TEXT CLASSIFICATION

Dosen Pengampu : Grezio Arifiyan Primajaya S.Kom, M.Kom

EKY FERNANDA | 3322600025

Text Classification with PySpark

MultiClass Text Classification

Task

- predict the subject category given a course title or text

Pyspark

- pipenv install pyspark

```
# Load Pkgs
from pyspark import SparkContext
```

```
sc = SparkContext(master="local[2]")
```

```
# Launch UI
sc
```

SparkContext

Spark UI

Version
v3.5.3
Master
local[2]
AppName
pyspark-shell

```
# Create A Spark Session
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName("TextClassifierwithPySpark").getOrCreate()
```

```
# Load Our Dataset
df = spark.read.csv("udemy_courses_clean.csv",header=True,inferSchema=True)
```

```
df.show()
```

```
+---+---------+--------------------+--------------------+-------+-----+--------------
|_c0|course_id|        course_title|                 url|is_paid|price|num_subscriber
+---+---------+--------------------+--------------------+-------+-----+--------------
```

```
|  0|   1070968|Ultimate Investme...|https://www.udemy...|   True|  200|             214
|  1|   1113822|Complete GST Cour...|https://www.udemy...|   True|   75|             279
|  2|   1006314|Financial Modelin...|https://www.udemy...|   True|   45|             217
|  3|   1210588|Beginner to Pro -...|https://www.udemy...|   True|   95|             245
|  4|   1011058|How To Maximize Y...|https://www.udemy...|   True|  200|             127
|  5|    192870|Trading Penny Sto...|https://www.udemy...|   True|  150|             922
|  6|    739964|Investing And Tra...|https://www.udemy...|   True|   65|             154
|  7|    403100|Trading Stock Cha...|https://www.udemy...|   True|   95|             291
|  8|    476268|Options Trading 3...|https://www.udemy...|   True|  195|             517
|  9|   1167710|The Only Investme...|https://www.udemy...|   True|  200|              82
| 10|    592338|Forex Trading Sec...|https://www.udemy...|   True|  200|             428
| 11|    975046|Trading Options W...|https://www.udemy...|   True|  200|             138
| 12|    742602|Financial Managem...|https://www.udemy...|   True|   30|             360
| 13|    794151|Forex Trading Cou...|https://www.udemy...|   True|  195|             406
| 14|   1196544|Python Algo Tradi...|https://www.udemy...|   True|  200|              29
| 15|    504036|Short Selling: Le...|https://www.udemy...|   True|   75|             227
| 16|    719698|Basic Technical A...|https://www.udemy...|   True|   20|             491
| 17|    564966|The Complete Char...|https://www.udemy...|   True|  200|             266
| 18|    606928|7 Deadly Mistakes...|https://www.udemy...|   True|   50|             535
| 19|     58977|Financial Stateme...|https://www.udemy...|   True|   95|             809
+---+----------+--------------------+--------------------+-------+-----+--------------
only showing top 20 rows
```

```
# Columns
df.columns
```

```
['_c0',
 'course_id',
 'course_title',
 'url',
 'is_paid',
 'price',
 'num_subscribers',
 'num_reviews',
 'num_lectures',
 'level',
 'content_duration',
 'published_timestamp',
 'subject',
 'clean_course_title']
```

```
# Select Columns
df.select('course_title','subject').show()
```

```
+--------------------+---------------+
|        course_title|        subject|
+--------------------+---------------+
|Ultimate Investme...|Business Finance|
|Complete GST Cour...|Business Finance|
|Financial Modelin...|Business Finance|
|Beginner to Pro -...|Business Finance|
|How To Maximize Y...|Business Finance|
```

```
 Trading Penny Sto...|Business Finance|
|Investing And Tra...|Business Finance|
|Trading Stock Cha...|Business Finance|
|Options Trading 3...|Business Finance|
|The Only Investme...|Business Finance|
|Forex Trading Sec...|Business Finance|
|Trading Options W...|Business Finance|
|Financial Managem...|Business Finance|
|Forex Trading Cou...|Business Finance|
|Python Algo Tradi...|Business Finance|
|Short Selling: Le...|Business Finance|
|Basic Technical A...|Business Finance|
|The Complete Char...|Business Finance|
|7 Deadly Mistakes...|Business Finance|
|Financial Stateme...|Business Finance|
+--------------------+---------------+
only showing top 20 rows
```

```
df = df.select('course_title','subject')
```

```
df.show(5)
```

```
+--------------------+---------------+
|        course_title|         subject|
+--------------------+---------------+
|Ultimate Investme...|Business Finance|
|Complete GST Cour...|Business Finance|
|Financial Modelin...|Business Finance|
|Beginner to Pro -...|Business Finance|
|How To Maximize Y...|Business Finance|
+--------------------+---------------+
only showing top 5 rows
```

```
# Value Counts
df.groupBy('subject').count().show()
```

```
+--------------------+-----+
|             subject|count|
+--------------------+-----+
|play Electric Gui...|    1|
|Multiply returns ...|    1|
|                NULL|    6|
|    Business Finance| 1198|
|Introduction Guit...|    1|
|Learn Play Fernan...|    1|
|      Graphic Design|  603|
|Aprende tocar el ...|    1|
|     Web Development| 1200|
|Learn Classical G...|    1|
| Musical Instruments|  676|
+--------------------+-----+
```

```
# Value Counts via pandas
df.toPandas()['subject'].value_counts()
```

