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 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("data/udemy_courses_clean.csv",header=True,inferSchema=
True)

df.show()
```

```
+---+---------+--------------------+--------------------+-------+-----
+---------------+-----------+------------+------------------
+----------------+--------------------+----------------
+--------------------+
|_c0|course_id|        course_title|                 url|is_paid|
price|num_subscribers|num_reviews|num_lectures|             level|
content_duration| published_timestamp|         subject|
clean_course_title|
+---+---------+--------------------+--------------------+-------+-----
+---------------+-----------+------------+------------------
+----------------+--------------------+----------------
+--------------------+
|  0|  1070968|Ultimate Investme...|https://www.udemy...|   True|
200|           2147|         23|          51|        All Levels|
1.5 hours|2017-01-18T20:58:58Z|Business Finance|Ultimate Investme...|
|  1|  1113822|Complete GST Cour...|https://www.udemy...|   True|
75|           2792|        923|         274|        All Levels|
```

```
39 hours|2017-03-09T16:34:20Z|Business Finance|Complete GST Cour...|
|  2|  1006314|Financial Modelin...|https://www.udemy...|   True|
45|           2174|         74|          51|Intermediate Level|
2.5 hours|2016-12-19T19:26:30Z|Business Finance|Financial Modelin...|
|  3|  1210588|Beginner to Pro -...|https://www.udemy...|   True|
95|           2451|         11|          36|        All Levels|
3 hours|2017-05-30T20:07:24Z|Business Finance|Beginner Pro  Fin...|
|  4|  1011058|How To Maximize Y...|https://www.udemy...|   True|
200|           1276|         45|          26|Intermediate Level|
2 hours|2016-12-13T14:57:18Z|Business Finance|Maximize Profits ...|
|  5|   192870|Trading Penny Sto...|https://www.udemy...|   True|
150|           9221|        138|          25|        All Levels|
3 hours|2014-05-02T15:13:30Z|Business Finance|Trading Penny Sto...|
|  6|   739964|Investing And Tra...|https://www.udemy...|   True|
65|           1540|        178|          26|    Beginner Level|
1 hour|2016-02-21T18:23:12Z|Business Finance|Investing Trading...|
|  7|   403100|Trading Stock Cha...|https://www.udemy...|   True|
95|           2917|        148|          23|        All Levels|
2.5 hours|2015-01-30T22:13:03Z|Business Finance|Trading Stock Cha...|
|  8|   476268|Options Trading 3...|https://www.udemy...|   True|
195|           5172|         34|          38|      Expert Level|
2.5 hours|2015-05-28T00:14:03Z|Business Finance|Options Trading 3...|
|  9|  1167710|The Only Investme...|https://www.udemy...|   True|
200|            827|         14|          15|        All Levels|
1 hour|2017-04-18T18:13:32Z|Business Finance|Investment Strate...|
| 10|   592338|Forex Trading Sec...|https://www.udemy...|   True|
200|           4284|         93|          76|        All Levels|
5 hours|2015-09-11T16:47:02Z|Business Finance|Forex Trading Sec...|
| 11|   975046|Trading Options W...|https://www.udemy...|   True|
200|           1380|         42|          17|        All Levels|
1 hour|2016-10-18T22:52:31Z|Business Finance|Trading Options M...|
| 12|   742602|Financial Managem...|https://www.udemy...|   True|
30|           3607|         21|          19|        All Levels|
1.5 hours|2016-02-03T18:04:01Z|Business Finance|Financial Managem...|
| 13|   794151|Forex Trading Cou...|https://www.udemy...|   True|
195|           4061|         52|          16|        All Levels|
2 hours|2016-03-16T15:40:19Z|Business Finance|Forex Trading Cou...|
| 14|  1196544|Python Algo Tradi...|https://www.udemy...|   True|
200|            294|         19|          42|        All Levels|
7 hours|2017-04-28T16:41:44Z|Business Finance|Python Algo Tradi...|
| 15|   504036|Short Selling: Le...|https://www.udemy...|   True|
75|           2276|        106|          19|Intermediate Level|
1.5 hours|2015-06-22T21:18:35Z|Business Finance|Short Selling Lea...|
| 16|   719698|Basic Technical A...|https://www.udemy...|   True|
20|           4919|         79|          16|    Beginner Level|
1.5 hours|2016-01-08T17:21:26Z|Business Finance|Basic Technical A...|
| 17|   564966|The Complete Char...|https://www.udemy...|   True|
200|           2666|        115|          52|        All Levels|
4 hours|2015-08-10T21:07:35Z|Business Finance|Complete Chart Pa...|
```

