Big Data Framework

Application of Machine Learning algorithm and Multi-Node Cluster

Reg. No: 18MCB1003, 18MCB1009, 18MCB1015, 18MCC1004, 18MCB1005

School: School of Computer Science and Engineering

Program: MTech Big Data

Date of Submission: 17-04-2019

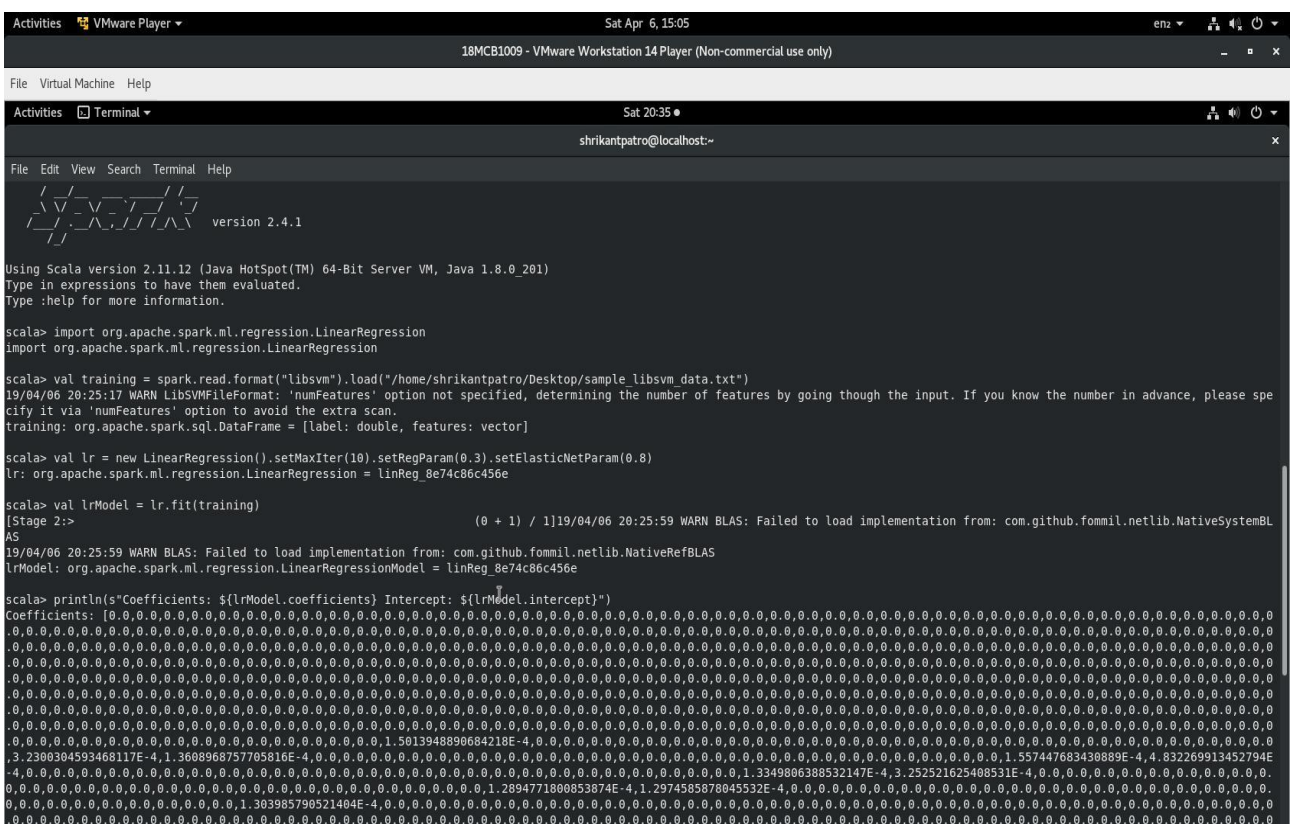
Aim: To apply Machine Learning Algorithm in Hadoop and Spark Framework.

Dataset: Fossils Dataset

Linear Regression

```
import org.apache.spark.ml.regression.LinearRegression
val training =

spark.read.format("libsvm").load("/home/shrikantpatro/Desktop/sample_libsvm_data.txt")

val lr = new LinearRegression().setMaxIter(10).setRegParam(0.3).setElasticNetParam(0.8)

val lrModel = lr.fit(training)

println(s"Coefficients: ${lrModel.coefficients} Intercept: ${lrModel.intercept}")
val trainingSummary = lrModel.summary println(s"numIterations:
${trainingSummary.totalIterations}")

println(s"objectiveHistory: [${trainingSummary.objectiveHistory.mkString(",")}]")

trainingSummary.residuals.show()

println(s"RMSE: ${trainingSummary.rootMeanSquaredError}")

println(s"r2: ${trainingSummary.r2}")
```

Output



```
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.1
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_201)
Type in expressions to have them evaluated.
Type :help for more information.

scala> import org.apache.spark.ml.regression.LinearRegression
import org.apache.spark.ml.regression.LinearRegression

scala> val training = spark.read.format("libsvm").load("/home/shrikantpatro/Desktop/sample_libsvm_data.txt")
19/04/06 20:25:17 WARN LibSVMFileFormat: 'numFeatures' option not specified, determining the number of features by going though the input. If you know the number in advance, please spe
cify it via 'numFeatures' option to avoid the extra scan.
training: org.apache.spark.sql.DataFrame = [label: double, features: vector]

scala> val lr = new LinearRegression().setMaxIter(10).setRegParam(0.3).setElasticNetParam(0.8)
lr: org.apache.spark.ml.regression.LinearRegression = linReg_8e74c86c456e

scala> val lrModel = lr.fit(training)
[Stage 2:>                                                          (0 + 1) / 1]19/04/06 20:25:59 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBL
AS
19/04/06 20:25:59 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
lrModel: org.apache.spark.ml.regression.LinearRegressionModel = linReg_8e74c86c456e

scala> println(s"Coefficients: ${lrModel.coefficients} Intercept: ${lrModel.intercept}")
Coefficients: [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.5013948890684218E-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,3.2300304593468117E-4,1.3608968757705816E-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.557447683430889E-4,4.832269913452794E
-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.3349806388532147E-4,3.252521625408531E-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.2894771800853874E-4,1.2974585878045532E-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.303985790521404E-4,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
```