subject
Web Development
Business Finance
Musical Instruments
Graphic Design
Multiply returns Value Investinghttpswwwudemycommultiplyyourreturnsusingvalueinvestingtrue2019421963All Levels45 hours20150723T000833Z 874284Weekly Forex Analysis Baraq FX
Learn Play Fernando Sors Study B minorhttpswwwudemycomstudyinbminortrue115140359Intermediate Level43 mins20140127T205816Z 398746Piano Chord Based System Learn Play Pros Do
play Electric Guitarhttpswwwudemycomelectricguitarbeginnersmethodtrue501105520Beginner Level2 hours20161229T002406Z 42038Learn Piano Today Play Piano Course Quick Lessons
Learn Classical Guitar Technique play Spanish Romancehttpswwwudemycomguitartechniquestrue19513164643All Levels5 hours20131118T175959Z 265888Learn Guitar Worship Learn 4 Songs unlock 1

```
# Check For Missing Values
df.toPandas()['subject'].isnull().sum()
```

```
6
```

```
# Drop Missing Values
df = df.dropna(subset=('subject'))
```

```
# Check For Missing Values
df.toPandas()['subject'].isnull().sum()
```

```
0
```

```
df.show(5)
```

```
+--------------------+---------------+
|        course_title|         subject|
+--------------------+---------------+
|Ultimate Investme...|Business Finance|
```

```
|Complete GST Cour...|Business Finance|
|Financial Modelin...|Business Finance|
|Beginner to Pro -...|Business Finance|
|How To Maximize Y...|Business Finance|
+--------------------+ --------------- +
only showing top 5 rows
```