```
| 18|   606928|7 Deadly Mistakes...|https://www.udemy...|   True|
50|          5354|         24|          23|        All Levels|
1.5 hours|2015-09-21T18:10:34Z|Business Finance|7 Deadly Mistakes...|
| 19|    58977|Financial Stateme...|https://www.udemy...|   True|
95|          8095|        249|          12|    Beginner Level|
35 mins|2013-06-09T00:21:26Z|Business Finance|Financial Stateme...|
+---+---------+--------------------+--------------------+-------+-----
+--------------+----------+-----------+-----------------
+----------------+--------------------+---------------
+--------------------+
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()
```

```
Web Development
1200
Business Finance
1198
Musical Instruments
676
Graphic Design
603
Introduction Guitar A Course
BeginnershttpswwwudemycomintroductiontoguitartrueÂ 251631156Beginner
Level25 hours20141030T155939Z 650804Guitar Master Class Learning Play
Guitar Z                                    1
Learn Play Fernando Sors Study B
minorhttpswwwudemycomstudyinbminortrueÂ 115140359Intermediate Level43
mins20140127T205816Z 398746Piano Chord Based System  Learn Play Pros
Do                                        1
Multiply returns Value
Investinghttpswwwudemycommultiplyyourreturnsusingvalueinvestingtrue201
9421963All Levels45 hours20150723T000833Z 874284Weekly Forex Analysis
Baraq FX                                        1
Aprende tocar el Acorden de odo con
tcnicahttpswwwudemycomaprendeatocarelacordeondeoidoycontecnicatrue2593
2134Beginner Level4 hours20140916T195145Z 263432Aprende los Secretos
de la Armnica con HARPSOUL       1
Learn Classical Guitar Technique play Spanish
RomancehttpswwwudemycomguitartechniquestrueÂ 19513164643All Levels5
hours20131118T175959Z 265888Learn Guitar Worship Learn 4 Songs unlock
1                          1
play Electric
GuitarhttpswwwudemycomelectricguitarbeginnersmethodtrueÂ 501105520Beginn
er Level2 hours20161229T002406Z 42038Learn Piano Today Play Piano
Course Quick Lessons                                      1
Name: subject, dtype: int64

# 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)
```

```
['Binarizer',
 'BucketedRandomProjectionLSH',
 'BucketedRandomProjectionLSHModel',
 'Bucketizer',
 'ChiSqSelector',
 'ChiSqSelectorModel',
 'CountVectorizer',
 'CountVectorizerModel',
 'DCT',
 'ElementwiseProduct',
 'FeatureHasher',
 'HasAggregationDepth',
 'HasBlockSize',
 'HasCheckpointInterval',
 'HasCollectSubModels',
 'HasDistanceMeasure',
 'HasElasticNetParam',
 'HasFeaturesCol',
 'HasFitIntercept',
 'HasHandleInvalid',
 'HasInputCol',
 'HasInputCols',
 'HasLabelCol',
 'HasLoss',
 'HasMaxIter',
```