Activities VMware Player Sat Apr 6, 15:05 en2

18MCB1009 - VMware Workstation 14 Player (Non-commercial use only)

File Virtual Machine Help

Activities Terminal Sat 20:35

shrikantpatro@localhost:~

File Edit View Search Terminal Help

```
scala> val trainingSummary = lrModel.summary
trainingSummary: org.apache.spark.ml.regression.LinearRegressionTrainingSummary = org.apache.spark.ml.regression.LinearRegressionTrainingSummary@352969e7

scala> println(s"numIterations: ${trainingSummary.totalIterations}")
numIterations: 11

scala> println(s"objectiveHistory: [${trainingSummary.objectiveHistory.mkString(",")}]")
objectiveHistory: [0.5000000000000001,0.46812248072553964,0.4373389815256357,0.4297108031755194,0.41639320086308157,0.40838618107748603,0.4048882753551667,0.3980443325208099,0.39623068
62562947,0.393309268421889 5,0.39221349422695734]

scala> trainingSummary.residuals.show()
+-------------------+
|          residuals|
+-------------------+
| -0.294554000642913|
|0.20337546495076442|
|0.18966293033659576|
| 0.2611001145786045|
|0.19278322750469878|
|-0.2999590222435593|
| 0.2069812030941396|
| 0.6573359136392394|
|-0.3116739652739924|
| -0.294554000642913|
|0.17759588354678657|
| -0.294554000642913|
| -0.294554000642913|
| 0.2521446212089007|
|-0.3722226451008439|
|0.18743032400750725|
| -0.294554000642913|
| -0.294554000642913|
|0.20344597554231925|
|  0.209421990994911|
+-------------------+
only showing top 20 rows


scala> println(s"RMSE: ${trainingSummary.rootMeanSquaredError}")
RMSE: 0.2650332068205178

scala> println(s"r2: ${trainingSummary.r2}")
r2: 0.713412481772471

scala>
```

To release input, press Ctrl+Alt

Logistic Regression

```
[18mcb1003@localhost ~]$ jps
10243 Jps
4072 NodeManager
3481 SecondaryNameNode
2985 NameNode
3212 DataNode
3709 ResourceManager
[18mcb1003@localhost ~]$ spark-shell
2019-04-06 22:14:51 WARN  Utils:66 - Your hostname, localhost.localdomain resolves to a
loopback address: 127.0.0.1; using 172.16.243.140 instead (on interface ens33)
2019-04-06 22:14:51 WARN  Utils:66 - Set SPARK_LOCAL_IP if you need to bind to another
address
2019-04-06 22:14:52 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://172.16.243.140:4040
Spark context available as 'sc' (master = local[*], app id = local-1554569109943).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_201)
Type in expressions to have them evaluated.
Type :help for more information.

scala> import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.classification.LogisticRegression

scala> val training =
spark.read.format("libsvm").load("/home/18mcb1003/Downloads/datachi.txt")
2019-04-06 22:17:45 WARN  LibSVMFileFormat:66 - 'numFeatures' option not specified,
determining the number of features by going though the input. If you know the number in
advance, please specify it via 'numFeatures' option to avoid the extra scan.
training: org.apache.spark.sql.DataFrame = [label: double, features: vector]
```

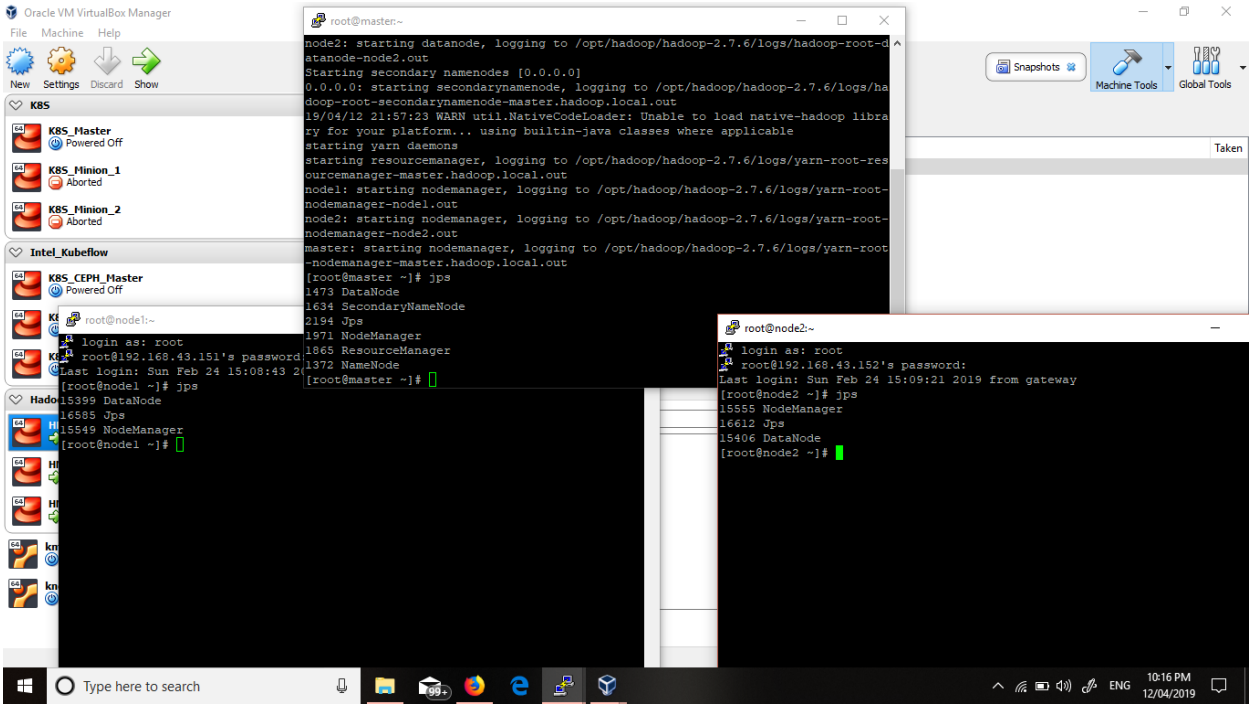
```
scala> val lr = new
LogisticRegression().setMaxIter(10).setRegParam(0.3).setElasticNetParam(0.8)
lr: org.apache.spark.ml.classification.LogisticRegression = logreg_706fef944da9

scala> val lrModel = lr.fit(training)
2019-04-06 22:18:11 WARN  BLAS:61 - Failed to load implementation from:
com.github.fommil.netlib.NativeSystemBLAS
2019-04-06 22:18:11 WARN  BLAS:61 - Failed to load implementation from:
com.github.fommil.netlib.NativeRefBLAS
lrModel: org.apache.spark.ml.classification.LogisticRegressionModel =
LogisticRegressionModel: uid = logreg_706fef944da9, numClasses = 2, numFeatures = 692

scala> println(s"Coefficients: ${lrModel.coefficients} Intercept: ${lrModel.intercept}")
Coefficients:
(692,[244,263,272,300,301,328,350,351,378,379,405,406,407,428,433,434,455,456,461,462,483
,484,489,490,496,511,512,517,539,540,568],[-7.353983524188197E-5,-9.102738505589466E-
5,-1.9467430546904298E-4,-2.0300642473486668E-4,-3.1476183314863995E-5,-
6.842977602660743E-5,1.5883626898239883E-5,1.4023497091372047E-
5,3.5432047524968605E-4,1.1443272898171087E-4,1.0016712383666666E-
4,6.014109303795481E-4,2.840248179122762E-4,-1.1541084736508837E-
4,3.85996886312906E-4,6.35019557424107E-4,-1.1506412384575676E-4,-
1.5271865864986808E-4,2.804933808994214E-4,6.070117471191634E-4,-
2.008459663247437E-4,-1.421075579290126E-4,2.739010341160883E-
4,2.7730456244968115E-4,-9.838027027269332E-5,-3.808522443517704E-4,-
2.5315198008555033E-4,2.7747714770754307E-4,-2.443619763919199E-4,-
0.0015394744687597765,-2.3073328411331293E-4]) Intercept: 0.22456315961250325

scala> val mlr = new
LogisticRegression().setMaxIter(10).setRegParam(0.3).setElasticNetParam(0.8).setFamily("mult
inomial")
mlr: org.apache.spark.ml.classification.LogisticRegression = logreg_c7296343cc34

scala> val mlrModel = mlr.fit(training)
mlrModel: org.apache.spark.ml.classification.LogisticRegressionModel =
LogisticRegressionModel: uid = logreg_c7296343cc34, numClasses = 2, numFeatures = 692

scala> println(s"Multinomial coefficients: ${mlrModel.coefficientMatrix}")
Multinomial coefficients: 2 x 692 CSCMatrix
(0,244) 4.290365458958277E-5
(1,244) -4.290365458958294E-5
(0,263) 6.488313287833108E-5
(1,263) -6.488313287833092E-5
(0,272) 1.2140666790834663E-4
(1,272) -1.2140666790834657E-4
(0,300) 1.3231861518665612E-4
(1,300) -1.3231861518665607E-4
```

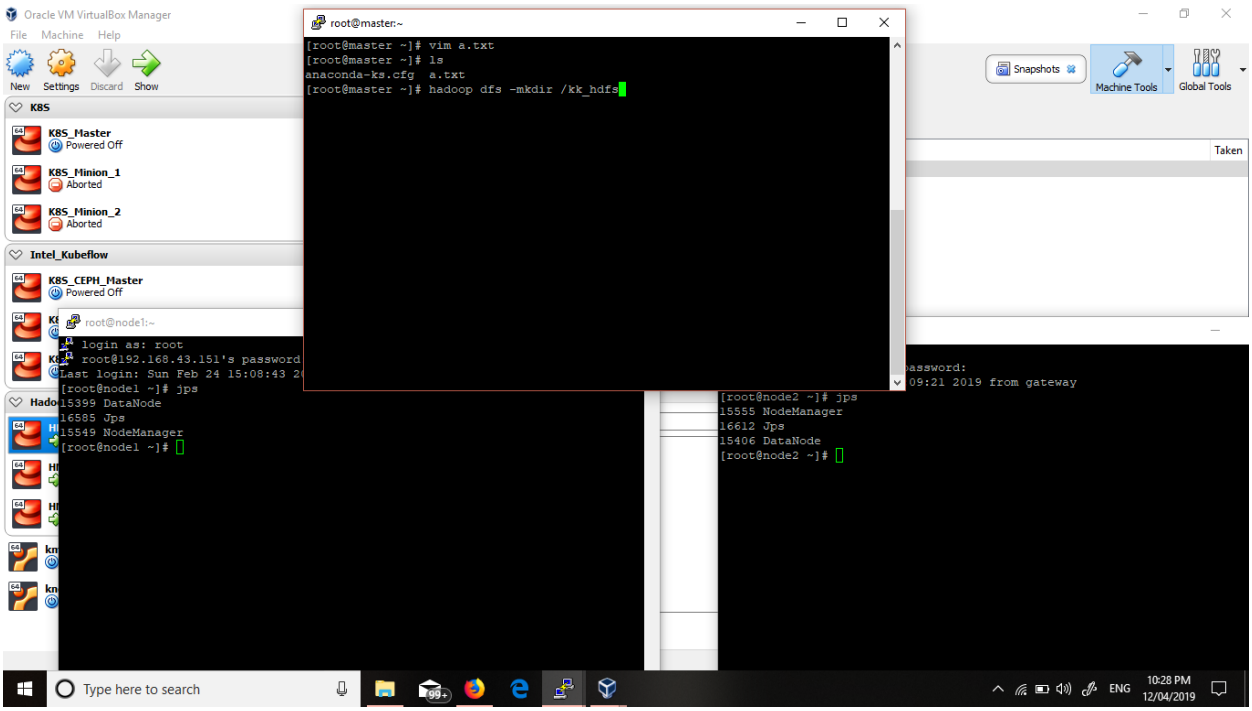
(0,350) -6.775444746760509E-7
(1,350) 6.775444746761932E-7
(0,351) -4.899237909429297E-7
(1,351) 4.899237909430322E-7
(0,378) -3.5812102770679596E-5
(1,378) 3.581210277067968E-5
(0,379) -2.3539704331222065E-5
(1,379) 2.353970433122204E-5
(0,405) -1.90295199030314E-5
(1,405) 1.90295199030314E-5
(0,406) -5.626696935778909E-4
(1,406) 5.626696935778912E-4
(0,407) -5.121519619099504E-5
(1,407) 5.1215196190995074E-5
(0,428) 8.080614545413342E-5
(1,428) -8.080614545413331E-5
(0,433) -4.256734915330487E-5
(1,433) 4.256734915330495E-5
(0,434) -7.080191510151425E-4
(1,434) 7.080191510151435E-4
(0,455) 8.094482475733589E-5
(1,455) -8.094482475733582E-5
(0,456) 1.0433687128309833E-4
(1,456) -1.0433687128309814E-4
(0,461) -5.4466605046259246E-5
(1,461) 5.4466605046259286E-5
(0,462) -5.667133061990392E-4
(1,462) 5.667133061990392E-4
(0,483) 1.2495896045528374E-4
(1,483) -1.249589604552838E-4
(0,484) 9.810519424784944E-5
(1,484) -9.810519424784941E-5
(0,489) -4.88440907254626E-5
(1,489) 4.8844090725462606E-5
(0,490) -4.324392733454803E-5
(1,490) 4.324392733454811E-5
(0,496) 6.903351855620161E-5
(1,496) -6.90335185562012E-5
(0,511) 3.946505594172827E-4
(1,511) -3.946505594172831E-4
(0,512) 2.621745995919226E-4
(1,512) -2.621745995919226E-4
(0,517) -4.459475951170906E-5
(1,517) 4.459475951170901E-5
(0,539) 2.5417562428184555E-4
(1,539) -2.5417562428184555E-4

