Amazon SageMaker Studio File Edit View Run Kernel Git Tabs Settings Help Feedback

/

Name	Last Modified
balanced	16 minutes ago
images	an hour ago
src	33 minutes ago
C2_W1_Assignment.ipynb	a minute ago
feature_store_export.tsv	9 minutes ago

Terminal 1 | C2_W1_Assignment.ipynb | prepare_data.py

Markdown | 2 vCPU + 4 GiB | Python 3 (Data Science) | Share

Feature transformation with Amazon SageMaker processing job and Feature Store

Introduction

In this lab you will start with the raw Women's Clothing Reviews dataset and prepare it to train a BERT-based natural language processing (NLP) model. The model will be used to classify customer reviews into positive (1), neutral (0) and negative (-1) sentiment.

You will convert the original review text into machine-readable features used by BERT. To perform the required feature transformation you will configure an Amazon SageMaker processing job, which will be running a custom Python script.

Table of Contents

[1]:
```
# please ignore warning messages during the installation
!pip install --disable-pip-version-check -q sagemaker==2.35.0
!conda install -q -y pytorch==1.6.0 -c pytorch
!pip install --disable-pip-version-check -q transformers==3.5.1
```

```
/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

## Package Plan ##

  environment location: /opt/conda

  added / updated specs:
    - pytorch==1.6.0


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    ca-certificates-2021.7.5   |       h06a4308_1         113 KB
    certifi-2021.5.30          |   py37h06a4308_0         139 KB
    cudatoolkit-10.2.89        |       hfd86e86_1       365.1 MB
    ninja-1.10.2               |       hff7bd54_1         1.4 MB
    pytorch-1.6.0              |py3.7_cuda10.2.89_cudnn7.6.5_0       537.7 MB  pytorch
    ------------------------------------------------------------
                                           Total:       904.5 MB

The following NEW packages will be INSTALLED:

  cudatoolkit        pkgs/main/linux-64::cudatoolkit-10.2.89-hfd86e86_1
  ninja              pkgs/main/linux-64::ninja-1.10.2-hff7bd54_1
  pytorch            pytorch/linux-64::pytorch-1.6.0-py3.7_cuda10.2.89_cudnn7.6.5_0

The following packages will be UPDATED:

  ca-certificates    conda-forge::ca-certificates-2021.5.3~ --> pkgs/main::ca-certificates-2021.7.5-h06a4308_1

The following packages will be SUPERSEDED by a higher-priority channel:

  certifi            conda-forge::certifi-2021.5.30-py37h8~ --> pkgs/main::certifi-2021.5.30-py37h06a4308_0


Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done
/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

[2]:
```
import boto3
import sagemaker
import botocore

config = botocore.config.Config(user_agent_extra='dlai-pds/c2/w1')

# low-level service client of the boto3 session
sm = boto3.client(service_name='sagemaker',
                  config=config)

featurestore_runtime = boto3.client(service_name='sagemaker-featurestore-runtime',
                                    config=config)

sess = sagemaker.Session(sagemaker_client=sm,
                         sagemaker_featurestore_runtime_client=featurestore_runtime)
```

```
bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = sess.boto_region_name
```

1. Configure the SageMaker Feature Store

1.1. Configure dataset

The raw dataset is in the public S3 bucket. Let's start by specifying the S3 location of it:

[3]:
```
raw_input_data_s3_uri = 's3://dlai-practical-data-science/data/raw/'
print(raw_input_data_s3_uri)
```

```
s3://dlai-practical-data-science/data/raw/
```

List the files in the S3 bucket (in this case it will be just one file):

[4]:
```
!aws s3 ls $raw_input_data_s3_uri
```

```
2021-04-30 02:21:06    8457214 womens_clothing_ecommerce_reviews.csv
```

1.2. Configure the SageMaker feature store

As the result of the transformation, in addition to generating files in S3 bucket, you will also save the transformed data in the **Amazon SageMaker Feature Store** to be used by others in your organization, for example.

To configure a Feature Store you need to setup a **Feature Group**. This is the main resource containing all of the metadata related to the data stored in the Feature Store. A Feature Group should contain a list of **Feature Definitions**. A Feature Definition consists of a name and the data type. The Feature Group also contains an online store configuration and an offline store configuration controlling where the data is stored. Enabling the online store allows quick access to the latest value for a record via the GetRecord API. The offline store allows storage of the data in your S3 bucket. You will be using the offline store in this lab.

Let's setup the Feature Group name and the Feature Store offline prefix in S3 bucket (you will use those later in the lab):

[5]:
```
import time
timestamp = int(time.time())

feature_group_name = 'reviews-feature-group-' + str(timestamp)
feature_store_offline_prefix = 'reviews-feature-store-' + str(timestamp)

print('Feature group name: {}'.format(feature_group_name))
print('Feature store offline prefix in S3: {}'.format(feature_store_offline_prefix))
```

```
Feature group name: reviews-feature-group-1628645791
Feature store offline prefix in S3: reviews-feature-store-1628645791
```

Taking two features from the original raw dataset (`Review Text` and `Rating`), you will transform it preparing to be used for the model training and then to be saved in the Feature Store. Here you will define the related features to be stored as a list of `FeatureDefinition` .

[6]:
```
from sagemaker.feature_store.feature_definition import (
    FeatureDefinition,
    FeatureTypeEnum,
)

feature_definitions= [
    # unique ID of the review
    FeatureDefinition(feature_name='review_id', feature_type=FeatureTypeEnum.STRING),
    # ingestion timestamp
    FeatureDefinition(feature_name='date', feature_type=FeatureTypeEnum.STRING),
    # sentiment: -1 (negative), 0 (neutral) or 1 (positive). It will be found the Rating values (1, 2, 3, 4, 5)
    FeatureDefinition(feature_name='sentiment', feature_type=FeatureTypeEnum.STRING),
    # label ID of the target class (sentiment)
    FeatureDefinition(feature_name='label_id', feature_type=FeatureTypeEnum.STRING),
    # reviews encoded with the BERT tokenizer
    FeatureDefinition(feature_name='input_ids', feature_type=FeatureTypeEnum.STRING),
    # original Review Text
    FeatureDefinition(feature_name='review_body', feature_type=FeatureTypeEnum.STRING),
    # train/validation/test label
    FeatureDefinition(feature_name='split_type', feature_type=FeatureTypeEnum.STRING)
]
```

Exercise 1

Create the feature group using the feature definitions defined above.

Instructions: Use the `FeatureGroup` function passing the defined above feature group name and the feature definitions.