```
 'HasNumFeatures',
 'HasOutputCol',
 'HasOutputCols',
 'HasParallelism',
 'HasPredictionCol',
 'HasProbabilityCol',
 'HasRawPredictionCol',
 'HasRegParam',
 'HasRelativeError',
 'HasSeed',
 'HasSolver',
 'HasStandardization',
 'HasStepSize',
 'HasThreshold',
 'HasThresholds',
 'HasTol',
 'HasValidationIndicatorCol',
 'HasVarianceCol',
 'HasWeightCol',
 'HashingTF',
 'IDF',
 'IDFModel',
 'Imputer',
 'ImputerModel',
 'IndexToString',
 'Interaction',
 'JavaEstimator',
 'JavaMLReadable',
 'JavaMLWritable',
 'JavaModel',
 'JavaParams',
 'JavaTransformer',
 'MaxAbsScaler',
 'MaxAbsScalerModel',
 'MinHashLSH',
 'MinHashLSHModel',
 'MinMaxScaler',
 'MinMaxScalerModel',
 'NGram',
 'Normalizer',
 'OneHotEncoder',
 'OneHotEncoderModel',
 'PCA',
 'PCAModel',
 'Param',
 'Params',
 'PolynomialExpansion',
 'QuantileDiscretizer',
 'RFormula',
```

```
 'RFormulaModel',
 'RegexTokenizer',
 'RobustScaler',
 'RobustScalerModel',
 'SQLTransformer',
 'SparkContext',
 'StandardScaler',
 'StandardScalerModel',
 'StopWordsRemover',
 'StringIndexer',
 'StringIndexerModel',
 'Tokenizer',
 'TypeConverters',
 'VectorAssembler',
 'VectorIndexer',
 'VectorIndexerModel',
 'VectorSizeHint',
 'VectorSlicer',
 'Word2Vec',
 'Word2VecModel',
 '_BucketedRandomProjectionLSHParams',
 '_ChiSqSelectorParams',
 '_CountVectorizerParams',
 '_IDFParams',
 '_ImputerParams',
 '_LSH',
 '_LSHModel',
 '_LSHParams',
 '_MaxAbsScalerParams',
 '_MinMaxScalerParams',
 '_OneHotEncoderParams',
 '_PCAParams',
 '_RFormulaParams',
 '_RobustScalerParams',
 '_StandardScalerParams',
 '_StringIndexerParams',
 '_VectorIndexerParams',
 '_Word2VecParams',
 '__all__',
 '__builtins__',
 '__cached__',
 '__doc__',
 '__file__',
 '__loader__',
 '__name__',
 '__package__',
 '__spec__',
 '_convert_to_vector',
 '_jvm',
```

```
 'basestring',
 'ignore_unicode_prefix',
 'inherit_doc',
 'keyword_only',
 'since',
 'sys']
```

```
# 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
tcnicahttpswwwudemycomaprendeatocarelacordeondeoidoycontecnicatrue2593
2134Beginner Level4 hours20140916T195145Z 263432Aprende los Secretos
de la Armnica con HARPSOUL',
 'Introduction Guitar A Course
BeginnershttpswwwudemycomintroductiontoguitartrueÂ251631156Beginner
Level25 hours20141030T155939Z 650804Guitar Master Class Learning Play
Guitar Z',
 'Learn Classical Guitar Technique play Spanish
RomancehttpswwwudemycomguitartechniquestrueÂ19513164643All Levels5
hours20131118T175959Z 265888Learn Guitar Worship Learn 4 Songs unlock
1',
 'Learn Play Fernando Sors Study B
minorhttpswwwudemycomstudyinbminortrueÂ115140359Intermediate Level43
mins20140127T205816Z 398746Piano Chord Based System  Learn Play Pros
Do',
 'Multiply returns Value
Investinghttpswwwudemycommultiplyyourreturnsusingvalueinvestingtrue201
9421963All Levels45 hours20150723T000833Z 874284Weekly Forex Analysis
Baraq FX',
 'play Electric
GuitarhttpswwwudemycomelectricguitarbeginnersmethodtrueÂ501105520Beginn
er 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_b97e00946095

pipeline.stages

Param(parent='Pipeline_b97e00946095', name='stages', doc='a list of
pipeline stages')