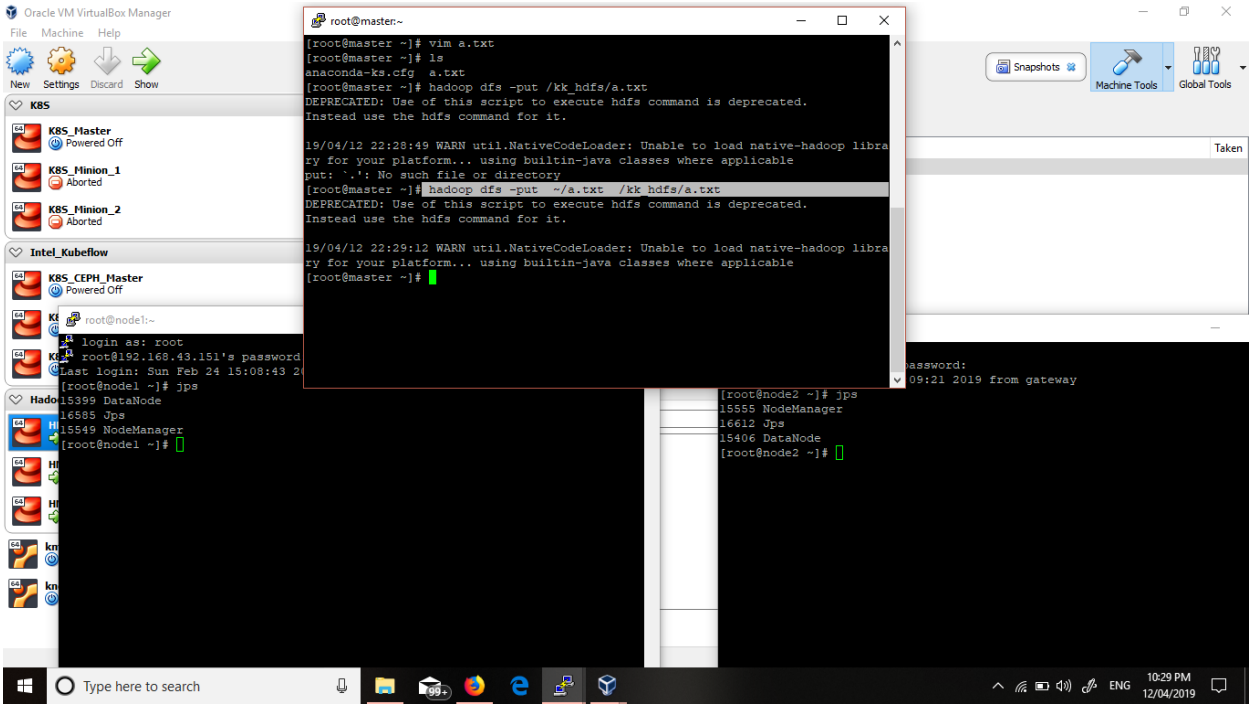
(0,540) 5.271781246228031E-4
(1,540) -5.271781246228032E-4
(0,568) 1.860255150352447E-4
(1,568) -1.8602551503524485E-4

scala> println(s"Multinomial intercepts: ${mlrModel.interceptVector}")
Multinomial intercepts: [-0.12065879445860686,0.12065879445860686]

Multi-node Cluster 1-master and 2 Slave




Oracle VM VirtualBox Manager
File Machine Help
New Settings Discard Show
K8S
K8S_Master
Powered Off
K8S_Minion_1
Aborted
K8S_Minion_2
Aborted
Intel_Kubeflow
K8S_CEPH_Master
Powered Off
root@master:~
[root@master ~]# vim a.txt
[root@master ~]# ls
anaconda-ks.cfg a.txt
[root@master ~]# hadoop dfs -mkdir /kk_hdfs
root@node1:~
login as: root
root@192.168.43.151's password
Last login: Sun Feb 24 15:08:43 2
[root@node1 ~]# jps
15399 DataNode
16585 Jps
15549 NodeManager
[root@node1 ~]#
password:
09:21 2019 from gateway
[root@node2 ~]# jps
15555 NodeManager
16612 Jps
15406 DataNode
[root@node2 ~]#
Snapshots
Machine Tools
Global Tools
Taken
Type here to search
ENG
10:28 PM
12/04/2019