Feature Extraction

Build Features From Text

- CountVectorizer
- TFIDF
- WordEmbedding
- HashingTF
- etc

```
# Load Our Pkgs
import pyspark.ml.feature
```

```
dir(pyspark.ml.feature)
```

```
['Any',
 'Binarizer',
 'BucketedRandomProjectionLSH',
 'BucketedRandomProjectionLSHModel',
 'Bucketizer',
 'ChiSqSelector',
 'ChiSqSelectorModel',
 'CountVectorizer',
 'CountVectorizerModel',
 'DCT',
 'DataFrame',
 'DenseMatrix',
 'DenseVector',
 'Dict',
 'ElementwiseProduct',
 'FeatureHasher',
 'Generic',
 'HasFeaturesCol',
 'HasHandleInvalid',
 'HasInputCol',
 'HasInputCols',
 'HasLabelCol',
 'HasMaxIter',
 'HasNumFeatures',
 'HasOutputCol',
 'HasOutputCols',
```

```
 'HasRelativeError',
 'HasSeed',
 'HasStepSize',
 'HasThreshold',
 'HasThresholds',
 'HashingTF',
 'IDF',
 'IDFModel',
 'Imputer',
 'ImputerModel',
 'IndexToString',
 'Interaction',
 'JM',
 'JavaEstimator',
 'JavaMLReadable',
 'JavaMLWritable',
 'JavaModel',
 'JavaParams',
 'JavaTransformer',
 'List',
 'MaxAbsScaler',
 'MaxAbsScalerModel',
 'MinHashLSH',
 'MinHashLSHModel',
 'MinMaxScaler',
 'MinMaxScalerModel',
 'NGram',
 'Normalizer',
 'OneHotEncoder',
 'OneHotEncoderModel',
 'Optional',
 'P',
```

```
# Load Our Transformer & Extractor Pkgs
from pyspark.ml.feature import Tokenizer,StopWordsRemover,CountVectorizer,IDF
from pyspark.ml.feature import StringIndexer
```

```
df.show(5)
```

```
+--------------------+---------------+
|        course_title|         subject|
+--------------------+---------------+
|Ultimate Investme...|Business Finance|
|Complete GST Cour...|Business Finance|
|Financial Modelin...|Business Finance|
|Beginner to Pro -...|Business Finance|
|How To Maximize Y...|Business Finance|
+--------------------+---------------+
only showing top 5 rows
```

```
# Stages For the Pipeline
tokenizer = Tokenizer(inputCol= course_title  outputCol= mytokens )
```

```
stopwords_remover = StopWordsRemover(inputCol='mytokens',outputCol='filtered_tokens')
vectorizer = CountVectorizer(inputCol='filtered_tokens',outputCol='rawFeatures')
idf = IDF(inputCol='rawFeatures',outputCol='vectorizedFeatures')
```

```
# LabelEncoding/LabelIndexing
labelEncoder = StringIndexer(inputCol='subject',outputCol='label').fit(df)
```

```
labelEncoder.transform(df).show(5)
```

```
+--------------------+----------------+-----+
|        course_title|         subject|label|
+--------------------+----------------+-----+
|Ultimate Investme...|Business Finance|  1.0|
|Complete GST Cour...|Business Finance|  1.0|
|Financial Modelin...|Business Finance|  1.0|
|Beginner to Pro -...|Business Finance|  1.0|
|How To Maximize Y...|Business Finance|  1.0|
+--------------------+----------------+-----+
only showing top 5 rows
```

```
labelEncoder.labels
```

```
['Web Development',
 'Business Finance',
 'Musical Instruments',
 'Graphic Design',
 'Aprende tocar el Acorden de odo con
tcnicahttpswwwudemycomaprendeatocarelacordeondeoidoycontecnicatrue25932134Beginner
Level4 hours20140916T195145Z 263432Aprende los Secretos de la Armnica con HARPSOUL',
'Introduction Guitar A Course
Beginnershttpswwwudemycomintroductiontoguitartrue251631156Beginner Level25
hours20141030T155939Z 650804Guitar Master Class Learning Play Guitar Z',
 'Learn Classical Guitar Technique play Spanish
RomancehttpswwwudemycomguitartechniquestrueI9513164643All Levels5
hours20131118T175959Z 265888Learn Guitar Worship Learn 4 Songs unlock 1',
'Learn Play Fernando Sors Study B
minorhttpswwwudemycomstudyinbminortrue115140359Intermediate Level43
mins20140127T205816Z 398746Piano Chord Based System  Learn Play Pros Do',
'Multiply returns Value
Investinghttpswwwudemycommultiplyyourreturnsusingvalueinvestingtrue2019421963All
Levels45 hours20150723T000833Z 874284Weekly Forex Analysis Baraq FX',
 'play Electric
Guitarhttpswwwudemycomelectricguitarbeginnersmethodtrue501105520Beginner Level2
hours20161229T002406Z 42038Learn Piano Today Play Piano Course Quick Lessons']
```

```
# Dict of Labels
label_dict = {'Web Development':0.0,
```

```
 'Business Finance':1.0,
 'Musical Instruments':2.0,
 'Graphic Design':3.0}
```

```
df.show()
```

```
+--------------------+---------------+
|        course_title|         subject|
+--------------------+---------------+
|Ultimate Investme...|Business Finance|
|Complete GST Cour...|Business Finance|
|Financial Modelin...|Business Finance|
|Beginner to Pro -...|Business Finance|
|How To Maximize Y...|Business Finance|
|Trading Penny Sto...|Business Finance|
|Investing And Tra...|Business Finance|
|Trading Stock Cha...|Business Finance|
|Options Trading 3...|Business Finance|
|The Only Investme...|Business Finance|
|Forex Trading Sec...|Business Finance|
|Trading Options W...|Business Finance|
|Financial Managem...|Business Finance|
|Forex Trading Cou...|Business Finance|
|Python Algo Tradi...|Business Finance|
|Short Selling: Le...|Business Finance|
|Basic Technical A...|Business Finance|
|The Complete Char...|Business Finance|
|7 Deadly Mistakes...|Business Finance|
|Financial Stateme...|Business Finance|
+--------------------+---------------+
only showing top 20 rows
```

```
df = labelEncoder.transform(df)
```

```
df.show(5)
```

```
+--------------------+---------------+-----+
|        course_title|         subject|label|
+--------------------+---------------+-----+
|Ultimate Investme...|Business Finance|  1.0|
|Complete GST Cour...|Business Finance|  1.0|
|Financial Modelin...|Business Finance|  1.0|
|Beginner to Pro -...|Business Finance|  1.0|
|How To Maximize Y...|Business Finance|  1.0|
+--------------------+---------------+-----+
only showing top 5 rows
```

```
### Split Dataset
(trainDF,testDF) = df.randomSplit((0.7,0.3),seed=42)
```

```
trainDF.show()
```

```
+--------------------+--------------------+-----+
|        course_title|             subject|label|
+--------------------+--------------------+-----+
|#1 Piano Hand Coo...| Musical Instruments|  2.0|
|#10 Hand Coordina...| Musical Instruments|  2.0|
|#4 Piano Hand Coo...| Musical Instruments|  2.0|
|#5  Piano Hand Co...| Musical Instruments|  2.0|
|#6 Piano Hand Coo...| Musical Instruments|  2.0|
|'Geometry Of Chan...|    Business Finance|  1.0|
|              000!""|Learn Classical G...|  6.0|
|1 - Concepts of S...|    Business Finance|  1.0|
|          1 Hour CSS|     Web Development|  0.0|
|1. Principles of ...|    Business Finance|  1.0|
|10 Numbers Every ...|    Business Finance|  1.0|
|10.  Bonds and Bo...|    Business Finance|  1.0|
|101 Blues riffs -...| Musical Instruments|  2.0|
|15 Mandamientos p...|    Business Finance|  1.0|
|17 Complete JavaS...|     Web Development|  0.0|
|188% Profit in 1Y...|    Business Finance|  1.0|
|2 Easy Steps To I...|    Business Finance|  1.0|
|3 step formula fo...| Musical Instruments|  2.0|
|30 Day Guitar Jum...| Musical Instruments|  2.0|
|3DS MAX - Learn 3...|      Graphic Design|  3.0|
+--------------------+--------------------+-----+
only showing top 20 rows
```

```
### Estimator
from pyspark.ml.classification import LogisticRegression
```

```
lr = LogisticRegression(featuresCol='vectorizedFeatures',labelCol='label')
```