# Building MOdel
lr_model = pipeline.fit(trainDF)

lr_model

PipelineModel_1875f1057964

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

predictions.show()

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

```
+--------------------+--------------------+----------+
|        course_title|             subject|label|            mytokens|
filtered_tokens|         rawFeatures|  vectorizedFeatures|
rawPrediction|         probability|prediction|
+--------------------+--------------------+-----+--------------------
+--------------------+--------------------+--------------------
+--------------------+--------------------+----------+
|#12 Hand Coordina...|Musical Instruments|  2.0|[#12, hand, coord...|
[#12, hand, coord...|(3670,[394,491,60...|(3670,[394,491,60...|
[8.22575678849003...|[0.86083740538013...|       0.0|
|#7 Piano Hand Coo...|Musical Instruments|  2.0|[#7, piano, hand,...|
[#7, piano, hand,...|(3670,[9,13,60,23...|(3670,[9,13,60,23...|[-
1.5816511969981...|[6.40379189870091...|       2.0|
|'Greensleeves' Cr...|Musical Instruments|  2.0|['greensleeves', ...|
['greensleeves', ...|(3670,[6,9,45,375...|(3670,[6,9,45,375...|
[0.38747123626564...|[1.29430064456987...|       2.0|
|* An Integrated A...|   Business Finance|  1.0|[*, an, integrate...|
[*, integrated, a...|(3670,[23,75,435,...|(3670,[23,75,435,...|[-
2.0540053505355...|[3.67476794956146...|       1.0|
|         1 Hour HTML|    Web Development|  0.0|     [1, hour, html]|
[1, hour, html]|(3670,[24,36,110]...|(3670,[24,36,110]...|
[24.7266193282529...|[0.99999999908079...|       0.0|
|   1 Hour JavaScript|    Web Development|  0.0|[1, hour, javascr...|
[1, hour, javascr...|(3670,[18,36,110]...|(3670,[18,36,110]...|
[22.2213462251437...|[0.99999999175336...|       0.0|
|       1 hour jQuery|    Web Development|  0.0|   [1, hour, jquery]|
[1, hour, jquery]|(3670,[36,62,110]...|(3670,[36,62,110]...|
[20.1005546377385...|[0.99999995838555...|       0.0|
|101 Awesome Rocka...|Musical Instruments|  2.0|[101, awesome, ro...|
[101, awesome, ro...|(3670,[7,233,291,...|(3670,[7,233,291,...|[-
5.9910327938499...|[2.64083766762944...|       2.0|
|15  Motion Graphi...|     Graphic Design|  3.0|[15, , motion, gr...|
[15, , motion, gr...|(3670,[35,90,434,...|(3670,[35,90,434,...|[-
19.729920863390...|[4.16984026967754...|       3.0|
|150 Rock Guitar L...|Musical Instruments|  2.0|[150, rock, guita...|
[150, rock, guita...|(3670,[7,145,175,...|(3670,[7,145,175,...|[-
2.6725325296694...|[9.29048167255554...|       2.0|
|16 Guitar Chords ...|Musical Instruments|  2.0|[16, guitar, chor...|
[16, guitar, chor...|(3670,[0,7,129,17...|(3670,[0,7,129,17...|[-
4.2209408441671...|[6.16872666903649...|       2.0|
|2. Principles of ...|   Business Finance|  1.0|[2., principles, ...|
[2., principles, ...|(3670,[0,41,102,3...|(3670,[0,41,102,3...|
[0.30936773295917...|[4.12860070994016...|       1.0|