Oracle VM VirtualBox Manager
File Machine Help
New Settings Discard Show
K8S
K8S_Master
Powered Off
K8S_Minion_1
Aborted
K8S_Minion_2
Aborted
Intel_Kubeflow
K8S_CEPH_Master
Powered Off
root@master:~
[root@master ~]# vim a.txt
[root@master ~]# ls
anaconda-ks.cfg a.txt
[root@master ~]# hadoop dfs -put /kk_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
19/04/12 22:28:49 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
put: `.': No such file or directory
[root@master ~]# hadoop dfs -put ~/a.txt /kk_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
19/04/12 22:29:12 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[root@master ~]#
root@node1:~
login as: root
root@192.168.43.151's password
Last login: Sun Feb 24 15:08:43 2
[root@node1 ~]# jps
15399 DataNode
16585 Jps
15549 NodeManager
[root@node1 ~]#
password:
09:21 2019 from gateway
[root@node2 ~]# jps
15555 NodeManager
16612 Jps
15406 DataNode
[root@node2 ~]#
Snapshots
Machine Tools
Global Tools
Taken
Type here to search
ENG
10:29 PM
12/04/2019

root@master:~

```
[root@master ~]# git clone git@github.com:khushalkunjir/WordCountJar.git
Cloning into 'WordCountJar'...
The authenticity of host 'github.com (192.30.253.112)' can't be established.
RSA key fingerprint is SHA256:nThbg6kXUpJWG17E1IGOCspRomTxdCARLviKw6E5SY8.
RSA key fingerprint is MD5:16:27:ac:a5:76:28:2d:36:63:1b:56:4d:eb:df:a6:48.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'github.com,192.30.253.112' (RSA) to the list of know
n hosts.
Permission denied (publickey).
fatal: Could not read from remote repository.

Please make sure you have the correct access rights
and the repository exists.
[root@master ~]# git clone https://github.com/khushalkunjir/WordCountJar.git
Cloning into 'WordCountJar'...
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (2/2), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (3/3), done.
[root@master ~]# ls
anaconda-ks.cfg  a.txt  WordCountJar  WordCountJar.git
[root@master ~]# WordCountJar/
-bash: WordCountJar/: Is a directory
[root@master ~]# wget https://raw.githubusercontent.com/khushalkunjir/WordCountJ
ar/master/WordCount.jar
--2019-04-12 23:21:43--  https://raw.githubusercontent.com/khushalkunjir/WordCou
ntJar/master/WordCount.jar
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.0.133
, 151.101.64.133, 151.101.128.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.0.13
3|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3108 (3.0K) [application/octet-stream]
Saving to: 'WordCount.jar'

100%[===================================>] 3,108       --.-K/s   in 0s

2019-04-12 23:21:43 (6.72 MB/s) - 'WordCount.jar' saved [3108/3108]

[root@master ~]# ls
anaconda-ks.cfg  a.txt  WordCount.jar  WordCountJar  WordCountJar.git
[root@master ~]# hadoop jar ~/WordCount.jar WordCountDriver /kk_hdfs/a.txt /kk_hdfs/outputwordcount
```

Type here to search ENG 11:22 PM 12/04/2019

root@master:~

```
doop-root-secondarynamenode-master.hadoop.local.out
19/04/13 00:46:37 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-res
ourcemanager-master.hadoop.local.out
node1: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-
nodemanager-node1.out
node2: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-
nodemanager-node2.out
master: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root
-nodemanager-master.hadoop.local.out
[root@master ~]# jps
4000 ResourceManager
3604 DataNode
4106 NodeManager
4138 Jps
3835 SecondaryNameNode
3502 NameNode
[root@master ~]# ls
anaconda-ks.cfg  WordCount.jar
[root@master ~]# vim a.txt
[root@master ~]# hadoop dfs -mkdir /vit_hdfs
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:48:14 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[root@master ~]# hadoop dfs -put ~/a
anaconda-ks.cfg  a.txt
[root@master ~]# hadoop dfs -put ~/a
anaconda-ks.cfg  a.txt
[root@master ~]# hadoop dfs -put ~/a.txt /vit_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:48:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[root@master ~]# hadoop dfs -put ~/a.txt /vit_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:49:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: `/vit_hdfs/a.txt': File exists
[root@master ~]#
```

Type here to search ENG 12:49 AM 13/04/2019

root@master:~

```
[root@master ~]# start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
19/04/13 00:45:27 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [master]
master: starting namenode, logging to /opt/hadoop/hadoop-2.7.6/logs/hadoop-root-
namenode-master.hadoop.local.out
master: starting datanode, logging to /opt/hadoop/hadoop-2.7.6/logs/hadoop-root-
datanode-master.hadoop.local.out
node2: starting datanode, logging to /opt/hadoop/hadoop-2.7.6/logs/hadoop-root-d
atanode-node2.out
node1: starting datanode, logging to /opt/hadoop/hadoop-2.7.6/logs/hadoop-root-d
atanode-node1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/hadoop-2.7.6/logs/ha
doop-root-secondarynamenode-master.hadoop.local.out
19/04/13 00:46:37 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-res
ourcemanager-master.hadoop.local.out
node1: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-
nodemanager-node1.out
node2: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root-
nodemanager-node2.out
master: starting nodemanager, logging to /opt/hadoop/hadoop-2.7.6/logs/yarn-root
-nodemanager-master.hadoop.local.out
[root@master ~]# jps
4000 ResourceManager
3604 DataNode
4106 NodeManager
4138 Jps
3835 SecondaryNameNode
3502 NameNode
[root@master ~]# ls
anaconda-ks.cfg  WordCount.jar
[root@master ~]# vim a.txt
[root@master ~]# hadoop dfs -mkdir /vit_hdfs
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:48:14 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[root@master ~]#
```

Type here to search ENG 12:48 AM 13/04/2019

root@master:~

```
[root@master ~]# ls
anaconda-ks.cfg  a.txt  WordCount.jar
[root@master ~]# hadoop jar ~/WordCount.jar WordCountDriver /vit_hdfs/a.txt /vit_hdfs/outputWordCount
19/04/13 00:50:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Type here to search ENG 12:50 AM 13/04/2019

root@master:~