```
feature_group = FeatureGroup(
    name=..., # Feature Group name
    feature_definitions=..., # a list of Feature Definitions
    sagemaker_session=sess # SageMaker session
)
```

[7]:
```
from sagemaker.feature_store.feature_group import FeatureGroup

feature_group = FeatureGroup(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    name='reviews-feature-group-1628645791', # Replace None
    feature_definitions=feature_definitions, # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
    sagemaker_session=sess
)

print(feature_group)
```

```
FeatureGroup(name='reviews-feature-group-1628645791', sagemaker_session=<sagemaker.session.Session object at 0x7f765a434090>, feature_definitions=
[FeatureDefinition(feature_name='review_id', feature_type=<FeatureTypeEnum.STRING: 'String'>), FeatureDefinition(feature_name='date', feature_type
=<FeatureTypeEnum.STRING: 'String'>), FeatureDefinition(feature_name='sentiment', feature_type=<FeatureTypeEnum.STRING: 'String'>), FeatureDefinit
ion(feature_name='label_id', feature_type=<FeatureTypeEnum.STRING: 'String'>), FeatureDefinition(feature_name='input_ids', feature_type=<FeatureTy
peEnum.STRING: 'String'>), FeatureDefinition(feature_name='review_body', feature_type=<FeatureTypeEnum.STRING: 'String'>), FeatureDefinition(featu
re_name='split_type', feature_type=<FeatureTypeEnum.STRING: 'String'>)])
```

You will use the defined Feature Group later in this lab, the actual creation of the Feature Group will take place in the processing job. Now let's move into the setup of the processing job to transform the dataset.

2. Transform the dataset

You will configure a SageMaker processing job to run a custom Python script to balance and transform the raw data into a format used by BERT model.

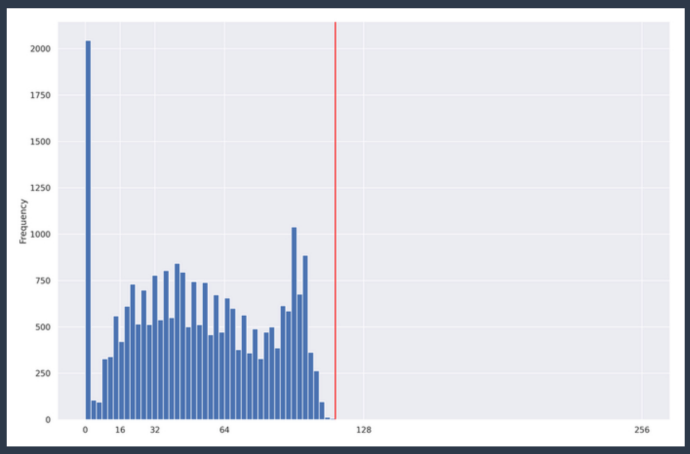
Set the transformation parameters including the instance type, instance count, and train/validation/test split percentages. For the purposes of this lab, you will use a relatively small instance type. Please refer to this link for additional instance types that may work for your use case outside of this lab.

You can also choose whether you want to balance the dataset or not. In this case, you will balance the dataset to avoid class imbalance in the target variable, `sentiment`.

Another important parameter of the model is the `max_seq_length`, which specifies the maximum length of the classified reviews for the RoBERTa model. If the sentence is shorter than the maximum length parameter, it will be padded. In another case, when the sentence is longer, it will be truncated from the right side.

Since a smaller `max_seq_length` leads to faster training and lower resource utilization, you want to find the smallest power-of-2 that captures `100%` of our reviews. For this dataset, the `100th` percentile is `115`. However, it's best to stick with powers-of-2 when using BERT. So let's choose `128` as this is the smallest power-of-2 greater than `115`. You will see below how the shorter sentences will be padded to a maximum length.

```
mean        52.512374
std         31.387048
min          1.000000
10%         10.000000
20%         22.000000
30%         32.000000
40%         41.000000
50%         51.000000
60%         61.000000
70%         73.000000
80%         88.000000
90%         97.000000
100%       115.000000
max        115.000000
```



[8]:
```
processing_instance_type='ml.c5.xlarge'
processing_instance_count=1
train_split_percentage=0.90
validation_split_percentage=0.05
test_split_percentage=0.05
balance_dataset=True
max_seq_length=128
```

To balance and transform our data, you will use a scikit-learn-based processing job. This is essentially a generic Python processing job with scikit-learn pre-installed. You can specify the version of scikit-learn you wish to use. Also pass the SageMaker execution role, processing instance type and instance count.

[9]:
```
from sagemaker.sklearn.processing import SKLearnProcessor

processor = SKLearnProcessor(
    framework_version='0.23-1',
    role=role,
    instance_type=processing_instance_type,
    instance_count=processing_instance_count,
    env={'AWS_DEFAULT_REGION': region},
    max_runtime_in_seconds=7200
)
```

The processing job will be running the Python code from the file `src/prepare_data.py`. In the following exercise you will review the contents of the file and familiarize yourself with main parts of it.

Exercise 2

1. Open the file src/prepare_data.py. Go through the comments to understand its content.
2. Find and review the `convert_to_bert_input_ids()` function, which contains the RoBERTa `tokenizer` configuration.
3. Complete method `encode_plus` of the RoBERTa `tokenizer`. Pass the `max_seq_length` as a value for the argument `max_length`. It defines a pad to a maximum length specified.
4. Save the file src/prepare_data.py (with the menu command File -> Save Python File).

This cell will take approximately 1-2 minutes to run.

[13]:
```
import sys, importlib
sys.path.append('src/')

# import the `prepare_data.py` module
import prepare_data

# reload the module if it has been previously loaded
if 'prepare_data' in sys.modules:
    importlib.reload(prepare_data)

input_ids = prepare_data.convert_to_bert_input_ids("this product is great!", max_seq_length)

updated_correctly = False

if len(input_ids) != max_seq_length:
    print('##########################################################################################################')
    print('Please check that the function \'convert_to_bert_input_ids\' in the file src/prepare_data.py is complete.')
    print('##########################################################################################################')
    raise Exception('Please check that the function \'convert_to_bert_input_ids\' in the file src/prepare_data.py is complete.')
else:
```

```
    print('#################')
    print('Updated correctly!')
    print('#################')

    updated_correctly = True
```

```
#################
Updated correctly!
#################
```

Review the results of tokenization for the given example (*"this product is great!"*):

[14]:
```
input_ids = prepare_data.convert_to_bert_input_ids("this product is great!", max_seq_length)

print(input_ids)
print('Length of the sequence: {}'.format(len(input_ids)))
```

```
[0, 9226, 1152, 16, 372, 328, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Length of the sequence: 128
```

Launch the processing job with the custom script passing defined above parameters.

[15]:
```
from sagemaker.processing import ProcessingInput, ProcessingOutput

if (updated_correctly):

    processor.run(code='src/prepare_data.py',
              inputs=[
                    ProcessingInput(source=raw_input_data_s3_uri,
                                    destination='/opt/ml/processing/input/data/',
                                    s3_data_distribution_type='ShardedByS3Key')
              ],
              outputs=[
                    ProcessingOutput(output_name='sentiment-train',
                                     source='/opt/ml/processing/output/sentiment/train',
                                     s3_upload_mode='EndOfJob'),
                    ProcessingOutput(output_name='sentiment-validation',
                                     source='/opt/ml/processing/output/sentiment/validation',
                                     s3_upload_mode='EndOfJob'),
                    ProcessingOutput(output_name='sentiment-test',
                                     source='/opt/ml/processing/output/sentiment/test',
                                     s3_upload_mode='EndOfJob')
              ],
              arguments=['--train-split-percentage', str(train_split_percentage),
                         '--validation-split-percentage', str(validation_split_percentage),
                         '--test-split-percentage', str(test_split_percentage),
                         '--balance-dataset', str(balance_dataset),
                         '--max-seq-length', str(max_seq_length),
                         '--feature-store-offline-prefix', str(feature_store_offline_prefix),
                         '--feature-group-name', str(feature_group_name)
              ],
              logs=True,
              wait=False)

else:
    print('#######################################')
    print('Please update the code correctly above.')
    print('#######################################')
```

```
Job Name:  sagemaker-scikit-learn-2021-08-11-02-29-20-395
Inputs:  [{'InputName': 'input-1', 'AppManaged': False, 'S3Input': {'S3Uri': 's3://dlai-practical-data-science/data/raw/', 'LocalPath': '/opt/ml/processing/input/data/', 'S3DataType': 'S3Prefix', 'S3InputMode': 'File', 'S3DataDistributionType': 'ShardedByS3Key', 'S3CompressionType': 'None'}}, {'InputName': 'code', 'AppManaged': False, 'S3Input': {'S3Uri': 's3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/input/code/prepare_data.py', 'LocalPath': '/opt/ml/processing/input/code', 'S3DataType': 'S3Prefix', 'S3InputMode': 'File', 'S3DataDistributionType': 'FullyReplicated', 'S3CompressionType': 'None'}}]
Outputs:  [{'OutputName': 'sentiment-train', 'AppManaged': False, 'S3Output': {'S3Uri': 's3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-train', 'LocalPath': '/opt/ml/processing/output/sentiment/train', 'S3UploadMode': 'EndOfJob'}}, {'OutputName': 'sentiment-validation', 'AppManaged': False, 'S3Output': {'S3Uri': 's3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-validation', 'LocalPath': '/opt/ml/processing/output/sentiment/validation', 'S3UploadMode': 'EndOfJob'}}, {'OutputName': 'sentiment-test', 'AppManaged': False, 'S3Output': {'S3Uri': 's3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-test', 'LocalPath': '/opt/ml/processing/output/sentiment/test', 'S3UploadMode': 'EndOfJob'}}]
```

You can see the information about the processing jobs using the `describe` function. The result is in dictionary format. Let's pull the processing job name:

[16]:
```
scikit_processing_job_name = processor.jobs[-1].describe()['ProcessingJobName']

print('Processing job name: {}'.format(scikit_processing_job_name))
```

```
Processing job name: sagemaker-scikit-learn-2021-08-11-02-29-20-395
```

Exercise 3

Pull the processing job status from the processing job description.

Instructions: Print the keys of the processing job description dictionary, choose the one related to the status of the processing job and print the value of it.

[17]:
```
print(processor.jobs[-1].describe().keys())
```

```
dict_keys(['ProcessingInputs', 'ProcessingOutputConfig', 'ProcessingJobName', 'ProcessingResources', 'StoppingCondition', 'AppSpecification', 'Environment', 'RoleArn', 'ProcessingJobArn', 'ProcessingJobStatus', 'LastModifiedTime', 'CreationTime', 'ResponseMetadata'])
```

[20]:
```
### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
scikit_processing_job_status = processor.jobs[-1].describe()['ProcessingJobStatus'] # Replace None
### END SOLUTION - DO NOT delete this comment for grading purposes
print('Processing job status: {}'.format(scikit_processing_job_status))
```

```
Processing job status: InProgress
```

Review the created processing job in the AWS console.

Instructions:

- open the link
- notice that you are in the section `Amazon SageMaker` -> `Processing jobs`
- check the name of the processing job, its status and other available information

[21]:
```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="blank" href="https://console.aws.amazon.com/sagemaker/home?region={}#/processing-jobs/{}">processing job</a></b
```

Review processing job

Wait for about 5 minutes to review the CloudWatch Logs. You may open the file src/prepare_data.py again and examine the outputs of the code in the CloudWatch logs.

[22]:
```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="blank" href="https://console.aws.amazon.com/cloudwatch/home?region={}#logStream:group=/aws/sagemaker/Processing
```

Review CloudWatch logs after about 5 minutes

After the completion of the processing job you can also review the output in the S3 bucket.

[23]:
```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/{}/{}/?region={}&tab=overview">S3 output data</a> aft
```

Review S3 output data after the processing job has completed

Wait for the processing job to complete.

This cell will take approximately 15 minutes to run.

[24]:
```
%%time

running_processor = sagemaker.processing.ProcessingJob.from_processing_name(
    processing_job_name=scikit_processing_job_name,
    sagemaker_session=sess
)

running_processor.wait(logs=False)
```

```
.....................................................................................................!CPU times: user 542 ms,
sys: 86.8 ms, total: 629 ms
Wall time: 10min 13s
```

Please wait until ^^ Processing Job ^^ completes above

Inspect the transformed and balanced data in the S3 bucket.

[25]:
```
processing_job_description = running_processor.describe()

output_config = processing_job_description['ProcessingOutputConfig']
for output in output_config['Outputs']:
    if output['OutputName'] == 'sentiment-train':
        processed_train_data_s3_uri = output['S3Output']['S3Uri']
    if output['OutputName'] == 'sentiment-validation':
        processed_validation_data_s3_uri = output['S3Output']['S3Uri']
    if output['OutputName'] == 'sentiment-test':
        processed_test_data_s3_uri = output['S3Output']['S3Uri']

print(processed_train_data_s3_uri)
print(processed_validation_data_s3_uri)
print(processed_test_data_s3_uri)
```

```
s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-train
s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-validation
s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-test
```

[26]:
```
!aws s3 ls $processed_train_data_s3_uri/
```

```
2021-08-11 02:43:15    4878637 part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

[27]:
```
!aws s3 ls $processed_validation_data_s3_uri/
```

```
2021-08-11 02:43:15     269423 part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

[28]:
```
!aws s3 ls $processed_test_data_s3_uri/
```

```
2021-08-11 02:43:16     273697 part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

Copy the data into the folder `balanced`.

[29]:
```
!aws s3 cp $processed_train_data_s3_uri/part-algo-1-womens_clothing_ecommerce_reviews.tsv ./balanced/sentiment-train/
!aws s3 cp $processed_validation_data_s3_uri/part-algo-1-womens_clothing_ecommerce_reviews.tsv ./balanced/sentiment-validation/
!aws s3 cp $processed_test_data_s3_uri/part-algo-1-womens_clothing_ecommerce_reviews.tsv ./balanced/sentiment-test/
```

```
download: s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-train/part-algo-1-womens_clothing_
ecommerce_reviews.tsv to balanced/sentiment-train/part-algo-1-womens_clothing_ecommerce_reviews.tsv
download: s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-validation/part-algo-1-womens_cloth
ing_ecommerce_reviews.tsv to balanced/sentiment-validation/part-algo-1-womens_clothing_ecommerce_reviews.tsv
download: s3://sagemaker-us-east-1-512205290621/sagemaker-scikit-learn-2021-08-11-02-29-20-395/output/sentiment-test/part-algo-1-womens_clothing_e
commerce_reviews.tsv to balanced/sentiment-test/part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

Review the training, validation and test data outputs:

[30]:
```
!head -n 5 ./balanced/sentiment-train/part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

```
review_id       sentiment       label_id        input_ids       review_body     date
9390    -1      0       [0, 387, 783, 4567, 66, 7586, 42, 16, 2422, 34358, 156, 8, 182, 3214, 734, 5, 1521, 16, 430, 1437, 1437, 5, 3195, 16, 272
1, 1437, 53, 5, 1468, 16, 10, 430, 527, 7586, 24, 197, 28, 7663, 23, 5, 144, 389, 22574, 7586, 45, 13950, 7586, 939, 437, 3357, 106, 4, 67, 1437,
51, 422, 182, 380, 7586, 328, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]  Bummed out.. this is super cheaply made and very expensiv
e... the design is different  the color is beautiful  but the material is a different story.. it should be priced at the most 30 bucks.. not 12
8.. i'm returning them. also  they run very big..!      2021-08-11T02:37:14Z
967     0       1       [0, 100, 21, 5779, 19, 42, 13855, 6439, 4, 77, 939, 78, 794, 24, 11, 5, 1400, 1437, 939, 802, 939, 21, 164, 7, 657, 24, 4,
77, 939, 1381, 24, 15, 1437, 5, 78, 696, 16, 5, 10199, 16, 350, 11708, 4, 939, 115, 2198, 192, 5, 17475, 9, 127, 20745, 8, 11689, 4, 24, 67, 1237,
169, 350, 380, 328, 939, 216, 5, 2576, 16, 3518, 7, 28, 7082, 8, 3041, 219, 1437, 53, 5, 13977, 782, 7, 28, 55, 7898, 2050, 7, 160, 278, 14, 1437,
8, 24, 18, 98, 3298, 4740, 4, 5, 3124, 6538, 32, 67, 847, 182, 614, 1437, 98, 77, 47, 5258, 110, 3701, 47, 64, 192, 144, 9, 110, 11689, 15, 5, 238
0, 1437, 1437, 939, 236, 242, 2, 1]     I was disappointed with this jumpsuit. when i first saw it in the store  i thought i was going to love it.
when i tried it on  the first issue is the fabric is too sheer. i could completely see the outline of my underwear and bra. it also runs way too b
ig! i know the bottom is supposed to be loose and flowy  but the waist needs to be more tapered to off set that  and it's so baggy. the arm holes
are also cut very low  so when you lift your arms you can see most of your bra on the sides   i wante    2021-08-11T02:37:14Z
2304    -1      0       [0, 133, 299, 16, 2579, 1437, 205, 1318, 1437, 4753, 17220, 53, 141, 109, 47, 489, 5, 128, 1529, 5, 4793, 108, 160, 5, 479
3, 116, 4318, 4719, 11, 7968, 62, 454, 24, 18, 95, 10, 15305, 1104, 299, 19, 10, 7782, 7082, 5397, 1902, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]     The top is nice  good quality  etc.... but how do you keep the 'off the shoulder'
off the shoulder? mine keeps inching up until it's just a boring white top with a strange loose neckline.       2021-08-11T02:37:14Z
17231   1       2       [0, 25101, 251, 1437, 98, 189, 240, 7, 23385, 114, 110, 11, 1090, 424, 16, 223, 14434, 3079, 845, 2721, 20580, 328, 2, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1]       "Very long  so may need to hem if your inseam is under ~34"". beautiful trousers!"      2021-08-11T02:37:14Z
```