|3 Little Pigs: A ...|   Business Finance|  1.0|[3, little, pigs:...|
[3, little, pigs:...|(3670,[2,11,60,14...|(3670,[2,11,60,14...|[-
7.0300584542586...|[1.45078790165638...|       1.0|
|3 documentos clav...|   Business Finance|  1.0|[3, documentos, c...|
[3, documentos, c...|(3670,[60,89,165,...|(3670,[60,89,165,...|
[5.45115805766104...|[0.06838618804257...|       1.0|
```

```
|3. Compound Inter...|   Business Finance|  1.0|[3., compound, in...|
[3., compound, in...| (3670,[1092],[1.0])|(3670,[1092],[6.7...|
[2.27499356707493...|[1.84395934235043...|       1.0|
|31 Day Guitar Cha...|Musical Instruments|  2.0|[31, day, guitar,...|
[31, day, guitar,...|(3670,[7,112,1870...|(3670,[7,112,1870...|[-
7.2943613577218...|[3.39187169125666...|       2.0|
|3D Programming wi...|    Web Development|  0.0|[3d, programming,...|
[3d, programming,...|(3670,[4,87,339],...|(3670,[4,87,339],...|
[10.9590754768583...|[0.92279982494833...|       0.0|
|4. Ordinary Simpl...|   Business Finance|  1.0|[4., ordinary, si...|
[4., ordinary, si...|(3670,[38,102],[1...|(3670,[38,102],[3...|
[3.28529702359769...|[0.00187020497933...|       1.0|
|5 lecciones que t...|Musical Instruments|  2.0|[5, lecciones, qu...|
[5, lecciones, qu...|(3670,[82,3515],[...|(3670,[82,3515],[...|
[6.8754950036276,...|[0.90133433872712...|       0.0|
|6 Must Know Trick...|Musical Instruments|  2.0|[6, must, know, t...|
[6, must, know, t...|(3670,[145,255,32...|(3670,[145,255,32...|
[14.2460416600579...|[0.99572595700419...|       0.0|
+--------------------+--------------------+-----+--------------------
+--------------------+--------------------+----------
+--------------------+--------------------+----------+
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','pr
ediction').show(10)

+--------------------+--------------------+--------------------+-----
+----------+
|       rawPrediction|         probability|             subject|label|
prediction|
+--------------------+--------------------+--------------------+-----
+----------+
|[8.22575678849003...|[0.86083740538013...|Musical Instruments|  2.0|
0.0|
|[-1.5816511969981...|[6.40379189870091...|Musical Instruments|  2.0|
```

```
2.0|
|[0.38747123626564...|[1.29430064456987...|Musical Instruments|  2.0|
2.0|
|[-2.0540053505355...|[3.67476794956146...|   Business Finance|  1.0|
1.0|
|[24.7266193282529...|[0.99999999908079...|    Web Development|  0.0|
0.0|
|[22.2213462251437...|[0.99999999175336...|    Web Development|  0.0|
0.0|
|[20.1005546377385...|[0.99999995838555...|    Web Development|  0.0|
0.0|
|[-5.9910327938499...|[2.64083766762944...|Musical Instruments|  2.0|
2.0|
|[-19.729920863390...|[4.16984026967754...|     Graphic Design|  3.0|
3.0|
|[-2.6725325296694...|[9.29048167255554...|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='pred
iction',metricName='accuracy')

accuracy = evaluator.evaluate(predictions)

accuracy
```

```
0.9163498098859315
```

Method 2: Precision. F1Score (Classification Report)

```
from pyspark.mllib.evaluation import MulticlassMetrics

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

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: 0.9163498098859315
Precision: 0.9544159544159544
```

```
Recall: 0.8792650918635171
F1Score: 0.9153005464480874
```

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([[317,  14,   1,   4,   0,   0],
       [ 11, 335,   3,   2,   0,   0],
       [  8,  13, 156,   1,   0,   0],
       [  8,  17,   4, 156,   0,   0],
       [  0,   1,   0,   0,   0,   0],
       [  0,   1,   0,   0,   0,   0]])
```

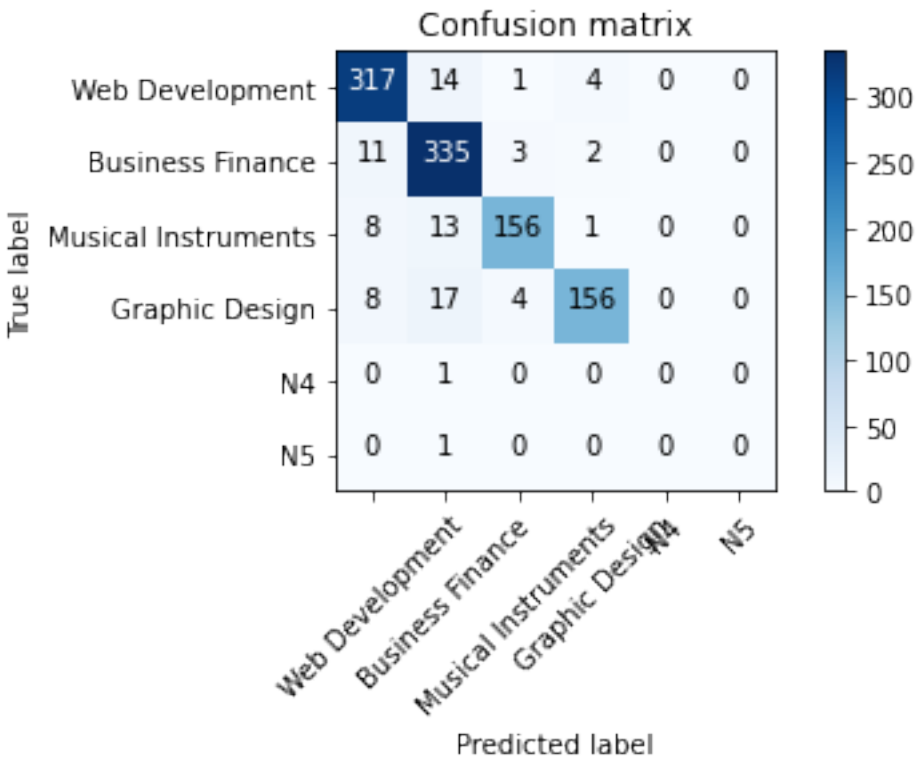
```
import matplotlib.pyplot as plt
import numpy as np
import itertools

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 Design','N4','N5']

plot_confusion_matrix(cm,class_names)
```

```
Confusion matrix, without normalization
[[317  14   1   4   0   0]
 [ 11 335   3   2   0   0]
 [  8  13 156   1   0   0]
 [  8  17   4 156   0   0]
 [  0   1   0   0   0   0]
 [  0   1   0   0   0   0]]
```




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

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

```
              precision    recall  f1-score   support

         0.0       0.92      0.94      0.93       336
         1.0       0.88      0.95      0.92       351
         2.0       0.95      0.88      0.91       178
         3.0       0.96      0.84      0.90       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       0.92      0.94      0.93       336
   Business Finance       0.88      0.95      0.92       351
```

```
Musical Instruments       0.95      0.88      0.91       178
     Graphic Design       0.96      0.84      0.90       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
```

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|
rawFeatures|  vectorizedFeatures|       rawPrediction|
probability|prediction|
+--------------------+---+--------------------+--------------------
+--------------------+--------------------+--------------------
+--------------------+----------+
|Building Machine ...| []|[building, machin...|[building, machin...|
(3670,[57,79,115,...|(3670,[57,79,115,...|[14.6893212262828...|
[0.99999805300087...|       0.0|
+--------------------+---+--------------------+--------------------
+--------------------+--------------------+--------------------
+--------------------+----------+


pred_ex1.columns

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

pred_ex1.select('course_title','rawPrediction','probability','predicti
on').show()

+--------------------+--------------------+--------------------
+----------+
|        course_title|       rawPrediction|         probability|
prediction|
+--------------------+--------------------+--------------------
+----------+
|Building Machine ...|[14.6893212262828...|[0.99999805300087...|
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)
```

```
#### Thanks For Your Time
#### Jesus Saves @JCharisTech
#### Jesse E.Agbe(JCharis)
#### Feb 26 2021
```