* Building the Pipeline

```
from pyspark.ml import Pipeline
```

```
pipeline = Pipeline(stages=[tokenizer,stopwords_remover,vectorizer,idf,lr])
```

```
pipeline
```

```
Pipeline_0db9131831a6
```

```
pipeline.stages
```

```
Param(parent='Pipeline_0db9131831a6', name='stages', doc='a list of pipeline
stages')
```

```
# Building MOdel
lr_model = pipeline.fit(trainDF)
```

```
lr_model
```

```
PipelineModel_393d44cb9f8b
```

```
# Predictions on our Test Dataset
predictions = lr_model.transform(testDF)
```

```
predictions.show()
```

```
+--------------------+--------------------+-----+--------------------+----------------
|        course_title|             subject|label|            mytokens|     filtered_to
+--------------------+--------------------+-----+--------------------+----------------
|#12 Hand Coordina...|Musical Instruments|  2.0|[#12, hand, coord...|[#12, hand, coor
|#7 Piano Hand Coo...|Musical Instruments|  2.0|[#7, piano, hand,...|[#7, piano, hand
|'Greensleeves' Cr...|Musical Instruments|  2.0|['greensleeves', ...|['greensleeves',
|* An Integrated A...|   Business Finance|  1.0|[*, an, integrate...|[*, integrated,
|         1 Hour HTML|    Web Development|  0.0|     [1, hour, html]|     [1, hour, h
|   1 Hour JavaScript|    Web Development|  0.0|[1, hour, javascr...|[1, hour, javasc
|       1 hour jQuery|    Web Development|  0.0|   [1, hour, jquery]|   [1, hour, jqu
|101 Awesome Rocka...|Musical Instruments|  2.0|[101, awesome, ro...|[101, awesome, r
|15  Motion Graphi...|     Graphic Design|  3.0|[15, , motion, gr...|[15, , motion, g
|150 Rock Guitar L...|Musical Instruments|  2.0|[150, rock, guita...|[150, rock, guit
|16 Guitar Chords ...|Musical Instruments|  2.0|[16, guitar, chor...|[16, guitar, cho
|2. Principles of ...|   Business Finance|  1.0|[2., principles, ...|[2., principles,
|3 Little Pigs: A ...|   Business Finance|  1.0|[3, little, pigs:...|[3, little, pigs
|3 documentos clav...|   Business Finance|  1.0|[3, documentos, c...|[3, documentos,
|3. Compound Inter...|   Business Finance|  1.0|[3., compound, in...|[3., compound, i
|31 Day Guitar Cha...|Musical Instruments|  2.0|[31, day, guitar,...|[31, day, guitar
|3D Programming wi...|    Web Development|  0.0|[3d, programming,...|[3d, programming
|4. Ordinary Simpl...|   Business Finance|  1.0|[4., ordinary, si...|[4., ordinary, s
|5 lecciones que t...|Musical Instruments|  2.0|[5, lecciones, qu...|[5, lecciones, q
|6 Must Know Trick...|Musical Instruments|  2.0|[6, must, know, t...|[6, must, know,
+--------------------+--------------------+-----+--------------------+----------------
only showing top 20 rows
```

```
# Select Columns
predictions.columns
```

```
['course_title',
 'subject',
 'label',
 'mytokens',
 'filtered_tokens',
```

```
 'rawFeatures',
 'vectorizedFeatures',
 'rawPrediction',
 'probability',
 'prediction']
```

```
predictions.select('rawPrediction','probability','subject','label','prediction').show(10)
```

```
+--------------------+--------------------+-------------------+-----+----------+
|       rawPrediction|         probability|            subject|label|prediction|
+--------------------+--------------------+-------------------+-----+----------+
|[5.54102657075933...|[0.28268355883716...|Musical Instruments|  2.0|       2.0|
|[-6.0823322910413...|[1.03215808700201...|Musical Instruments|  2.0|       2.0|
|[-1.0421312508398...|[1.23270089236449...|Musical Instruments|  2.0|       2.0|
|[-2.8211817263258...|[4.77760514763108...|    Business Finance|  1.0|       1.0|
|[21.7088196951149...|[0.99999998997280...|     Web Development|  0.0|       0.0|
|[20.0054038056691...|[0.99999996031555...|     Web Development|  0.0|       0.0|
|[18.0923732218416...|[0.99999972290786...|     Web Development|  0.0|       0.0|
|[-8.4472424793961...|[4.30624170780055...|Musical Instruments|  2.0|       2.0|
|[-24.070755043493...|[7.95180496786136...|     Graphic Design|  3.0|       3.0|
|[-6.0537484580850...|[8.42550607102213...|Musical Instruments|  2.0|       2.0|
+--------------------+--------------------+-------------------+-----+----------+
only showing top 10 rows
```

```
# ### Model Evaluation
# + Accuracy
# # + Precision
# + F1score
# + etc
```