[31]:
```
!head -n 5 ./balanced/sentiment-validation/part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

```
review_id       sentiment       label_id        input_ids       review_body     date
10284   0       1       [0, 100, 5940, 42, 7, 5961, 53, 19, 13747, 4, 24, 16, 2721, 546, 1437, 941, 5, 22990, 4745, 4, 1437, 1437, 24, 16, 67, 18
2, 1810, 4, 5, 4366, 4745, 16, 10, 1498, 1521, 5998, 1437, 24, 22475, 62, 235, 874, 5, 11044, 1437, 442, 65, 356, 1810, 8, 1256, 10861, 4, 24, 18,
4678, 14, 5, 1421, 16, 9646, 5, 4366, 4202, 1116, 30, 9405, 69, 3701, 593, 7, 69, 809, 4, 53, 939, 1467, 42, 31, 5, 97, 6173, 1278, 259, 98, 24, 2
1, 117, 2755, 4, 1437, 1437, 939, 2740, 1836, 579, 1437, 5, 15654, 577, 1437, 8, 24, 2564, 182, 1810, 55, 101, 10, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1]      I recommend this topper but with reservations. it is beautiful looking  especially the lace portion.   it is also
very wide. the gathered portion is a complete design fail  it balloons up right below the bust  making one look wide and pretty ridiculous. it's o
bvious that the model is hiding the gathered poof by placing her arms close to her body. but i knew this from the other reviews posted here so it
was no surprise.   i ordered size s  the smallest available  and it fit very wide more like a   2021-08-11T02:37:14Z
118     0       1       [0, 133, 299, 25, 19, 144, 9, 6256, 18, 13657, 16, 157, 1690, 17670, 4, 1468, 16, 182, 9800, 4, 114, 47, 33, 739, 11044, 2
```

```
4, 16, 10, 410, 26801, 3923, 4, 42, 189, 14302, 47, 7, 3568, 402, 12213, 7, 356, 6473, 8, 464, 5, 3989, 9, 5, 299, 328, 2, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]        The top as with most of ap's tops is well stitched. material is very uncomfortabl
e. if you have large bust it is a little divulging. this may prompt you to wear something underneath to look modest and change the shape of the to
p!      2021-08-11T02:37:14Z
19076   1       2       [0, 5975, 1437, 939, 2740, 258, 29874, 7952, 8, 939, 218, 75, 216, 61, 7, 489, 734, 5, 1468, 16, 10, 2579, 3793, 29716, 4,
5, 5397, 516, 16, 818, 4293, 12, 20020, 1437, 19, 410, 14893, 1437, 98, 11962, 4, 8089, 32, 269, 2579, 350, 4, 939, 1381, 24, 15, 19, 2972, 5246,
918, 8, 20580, 10, 295, 417, 12369, 1415, 372, 19, 258, 4, 939, 439, 19, 5, 4716, 1459, 1732, 1437, 13, 167, 10941, 821, 1536, 66, 89, 734, 37863,
29, 4716, 1459, 2564, 162, 1437, 939, 2489, 705, 858, 75, 1381, 1021, 5, 3023, 29, 4716, 1459, 648, 4, 36, 245, 108, 134, 4, 245, 113, 8, 12312, 2
3246, 322, 939, 206, 24, 74, 784, 5361, 2579, 19, 10, 21451, 16576, 350, 1437, 2]       "Now  i ordered both striped versions and i don't know whi
ch to keep... the material is a nice soft knit. the neck line is almost boat-neck  with little buttons  so cute. colors are really nice too. i tri
ed it on with skinnies and trousers a nd ti looked great with both. i went with the petite version  for those shorter gals out there... xxs petite
fit me  i havne't tried o the xs petite yet. (5'1.5"" and 115 lbs). i think it would loke nice with a pencil skirt too  if i only had one that wen
t wit"  2021-08-11T02:37:14Z
9136    1       2       [0, 100, 56, 57, 20087, 9798, 42, 3588, 13, 81, 10, 76, 8, 1747, 3032, 2185, 7, 24, 13, 5, 6875, 4, 78, 1437, 24, 16, 1205
8, 8, 182, 802, 6459, 1887, 7, 146, 47, 356, 2770, 4, 147, 383, 1136, 160, 5, 26717, 13, 162, 16, 5, 10199, 2031, 13, 5, 21764, 734, 10010, 32, 9
8, 3036, 15651, 8, 32156, 29650, 444, 350, 2773, 4, 5, 3588, 14, 939, 829, 11, 5, 7107, 56, 80, 739, 4543, 8299, 36, 1264, 11, 349, 24150, 43, 14,
58, 416, 878, 10, 828, 4, 7, 162, 14, 21, 10, 1307, 1275, 3794, 14, 939, 4395, 75, 28, 1408, 98, 203, 418, 1021, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1]    I had been admiring this dress for over a year and finally treated myself to it for the holidays. first  it is gorgeous an
d very thoughtfully designed to make you look amazing. where things fall off the rails for me is the fabric choice for the sleeves...they are sooo
delicate and evidently snag far too easily. the dress that i received in the mail had two large snags (one in each sleeve) that were already runni
ng a bit. to me that was a huge red flag that i shouldn't be spending so much money o   2021-08-11T02:37:14Z
```

[32]:
```
!head -n 5 ./balanced/sentiment-test/part-algo-1-womens_clothing_ecommerce_reviews.tsv
```

