```
%pyspark

# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://amazonratingtraininginstruments.s3-us-west-1.amazonaws.com/musical_instruments_training.csv"
spark.sparkContext.addFile(url)
start_data = spark.read.csv(SparkFiles.get("musical_instruments_training.csv"), sep=",", header=True)
# start_data = spark.read.csv(SparkFiles.get("amazon_video_training.csv"), sep=",", header=True)
start_data = start_data.dropna()

# Show DataFrame
start_data.show()
```

```
+----------+-----------+--------------------+--------------------+
|      ASIN|Star_Rating|             Summary|        Clean_Review|
+----------+-----------+--------------------+--------------------+
|1384719342|          5|                good|Not much to write...|
|1384719342|          5|                Jake|The product does ...|
|1384719342|          5|It Does The Job Well|The primary job o...|
|1384719342|          5|GOOD WINDSCREEN F...|Nice windscreen p...|
|1384719342|          5|No more pops when...|This pop filter i...|
|B00004Y2UT|          5|      The Best Cable|So good that I bo...|
|B00004Y2UT|          5|Monster Standard ...|I have used monst...|
|B00004Y2UT|          3|Didn't fit my 199...|I now use this ca...|
|B00004Y2UT|          5|         Great cable|Perfect for my Ep...|
|B00004Y2UT|          5|Best Instrument C...|Monster makes the...|
|B00004Y2UT|          5|One of the best i...|Monster makes a w...|
|B00005ML71|          4|It works great bu...|I got it to have ...|
|B00005ML71|          3|HAS TO GET USE TO...|If you are not us...|
|B00005ML71|          5|             awesome|I love it I used ...|
|B00005ML71|          5|           It works!|I bought this to ...|
|B00005ML71|          2|Definitely Not Fo...|I bought this to ...|
|B000068NSX|          4|Durable Instrumen...|This Fender cable...|
|B000068NSX|          5|fender 18 ft. Cal...|wanted it just on...|
|B000068NSX|          5|So far so good.  ...|Ive been using th...|
|B000068NSX|          5|Add California to...|Fender cords look...|
+----------+-----------+--------------------+--------------------+
only showing top 20 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 3 sec 876 millisec. Updated by danconde on October 28 2019, 7:55:54 PM (PDT)

```
%pyspark

from pyspark.sql.functions import length
# Create a length column to be used as a future feature
data_df = start_data.withColumn('length(Summary)', length(start_data['Summary']))
data_df.show()
```

```
+----------+-----------+--------------------+--------------------+--------------+
|      ASIN|Star_Rating|             Summary|        Clean_Review|length(Summary)|
+----------+-----------+--------------------+--------------------+--------------+
|1384719342|          5|                good|Not much to write...|             4|
|1384719342|          5|                Jake|The product does ...|             4|
|1384719342|          5|It Does The Job Well|The primary job o...|            20|
|1384719342|          5|GOOD WINDSCREEN F...|Nice windscreen p...|            29|
|1384719342|          5|No more pops when...|This pop filter i...|            37|
|B00004Y2UT|          5|      The Best Cable|So good that I bo...|            14|
|B00004Y2UT|          5|Monster Standard ...|I have used monst...|            43|
|B00004Y2UT|          3|Didn't fit my 199...|I now use this ca...|            34|
|B00004Y2UT|          5|         Great cable|Perfect for my Ep...|            11|
|B00004Y2UT|          5|Best Instrument C...|Monster makes the...|            36|
|B00004Y2UT|          5|One of the best i...|Monster makes a w...|            50|
|B00005ML71|          4|It works great bu...|I got it to have ...|            35|
|B00005ML71|          3|HAS TO GET USE TO...|If you are not us...|            26|
|B00005ML71|          5|             awesome|I love it I used ...|             7|
|B00005ML71|          5|           It works!|I bought this to ...|             9|
|B00005ML71|          2|Definitely Not Fo...|I bought this to ...|            44|
|B000068NSX|          4|Durable Instrumen...|This Fender cable...|            24|
|B000068NSX|          5|fender 18 ft. Cal...|wanted it just on...|            27|
|B000068NSX|          5|So far so good.  ...|Ive been using th...|            52|
|B000068NSX|          5|Add California to...|Fender cords look...|            38|
+----------+-----------+--------------------+--------------------+--------------+
only showing top 20 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 213 millisec. Updated by danconde on October 28 2019, 7:55:54 PM (PDT)

```
%pyspark

from pyspark.ml.feature import Tokenizer, StopWordsRemover, HashingTF, IDF

# Create all the features to the data set
tokenizer = Tokenizer(inputCol='Summary', outputCol='token_text')
stopremove = StopWordsRemover(inputCol='token_text',outputCol='stop_tokens')
hashingTF = HashingTF(inputCol='token_text', outputCol='hash_token')
idf = IDF(inputCol='hash_token', outputCol='idf_token')
```

Interpreter: spark.pyspark. **FINISHED** Took 215 millisec. Updated by danconde on October 28 2019, 7:55:54 PM (PDT)

```
%pyspark

from pyspark.ml.feature import VectorAssembler
from pyspark.ml.linalg import Vector

# Create feature vectors
clean_up = VectorAssembler(inputCols=['idf_token', 'length(Summary)'], outputCol='features')
```

Interpreter: spark.pyspark. **FINISHED** Took 109 millisec. Updated by danconde on October 28 2019, 7:55:54 PM (PDT)

```
%pyspark

# Create a and run a data processing Pipeline
from pyspark.ml import Pipeline
```

```
from pyspark.ml import Pipeline
data_prep_pipeline = Pipeline(stages=[tokenizer, stopremove, hashingTF, idf, clean_up])
```

Interpreter: spark.pyspark. **FINISHED** Took 110 millisec. Updated by danconde on October 28 2019, 7:55:54 PM (PDT)

```
%pyspark
from pyspark.sql.functions import col, size

# Fit and transform the pipeline
cleaner = data_prep_pipeline.fit(data_df)
cleaned = cleaner.transform(data_df)
cleaned.show(500)
```

```
+----------+-----------+--------------------+--------------------+--------------+--------------------+--------------------+--------------------
+-+--------------------+
|      ASIN|Star_Rating|             Summary|        Clean_Review|length(Summary)|          token_text|         stop_tokens|          hash_token|           idf_tok
en|            features|
+----------+-----------+--------------------+--------------------+--------------+--------------------+--------------------+--------------------
+-+--------------------+
|1384719342|          5|                good|Not much to write...|             4|              [good]|              [good]|(262144,[113432],...|(262144,[11343
2],...|(262145,[113432,2...|
|1384719342|          5|                Jake|The product does ...|             4|              [jake]|              [jake]|(262144,[74472],[...|(262144,[74472],
[...|(262145,[74472,26...|
|1384719342|          5|It Does The Job Well|The primary job o...|            20|[it, does, the, j...|         [job, well]|(262144,[58162,86...|(262144,[58162,8
6...|(262145,[58162,86...|
|1384719342|          5|GOOD WINDSCREEN F...|Nice windscreen p...|            29|[good, windscreen...|[good, windscreen...|(262144,[16332,10...|(262144,[16332,1
0...|(262145,[16332,10...|
|1384719342|          5|No more pops when...|This pop filter i...|            37|[no, more, pops, ...|[pops, record, vo...|(262144,[24417,24...|(262144,[24417,2
4...|(262145,[24417,24...|
|B00004Y2UT|          5|      The Best Cable|So good that I bo...|            14|  [the, best, cable]|       [best, cable]|(262144,[103838,1...|(262144,[103838,
1...|(262145,[103838,1...|
|B00004Y2UT|          5|Monster Standard ...|I have used monst...|            43|[monster, standar...|[monster, standar...|(262144,[45531,84...|(262144,[45531,8
4...|(262145,[45531,84...|
|B00004Y2UT|          3|Didn't fit my 199...|I now use this ca...|            34|[didn't, fit, my,...|[fit, 1996, fende...|(262144,[12084,37...|(262144,[12084,3
7...|(262145,[12084,37...|
|B00004Y2UT|          5|         Great cable|Perfect for my Ep...|            11|      [great, cable]|      [great, cable]|(262144,[107306,1...|(262144,[107306,
1...|(262145,[107306,1...|
|B00004Y2UT|          5|Best Instrument C...|Monster makes the...|            36|[best, instrument...|[best, instrument...|(262144,[100258,1...|(262144,[100258,
1...|(262145,[100258,1...|
|B00004Y2UT|          5|One of the best i...|Monster makes a w...|            50|[one, of, the, be...|[one, best, instr...|(262144,[9639,103...|(262144,[9639,10
3...|(262145,[9639,103...|
|B00005ML71|          4|It works great bu...|I got it to have ...|            35|[it, works, great...|[works, great, ha...|(262144,[12888,24...|(262144,[12888,2
```

Interpreter: spark.pyspark. **FINISHED** Took 2 sec 937 millisec. Updated by danconde on October 28 2019, 7:55:57 PM (PDT)

```
%pyspark

from pyspark.ml.classification import NaiveBayes

# Create second dataframe containing 'label' and 'features' and converting 'label' column to integers
concise_cleaned = cleaned.select(cleaned.Star_Rating.cast('int').alias('label'), col('features'))
concise_cleaned.show()
concise_cleaned.dtypes
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|    5|(262145,[113432,2...|
|    5|(262145,[74472,26...|
|    5|(262145,[58162,86...|
|    5|(262145,[16332,10...|
|    5|(262145,[24417,24...|
|    5|(262145,[103838,1...|
|    5|(262145,[45531,84...|
|    3|(262145,[12084,37...|
|    5|(262145,[107306,1...|
|    5|(262145,[100258,1...|
|    5|(262145,[9639,103...|
|    4|(262145,[12888,24...|
|    3|(262145,[99895,10...|
|    5|(262145,[82495,26...|
|    5|(262145,[86175,21...|
|    2|(262145,[16332,75...|
|    4|(262145,[107306,1...|
|    5|(262145,[67959,83...|
|    5|(262145,[3726,186...|
|    5|(262145,[24417,35...|
+-----+--------------------+
only showing top 20 rows

[('label', 'int'), ('features', 'vector')]
```

Interpreter: spark.pyspark. **FINISHED** Took 3 sec 622 millisec. Updated by danconde on October 28 2019, 7:56:01 PM (PDT)

```
%pyspark

# Break data down into a training set and a testing set
training, testing = concise_cleaned.randomSplit([0.9, 0.1])

# Create a Naive Bayes model and fit training data
nb = NaiveBayes()
predictor = nb.fit(training)
```

Interpreter: spark.pyspark. **FINISHED** Took 10 sec 890 millisec. Updated by danconde on October 28 2019, 7:56:12 PM (PDT)

```
%pyspark

# Tranform the model with the testing data
test_results = predictor.transform(testing)
test_results.show(20)
```

```
+-----+--------------------+--------------------+--------------------+----------+
|label|            features|       rawPrediction|         probability|prediction|
+-----+--------------------+--------------------+--------------------+----------+
|    1|(262145,[14,16332...|[-1018.9211919369...|[1.23791686380733...|       4.0|
```

```
|    1|(262145,[14,16332...|[-1018.9211919369...|[1.23791686380733...|       4.0|
|    1|(262145,[6981,129...|[-561.75782010116...|[2.77331040148778...|       4.0|
|    1|(262145,[9639,113...|[-575.04499984273...|[1.15311697280772...|       4.0|
|    1|(262145,[12888,55...|[-401.45400711086...|[2.41261706563998...|       4.0|
|    1|(262145,[14244,15...|[-273.39083881795...|[7.46727247566329...|       4.0|
|    1|(262145,[17129,52...|[-544.69206374652...|[1.71656572392885...|       4.0|
|    1|(262145,[17252,17...|[-331.37619286209...|[3.23175457865934...|       4.0|
|    1|(262145,[31986,26...|[-84.699448274116...|[0.51343422176181...|       0.0|
|    1|(262145,[34343,58...|[-159.05809814639...|[2.82831062375887...|       4.0|
|    1|(262145,[45245,25...|[-238.51066411638...|[4.97089622964174...|       4.0|
|    1|(262145,[50940,11...|[-399.30385567829...|[2.92122615853076...|       4.0|
|    1|(262145,[68707,75...|[-179.83233780261...|[9.41666850145893...|       4.0|
|    1|(262145,[74200,26...|[-90.635433139915...|[0.99999976581130...|       0.0|
|    1|(262145,[138356,2...|[-176.63276594647...|[3.03496210873983...|       4.0|
|    2|(262145,[2437,802...|[-809.27990727254...|[7.40872641381802...|       4.0|
|    2|(262145,[4980,933...|[-309.49000499415...|[7.02209097620616...|       4.0|
|    2|(262145,[5381,120...|[-325.69025083024...|[2.62597674848585...|       4.0|
|    2|(262145,[5777,880...|[-980.57208783423...|[1.11015305592609...|       4.0|
|    2|(262145,[13114,14...|[-1801.6403056743...|[2.37334022455177...|       4.0|
|    2|(262145,[15889,91...|[-699.28824075262...|[1.83923065671138...|       4.0|
+-----+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 2 sec 668 millisec. Updated by danconde on October 28 2019, 7:56:14 PM (PDT)

```
%pyspark
# Use the Class Evaluator for a cleaner description
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

acc_eval = MulticlassClassificationEvaluator()
acc = acc_eval.evaluate(test_results)
print("Accuracy of model at predicting reviews was: %f" % acc)
```

```
Accuracy of model at predicting reviews was: 0.097475
```

Interpreter: spark.pyspark. **FINISHED** Took 8 sec 698 millisec. Updated by danconde on October 28 2019, 7:56:23 PM (PDT)

```
%pyspark

testing.show(20)
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|    1|(262145,[14,16332...|
|    1|(262145,[6981,129...|
|    1|(262145,[9639,113...|
|    1|(262145,[12888,55...|
|    1|(262145,[14244,15...|
|    1|(262145,[17129,52...|
|    1|(262145,[17252,17...|
|    1|(262145,[31986,26...|
|    1|(262145,[34343,58...|
|    1|(262145,[45245,25...|
|    1|(262145,[50940,11...|
|    1|(262145,[68707,75...|
|    1|(262145,[74200,26...|
|    1|(262145,[138356,2...|
|    2|(262145,[2437,802...|
|    2|(262145,[4980,933...|
|    2|(262145,[5381,120...|
|    2|(262145,[5777,880...|
|    2|(262145,[13114,14...|
|    2|(262145,[15889,91...|
+-----+--------------------+
only showing top 20 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 2 sec 66 millisec. Updated by danconde on October 28 2019, 7:56:25 PM (PDT)

Interpreter: spark.