CS5542 Big Data Apps and Analytics
LAB ASSIGNMENT #4

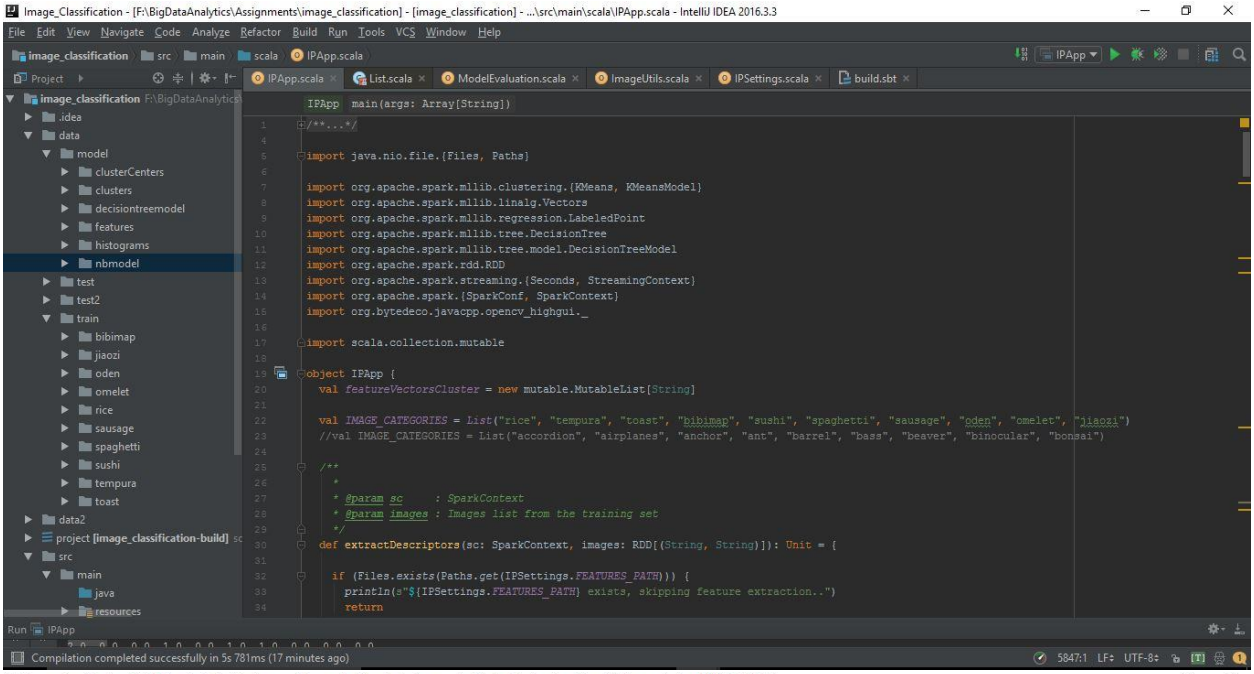
The image classification program is run using decision tree algorithm instead of random forest algorithm. The data set used is the data set provided in the source code, since the other data set is not suitable/ giving no proper output. The decision tree algorithm is a classification algorithm, starting from top to bottom, the train examples are always at the root of the tree.In the decision tree algorithm, the partitioning is done using selected attributes. Here, for this lab tutorial, the data set is the default data set provided, the categories are:
bibimap
jiaozi
oden
omelet
rice
sausage
spaghetti
sushi
tempura
toast

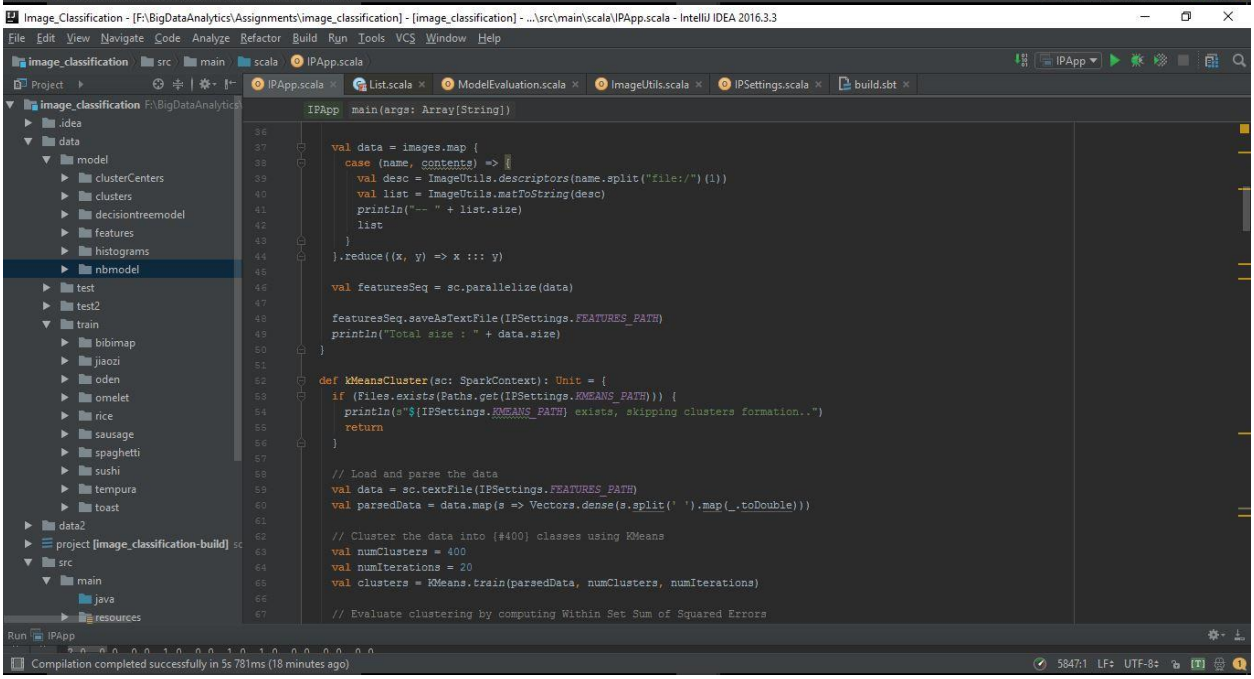
these are divided into 70% train and 30% test. The image classification program is run on this data set and the resulting output is the accuracy obtained by running the program: 0.22

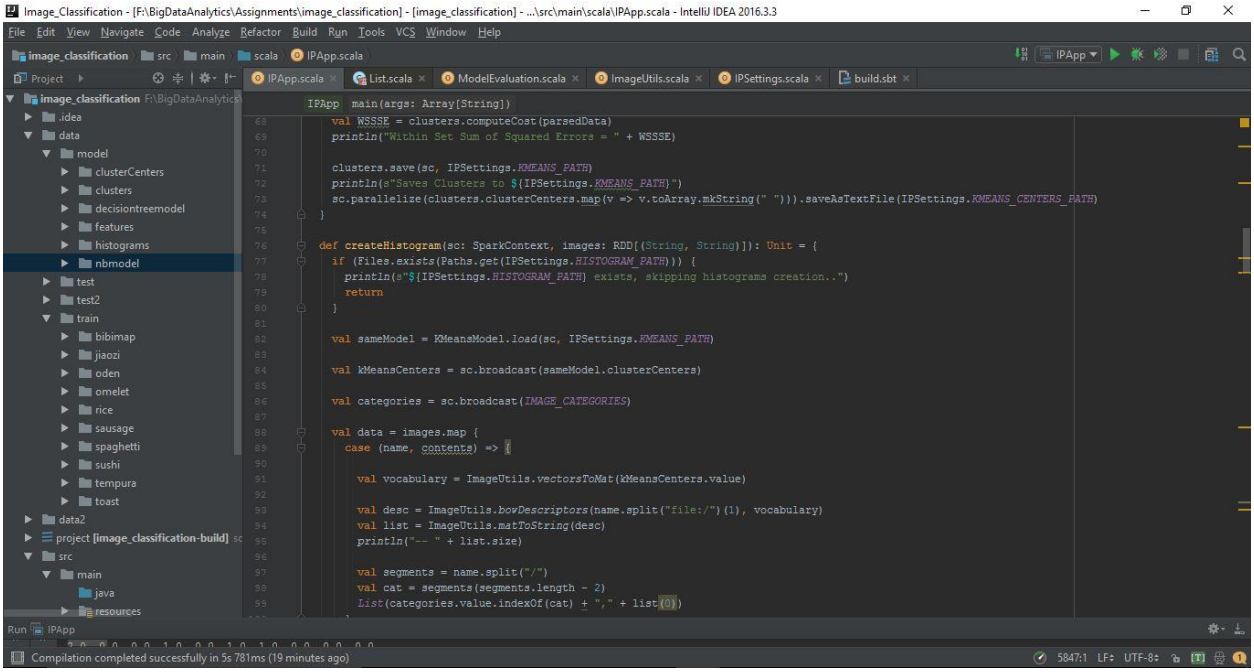
The program is also run on other data set(DataLab4) with 5 categories:
airplane
apples
cats
eggs
wolves

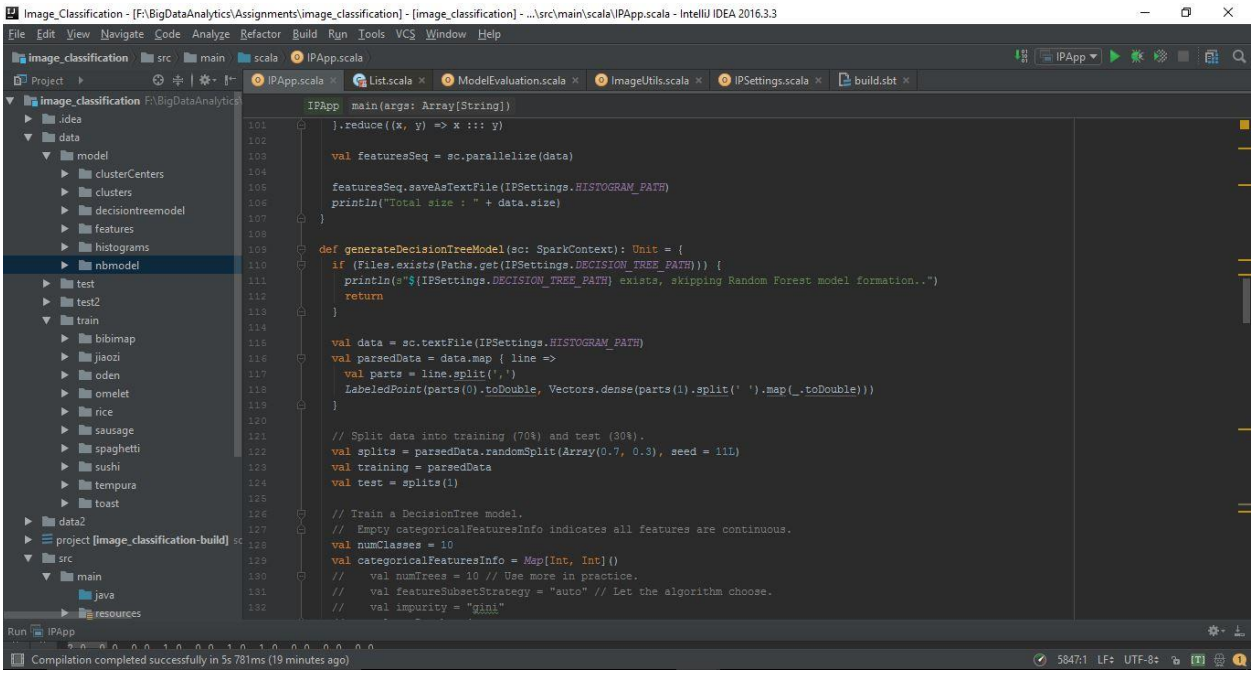
the correct/desired output was not obtained, there was a problem with the images/ image format, so the original data set was used to get the output.