```
review_id       sentiment       label_id        input_ids       review_body     date
3776    -1      0       [0, 713, 299, 21, 12058, 8, 2564, 6683, 235, 66, 9, 5, 3298, 1437, 19, 129, 10, 2119, 6440, 154, 4, 9574, 1437, 5, 78, 86,
939, 15158, 24, 36, 33912, 514, 1437, 6713, 3841, 43, 1437, 24, 28704, 98, 14, 24, 117, 1181, 10698, 420, 5, 11044, 396, 7898, 147, 24, 21, 683, 9
29, 219, 1437, 16, 203, 10941, 1437, 8, 2653, 16799, 571, 1843, 7, 5, 2842, 4, 101, 277, 37102, 26, 36, 879, 45779, 1437, 71, 939, 56, 3584, 8, 15
158, 13278, 1437, 939, 2813, 5, 6929, 74, 33, 95, 26, 7, 3841, 2382, 129, 1437, 142, 939, 9917, 14784, 24, 8, 74, 33, 626, 3680, 114, 939, 56, 112
69, 139, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]      This top was gorgeous and fit perfectly right out of the bag  with only a quick ironing. u
nfortunately  the first time i washed it (cold water  hang dry)  it shrunk so that it no longer fits across the bust without tape where it was onc
e roomy  is much shorter  and feels rougher to the touch. like another reviewer said (unfortunately  after i had purchased and washed!)  i wish th
e label would have just said to dry clean only  because i regret washing it and would have done otherwise if i had kno  2021-08-11T02:37:14Z
4053    1       2       [0, 100, 657, 209, 9304, 8, 3568, 106, 70, 5, 86, 4, 939, 2813, 51, 5844, 75, 57, 27251, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
I love these pants and wear them all the time. i wish they hadn't been discontinued     2021-08-11T02:37:14Z
4587    1       2       [0, 100, 524, 1622, 7758, 59, 42, 3588, 4, 939, 524, 2735, 15, 299, 19, 2514, 28097, 4, 127, 11044, 16, 10, 34384, 1722, 7
41, 53, 939, 115, 192, 42, 447, 15, 10, 50, 740, 12988, 67, 4, 42, 3588, 16, 2778, 34203, 13, 127, 809, 1907, 8, 5, 28072, 2463, 10698, 101, 24, 2
1, 156, 13, 162, 4, 24, 10698, 2051, 396, 10, 11689, 8, 95, 6683, 19, 10, 4126, 37342, 18052, 16979, 4, 939, 300, 42, 3588, 7, 3568, 7, 10, 1035,
3312, 4, 5, 18632, 8, 3195, 32, 2721, 8, 939, 40, 1763, 19, 2007, 11449, 8, 16109, 8872, 4, 24, 18, 10, 372, 465, 8, 4206, 1318, 4, 26606, 2, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]  I am simply mad about this dress. i am smaller on top with larger hips. my bust is a mellow b but i could
see this working on a or c cups also. this dress is extremely flattering for my body type and the bodice fits like it was made for me. it fits fin
e without a bra and just perfectly with a semi padded strapless. i got this dress to wear to a summer wedding. the texture and color are beautiful
and i will pair with simple jewelry and nude heels. it's a great find and excellent quality. definite   2021-08-11T02:37:14Z
12091   0       1       [0, 100, 15655, 697, 11, 18437, 9304, 1437, 8, 939, 101, 5, 2216, 2496, 9, 42, 1810, 12, 4308, 1732, 4, 939, 437, 3489, 1
0, 231, 73, 2517, 11, 9304, 8, 2740, 5, 650, 61, 2564, 157, 1437, 114, 10, 828, 380, 4, 5, 1374, 2564, 16, 372, 111, 11875, 23, 5, 6605, 19, 10, 2
579, 1810, 13977, 9484, 1437, 8, 2342, 7933, 7790, 12, 12459, 1882, 12189, 4, 5, 9304, 269, 11875, 127, 28097, 8, 694, 10, 2579, 33585, 19, 5, 181
0, 5856, 4, 5, 10199, 16, 3793, 8, 7964, 8, 939, 222, 45, 676, 143, 11708, 1825, 4, 25, 13, 3195, 111, 939, 74, 486, 209, 10, 1844, 3035, 6219, 14
37, 2299, 45, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]    I practically live in lounge pants  and i like the unique style of this wide-leg version.
i'm generally a 6/28 in pants and ordered the small which fit well  if a bit big. the overall fit is great - slim at the hip with a nice wide wais
tband  and flatteringly-placed dropped pockets. the pants really slim my hips and provide a nice silhouette with the wide legs. the fabric is soft
and flows and i did not experience any sheerness. as for color - i would call these a deep cool brown  definitely not   2021-08-11T02:37:14Z
```

3. Query the Feature Store

In addition to transforming the data and saving in S3 bucket, the processing job populates the feature store with the transformed and balanced data. Let's query this data using Amazon Athena.

3.1. Export training, validation, and test datasets from the Feature Store

Here you will do the export only for the training dataset, as an example.

Use `athena_query()` function to create an Athena query for the defined above Feature Group. Then you can pull the table name of the Amazon Glue Data Catalog table which is auto-generated by Feature Store.

[33]:
```
feature_store_query = feature_group.athena_query()

feature_store_table = feature_store_query.table_name

query_string = """
    SELECT date,
        review_id,
        sentiment,
        label_id,
        input_ids,
        review_body
    FROM "{}"
    WHERE split_type='train'
    LIMIT 5
""".format(feature_store_table)

print('Glue Catalog table name: {}'.format(feature_store_table))
print('Running query: {}'.format(query_string))
```