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

```
evaluator = MulticlassClassificationEvaluator(labelCol='label',predictionCol='prediction'
```

```
accuracy = evaluator.evaluate(predictions)
```

```
accuracy
```

```
0.9173003802281369
```

```
#### Method 2: Precision. F1Score (Classification Report)
from pyspark.mllib.evaluation import MulticlassMetrics
```

```
lr_metric = MulticlassMetrics(predictions['label','prediction'].rdd)
```

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/context.py:158: FutureWarning: D
```

```
  warnings.warn(
```

```
print("Accuracy:",lr_metric.accuracy)
print("Precision:",lr_metric.precision(1.0))
print("Recall:",lr_metric.recall(1.0))
print("F1Score:",lr_metric.fMeasure(1.0))
```

```
Accuracy: 0.9173003802281369
Precision: 0.9686609686609686
Recall: 0.8717948717948718
F1Score: 0.9176788124156545
```

Confusion Matrix

- convert to pandas
- sklearn

```
y_true = predictions.select('label')
y_true = y_true.toPandas()
y_pred = predictions.select('prediction')
y_pred = y_pred.toPandas()
```

```
from sklearn.metrics import confusion_matrix,classification_report
```

```
cm = confusion_matrix(y_true,y_pred)
```

```
cm
```

```
array([[321,  10,   1,   4,   0,   0],
       [  9, 340,   1,   1,   0,   0],
       [  6,  15, 156,   1,   0,   0],
       [  8,  23,   6, 148,   0,   0],
       [  0,   1,   0,   0,   0,   0],
       [  0,   1,   0,   0,   0,   0]])
```