```
anaconda-ks.cfg  a.txt  WordCount.jar
[root@master ~]# hadoop jar ~/WordCount.jar WordCountDriver /vit_hdfs/a.txt /vit_hdfs/outputWordCount
19/04/13 00:50:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/04/13 00:50:51 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.43.150:8032
19/04/13 00:50:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this.
19/04/13 00:50:55 INFO input.FileInputFormat: Total input paths to process : 1
19/04/13 00:50:57 INFO mapreduce.JobSubmitter: number of splits:1
19/04/13 00:50:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555096616980_0001
19/04/13 00:50:59 INFO impl.YarnClientImpl: Submitted application application_1555096616980_0001
19/04/13 00:50:59 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1555096616980_0001/
19/04/13 00:50:59 INFO mapreduce.Job: Running job: job_1555096616980_0001
19/04/13 00:51:32 INFO mapreduce.Job: Job job_1555096616980_0001 running in uber mode : false
19/04/13 00:51:32 INFO mapreduce.Job:  map 0% reduce 0%
19/04/13 00:51:48 INFO mapreduce.Job:  map 100% reduce 0%
19/04/13 00:52:12 INFO mapreduce.Job:  map 100% reduce 100%
19/04/13 00:52:13 INFO mapreduce.Job: Job job_1555096616980_0001 completed successfully
19/04/13 00:52:14 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=199
                FILE: Number of bytes written=245313
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=183
                HDFS: Number of bytes written=115
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=13243
                Total time spent by all reduces in occupied slots (ms)=20590
                Total time spent by all map tasks (ms)=13243
                Total time spent by all reduce tasks (ms)=20590
                Total vcore-milliseconds taken by all map tasks=13243
                Total vcore-milliseconds taken by all reduce tasks=20590
                Total megabyte-milliseconds taken by all map tasks=13560832
                Total megabyte-milliseconds taken by all reduce tasks=21084160
        Map-Reduce Framework
                Map input records=3
                Map output records=18
```

Type here to search ENG 12:52 AM 13/04/2019

root@master:~

```
anaconda-ks.cfg  a.txt
[root@master ~]# hadoop dfs -put ~/a.txt /vit_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:48:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[root@master ~]# hadoop dfs -put ~/a.txt /vit_hdfs/a.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/04/13 00:49:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: `/vit_hdfs/a.txt': File exists
[root@master ~]# clear
[root@master ~]# ls
anaconda-ks.cfg  a.txt  WordCount.jar
[root@master ~]# hadoop jar ~/WordCount.jar WordCountDriver /vit_hdfs/a.txt /vit_hdfs/outputWordCount
19/04/13 00:50:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/04/13 00:50:51 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.43.150:8032
19/04/13 00:50:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this.
19/04/13 00:50:55 INFO input.FileInputFormat: Total input paths to process : 1
19/04/13 00:50:57 INFO mapreduce.JobSubmitter: number of splits:1
19/04/13 00:50:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555096616980_0001
19/04/13 00:50:59 INFO impl.YarnClientImpl: Submitted application application_1555096616980_0001
19/04/13 00:50:59 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1555096616980_0001/
19/04/13 00:50:59 INFO mapreduce.Job: Running job: job_1555096616980_0001
19/04/13 00:51:32 INFO mapreduce.Job: Job job_1555096616980_0001 running in uber mode : false
19/04/13 00:51:32 INFO mapreduce.Job:  map 0% reduce 0%
19/04/13 00:51:48 INFO mapreduce.Job:  map 100% reduce 0%
19/04/13 00:52:12 INFO mapreduce.Job:  map 100% reduce 100%
19/04/13 00:52:13 INFO mapreduce.Job: Job job_1555096616980_0001 completed successfully
19/04/13 00:52:14 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=199
                FILE: Number of bytes written=245313
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=183
                HDFS: Number of bytes written=115
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
```

Type here to search ENG 12:52 AM 13/04/2019

root@master:~

```
                Map output records=18
                Map output bytes=157
                Map output materialized bytes=199
                Input split bytes=98
                Combine input records=0
                Combine output records=0
                Reduce input groups=17
                Reduce shuffle bytes=199
                Reduce input records=18
                Reduce output records=17
                Spilled Records=36
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=502
                CPU time spent (ms)=3890
                Physical memory (bytes) snapshot=298135552
                Virtual memory (bytes) snapshot=4159451136
                Total committed heap usage (bytes)=140091392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=85
        File Output Format Counters
                Bytes Written=115
[root@master ~]# stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
19/04/13 00:53:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Stopping namenodes on [master]
master: stopping namenode
node2: stopping datanode
node1: stopping datanode
master: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
19/04/13 00:53:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
stopping yarn daemons
stopping resourcemanager
```

Type here to search

ENG 12:53 AM 13/04/2019

KNN Classification on Iris Dataset.

Code:

Driver Class

```
import java.io.BufferedReader;
import java.io.IOException;
import java.io.InputStreamReader;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class Driver {
public static void main(String[] args) throws IOException, InterruptedException,
ClassNotFoundException
{
int num_features=0;
Configuration conf = new Configuration();
FileSystem hdfs = FileSystem.get(conf);
//args[0] is the path to the file which has features of the input waiting to be
classified.
BufferedReader br = new BufferedReader(new InputStreamReader(hdfs.open(new
Path(args[0]))));
String line=null;
while((line=br.readLine())!=null)
{
String[] feat=line.toString().split("\\ ");
for(int i=0;i<feat.length;i++)
conf.setFloat("feat"+i, Float.parseFloat(feat[i]));
num_features=feat.length;
break;
}
br.close();
hdfs.close();
conf.setInt("num_features",num_features);
//args[1] is the name of the entity to be classified.
conf.set("name",args[1]);
Job job = new Job(conf,"KNN Classification MapReduce");
job.setJarByClass(Driver.class);
//args[2] is the path to the input file which will be used for
FileInputFormat.setInputPaths(job, new Path(args[2])); //args[3] is the path to
the output file.
FileOutputFormat.setOutputPath(job, new Path(args[3]));
job.setMapperClass(Map.class); //job.setCombinerClass(Reduce.class);
```