```
Glue Catalog table name: reviews-feature-group-1628645791-1628649409
Running query:
    SELECT date,
        review_id,
        sentiment,
        label_id,
        input_ids,
        review_body
    FROM "reviews-feature-group-1628645791-1628649409"
    WHERE split_type='train'
    LIMIT 5
```

Configure the S3 location for the query results. This allows us to re-use the query results for future queries if the data has not changed. We can even share this S3 location between team members to improve query performance for common queries on data that does not change often.

[34]:
```
output_s3_uri = 's3://{}/query_results/{}/'.format(bucket, feature_store_offline_prefix)
print(output_s3_uri)
```

```
s3://sagemaker-us-east-1-512205290621/query_results/reviews-feature-store-1628645791/
```

Exercise 4

Query the feature store.

Instructions: Use `feature_store_query.run` function passing the constructed above query string and the location of the output S3 bucket.

```
    feature_store_query.run(
        query_string=..., # query string
```

```
        output_location=... # location of the output S3 bucket
    )
```

[37]:
```
feature_store_query.run(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    query_string=query_string, # Replace None
    output_location='s3://sagemaker-us-east-1-512205290621/query_results/reviews-feature-store-1628645791/' # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
)

feature_store_query.wait()
```

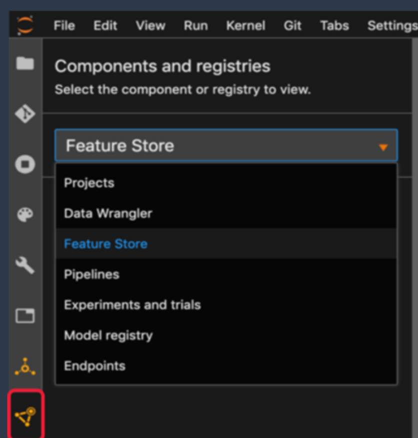
[38]:
```
import pandas as pd
pd.set_option("max_colwidth", 100)

df_feature_store = feature_store_query.as_dataframe()
df_feature_store
```

[38]:

	date	review_id	sentiment	label_id	input_ids	review_body
0	2021-08-11T02:37:14Z	12144	0	1	[0, 16587, 5, 299, 1521, 1437, 45, 98, 203, 5, 10199, 4, 939, 197, 348, 1166, 5, 299, 21, 727, 2...	Love the top design not so much the fabric. i should've read the top was 100% polyester. return...
1	2021-08-11T02:37:14Z	3419	0	1	[0, 713, 2551, 7, 422, 10, 410, 828, 650, 15, 162, 734, 55, 1026, 12365, 87, 15, 5, 1421, 11, 5,...	This seemed to run a little bit small on me... more form fitting than on the model in the photo....
2	2021-08-11T02:37:14Z	8006	-1	0	[0, 100, 3584, 42, 3588, 13, 127, 11412, 4, 5, 8089, 32, 203, 55, 5676, 8, 4520, 172, 51, 2082, ...	I purchased this dress for my graduation. the colors are much more intense and bright then they ...
3	2021-08-11T02:37:14Z	13012	0	1	[0, 100, 770, 7, 657, 42, 299, 328, 15, 162, 24, 21, 7735, 352, 2233, 219, 4, 24, 21, 350, 1810,...	I wanted to love this top! on me it was weirdly boxy. it was too wide and too short and it was ...
4	2021-08-11T02:37:14Z	19440	-1	0	[0, 713, 3588, 56, 98, 203, 801, 53, 5, 25490, 29716, 19750, 24, 4, 939, 1381, 15, 1836, 37863, ...	This dress had so much potential but the horizontal knit ruined it. i tried on size xxs in grey ...

Review the Feature Store in SageMaker Studio



3.2. Export TSV from Feature Store

Save the output as a TSV file:

[39]:
```
df_feature_store.to_csv('./feature_store_export.tsv',
                        sep='\t',
                        index=False,
                        header=True)
```

[40]:
```
!head -n 5 ./feature_store_export.tsv
```

```
date	review_id	sentiment	label_id	input_ids	review_body
2021-08-11T02:37:14Z	12144	0	1	[0, 16587, 5, 299, 1521, 1437, 45, 98, 203, 5, 10199, 4, 939, 197, 348, 1166, 5, 299, 21, 727, 20
7, 11424, 8939, 4, 3357, 8, 2609, 10, 357, 1318, 299, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	Love the top design  not so much the fabric. i should've read the top was 100% polyester.
returning and finding a better quality top.
2021-08-11T02:37:14Z	3419	0	1	[0, 713, 2551, 7, 422, 10, 410, 828, 650, 15, 162, 734, 55, 1026, 12365, 87, 15, 5, 1421, 11, 5, 1
345, 4, 67, 1437, 24, 18, 7174, 4, 939, 1017, 28, 1088, 15, 24, 13, 686, 114, 24, 58, 95, 10, 410, 1181, 350, 142, 939, 101, 5, 818, 12, 1527, 238
34, 12, 3341, 356, 4, 5, 22990, 45432, 5150, 16, 1256, 1437, 8, 114, 47, 2229, 24, 13, 14, 1437, 172, 24, 269, 924, 62, 357, 15, 5, 80, 13026, 808
9, 111, 24, 269, 1516, 685, 70, 11, 909, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1]	This seemed to run a little bit small on me... more form fitting than on the model in the photo. also  it's thin. i'd be sold on it for su
re if it were just a little longer too because i like the almost-lingerie-like look. the lace applique is pretty  and if you purchase it for that
then it really shows up better on the two lighter colors - it really gets lost all in black.
2021-08-11T02:37:14Z	8006	-1	0	[0, 100, 3584, 42, 3588, 13, 127, 11412, 4, 5, 8089, 32, 203, 55, 5676, 8, 4520, 172, 51, 2082, 80
4, 4, 939, 437, 761, 9, 2273, 14, 939, 356, 101, 10, 20781, 11221, 10178, 4, 939, 101, 5, 2496, 53, 5, 10199, 2653, 6162, 13, 5, 701, 9, 5, 3588,
4, 5, 10199, 16, 727, 207, 16807, 4, 627, 2564, 16, 372, 4, 5, 3588, 16, 1528, 7, 1836, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	I purchased this dress for
my graduation. the colors are much more intense and bright then they appear online. i'm kind of concerned that i look like a carnival tent. i like
the style but the fabric feels cheap for the cost of the dress. the fabric is 100% synthetic.the fit is great. the dress is true to size
2021-08-11T02:37:14Z	13012	0	1	[0, 100, 770, 7, 657, 42, 299, 328, 15, 162, 24, 21, 7735, 352, 2233, 219, 4, 24, 21, 350, 1810,
8, 350, 765, 1437, 8, 24, 21, 55, 11708, 87, 939, 421, 24, 4, 5, 10199, 16, 13719, 53, 36045, 269, 1365, 8, 13585, 3989, 1256, 1335, 4, 939, 67, 8
02, 5, 10199, 74, 28, 19351, 8, 33, 55, 3184, 87, 24, 34, 4, 939, 657, 5, 2850, 1250, 15911, 4617, 23, 5, 3701, 53, 71, 59, 292, 728, 9, 3568, 42,
1299, 2422, 3977, 783, 18331, 8, 95, 761, 9, 26654, 4, 9574, 24, 40, 28, 164, 124, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1]	I wanted to love this top! on me it was weirdly boxy. it was too wide and too short  and it was more sheer than i expected
it. the fabric is lightweight but wrinkles really easy and loses shape pretty quickly. i also thought the fabric would be heavier and have more st
ructure than it has. i love the scalloped detail at the arms but after about five minutes of wear this felt super crumpled and just kind of slopp
y. unfortunately it will be going back.
```

Upload TSV to the S3 bucket:

[41]:
```
!aws s3 cp ./feature_store_export.tsv s3://$bucket/feature_store/feature_store_export.tsv
```

```
upload: ./feature_store_export.tsv to s3://sagemaker-us-east-1-512205290621/feature_store/feature_store_export.tsv
```

Check the file in the S3 bucket:

[42]:
```
!aws s3 ls --recursive s3://$bucket/feature_store/feature_store_export.tsv
```

```
2021-08-11 02:51:28       4562 feature_store/feature_store_export.tsv
```

3.3. Check that the dataset in the Feature Store is balanced by sentiment

Now you can setup an Athena query to check that the stored dataset is balanced by the target class `sentiment`.

Exercise 5

Write an SQL query to count the total number of the reviews per `sentiment` stored in the Feature Group.

Instructions: Pass the SQL statement of the form

```
SELECT category_column, COUNT(*) AS new_column_name
FROM table_name
GROUP BY category_column
```

into the variable `query_string_count_by_sentiment`. Here you would need to use the column `sentiment` and give a name `count_reviews` to the new column with the counts.

[43]:
```
feature_store_query_2 = feature_group.athena_query()

# Replace all None
### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
query_string_count_by_sentiment = """
SELECT sentiment, COUNT(*) AS count_reviews
FROM "{}"
GROUP BY sentiment
""".format(feature_store_table)
### END SOLUTION - DO NOT delete this comment for grading purposes
```

Exercise 6

Query the feature store.

Instructions: Use `run` function of the Feature Store query, passing the new query string `query_string_count_by_sentiment`. The output S3 bucket will remain unchanged. You can follow the example above.

[44]:
```
feature_store_query_2.run(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    query_string=query_string_count_by_sentiment, # Replace None
    output_location='s3://sagemaker-us-east-1-512205290621/query_results/reviews-feature-store-1628645791/' # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
)

feature_store_query_2.wait()

df_count_by_sentiment = feature_store_query_2.as_dataframe()
df_count_by_sentiment
```

[44]:

	sentiment	count_reviews
0	0	2051
1	-1	2051
2	1	2051

Exercise 7

Visualize the result of the query in the bar plot, showing the count of the reviews by sentiment value.

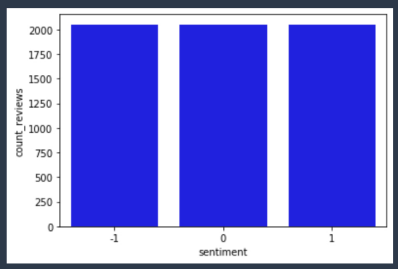
Instructions: Pass the resulting data frame `df_count_by_sentiment` into the `barplot` function of the `seaborn` library.

```
sns.barplot(
    data=...,
    x='...',
    y='...',
    color="blue"
)
```

[46]:
```
import seaborn as sns

sns.barplot(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    data=df_count_by_sentiment, # Replace None
    x='sentiment', # Replace None
    y='count_reviews', # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
    color="blue"
)
```

[46]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f75f16f4a90>
```



Upload the notebook and `prepare_data.py` file into S3 bucket for grading purposes.

Note: you may need to save the file before the upload.

[47]:
```
!aws s3 cp ./C2_W1_Assignment.ipynb s3://$bucket/C2_W1_Assignment_Learner.ipynb
!aws s3 cp ./src/prepare_data.py s3://$bucket/src/C2_W1_prepare_data_Learner.py
```

```
upload: ./C2_W1_Assignment.ipynb to s3://sagemaker-us-east-1-512205290621/C2_W1_Assignment_Learner.ipynb
upload: src/prepare_data.py to s3://sagemaker-us-east-1-512205290621/src/C2_W1_prepare_data_Learner.py
```

Please go to the main lab window and click on `Submit` button (see the `Finish the lab` section of the instructions).

[ ]:


1
1
Git: idle
Python 3 (Data Science) | Idle
Kernel: Idle | Instance MEM
Mode: Command
Ln 1, Col 1
C2_W1_Assignment.ipynb