```
import matplotlib.pyplot as plt
import numpy as np
import itertools

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    """
    This function prints and plots the confusion matrix.
```

```
    Normalization can be applied by setting `normalize=True`.
    """
    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    print(cm)

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    fmt = '.2f' if normalize else 'd'
    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, format(cm[i, j], fmt),
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
```

```
label_dict.keys()
```

```
dict_keys(['Web Development', 'Business Finance', 'Musical Instruments', 'Graphic
Design'])
```

```
class_names = ['Web Development', 'Business Finance', 'Musical Instruments', 'Graphic Des
```

```
plot_confusion_matrix(cm,class_names)
```

```
import warnings
warnings.filterwarnings('ignore')
```

```
# Classication Report
print(classification_report(y_true,y_pred))
```

```
                   precision    recall  f1-score   support

              0.0       0.93      0.96      0.94       336
              1.0       0.87      0.97      0.92       351
              2.0       0.95      0.88      0.91       178
              3.0       0.96      0.80      0.87       185
              5.0       0.00      0.00      0.00         1
              8.0       0.00      0.00      0.00         1

         accuracy                           0.92      1052
        macro avg       0.62      0.60      0.61      1052
     weighted avg       0.92      0.92      0.92      1052
```

```
# Classication Report
print(classification_report(y_true,y_pred,target_names=class_names))
```

```
                         precision    recall  f1-score   support

          Web Development       0.93      0.96      0.94       336
        Business Finance        0.87      0.97      0.92       351
     Musical Instruments        0.95      0.88      0.91       178
```

```
          Graphic Design       0.96      0.80      0.87       185
                      N4       0.00      0.00      0.00         1
                      N5       0.00      0.00      0.00         1

                accuracy                           0.92      1052
               macro avg       0.62      0.60      0.61      1052
            weighted avg       0.92      0.92      0.92      1052
```

```
class_temp = predictions.select("label").groupBy("label")\
                        .count().sort('count', ascending=False).toPandas()
class_temp = class_temp["label"].values.tolist()
class_names = map(str, class_temp)
# # # print(class_name)
class_names
```

```
<map at 0x780b59c9cdc0>
```

Making Single Prediction

- sample as DF
- apply pipeline

```
from pyspark.sql.types import StringType
```

```
ex1 = spark.createDataFrame([
    ("Building Machine Learning Apps with Python and PySpark",StringType())
],
# Column Name
["course_title"]

)
```

```
ex1.show()
```

```
+--------------------+---+
|        course_title| _2|
+--------------------+---+
|Building Machine ...| {}|
+--------------------+---+
```

```
# Show Full
ex1.show(truncate=False)
```

```
+--------------------------------------------------+---+
```

```
|course_title                                          |_2 |
+------------------------------------------------------+---+
|Building Machine Learning Apps with Python and PySpark|{} |
+------------------------------------------------------+---+
```

```
# Predict
pred_ex1 = lr_model.transform(ex1)
```

```
pred_ex1.show()
```

```
+--------------------+---+--------------------+--------------------+----------------
|        course_title| _2|            mytokens|     filtered_tokens|         rawFeatu
+--------------------+---+--------------------+--------------------+----------------
|Building Machine ...| {}|[building, machin...|[building, machin...|(3670,[56,80,113,
+--------------------+---+--------------------+--------------------+----------------
```

```
pred_ex1.columns
```

```
['course_title',
 '_2',
 'mytokens',
 'filtered_tokens',
 'rawFeatures',
 'vectorizedFeatures',
 'rawPrediction',
 'probability',
 'prediction']
```

```
pred_ex1.select('course_title','rawPrediction','probability','prediction').show()
```

```
+--------------------+--------------------+--------------------+ --------- +
|        course_title|       rawPrediction|         probability|prediction|
+--------------------+--------------------+--------------------+ --------- +
|Building Machine ...|[14.1592174201726...|[0.99999485486236...|       0.0|
+--------------------+--------------------+--------------------+ --------- +
```

```
label_dict
```

```
{'Web Development': 0.0,
 'Business Finance': 1.0,
 'Musical Instruments': 2.0,
 'Graphic Design': 3.0}
```

```
### Save and Load Model
```

```
# Saving Model
```

```
modelPath = "models/pyspark_lr_model_26_Feb_2021"
lr_model.save(modelPath)
```

```
# Loading pickled model via pipeline api
from pyspark.ml.pipeline import PipelineModel
persistedModel = PipelineModel.load(modelPath)
```