CSE4/587

Lab 3: Data Analytics Pipeline using Apache Spark

Contents

Collaborators

Utsav Mathur and Prajin Jonchhe have contributed equally towards the successful completion of Lab 3.

Environment

For this lab, Apache Spark has been used along with Scala in Ubuntu operating system(VM).

Pipeline

Naïve Bayes Classifier

- String Indexer
  StringIndexer encodes topic column to a column of label indices.
- Tokenizer
  Tokenizer takes text as input and breaks it into individual words.
- StopWordsRemover
  StopWordsRemover takes as input a sequence of strings and drops all the stop words from the input sequences.
- HashingTF
  HashingTF maps the terms to their frequencies using hash.
- IDF
  IDF calculates the inverse document frequency.
- Naïve Bayes Classifier
  Naive Bayes is a simple multiclass classification algorithm used in this pipeline.

Random Forest Classifier

- String Indexer
  StringIndexer encodes topic column to a column of label indices.
- Tokenizer
  Tokenizer takes text as input and breaks it into individual words.
- StopWordsRemover
  StopWordsRemover takes as input a sequence of strings and drops all the stop words from the input sequences.
- HashingTF
  HashingTF maps the terms to their frequencies using hash.
- IDF
  IDF calculates the inverse document frequency.
- Random Forest Classifier
  Random Forest is a classification technique used in this pipeline.

Run the program

The articles used to train the models are placed in the “articles” folder segregated by categories and the test articles are placed in “testarticles” folder.

The naivebayes.scala file contains the code for Naïve Bayes Classifier. To run it, first run the spark-shell using “spark-shell” command at the terminal. Then once spark is initialized, load this file using “:load naivebayes.scala” command and observe the output. Similarly, the randomforest.scala file contains the code for Random Forest Classifier. To run it, first run the spark-shell using “spark-shell” command at the terminal. Then once spark is initialized, load this file using “:load randomforest.scala” command and observe the output.

Output

The file named naivebayes.scala contains the code for training the model over the training article set using Naïve Bayes Classifier. The model is trained with 3-fold Cross Validation. The model is then used to predict the class/category of test articles. Based on prediction, the accuracy of the Naïve Bayes Classifier is observed to be 81%

Similarly, the file names randomforest.scala contains the code for training the model over the training article set using Naïve Bayes Classifier. The model is trained with 3-fold Cross Validation. The model is then used to predict the class/category of test articles. Based on prediction, the accuracy of the Random Forest Classifier is observed to be 63%

On comparing the accuracy, Naïve Bayes Classifier was observed to predict more accurately than Random Forest Classifier.

Screenshots

Naïve Bayes Classifier

```
training: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: string, text: string ... 1 more field]
test: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: string, text: string ... 1 more field]
indexer: org.apache.spark.ml.feature.StringIndexer = strIdx_48f4b768bc18
tokenizer: org.apache.spark.ml.feature.Tokenizer = tok_3db37bdab71d
remover: org.apache.spark.ml.feature.StopWordsRemover = stopWords_a657657e4efa
hashingTF: org.apache.spark.ml.feature.HashingTF = hashingTF_7d28f074d43b
idf: org.apache.spark.ml.feature.IDF = idf_4d2855c1b1d7
nb: org.apache.spark.ml.classification.NaiveBayes = nb_3c55b8c6e37d
pipeline: org.apache.spark.ml.Pipeline = pipeline_790421ff847a
model: org.apache.spark.ml.PipelineModel = pipeline_790421ff847a
predictions: org.apache.spark.sql.DataFrame = [id: string, text: string ... 9 more fields]
18/05/11 15:48:09 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
18/05/11 15:48:09 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|     id|                text|   topic|label|               words|            filtered|         rawFeatures|            features|       rawPrediction|         probability|prediction|
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|468.txt|WASHINGTON — Pres...|politics|  2.0|[washington, —, p...|[washington, —, p...|(1000,[0,1,5,6,9,...|(1000,[0,1,5,6,9,...|[-5430.9546312517...|[3.95979851626419...|       2.0|
|478.txt|When Donald J. Tr...|politics|  2.0|[when, donald, j....|[donald, j., trum...|(1000,[1,3,4,5,8,...|(1000,[1,3,4,5,8,...|[-11935.065897350...|[1.0,1.5802805900...|       0.0|
|467.txt|President Trump s...|politics|  2.0|[president, trump...|[president, trump...|(1000,[0,1,3,4,6,...|(1000,[0,1,3,4,6,...|[-21337.442045765...|[0.99989978964129...|       0.0|
|533.txt|The F.B.I. monito...|  sports|  1.0|[the, f.b.i., mon...|[f.b.i., monitore...|(1000,[4,14,15,19...|(1000,[4,14,15,19...|[-2154.4307964409...|[1.02817026566972...|       1.0|
|531.txt|HARTFORD — Women'...|  sports|  1.0|[hartford, —, wom...|[hartford, —, wom...|(1000,[0,4,6,7,8,...|(1000,[0,4,6,7,8,...|[-3216.4587323571...|[2.88464013630023...|       1.0|
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
only showing top 5 rows

evaluator: org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator = mcEval_29a83ebca68b
accuracy: Double = 0.8181818181818182
Test Error = 0.18181818181818177
```

Random Forest Classifier

```
training: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: string, text: string ... 1 more field]
test: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: string, text: string ... 1 more field]
indexer: org.apache.spark.ml.feature.StringIndexer = strIdx_3091137b9d2f
tokenizer: org.apache.spark.ml.feature.Tokenizer = tok_cc324b8fd147
remover: org.apache.spark.ml.feature.StopWordsRemover = stopWords_b6457ce91f4c
hashingTF: org.apache.spark.ml.feature.HashingTF = hashingTF_598775935331
idf: org.apache.spark.ml.feature.IDF = idf_d41d1dc07f53
rf: org.apache.spark.ml.classification.RandomForestClassifier = rfc_c908b8274255
pipeline: org.apache.spark.ml.Pipeline = pipeline_b6b0bfc97fdf
model: org.apache.spark.ml.PipelineModel = pipeline_b6b0bfc97fdf
predictions: org.apache.spark.sql.DataFrame = [id: string, text: string ... 9 more fields]
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|     id|                text|   topic|label|               words|            filtered|         rawFeatures|            features|       rawPrediction|         probability|prediction|
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
|468.txt|WASHINGTON — Pres...|politics|  2.0|[washington, —, p...|[washington, —, p...|(1000,[0,1,5,6,9,...|(1000,[0,1,5,6,9,...|[76.2075619721912...|[0.38103780986095...|       0.0|
|478.txt|When Donald J. Tr...|politics|  2.0|[when, donald, j....|[donald, j., trum...|(1000,[1,3,4,5,8,...|(1000,[1,3,4,5,8,...|[60.8820434986088...|[0.30441021749304...|       0.0|
|467.txt|President Trump s...|politics|  2.0|[president, trump...|[president, trump...|(1000,[0,1,3,4,6,...|(1000,[0,1,3,4,6,...|[42.9613285905267...|[0.21480664295263...|       3.0|
|533.txt|The F.B.I. monito...|  sports|  1.0|[the, f.b.i., mon...|[f.b.i., monitore...|(1000,[4,14,15,19...|(1000,[4,14,15,19...|[45.0813397358158...|[0.22540669867907...|       1.0|
|531.txt|HARTFORD — Women'...|  sports|  1.0|[hartford, —, wom...|[hartford, —, wom...|(1000,[0,4,6,7,8,...|(1000,[0,4,6,7,8,...|[46.9154565269606...|[0.23457728263480...|       1.0|
+-------+--------------------+--------+-----+--------------------+--------------------+--------------------+--------------------+--------------------+----------+
only showing top 5 rows

evaluator: org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator = mcEval_c55cddd87ef5
accuracy: Double = 0.6363636363636364
Test Error = 0.36363636363636365
```

Block Diagram