```
job.setReducerClass(Reduce.class); job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class); job.waitForCompletion(true);
}
}
```

Mapper Class

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.Collections;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class Map extends Mapper<LongWritable, Text, Text, Text>
{
public static long byteoffset = 0;
public static Float[] feat=null;
public static String species=null;
public static ArrayList<String> dists=new ArrayList<String>();
public static float min_dist=0;
public static int num_features=0;
public static float euc_dist(Float[] feat, Float[] test,int num){
float distance=0;
float val=0;
for(int i=0;i<num;i++)
{
val+=((feat[i]-test[i])*(feat[i]-test[i]));
}
distance=(float) Math.sqrt(val);
return distance;
}
@Override
public void setup(Context context) throws IOException, InterruptedException{
num_features=(context.getConfiguration().getInt("num_features",1));
feat=new Float[num_features];
for(int i=0;i<num_features;i++)
{
feat[i]=(context.getConfiguration().getFloat("feat"+i, 0)); }
18MCB1015 Divyansh Gupta MTech Big Data
}
public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {
String[] characteristics=value.toString().split("\\ ");
Float[] test=new Float[num_features];
for(int i=0;i<num_features;i++)
{
```

```
test[i]=Float.parseFloat(characteristics[i]);
}
species=characteristics[num_features].replace("\"", "");
dists.add(String.valueOf(euc_dist(feat,test,num_features))+species);
byteoffset=Long.parseLong(key.toString());
}
@Override
public void cleanup(Context context) throws IOException, InterruptedException{
Collections.sort(dists);
int iter=0;
String[] species=new String[3]; String str="";
for(int i=0;i<3;i++){ str=dists.get(i);
String spec=String.valueOf(str.replaceAll("[\\d.]", """)); species[iter]=spec;
iter++;
} Arrays.sort(species);
for(int i=0;i<species.length-1;i++){
if(species[i].equals(species[i+1])){
context.write(new Text("1"), new Text(species[i])); break;
} }
}
}
```

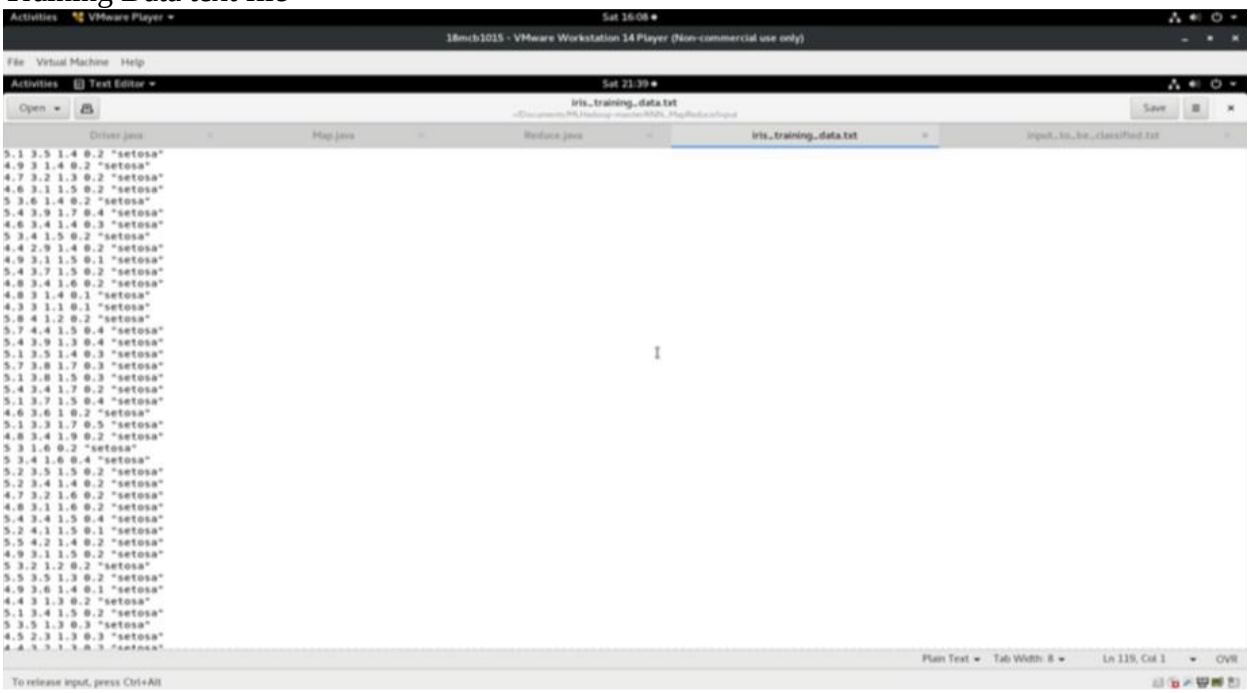
Reducer Class

```
import java.io.IOException;
import java.util.HashMap;
import java.util.Map.Entry;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class Reduce extends Reducer<Text, Text, Text, Text> { String
flower_name=null;
@Override
public void setup(Context context){
flower_name=String.valueOf(context.getConfiguration().get("name")); }
public void reduce(Text key, Iterable<Text> values, Context context) throws
IOException, InterruptedException{
HashMap<String,Integer> map=new HashMap<String,Integer>(); String
maxkey=null;int maxvalue=-1;
for(Text value:values){
if(!map.containsKey(value.toString())){ map.put(value.toString(), 1);
} else{
map.put(value.toString(), map.get(value.toString())+1);
}
} for(Entry<String, Integer> entry: map.entrySet()){
if(entry.getValue()>maxvalue){ maxkey=entry.getKey();
maxvalue=entry.getValue();
}
} context.write(null, new Text(flower_name+" belongs to the species of
```

