-1.0.1-py3-none-any.whl (13 kB)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-
packages (from rouge) (1.16.0)
Installing collected packages: rouge
Successfully installed rouge-1.0.1
```

```python
from rouge import Rouge
rouge = Rouge()
# Calculate ROUGE-1 scores for each row in the data frame
scores = df.apply(lambda row: rouge.get_scores(row["summary"],
row["abstract"])[0]["rouge-1"]["f"], axis=1)

# Print the average ROUGE-1 score
print("Average ROUGE-1 score: {:.2f}".format(scores.mean()))
```

```
Average ROUGE-1 score: 0.25
```