(/U4G66226D/spaces)

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://s3.amazonaws.com/dataviz-curriculum/day_2/yelp_reviews.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("yelp_reviews.csv"), sep=",", header=True)

# Show DataFrame
df.show()
```

```
+--------+--------------------+
|   class|                text|
+--------+--------------------+
|positive|Wow... Loved this...|
|negative|  Crust is not good.|
|negative|Not tasty and the...|
|positive|Stopped by during...|
|positive|The selection on ...|
|negative|Now I am getting ...|
|negative|Honeslty it didn'...|
|negative|The potatoes were...|
|positive|The fries were gr...|
|positive|      A great touch.|
|positive|Service was very ...|
|negative|  Would not go back.|
|negative|The cashier had n...|
|positive|I tried the Cape ...|
|negative|I was disgusted b...|
|negative|I was shocked bec...|
|positive| Highly recommended.|
|negative|Waitress was a li...|
|negative|This place is not...|
|negative|did not like at all.|
+--------+--------------------+
only showing top 20 rows
```

Run

Started Juno

Interpreter: spark.pyspark. **FINISHED** Took 28 sec 575 millisec. Updated by ars0107 on February 04 2019, 9:13:42 AM (CST)

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

```
+--------+--------------------+------+
|   class|                text|length|
+--------+--------------------+------+
|positive|Wow... Loved this...|    24|
|negative|  Crust is not good.|    18|
|negative|Not tasty and the...|    41|
|positive|Stopped by during...|    87|
|positive|The selection on ...|    59|
|negative|Now I am getting ...|    46|
|negative|Honeslty it didn'...|    37|
|negative|The potatoes were...|   111|
|positive|The fries were gr...|    25|
|positive|      A great touch.|    14|
|positive|Service was very ...|    24|
|negative|  Would not go back.|    18|
|negative|The cashier had n...|    99|
|positive|I tried the Cape ...|    59|
|negative|I was disgusted b...|    62|
|negative|I was shocked bec...|    50|
|positive| Highly recommended.|    19|
|negative|Waitress was a li...|    38|
|negative|This place is not...|    51|
|negative|did not like at all.|    20|
+--------+--------------------+------+
only showing top 20 rows
```

Feature Transformations

```
%pyspark
from pyspark.ml.feature import Tokenizer, StopWordsRemover, HashingTF, IDF, StringIndexer
# Create all the features to the data set
pos_neg_to_num = StringIndexer(inputCol='class',outputCol='label')
tokenizer = Tokenizer(inputCol="text", 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')
```

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

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

```
%pyspark
# Create a and run a data processing Pipeline
from pyspark.ml import Pipeline
data_prep_pipeline = Pipeline(stages=[pos_neg_to_num, tokenizer, stopremove, hashingTF, idf, clean_up])
```

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

```
%pyspark (/U4G66226D/spaces)
# Show label and resulting features
cleaned.select(['label', 'features']).show()
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|(262145,[33933,69...|
|  1.0|(262145,[15889,13...|
|  1.0|(262145,[25570,63...|
|  0.0|(262145,[6286,272...|
|  0.0|(262145,[6979,255...|
|  1.0|(262145,[24417,24...|
|  1.0|(262145,[12084,48...|
|  1.0|(262145,[3645,963...|
|  0.0|(262145,[53777,10...|
|  0.0|(262145,[138356,2...|
|  0.0|(262145,[24113,25...|
|  1.0|(262145,[68867,13...|
|  1.0|(262145,[24417,36...|
|  0.0|(262145,[18098,24...|
|  1.0|(262145,[24417,25...|
|  1.0|(262145,[24417,25...|
|  0.0|(262145,[31704,21...|
|  1.0|(262145,[25570,27...|
|  1.0|(262145,[12329,15...|
|  1.0|(262145,[8287,139...|
+-----+--------------------+
only showing top 20 rows
```

```
%pyspark
from pyspark.ml.classification import NaiveBayes
# Break data down into a training set and a testing set
training, testing = cleaned.randomSplit([0.7, 0.3])

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

Run Started Juno

```
%pyspark
# Tranform the model with the testing data
test_results = predictor.transform(testing)
test_results.show(5)
```

```
+--------+--------------------+------+-----+--------------------+--------------------+--------------------+---------
-----------+--------------------+--------------------+--------------------+----------+
|   class|                text|length|label|          token_text|         stop_tokens|          hash_token|
  idf_token|            features|       rawPrediction|         probability|prediction|
+--------+--------------------+------+-----+--------------------+--------------------+--------------------+---------
-----------+--------------------+--------------------+--------------------+----------+
|negative| "As for the ""mains|    19|  1.0|["as, for, the, "...|      ["as, ""mains]|(262144,[16332,10...|(262144,
[16332,10...|(262145,[16332,10...|[-234.83228868788...|[0.07944294469117...|       1.0|
|negative|"It was extremely...|    51|  1.0|["it, was, extrem...|["it, extremely, ...|(262144,[7388,255...|(262144,
[7388,255...|(262145,[7388,255...|[-494.12578915081...|[3.32950181068878...|       1.0|
|negative|(It wasn't busy e...|    61|  1.0|[(it, wasn't, bus...|[(it, wasn't, bus...|(262144,[329,2101...|(262144,
[329,2101...|(262145,[329,2101...|[-755.20341948045...|[1.36255367304869...|       1.0|
|negative|A FLY was in my a...|    43|  1.0|[a, fly, was, in,...|[fly, apple, juic...|(262144,[25570,37...|(262144,
[25570,37...|(262145,[25570,37...|[-507.67086658065...|[0.03102344148914...|       1.0|
|negative|A greasy, unhealt...|    25|  1.0|[a, greasy,, unhe...|[greasy,, unhealt...|(262144,[47281,16...|(262144,
[47281,16...|(262145,[47281,16...|[-304.95938124860...|[6.53946080952648...|       1.0|
+--------+--------------------+------+-----+--------------------+--------------------+--------------------+---------
-----------+--------------------+--------------------+--------------------+----------+
only showing top 5 rows
```

```
%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: 0.738621
```