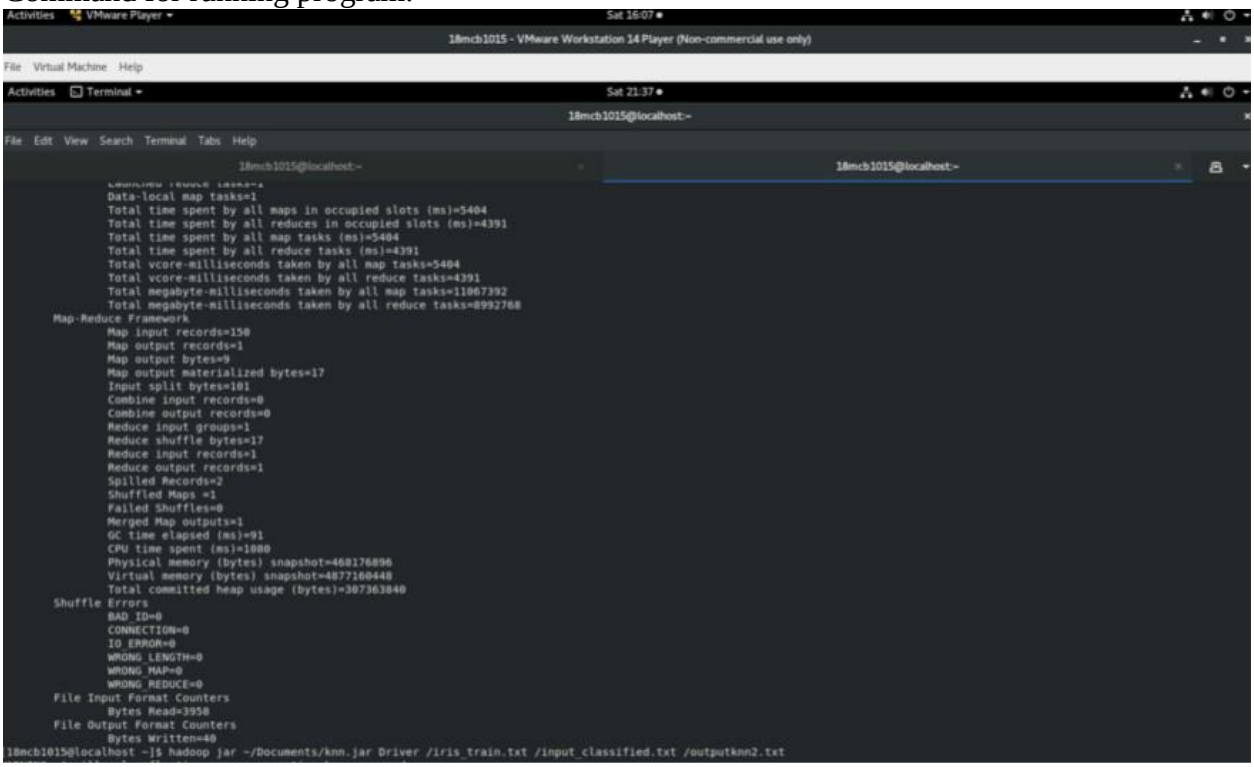
```
"+maxkey));
}
}
```

Output

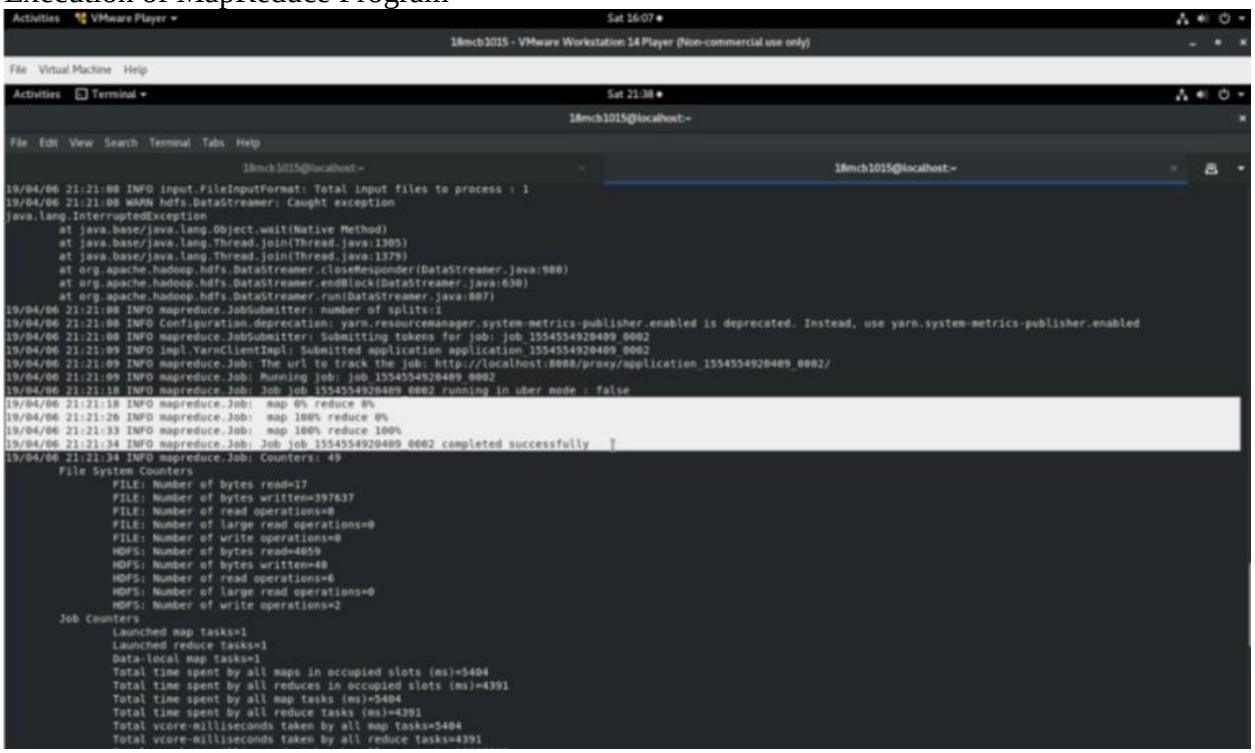
Training Data text file



Command for running program.



Execution of MapReduce Program



Output of KNN classification.