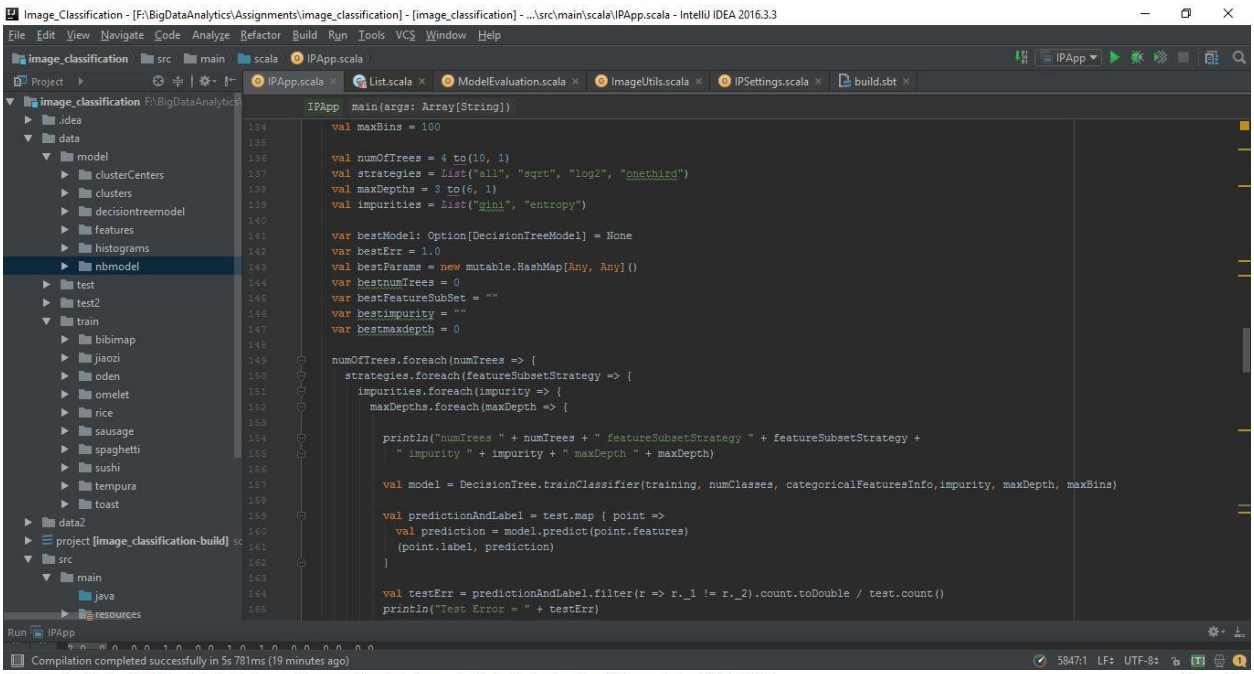
These are few screen shots from the program:

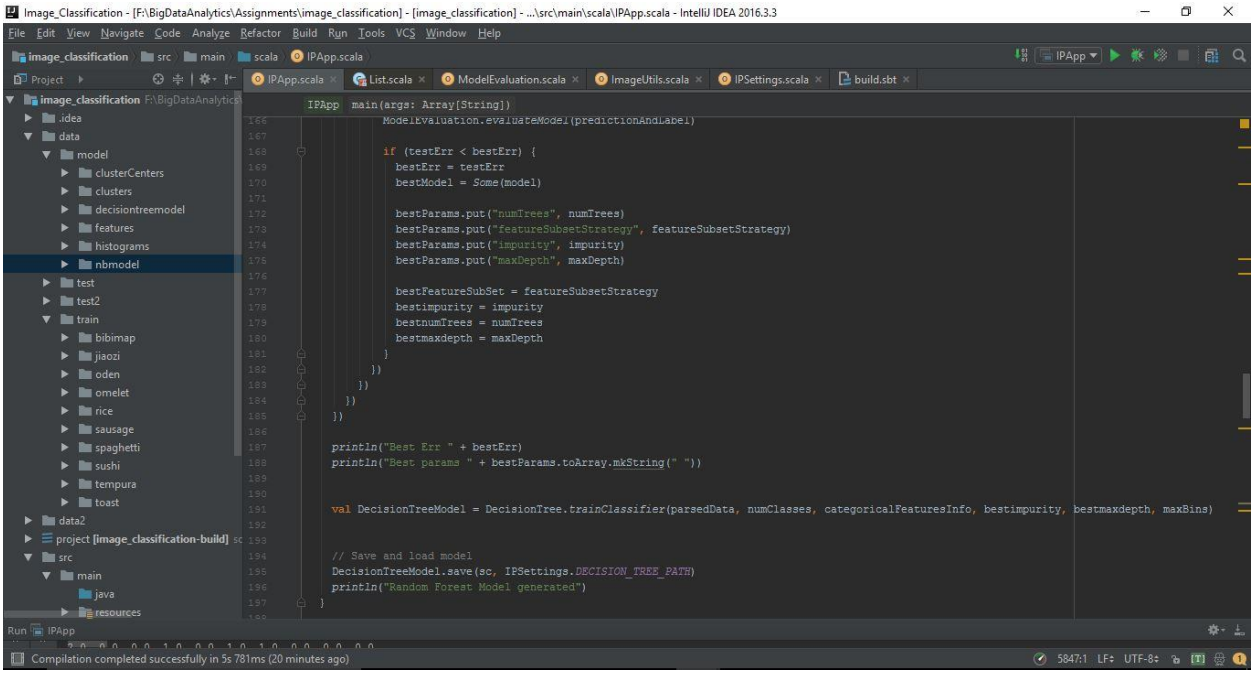

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
/**...*/
import java.nio.file.{Files, Paths}
import org.apache.spark.mllib.clustering.{KMeans, KMeansModel}
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.tree.DecisionTree
import org.apache.spark.mllib.tree.model.DecisionTreeModel
import org.apache.spark.rdd.RDD
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.{SparkConf, SparkContext}
import org.bytedeco.javacpp.opencv_highgui._
import scala.collection.mutable
object IPApp {
val featureVectorsCluster = new mutable.MutableList[String]
val IMAGE_CATEGORIES = List("rice", "tempura", "toast", "bibimap", "sushi", "spaghetti", "sausage", "oden", "omelet", "jiaozi")
//val IMAGE_CATEGORIES = List("accordion", "airplanes", "anchor", "ant", "barrel", "bass", "beaver", "binocular", "bonsai")
/**
*
* @param sc : SparkContext
* @param images : Images list from the training set
*/
def extractDescriptors(sc: SparkContext, images: RDD[(String, String)]): Unit = {
if (Files.exists(Paths.get(IPSettings.FEATURES_PATH))) {
println(s"${IPSettings.FEATURES_PATH} exists, skipping feature extraction..")
return
Run IPApp
Compilation completed successfully in 5s 781ms (17 minutes ago)
5847:1 LF: UTF-8:

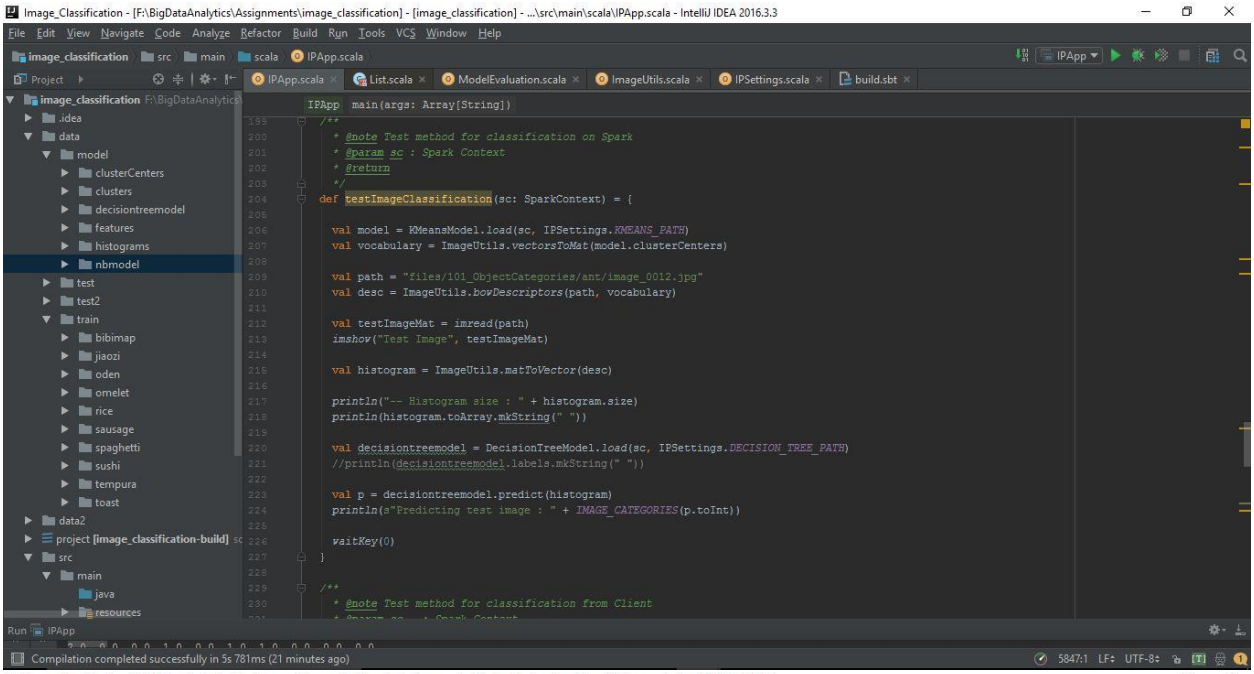

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
val data = images.map {
case (name, contents) => {
val desc = ImageUtils.descriptors(name.split("file:/")(1))
val list = ImageUtils.matToString(desc)
println("-- " + list.size)
list
}
}.reduce((x, y) => x ::: y)
val featuresSeq = sc.parallelize(data)
featuresSeq.saveAsTextFile(IPSettings.FEATURES_PATH)
println("Total size : " + data.size)
}
def kMeansCluster(sc: SparkContext): Unit = {
if (Files.exists(Paths.get(IPSettings.KMEANS_PATH))) {
println(s"${IPSettings.KMEANS_PATH} exists, skipping clusters formation..")
return
}
// Load and parse the data
val data = sc.textFile(IPSettings.FEATURES_PATH)
val parsedData = data.map(s => Vectors.dense(s.split(' ').map(_.toDouble)))
// Cluster the data into {#400} classes using KMeans
val numClusters = 400
val numIterations = 20
val clusters = KMeans.train(parsedData, numClusters, numIterations)
// Evaluate clustering by computing Within Set Sum of Squared Errors
Run IPApp
Compilation completed successfully in 5s 781ms (18 minutes ago)
5847:1 LF: UTF-8:

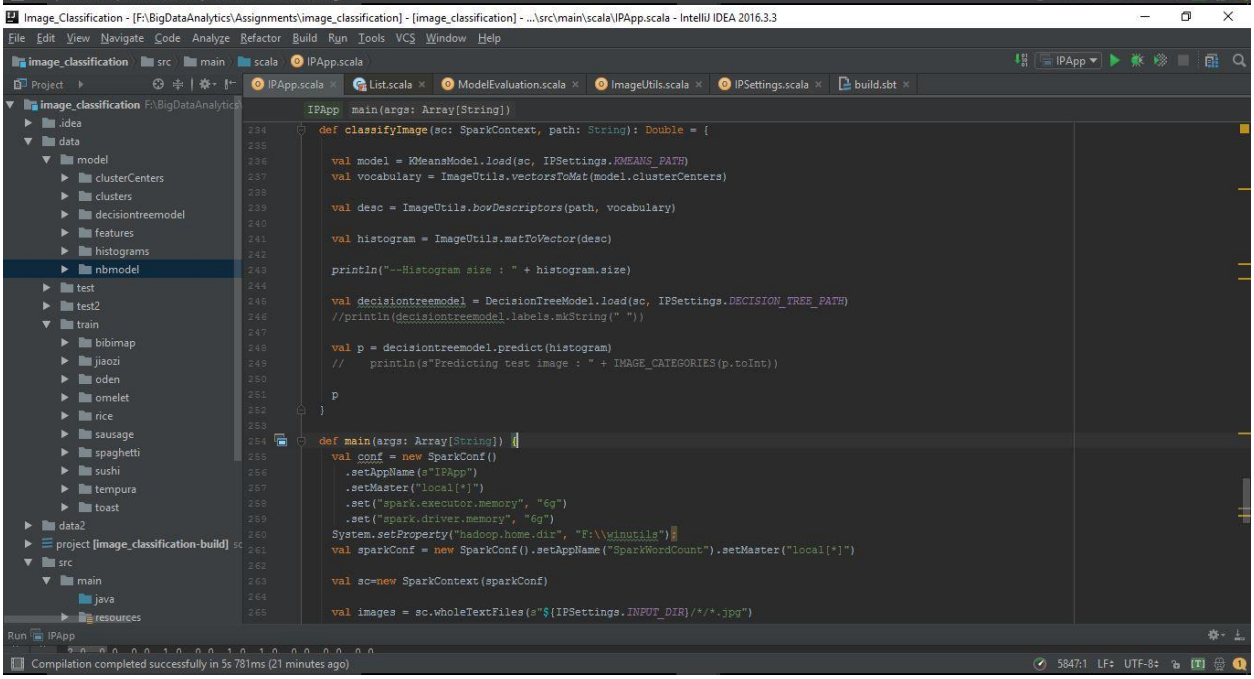

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
val WSSSE = clusters.computeCost(parsedData)
println("Within Set Sum of Squared Errors = " + WSSSE)
clusters.save(sc, IPSettings.KMEANS_PATH)
println(s"Saves Clusters to ${IPSettings.KMEANS_PATH}")
sc.parallelize(clusters.clusterCenters.map(v => v.toArray.mkString(" "))).saveAsTextFile(IPSettings.KMEANS_CENTERS_PATH)
def createHistogram(sc: SparkContext, images: RDD[(String, String)]): Unit = {
if (Files.exists(Paths.get(IPSettings.HISTOGRAM_PATH))) {
println(s"${IPSettings.HISTOGRAM_PATH} exists, skipping histograms creation..")
return
val sameModel = KMeansModel.load(sc, IPSettings.KMEANS_PATH)
val kMeansCenters = sc.broadcast(sameModel.clusterCenters)
val categories = sc.broadcast(IMAGE_CATEGORIES)
val data = images.map {
case (name, contents) => {
val vocabulary = ImageUtils.vectorsToMat(kMeansCenters.value)
val desc = ImageUtils.bowDescriptors(name.split("file:/")(1), vocabulary)
val list = ImageUtils.matToString(desc)
println("-- " + list.size)
val segments = name.split("/")
val cat = segments(segments.length - 2)
List(categories.value.indexOf(cat) + "," + list(0))
Run IPApp
Compilation completed successfully in 5s 781ms (19 minutes ago)
5847:1 LF: UTF-8:

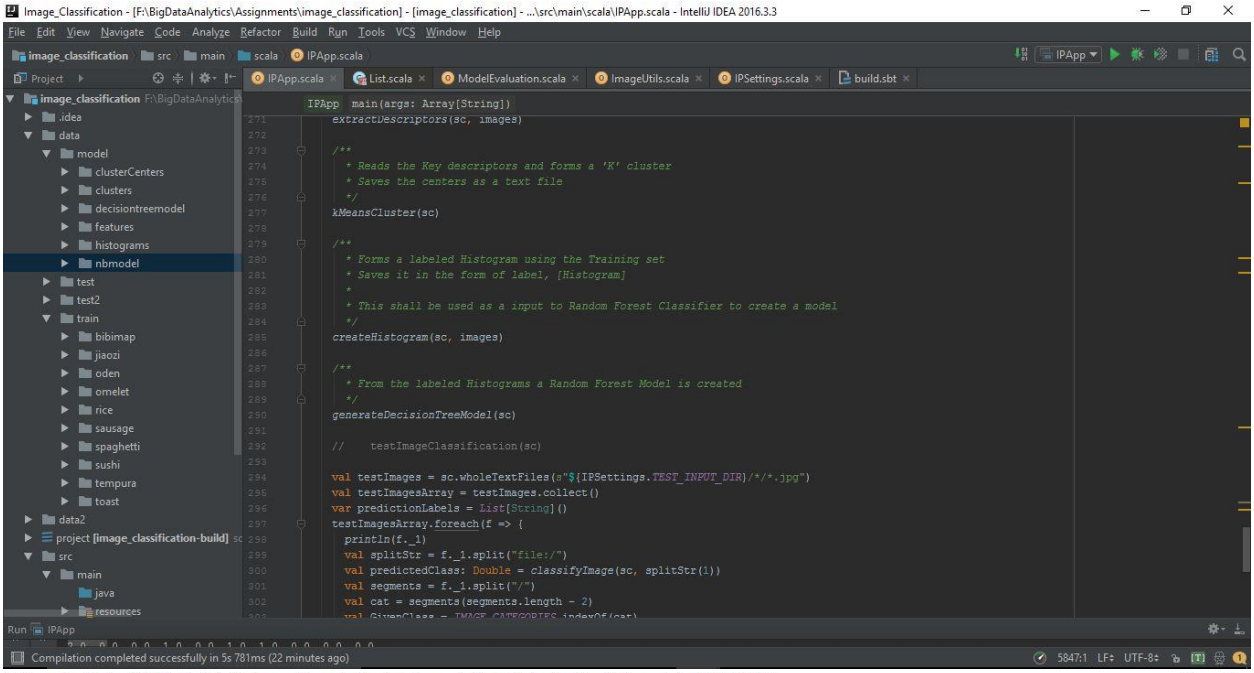

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
}.reduce((x, y) => x ::: y)
val featuresSeq = sc.parallelize(data)
featuresSeq.saveAsTextFile(IPSettings.HISTOGRAM_PATH)
println("Total size : " + data.size)
def generateDecisionTreeModel(sc: SparkContext): Unit = {
if (Files.exists(Paths.get(IPSettings.DECISION_TREE_PATH))) {
println(s"${IPSettings.DECISION_TREE_PATH} exists, skipping Random Forest model formation..")
return
val data = sc.textFile(IPSettings.HISTOGRAM_PATH)
val parsedData = data.map { line =>
val parts = line.split(',')
LabeledPoint(parts(0).toDouble, Vectors.dense(parts(1).split(' ').map(_.toDouble)))
// Split data into training (70%) and test (30%).
val splits = parsedData.randomSplit(Array(0.7, 0.3), seed = 11L)
val training = parsedData
val test = splits(1)
// Train a DecisionTree model.
// Empty categoricalFeaturesInfo indicates all features are continuous.
val numClasses = 10
val categoricalFeaturesInfo = Map[Int, Int]()
// val numTrees = 10 // Use more in practice.
// val featureSubsetStrategy = "auto" // Let the algorithm choose.
// val impurity = "gini"
Run IPApp
Compilation completed successfully in 5s 781ms (19 minutes ago)
5847:1 LF: UTF-8:

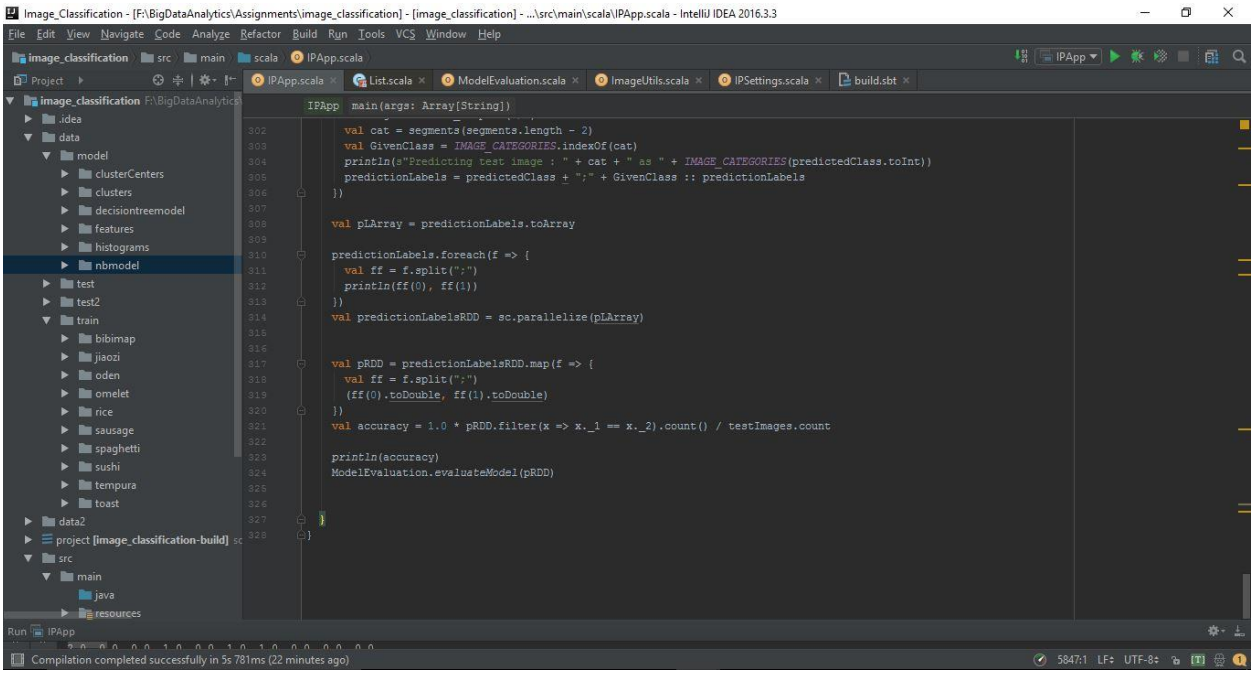

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala
List.scala
ModelEvaluation.scala
ImageUtils.scala
IPSettings.scala
build.sbt
IPApp main(args: Array[String])
val maxBins = 100
val numOfTrees = 4 to(10, 1)
val strategies = List("all", "sqrt", "log2", "onethird")
val maxDepths = 3 to(6, 1)
val impurities = List("gini", "entropy")
var bestModel: Option[DecisionTreeModel] = None
var bestErr = 1.0
val bestParams = new mutable.HashMap[Any, Any]()
var bestnumTrees = 0
var bestFeatureSubSet = ""
var bestimpurity = ""
var bestmaxdepth = 0
numOfTrees.foreach(numTrees => {
strategies.foreach(featureSubsetStrategy => {
impurities.foreach(impurity => {
maxDepths.foreach(maxDepth => {
println("numTrees " + numTrees + " featureSubsetStrategy " + featureSubsetStrategy +
" impurity " + impurity + " maxDepth " + maxDepth)
val model = DecisionTree.trainClassifier(training, numClasses, categoricalFeaturesInfo,impurity, maxDepth, maxBins)
val predictionAndLabel = test.map { point =>
val prediction = model.predict(point.features)
(point.label, prediction)
}
val testErr = predictionAndLabel.filter(r => r._1 != r._2).count.toDouble / test.count()
println("Test Error = " + testErr)
Compilation completed successfully in 5s 781ms (19 minutes ago)
5847:1 LF: UTF-8:

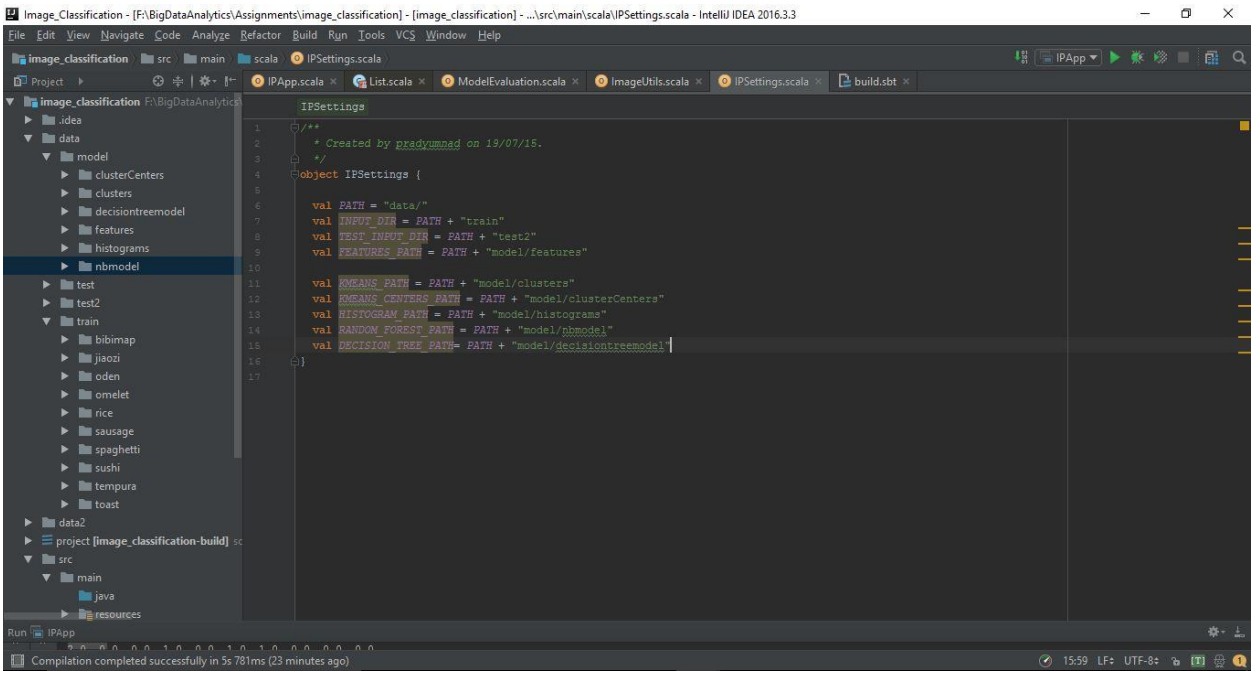

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala
List.scala
ModelEvaluation.scala
ImageUtils.scala
IPSettings.scala
build.sbt
IPApp main(args: Array[String])
ModelEvaluation.evaluateModel(predictionAndLabel)
if (testErr < bestErr) {
bestErr = testErr
bestModel = Some(model)
bestParams.put("numTrees", numTrees)
bestParams.put("featureSubsetStrategy", featureSubsetStrategy)
bestParams.put("impurity", impurity)
bestParams.put("maxDepth", maxDepth)
bestFeatureSubSet = featureSubsetStrategy
bestimpurity = impurity
bestnumTrees = numTrees
bestmaxdepth = maxDepth
println("Best Err " + bestErr)
println("Best params " + bestParams.toArray.mkString(" "))
val DecisionTreeModel = DecisionTree.trainClassifier(parsedData, numClasses, categoricalFeaturesInfo, bestimpurity, bestmaxdepth, maxBins)
// Save and load model
DecisionTreeModel.save(sc, IPSettings.DECISION_TREE_PATH)
println("Random Forest Model generated")
Compilation completed successfully in 5s 781ms (20 minutes ago)
5847:1 LF: UTF-8:

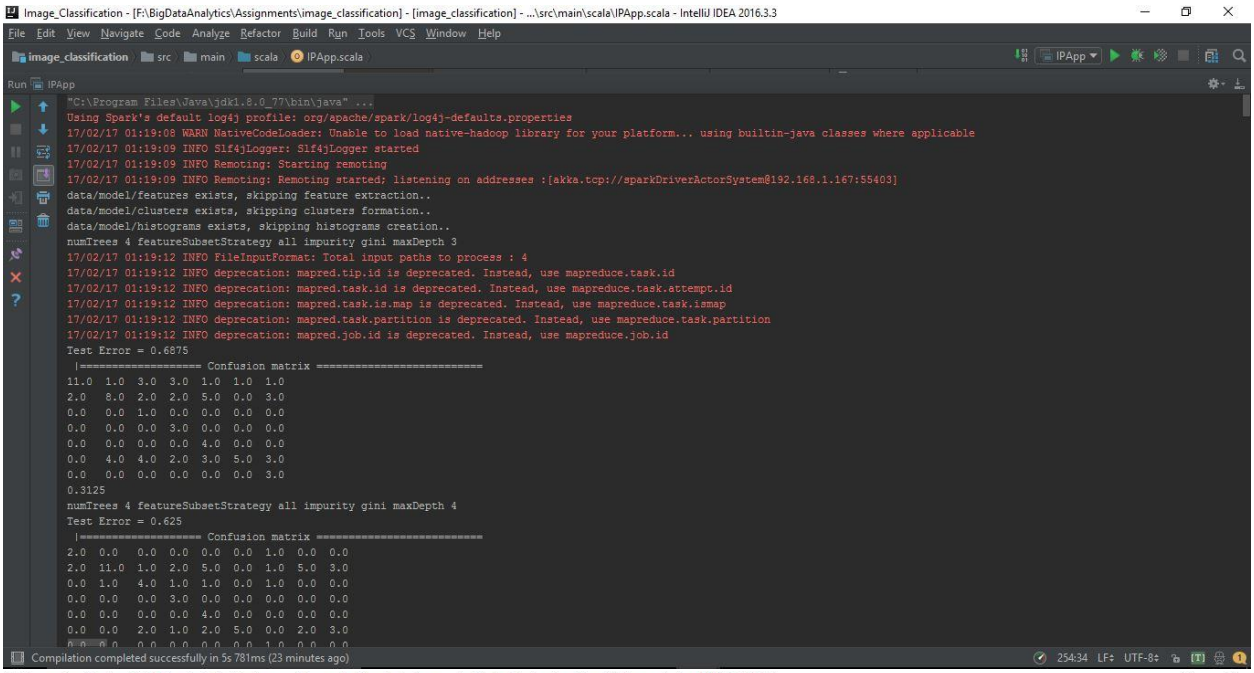

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
image_classification F:\BigDataAnalytics
.idea
data
model
clusterCenters
clusters
decisiontreemodel
features
histograms
nbmodel
test
test2
train
bibimap
jiaozi
oden
omelet
rice
sausage
spaghetti
sushi
tempura
toast
data2
project [image_classification-build]
src
main
java
resources
/**
* @note Test method for classification on Spark
* @param sc : Spark Context
* @return
*/
def testImageClassification(sc: SparkContext) = {
val model = KMeansModel.load(sc, IPSettings.KMEANS_PATH)
val vocabulary = ImageUtils.vectorsToMat(model.clusterCenters)
val path = "files/101_ObjectCategories/ant/image_0012.jpg"
val desc = ImageUtils.bowDescriptors(path, vocabulary)
val testImageMat = imread(path)
imshow("Test Image", testImageMat)
val histogram = ImageUtils.matToVector(desc)
println("-- Histogram size : " + histogram.size)
println(histogram.toArray.mkString(" "))
val decisiontreemodel = DecisionTreeModel.load(sc, IPSettings.DECISION_TREE_PATH)
//println(decisiontreemodel.labels.mkString(" "))
val p = decisiontreemodel.predict(histogram)
println(s"Predicting test image : " + IMAGE_CATEGORIES(p.toInt))
waitKey(0)
}
/**
* @note Test method for classification from Client
Run IPApp
Compilation completed successfully in 5s 781ms (21 minutes ago)
5847:1 LF: UTF-8:

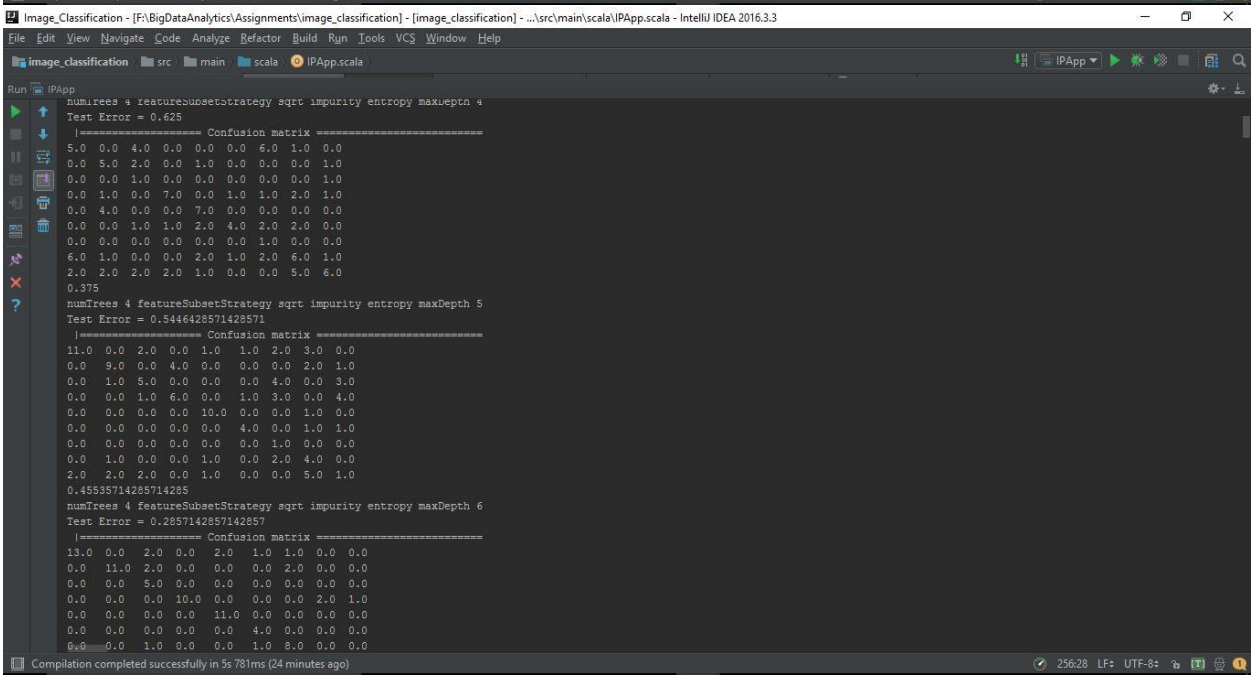

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPApp main(args: Array[String])
image_classification F:\BigDataAnalytics
.idea
data
model
clusterCenters
clusters
decisiontreemodel
features
histograms
nbmodel
test
test2
train
bibimap
jiaozi
oden
omelet
rice
sausage
spaghetti
sushi
tempura
toast
data2
project [image_classification-build]
src
main
java
resources
def classifyImage(sc: SparkContext, path: String): Double = {
val model = KMeansModel.load(sc, IPSettings.KMEANS_PATH)
val vocabulary = ImageUtils.vectorsToMat(model.clusterCenters)
val desc = ImageUtils.bowDescriptors(path, vocabulary)
val histogram = ImageUtils.matToVector(desc)
println("--Histogram size : " + histogram.size)
val decisiontreemodel = DecisionTreeModel.load(sc, IPSettings.DECISION_TREE_PATH)
//println(decisiontreemodel.labels.mkString(" "))
val p = decisiontreemodel.predict(histogram)
// println(s"Predicting test image : " + IMAGE_CATEGORIES(p.toInt))
p
}
def main(args: Array[String]) {
val conf = new SparkConf()
.setAppName(s"IPApp")
.setMaster("local[*]")
.set("spark.executor.memory", "6g")
.set("spark.driver.memory", "6g")
System.setProperty("hadoop.home.dir", "F:\\winutils");
val sparkConf = new SparkConf().setAppName("SparkWordCount").setMaster("local[*]")
val sc=new SparkContext(sparkConf)
val images = sc.wholeTextFiles(s"${IPSettings.INPUT_DIR}/*/*.jpg")
Run IPApp
Compilation completed successfully in 5s 781ms (21 minutes ago)
5847:1 LF: UTF-8:

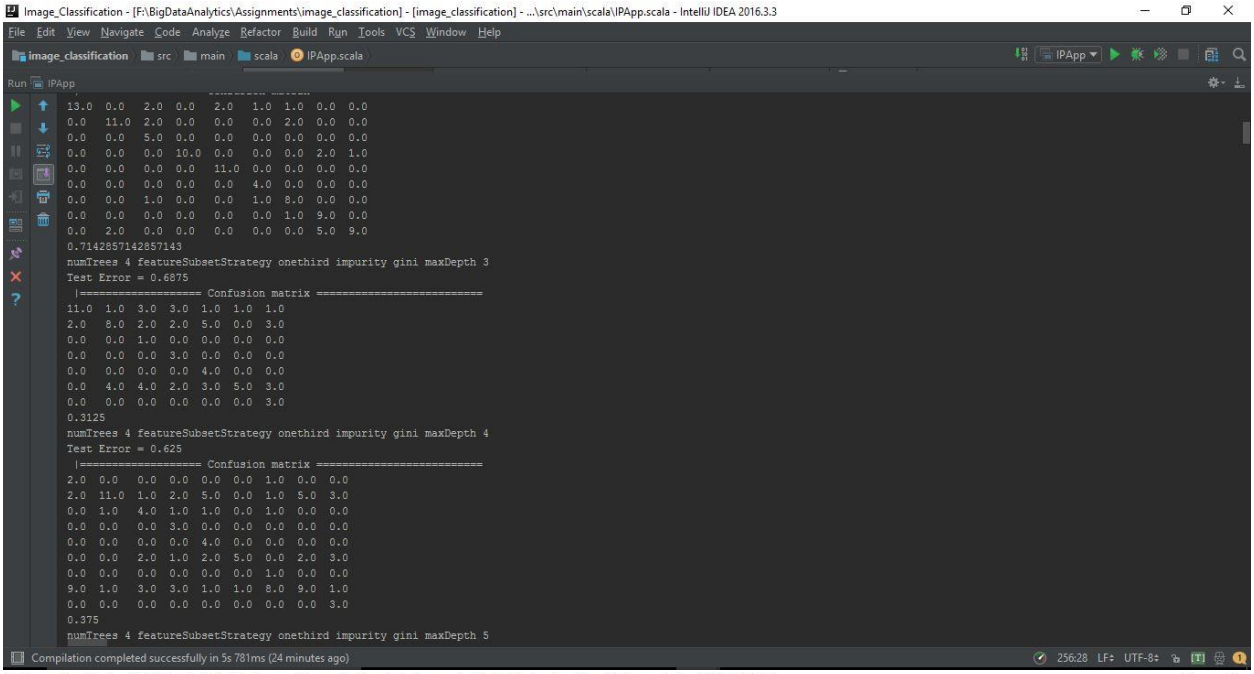

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala
List.scala
ModelEvaluation.scala
ImageUtils.scala
IPSettings.scala
build.sbt
IPApp main(args: Array[String])
extractDescriptors(sc, images)
/**
* Reads the Key descriptors and forms a 'K' cluster
* Saves the centers as a text file
*/
kMeansCluster(sc)
/**
* Forms a labeled Histogram using the Training set
* Saves it in the form of label, [Histogram]
*
* This shall be used as a input to Random Forest Classifier to create a model
*/
createHistogram(sc, images)
/**
* From the labeled Histograms a Random Forest Model is created
*/
generateDecisionTreeModel(sc)
// testImageClassification(sc)
val testImages = sc.wholeTextFiles(s"${IPSettings.TEST_INPUT_DIR}/*/*.jpg")
val testImagesArray = testImages.collect()
var predictionLabels = List[String]()
testImagesArray.foreach(f => {
println(f._1)
val splitStr = f._1.split("file:/")
val predictedClass: Double = classifyImage(sc, splitStr(1))
val segments = f._1.split("/")
val cat = segments(segments.length - 2)
Compilation completed successfully in 5s 781ms (22 minutes ago)
5847:1 LF: UTF-8:

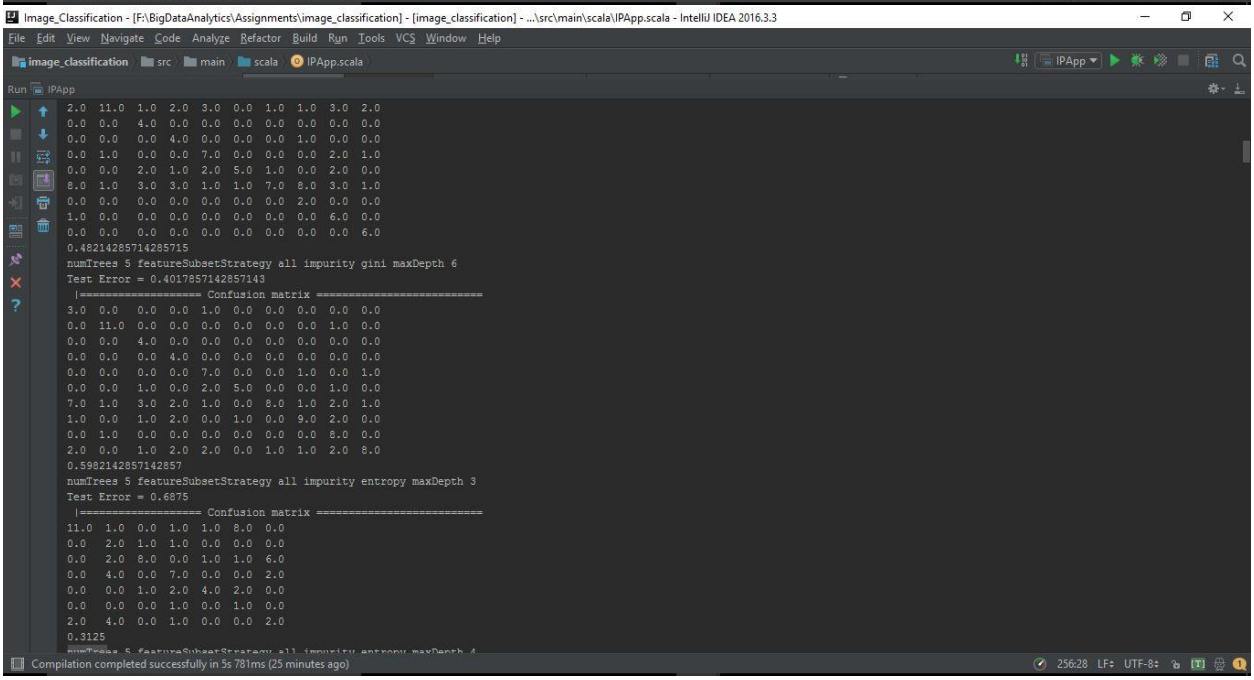

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IPApp.scala
List.scala
ModelEvaluation.scala
ImageUtils.scala
IPSettings.scala
build.sbt
IPApp main(args: Array[String])
val cat = segments(segments.length - 2)
val GivenClass = IMAGE_CATEGORIES.indexOf(cat)
println(s"Predicting test image : " + cat + " as " + IMAGE_CATEGORIES(predictedClass.toInt))
predictionLabels = predictedClass + ";" + GivenClass :: predictionLabels
})
val pLArray = predictionLabels.toArray
predictionLabels.foreach(f => {
val ff = f.split(";")
println(ff(0), ff(1))
})
val predictionLabelsRDD = sc.parallelize(pLArray)
val pRDD = predictionLabelsRDD.map(f => {
val ff = f.split(";")
(ff(0).toDouble, ff(1).toDouble)
})
val accuracy = 1.0 * pRDD.filter(x => x._1 == x._2).count() / testImages.count
println(accuracy)
ModelEvaluation.evaluateModel(pRDD)
}
}
Compilation completed successfully in 5s 781ms (22 minutes ago)
5847:1 LF: UTF-8:

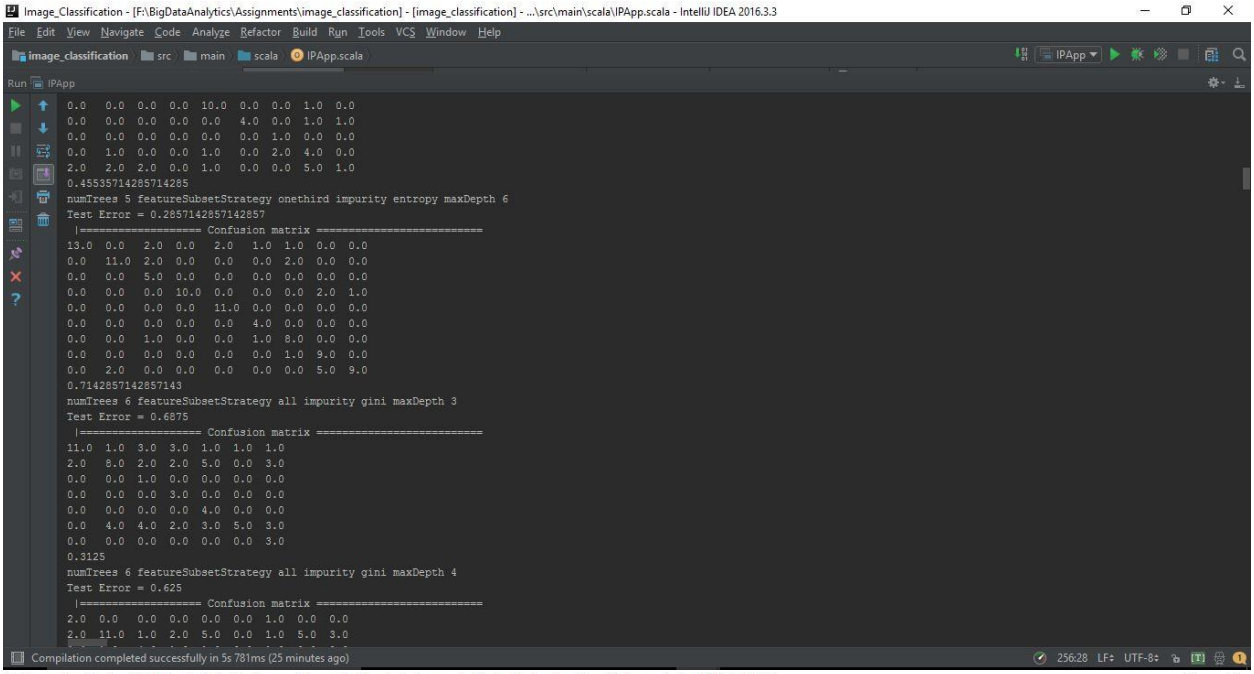

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPSettings.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPSettings.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
image_classification F:\BigDataAnalytics\
.idea
data
model
clusterCenters
clusters
decisiontreemodel
features
histograms
nbmodel
test
test2
train
bibimap
jiaozi
oden
omelet
rice
sausage
spaghetti
sushi
tempura
toast
data2
project [image_classification-build]
src
main
java
resources
IPSettings
/**
* Created by pradyumnad on 19/07/15.
*/
object IPSettings {
val PATH = "data/"
val INPUT_DIR = PATH + "train"
val TEST_INPUT_DIR = PATH + "test2"
val FEATURES_PATH = PATH + "model/features"
val KMEANS_PATH = PATH + "model/clusters"
val KMEANS_CENTERS_PATH = PATH + "model/clusterCenters"
val HISTOGRAM_PATH = PATH + "model/histograms"
val RANDOM_FOREST_PATH = PATH + "model/nbmodel"
val DECISION_TREE_PATH= PATH + "model/decisiontreemodel"
}
Run IPApp
Compilation completed successfully in 5s 781ms (23 minutes ago)
15:59 LF: UTF-8:

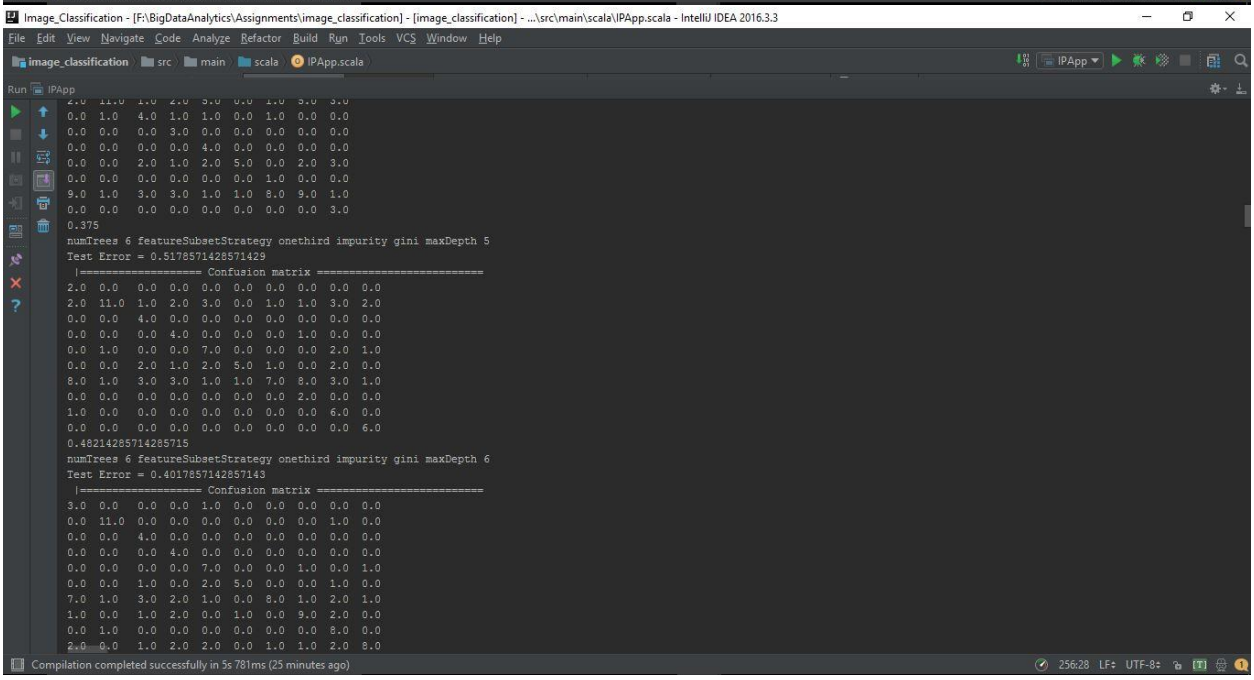

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
"C:\Program Files\Java\jdk1.8.0_77\bin\java" ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/02/17 01:19:08 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/02/17 01:19:09 INFO Slf4jLogger: Slf4jLogger started
17/02/17 01:19:09 INFO Remoting: Starting remoting
17/02/17 01:19:09 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.1.167:55403]
data/model/features exists, skipping feature extraction..
data/model/clusters exists, skipping clusters formation..
data/model/histograms exists, skipping histograms creation..
numTrees 4 featureSubsetStrategy all impurity gini maxDepth 3
17/02/17 01:19:12 INFO FileInputFormat: Total input paths to process : 4
17/02/17 01:19:12 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
17/02/17 01:19:12 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
17/02/17 01:19:12 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
17/02/17 01:19:12 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
17/02/17 01:19:12 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
Test Error = 0.6875
|================== Confusion matrix ==========================
11.0 1.0 3.0 3.0 1.0 1.0 1.0
2.0 8.0 2.0 2.0 5.0 0.0 3.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0
0.0 4.0 4.0 2.0 3.0 5.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.3125
numTrees 4 featureSubsetStrategy all impurity gini maxDepth 4
Test Error = 0.625
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
2.0 11.0 1.0 2.0 5.0 0.0 1.0 5.0 3.0
0.0 1.0 4.0 1.0 1.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0
0.0 0.0 2.0 1.0 2.0 5.0 0.0 2.0 3.0
Compilation completed successfully in 5s 781ms (23 minutes ago)
254:34 LF: UTF-8:

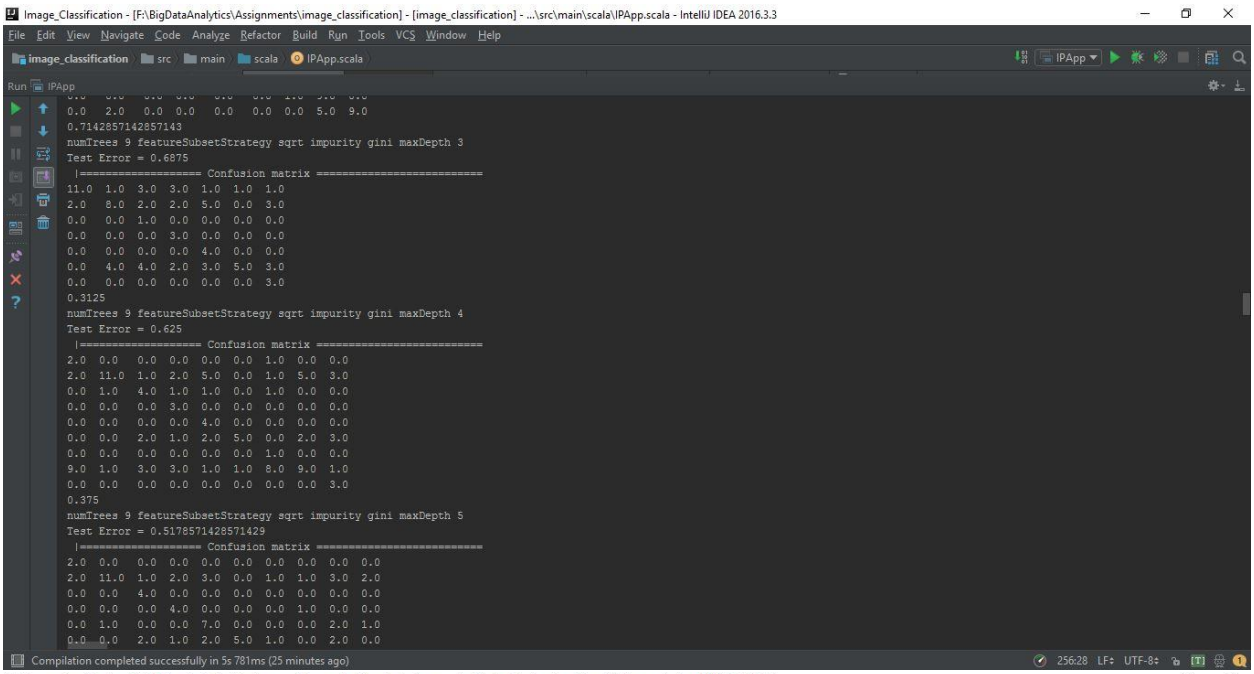

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
numTrees 4 featureSubsetStrategy sqrt impurity entropy maxDepth 4
Test Error = 0.625
|================== Confusion matrix ==========================
5.0 0.0 4.0 0.0 0.0 0.0 6.0 1.0 0.0
0.0 5.0 2.0 0.0 1.0 0.0 0.0 0.0 1.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0
0.0 1.0 0.0 7.0 0.0 1.0 1.0 2.0 1.0
0.0 4.0 0.0 0.0 7.0 0.0 0.0 0.0 0.0
0.0 0.0 1.0 1.0 2.0 4.0 2.0 2.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
6.0 1.0 0.0 0.0 2.0 1.0 2.0 6.0 1.0
2.0 2.0 2.0 2.0 1.0 0.0 0.0 5.0 6.0
0.375
numTrees 4 featureSubsetStrategy sqrt impurity entropy maxDepth 5
Test Error = 0.5446428571428571
|================== Confusion matrix ==========================
11.0 0.0 2.0 0.0 1.0 1.0 2.0 3.0 0.0
0.0 9.0 0.0 4.0 0.0 0.0 0.0 2.0 1.0
0.0 1.0 5.0 0.0 0.0 0.0 4.0 0.0 3.0
0.0 0.0 1.0 6.0 0.0 1.0 3.0 0.0 4.0
0.0 0.0 0.0 0.0 10.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.0 4.0 0.0 1.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 1.0 0.0 0.0 1.0 0.0 2.0 4.0 0.0
2.0 2.0 2.0 0.0 1.0 0.0 0.0 5.0 1.0
0.45535714285714285
numTrees 4 featureSubsetStrategy sqrt impurity entropy maxDepth 6
Test Error = 0.2857142857142857
|================== Confusion matrix ==========================
13.0 0.0 2.0 0.0 2.0 1.0 1.0 0.0 0.0
0.0 11.0 2.0 0.0 0.0 0.0 2.0 0.0 0.0
0.0 0.0 5.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 10.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 0.0 0.0 11.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 1.0 8.0 0.0 0.0
Compilation completed successfully in 5s 781ms (24 minutes ago)
256:28 LF: UTF-8:

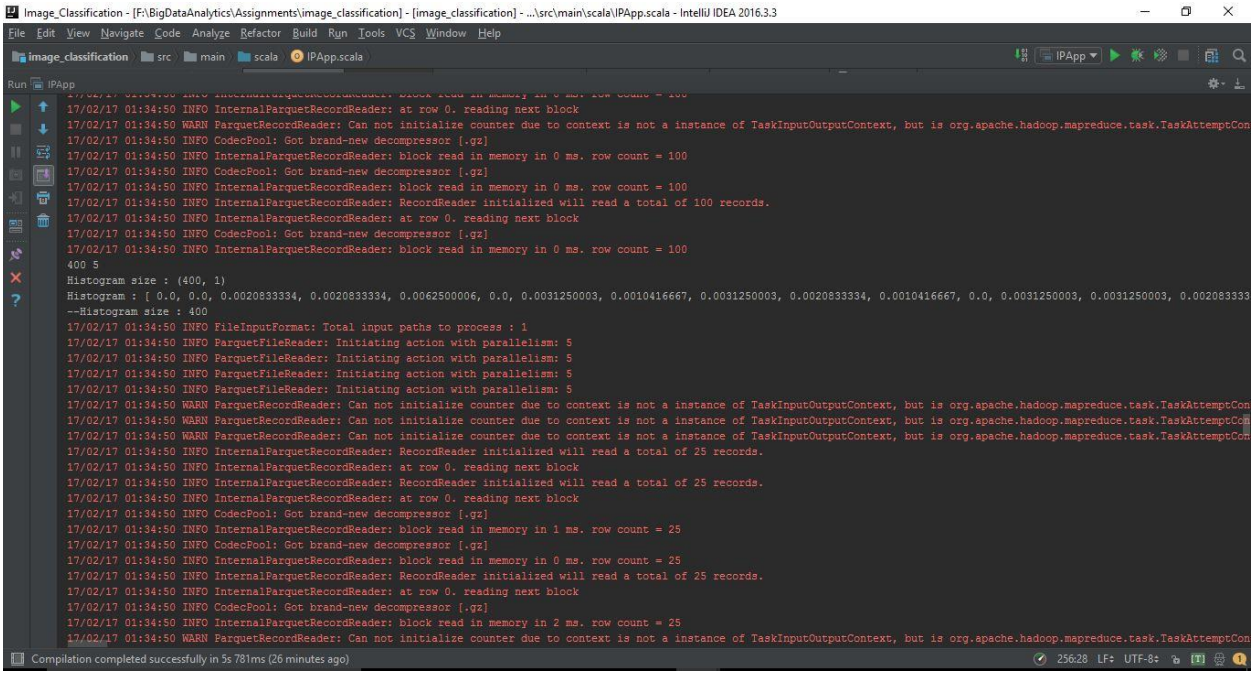

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
Run IPApp
13.0 0.0 2.0 0.0 2.0 1.0 1.0 0.0 0.0
0.0 11.0 2.0 0.0 0.0 0.0 2.0 0.0 0.0
0.0 0.0 5.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 10.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 0.0 0.0 11.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 1.0 8.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 9.0 0.0
0.0 2.0 0.0 0.0 0.0 0.0 0.0 5.0 9.0
0.7142857142857143
numTrees 4 featureSubsetStrategy onethird impurity gini maxDepth 3
Test Error = 0.6875
|================== Confusion matrix ==========================
11.0 1.0 3.0 3.0 1.0 1.0 1.0
2.0 8.0 2.0 2.0 5.0 0.0 3.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0
0.0 4.0 4.0 2.0 3.0 5.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.3125
numTrees 4 featureSubsetStrategy onethird impurity gini maxDepth 4
Test Error = 0.625
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
2.0 11.0 1.0 2.0 5.0 0.0 1.0 5.0 3.0
0.0 1.0 4.0 1.0 1.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0
0.0 0.0 2.0 1.0 2.0 5.0 0.0 2.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
9.0 1.0 3.0 3.0 1.0 1.0 8.0 9.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.375
numTrees 4 featureSubsetStrategy onethird impurity gini maxDepth 5
Compilation completed successfully in 5s 781ms (24 minutes ago)
256:28 LF: UTF-8:

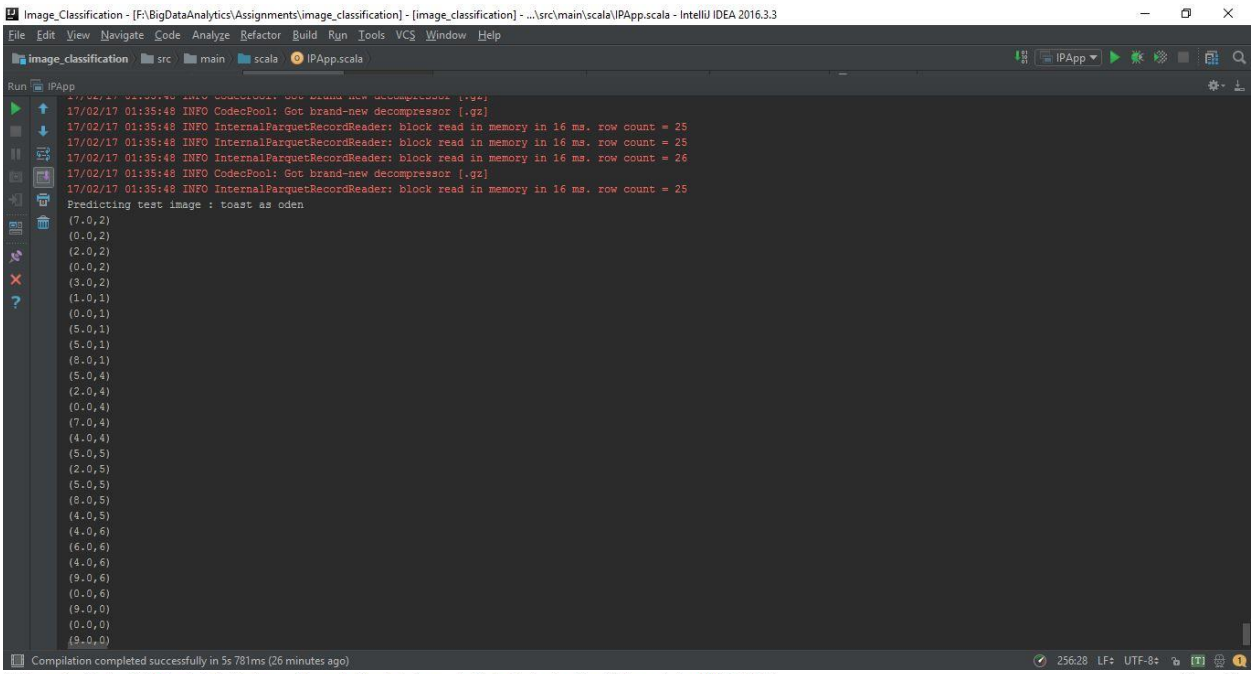

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
Run IPApp
2.0 11.0 1.0 2.0 3.0 0.0 1.0 1.0 3.0 2.0
0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 4.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 1.0 0.0 0.0 7.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 2.0 1.0 2.0 5.0 1.0 0.0 2.0 0.0
8.0 1.0 3.0 3.0 1.0 1.0 7.0 8.0 3.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0
1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 6.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 6.0
0.4821428571428571
numTrees 5 featureSubsetStrategy all impurity gini maxDepth 6
Test Error = 0.4017857142857143
|================== Confusion matrix ==========================
3.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0
0.0 11.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 7.0 0.0 0.0 1.0 0.0 1.0
0.0 0.0 1.0 0.0 2.0 5.0 0.0 0.0 1.0 0.0
7.0 1.0 3.0 2.0 1.0 0.0 8.0 1.0 2.0 1.0
1.0 0.0 1.0 2.0 0.0 1.0 0.0 9.0 2.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 8.0 0.0
2.0 0.0 1.0 2.0 2.0 0.0 1.0 1.0 2.0 8.0
0.598214285714285
numTrees 5 featureSubsetStrategy all impurity entropy maxDepth 3
Test Error = 0.6875
|================== Confusion matrix ==========================
11.0 1.0 0.0 1.0 1.0 8.0 0.0
0.0 2.0 1.0 1.0 0.0 0.0 0.0
0.0 2.0 8.0 0.0 1.0 1.0 6.0
0.0 4.0 0.0 7.0 0.0 0.0 2.0
0.0 0.0 1.0 2.0 4.0 2.0 0.0
0.0 0.0 0.0 1.0 0.0 1.0 0.0
2.0 4.0 0.0 1.0 0.0 0.0 2.0
0.3125
Compilation completed successfully in 5s 781ms (25 minutes ago)
256:28 LF: UTF-8:

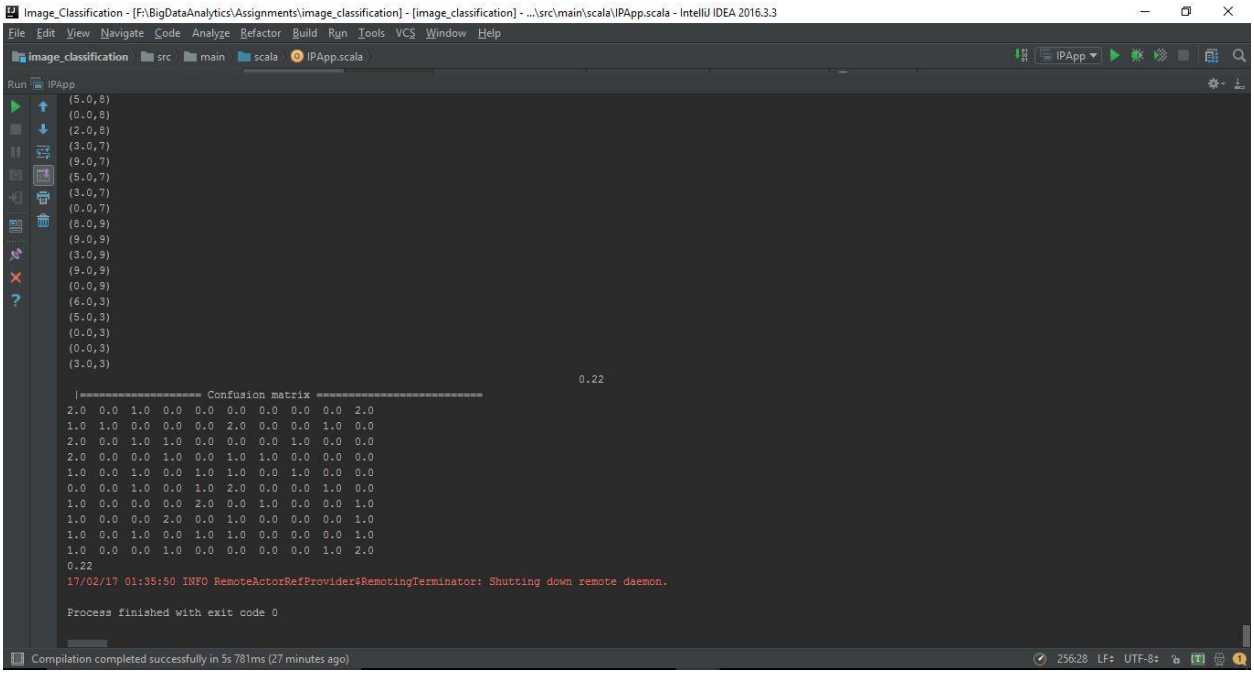

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
0.0 0.0 0.0 0.0 10.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.0 4.0 0.0 1.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 1.0 0.0 0.0 1.0 0.0 2.0 4.0 0.0
2.0 2.0 2.0 0.0 1.0 0.0 0.0 5.0 1.0
0.4553571428571285
numTrees 5 featureSubsetStrategy onethird impurity entropy maxDepth 6
Test Error = 0.2857142857142857
|================== Confusion matrix ==========================
13.0 0.0 2.0 0.0 2.0 1.0 1.0 0.0 0.0
0.0 11.0 2.0 0.0 0.0 0.0 2.0 0.0 0.0
0.0 0.0 5.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 10.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 0.0 0.0 11.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 1.0 8.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 9.0 0.0
0.0 2.0 0.0 0.0 0.0 0.0 0.0 5.0 9.0
0.7142857142857143
numTrees 6 featureSubsetStrategy all impurity gini maxDepth 3
Test Error = 0.6875
|================== Confusion matrix ==========================
11.0 1.0 3.0 3.0 1.0 1.0 1.0
2.0 8.0 2.0 2.0 5.0 0.0 3.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0
0.0 4.0 4.0 2.0 3.0 5.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.3125
numTrees 6 featureSubsetStrategy all impurity gini maxDepth 4
Test Error = 0.625
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
2.0 11.0 1.0 2.0 5.0 0.0 1.0 5.0 3.0
Compilation completed successfully in 5s 781ms (25 minutes ago)
256:28 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
0.0 1.0 4.0 1.0 1.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0
0.0 0.0 2.0 1.0 2.0 5.0 0.0 2.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
9.0 1.0 3.0 3.0 1.0 1.0 8.0 9.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.375
numTrees 6 featureSubsetStrategy onethird impurity gini maxDepth 5
Test Error = 0.5178571428571429
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
2.0 11.0 1.0 2.0 3.0 0.0 1.0 1.0 3.0 2.0
0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 4.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 1.0 0.0 0.0 7.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 2.0 1.0 2.0 5.0 1.0 0.0 2.0 0.0
8.0 1.0 3.0 3.0 1.0 1.0 7.0 8.0 3.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0
1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 6.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 6.0
0.4821428571428571
numTrees 6 featureSubsetStrategy onethird impurity gini maxDepth 6
Test Error = 0.4017857142857143
|================== Confusion matrix ==========================
3.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0
0.0 11.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 7.0 0.0 0.0 1.0 0.0 1.0
0.0 0.0 1.0 0.0 2.0 5.0 0.0 0.0 1.0 0.0
7.0 1.0 3.0 2.0 1.0 0.0 8.0 1.0 2.0 1.0
1.0 0.0 1.0 2.0 0.0 1.0 0.0 9.0 2.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 8.0 0.0
2.0 0.0 1.0 2.0 2.0 0.0 1.0 1.0 2.0 8.0
Compilation completed successfully in 5s 781ms (25 minutes ago)
256:28 LF: UTF-8:

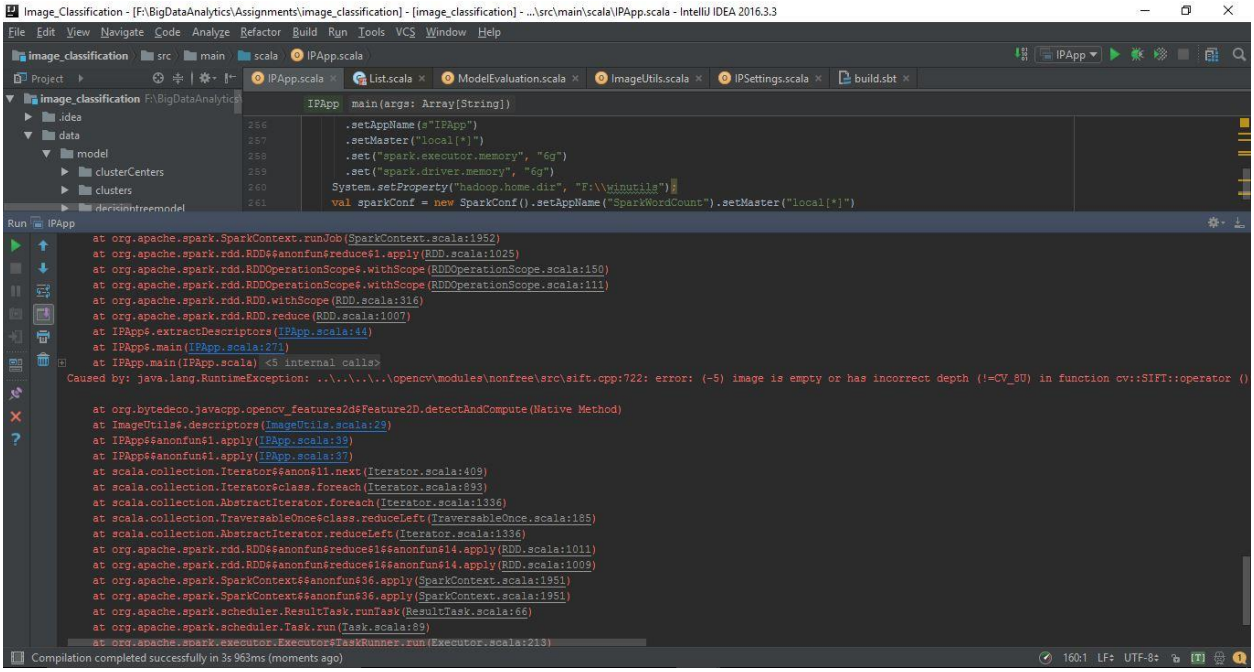

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
0.0 2.0 0.0 0.0 0.0 0.0 0.0 5.0 9.0
0.7142857142857143
numTrees 9 featureSubsetStrategy sqrt impurity gini maxDepth 3
Test Error = 0.6875
|================== Confusion matrix ==========================
11.0 1.0 3.0 3.0 1.0 1.0 1.0
2.0 8.0 2.0 2.0 5.0 0.0 3.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0
0.0 4.0 4.0 2.0 3.0 5.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.3125
numTrees 9 featureSubsetStrategy sqrt impurity gini maxDepth 4
Test Error = 0.625
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
2.0 11.0 1.0 2.0 5.0 0.0 1.0 5.0 3.0
0.0 1.0 4.0 1.0 1.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0
0.0 0.0 2.0 1.0 2.0 5.0 0.0 2.0 3.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
9.0 1.0 3.0 3.0 1.0 1.0 8.0 9.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0
0.375
numTrees 9 featureSubsetStrategy sqrt impurity gini maxDepth 5
Test Error = 0.5178571428571429
|================== Confusion matrix ==========================
2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
2.0 11.0 1.0 2.0 3.0 0.0 1.0 1.0 3.0 2.0
0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 4.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 1.0 0.0 0.0 7.0 0.0 0.0 0.0 2.0 1.0
0.0 0.0 2.0 1.0 2.0 5.0 1.0 0.0 2.0 0.0
Compilation completed successfully in 5s 781ms (25 minutes ago)
256:28 LF: UTF-8:

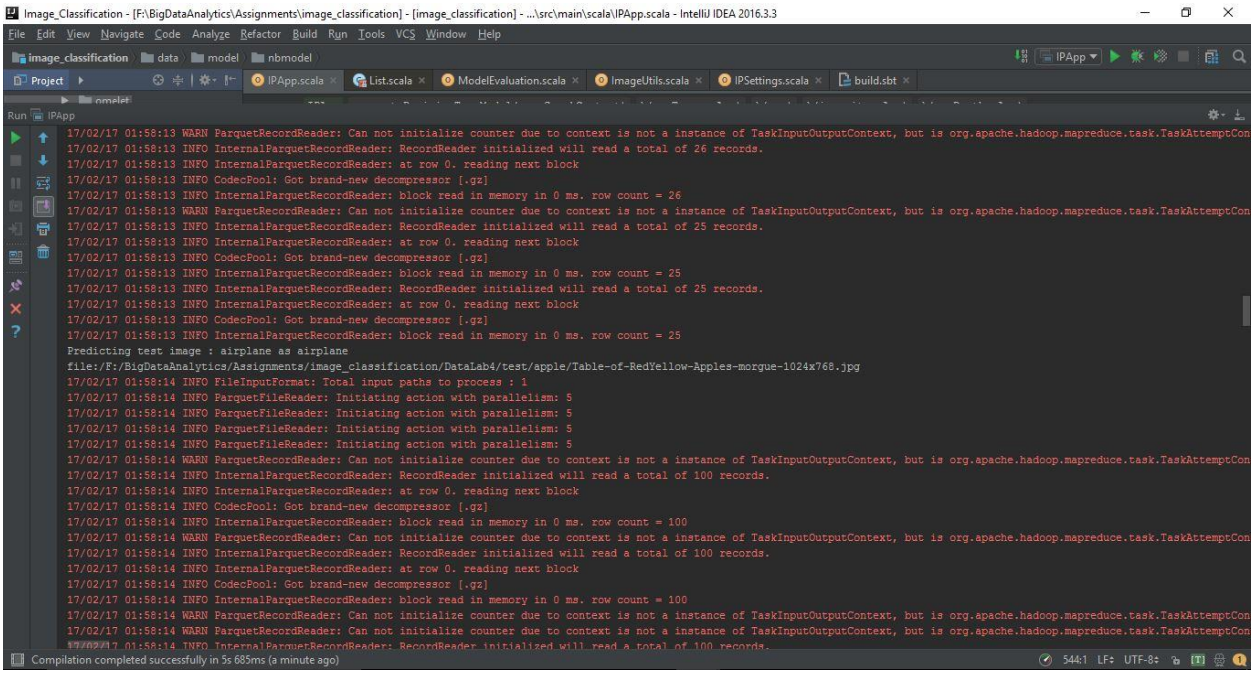

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Run IPApp
17/02/17 01:34:50 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:34:50 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:34:50 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:34:50 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
400 5
Histogram size : (400, 1)
Histogram : [0.0, 0.0, 0.0020833334, 0.0020833334, 0.0062500006, 0.0, 0.0031250003, 0.0010416667, 0.0031250003, 0.0020833334, 0.0010416667, 0.0, 0.0031250003, 0.0031250003, 0.002083333
--Histogram size : 400
17/02/17 01:34:50 INFO FileInputFormat: Total input paths to process : 1
17/02/17 01:34:50 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:34:50 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:34:50 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:34:50 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:34:50 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:34:50 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:34:50 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:34:50 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:34:50 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:34:50 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:34:50 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 1 ms. row count = 25
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
17/02/17 01:34:50 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:34:50 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:34:50 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:34:50 INFO InternalParquetRecordReader: block read in memory in 2 ms. row count = 25
17/02/17 01:34:50 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
Compilation completed successfully in 5s 781ms (26 minutes ago)
256:28 LF: UTF-8:

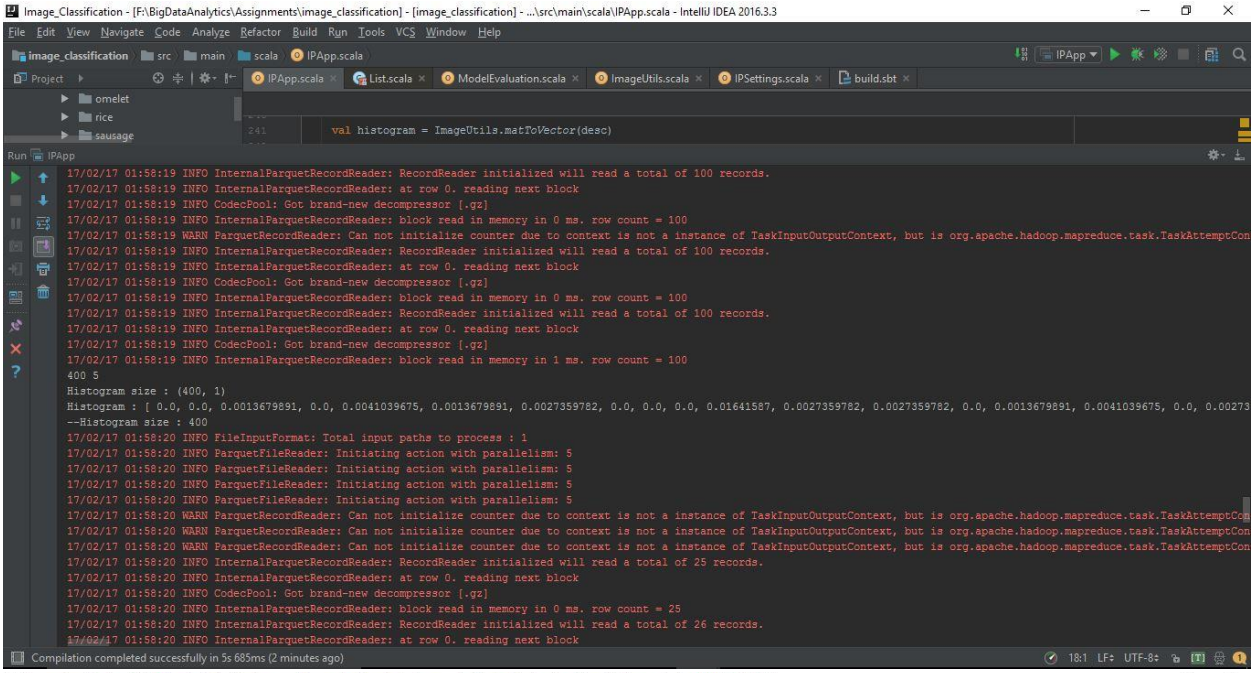


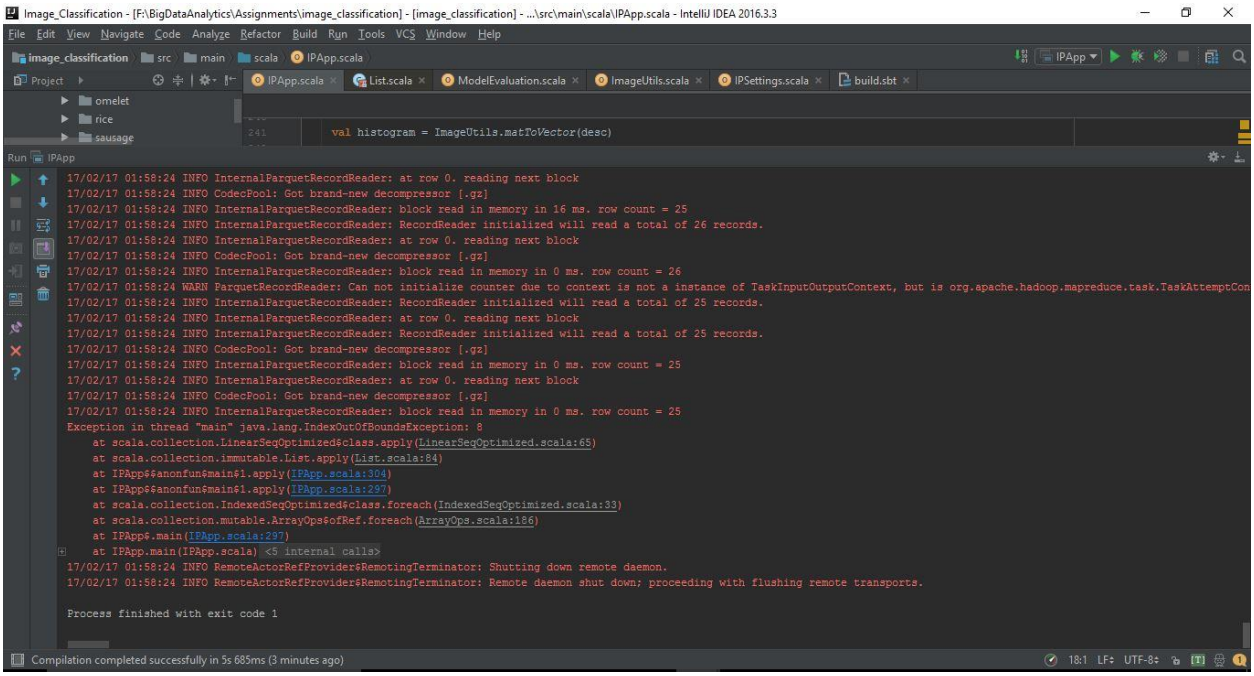


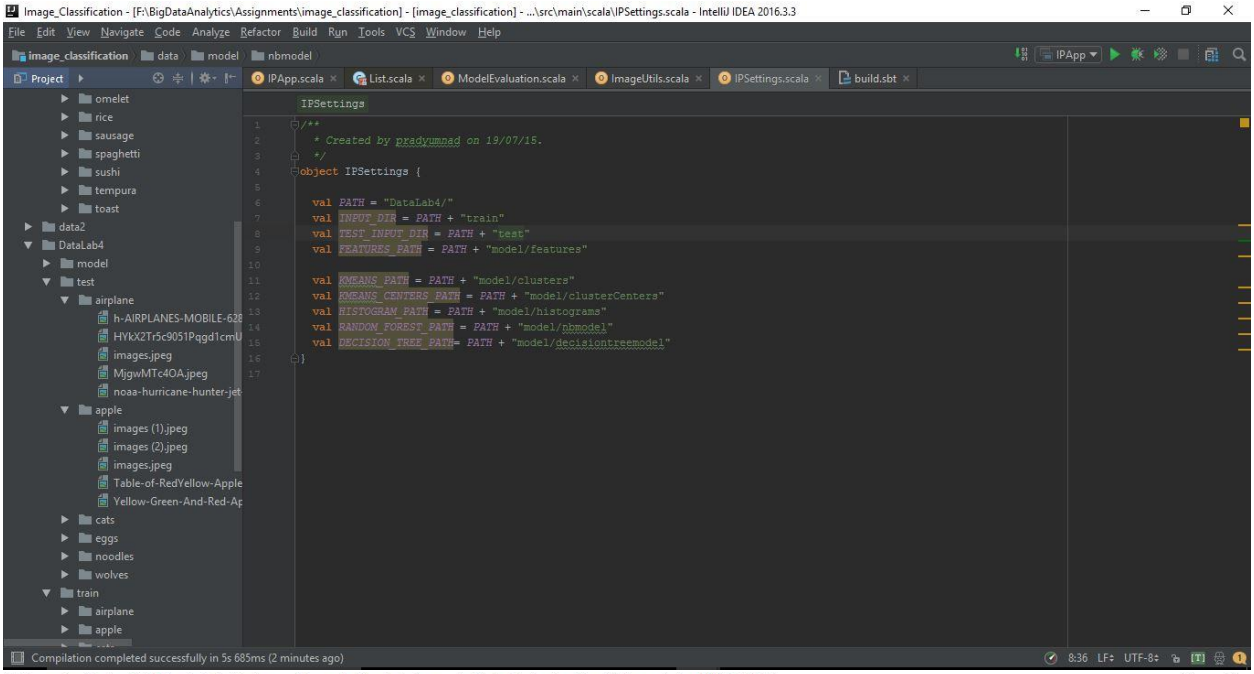
Output for other data set:

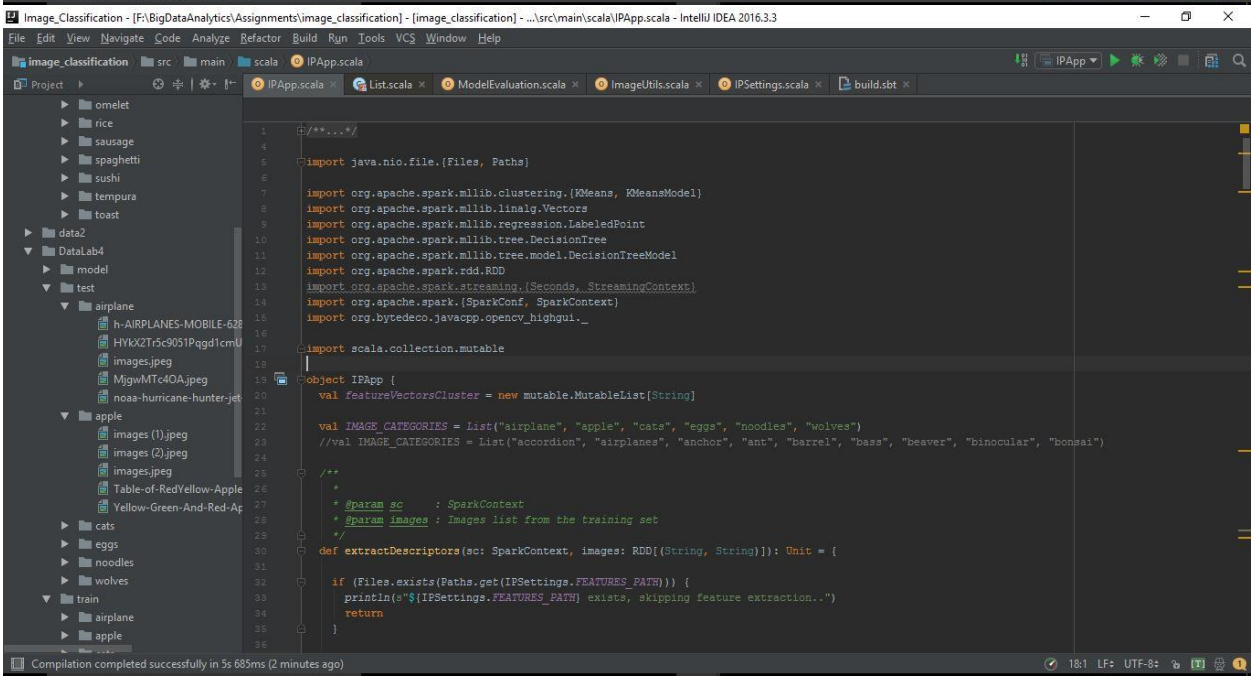

Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
image_classification F:\BigDataAnalytics\
.idea
data
model
clusterCenters
clusters
IPApp main(args: Array[String])
.setAppName(a"IPApp")
.setMaster("local[*]")
.set("spark.executor.memory", "6g")
.set("spark.driver.memory", "6g")
System.setProperty("hadoop.home.dir", "F:\\winutils")
val sparkConf = new SparkConf().setAppName("SparkWordCount").setMaster("local[*]")
Run IPApp
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1952)
at org.apache.spark.rdd.RDD$$anonfun$reduce$1.apply(RDD.scala:1025)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:111)
at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
at org.apache.spark.rdd.RDD.reduce(RDD.scala:1007)
at IPApp$.extractDescriptors(IPApp.scala:44)
at IPApp$.main(IPApp.scala:271)
at IPApp.main(IPApp.scala) <5 internal calls>
Caused by: java.lang.RuntimeException: ..\..\..\..\opencv\modules\nonfree\src\sift.cpp:722: error: (-5) image is empty or has incorrect depth (!=CV_8U) in function cv::SIFT::operator ()
at org.bytedeco.javacpp.opencv_features2d$Feature2D.detectAndCompute(Native Method)
at ImageUtils$.descriptors(ImageUtils.scala:29)
at IPApp$$anonfun$1.apply(IPApp.scala:39)
at IPApp$$anonfun$1.apply(IPApp.scala:37)
at scala.collection.Iterator$$anon$11.next(Iterator.scala:409)
at scala.collection.Iterator$class.foreach(Iterator.scala:893)
at scala.collection.AbstractIterator.foreach(Iterator.scala:1336)
at scala.collection.TraversableOnce$class.reduceLeft(TraversableOnce.scala:185)
at scala.collection.AbstractIterator.reduceLeft(Iterator.scala:1336)
at org.apache.spark.rdd.RDD$$anonfun$reduce$1$$anonfun$14.apply(RDD.scala:1011)
at org.apache.spark.rdd.RDD$$anonfun$reduce$1$$anonfun$14.apply(RDD.scala:1009)
at org.apache.spark.SparkContext$$anonfun$36.apply(SparkContext.scala:1951)
at org.apache.spark.SparkContext$$anonfun$36.apply(SparkContext.scala:1951)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
at org.apache.spark.scheduler.Task.run(Task.scala:89)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:213)
Compilation completed successfully in 3s 963ms (moments ago)
160:1 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification data model nbmodel
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
Run IPApp
17/02/17 01:58:13 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:13 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 26 records.
17/02/17 01:58:13 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:13 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:13 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 26
17/02/17 01:58:13 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:13 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:58:13 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:13 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:13 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
17/02/17 01:58:13 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:58:13 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:13 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:13 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
Predicting test image : airplane as airplane
file:/F:/BigDataAnalytics/Assignments/image_classification/DataLab4/test/apple/Table-of-RedYellow-Apples-morgue-1024x768.jpg
17/02/17 01:58:14 INFO FileInputFormat: Total input paths to process : 1
17/02/17 01:58:14 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:14 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:14 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:14 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:14 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:14 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:58:14 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:14 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:14 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:58:14 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:14 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:58:14 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:14 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:14 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:58:14 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:14 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
Compilation completed successfully in 5s 685ms (a minute ago)
544:1 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
omelet
rice
sausage
241 val histogram = ImageUtils.matToVector(desc)
Run IPApp
17/02/17 01:58:19 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:58:19 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:19 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:19 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:58:19 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:19 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:58:19 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:19 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:19 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/17 01:58:19 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/17 01:58:19 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:19 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:19 INFO InternalParquetRecordReader: block read in memory in 1 ms. row count = 100
400 5
Histogram size : (400, 1)
Histogram : [0.0, 0.0, 0.0013679891, 0.0, 0.0041039675, 0.0013679891, 0.0027359782, 0.0, 0.0, 0.0, 0.01641587, 0.0027359782, 0.0027359782, 0.0, 0.0013679891, 0.0041039675, 0.0, 0.00273
--Histogram size : 400
17/02/17 01:58:20 INFO FileInputFormat: Total input paths to process : 1
17/02/17 01:58:20 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:20 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:20 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:20 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/17 01:58:20 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:20 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:20 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:20 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:58:20 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:20 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:20 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
17/02/17 01:58:20 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 26 records.
17/02/17 01:58:20 INFO InternalParquetRecordReader: at row 0. reading next block
Compilation completed successfully in 5s 685ms (2 minutes ago)
18:1 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
omelet
rice
sausage
241 val histogram = ImageUtils.matToVector(desc)
Run IPApp
17/02/17 01:58:24 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:24 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:24 INFO InternalParquetRecordReader: block read in memory in 16 ms. row count = 25
17/02/17 01:58:24 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 26 records.
17/02/17 01:58:24 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:24 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:24 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 26
17/02/17 01:58:24 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.TaskAttemptCon
17/02/17 01:58:24 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:58:24 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:24 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 25 records.
17/02/17 01:58:24 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:24 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
17/02/17 01:58:24 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/17 01:58:24 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/17 01:58:24 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 25
Exception in thread "main" java.lang.IndexOutOfBoundsException: 8
at scala.collection.LinearSeqOptimized$class.apply(LinearSeqOptimized.scala:65)
at scala.collection.immutable.List.apply(List.scala:84)
at IPApp$$anonfun$main$1.apply(IPApp.scala:304)
at IPApp$$anonfun$main$1.apply(IPApp.scala:297)
at scala.collection.IndexedSeqOptimized$class.foreach(IndexedSeqOptimized.scala:33)
at scala.collection.mutable.ArrayOps$ofRef.foreach(ArrayOps.scala:186)
at IPApp$.main(IPApp.scala:297)
at IPApp.main(IPApp.scala) <5 internal calls>
17/02/17 01:58:24 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
17/02/17 01:58:24 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
Process finished with exit code 1
Compilation completed successfully in 5s 685ms (3 minutes ago)
18:1 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPSettings.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification data model nbmodel
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
IPSettings
/**
* Created by pradyumnad on 19/07/15.
*/
object IPSettings {
val PATH = "DataLab4/"
val INPUT_DIR = PATH + "train"
val TEST_INPUT_DIR = PATH + "test"
val FEATURES_PATH = PATH + "model/features"
val KMEANS_PATH = PATH + "model/clusters"
val KMEANS_CENTERS_PATH = PATH + "model/clusterCenters"
val HISTOGRAM_PATH = PATH + "model/histograms"
val RANDOM_FOREST_PATH = PATH + "model/nbmodel"
val DECISION_TREE_PATH= PATH + "model/decisiontreemodel"
}
omelet
rice
sausage
spaghetti
sushi
tempura
toast
data2
DataLab4
model
test
airplane
h-AIRPLANES-MOBILE-628
HYkX2Tr5c9051Pqgd1cmU
images.jpeg
MjgwMTc4OA.jpeg
noaa-hurricane-hunter-jet
apple
images (1).jpeg
images (2).jpeg
images.jpeg
Table-of-RedYellow-Apple
Yellow-Green-And-Red-Ap
cats
eggs
noodles
wolves
train
airplane
apple
Compilation completed successfully in 5s 685ms (2 minutes ago)
8:36 LF: UTF-8:


Image_Classification - [F:\BigDataAnalytics\Assignments\image_classification] - [image_classification] - ...\src\main\scala\IPApp.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
image_classification src main scala IPApp.scala
IPApp
Project
IPApp.scala List.scala ModelEvaluation.scala ImageUtils.scala IPSettings.scala build.sbt
/**...*/
import java.nio.file.{Files, Paths}
import org.apache.spark.mllib.clustering.{KMeans, KMeansModel}
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.tree.DecisionTree
import org.apache.spark.mllib.tree.model.DecisionTreeModel
import org.apache.spark.rdd.RDD
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.{SparkConf, SparkContext}
import org.bytedeco.javacpp.opencv_highgui._
import scala.collection.mutable
object IPApp {
val featureVectorsCluster = new mutable.MutableList[String]
val IMAGE_CATEGORIES = List("airplane", "apple", "cats", "eggs", "noodles", "wolves")
//val IMAGE_CATEGORIES = List("accordion", "airplanes", "anchor", "ant", "barrel", "bass", "beaver", "binocular", "bonsai")
/**
*
* @param sc : SparkContext
* @param images : Images list from the training set
*/
def extractDescriptors(sc: SparkContext, images: RDD[(String, String)]): Unit = {
if (Files.exists(Paths.get(IPSettings.FEATURES_PATH))) {
println(s"${IPSettings.FEATURES_PATH} exists, skipping feature extraction..")
return
}
omelet
rice
sausage
spaghetti
sushi
tempura
toast
data2
DataLab4
model
test
airplane
h-AIRPLANES-MOBILE-628
HYkX2Tr5c9051Pqgd1cmU
images.jpeg
MjgwMTc4OA.jpeg
noaa-hurricane-hunter-jet
apple
images (1).jpeg
images (2).jpeg
images.jpeg
Table-of-RedYellow-Apple
Yellow-Green-And-Red-Ap
cats
eggs
noodles
wolves
train
airplane
apple
Compilation completed successfully in 5s 685ms (2 minutes ago)
18:1 LF: UTF-8: