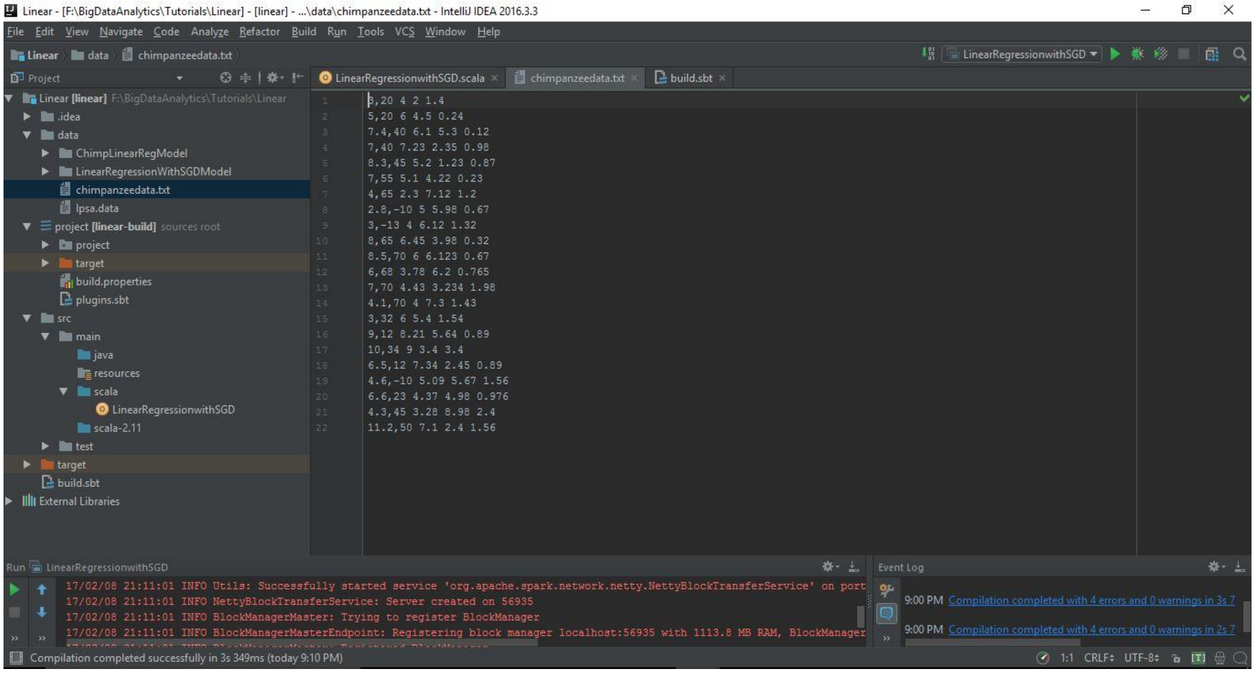
CS5542 Big Data Apps and Analytics

LAB ASSIGNMENT #3

Name: Venkata Raghava Kundavajjala
Class ID: 19

Spark Programming

Part1: Linear Regression Model on the Chimpanzee data set is built. There are several parameters that are used in the data for building a linear model. The data set contains five parameters, the outcome parameter is the amount sleep in hours for a chimpanzee based on the temperature, number of active hours, number of fighting hours, number of pounds of food consumed per day. The data is in a brief form, with minimum number of observations, less than 30. The class is the predefined class **LinerRegressionwithSGD( ),** the text file for the input data is as follows:


Linear - [F:\BigDataAnalytics\Tutorials\Linear] - [linear] - ...\data\chimpanzeedata.txt - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Linear data chimpanzeedata.txt
LinearRegressionwithSGD
Project
Linear [linear] F:\BigDataAnalytics\Tutorials\Linear
.idea
data
ChimpLinearRegModel
LinearRegressionWithSGDModel
chimpanzeedata.txt
lpsa.data
project [linear-build] sources root
project
target
build.properties
plugins.sbt
src
main
java
resources
scala
LinearRegressionwithSGD
scala-2.11
test
target
build.sbt
External Libraries
LinearRegressionwithSGD.scala
chimpanzeedata.txt
build.sbt
Run LinearRegressionwithSGD
Event Log
1:1 CRLF UTF-8

```
8,20 4 2 1.4
5,20 6 4.5 0.24
7.4,40 6.1 5.3 0.12
7,40 7.23 2.35 0.98
8.3,45 5.2 1.23 0.87
7,55 5.1 4.22 0.23
4,65 2.3 7.12 1.2
2.8,-10 5 5.98 0.67
3,-13 4 6.12 1.32
8,65 6.45 3.98 0.32
8.5,70 6 6.123 0.67
6,68 3.78 6.2 0.765
7,70 4.43 3.234 1.98
4.1,70 4 7.3 1.43
3,32 6 5.4 1.54
9,12 8.21 5.64 0.89
10,34 9 3.4 3.4
6.5,12 7.34 2.45 0.89
4.6,-10 5.09 5.67 1.56
6.6,23 4.37 4.98 0.976
4.3,45 3.28 8.98 2.4
11.2,50 7.1 2.4 1.56
```

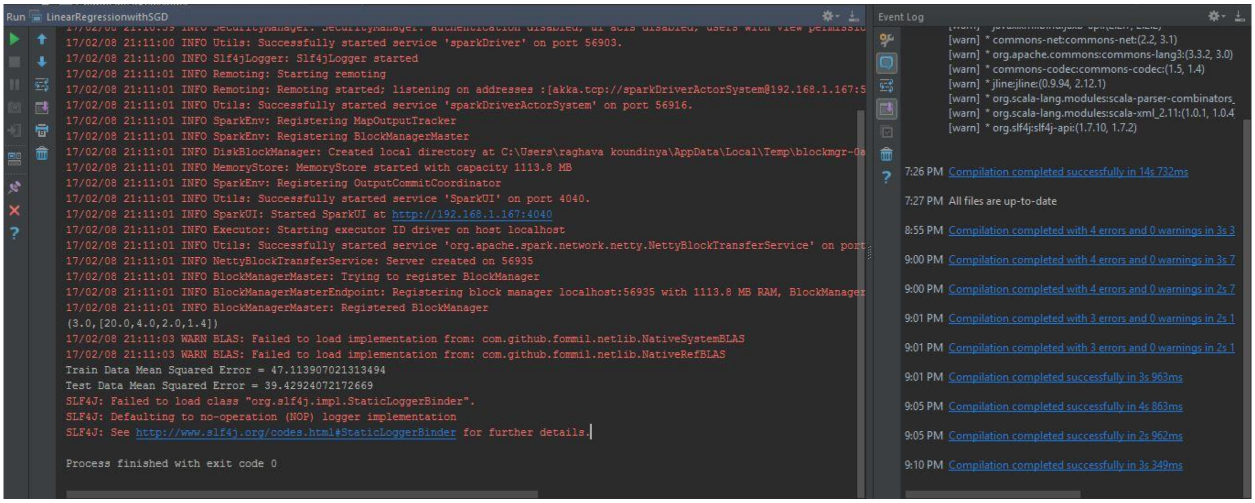
```
17/02/08 21:11:01 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port
17/02/08 21:11:01 INFO NettyBlockTransferService: Server created on 56935
17/02/08 21:11:01 INFO BlockManagerMaster: Trying to register BlockManager
17/02/08 21:11:01 INFO BlockManagerMasterEndpoint: Registering block manager localhost:56935 with 1113.8 MB RAM, BlockManager
```

```
9:00 PM Compilation completed with 4 errors and 0 warnings in 3s 7
9:00 PM Compilation completed with 4 errors and 0 warnings in 2s 7
```

```
Compilation completed successfully in 3s 349ms (today 9:10 PM)
```

The left most column is the prediction output i.e, the number of sleep hours of a chimpanzee. The next are temperature, food intake in pounds, number of active hours, number of hours fighting (from left to right in order). The linear model takes 65% of the data as training data and the remaining 35% of the data as test data. There are 110 iterations and the step size is taken as 0.00000002. The model is in the form

$Y=\beta_0+\beta 1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4$.The text file is **chimpanzeedata.txt,** the input file. The output of the program is as follows:

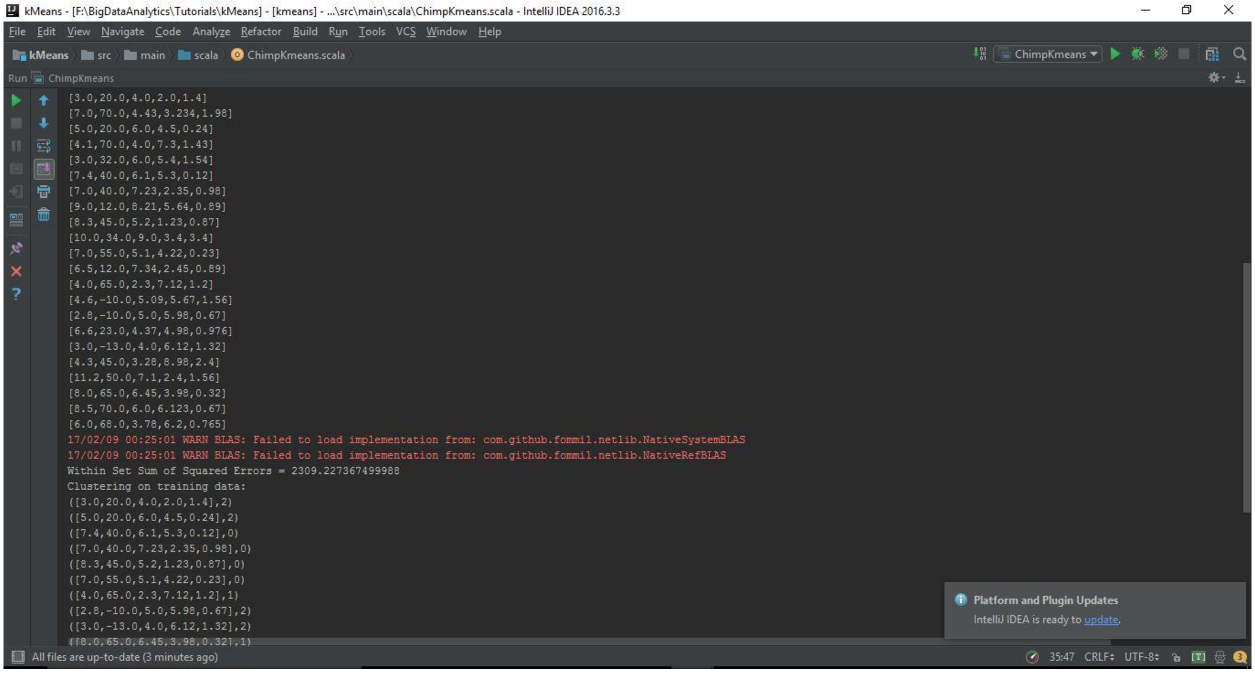


```
17/02/08 21:10:59 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissio
17/02/08 21:11:00 INFO Utils: Successfully started service 'sparkDriver' on port 56903.
17/02/08 21:11:00 INFO Slf4jLogger: Slf4jLogger started
17/02/08 21:11:01 INFO Remoting: Starting remoting
17/02/08 21:11:01 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.1.167:5
17/02/08 21:11:01 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 56916.
17/02/08 21:11:01 INFO SparkEnv: Registering MapOutputTracker
17/02/08 21:11:01 INFO SparkEnv: Registering BlockManagerMaster
17/02/08 21:11:01 INFO DiskBlockManager: Created local directory at C:\Users\raghava koundinya\AppData\Local\Temp\blockmgr-0a
17/02/08 21:11:01 INFO MemoryStore: MemoryStore started with capacity 1113.8 MB
17/02/08 21:11:01 INFO SparkEnv: Registering OutputCommitCoordinator
17/02/08 21:11:01 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/02/08 21:11:01 INFO SparkUI: Started SparkUI at http://192.168.1.167:4040
17/02/08 21:11:01 INFO Executor: Starting executor ID driver on host localhost
17/02/08 21:11:01 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port
17/02/08 21:11:01 INFO NettyBlockTransferService: Server created on 56935
17/02/08 21:11:01 INFO BlockManagerMaster: Trying to register BlockManager
17/02/08 21:11:01 INFO BlockManagerMasterEndpoint: Registering block manager localhost:56935 with 1113.8 MB RAM, BlockManager
17/02/08 21:11:01 INFO BlockManagerMaster: Registered BlockManager
(3.0,[20.0,4.0,2.0,1.4])
17/02/08 21:11:03 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
17/02/08 21:11:03 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Train Data Mean Squared Error = 47.113907021313494
Test Data Mean Squared Error = 39.42924072172669
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.

Process finished with exit code 0
```

```
[warn] * commons-net:commons-net:(2.2, 3.1)
[warn] * org.apache.commons:commons-lang3:(3.3.2, 3.0)
[warn] * commons-codec:commons-codec:(1.5, 1.4)
[warn] * jline:jline:(0.9.94, 2.12.1)
[warn] * org.scala-lang.modules:scala-parser-combinators_
[warn] * org.scala-lang.modules:scala-xml_2.11:(1.0.1, 1.0.4
[warn] * org.slf4j:slf4j-api:(1.7.10, 1.7.2)

7:26 PM Compilation completed successfully in 14s 732ms

7:27 PM All files are up-to-date

8:55 PM Compilation completed with 4 errors and 0 warnings in 3s 3

9:00 PM Compilation completed with 4 errors and 0 warnings in 3s 7

9:00 PM Compilation completed with 4 errors and 0 warnings in 2s 7

9:01 PM Compilation completed with 3 errors and 0 warnings in 2s 1

9:01 PM Compilation completed with 3 errors and 0 warnings in 2s 1

9:01 PM Compilation completed successfully in 3s 963ms

9:05 PM Compilation completed successfully in 4s 863ms

9:05 PM Compilation completed successfully in 2s 962ms

9:10 PM Compilation completed successfully in 3s 349ms
```

With 47.11% train MSE and 39.4% test MSE. The **ChimpLinearRegModel** contains the data and metadata of the output.

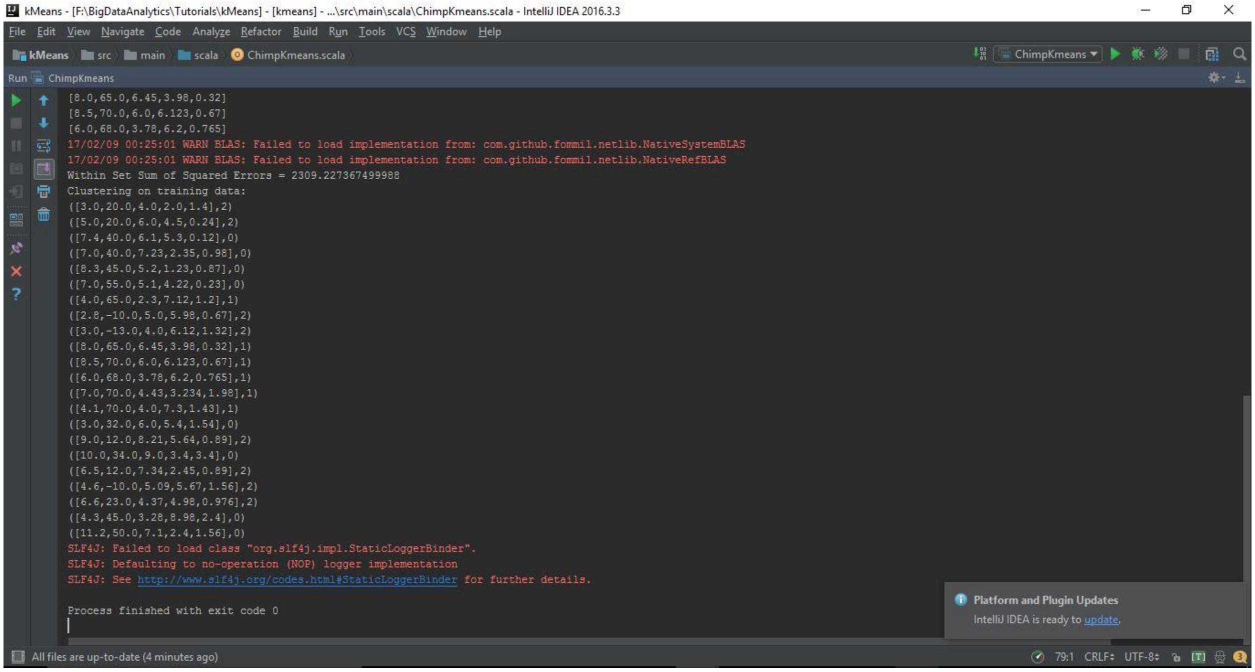
Part2: Running the K-means clustering algorithm on the data set. The k-means clustering is done with number of clusters as 3 and number of iterations as 15. The data is **chimpanzeedata2.txt.** The data is divided into 3 clusters 0,1 and 2, the WSSE is 2389.227. The output of the k-means is as follows:


kMeans - [F:\BigDataAnalytics\Tutorials\kMeans] - [kmeans] - ...\src\main\scala\ChimpKmeans.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
kMeans
src
main
scala
ChimpKmeans.scala
ChimpKmeans
Run
ChimpKmeans
35:47 CRLF UTF-8

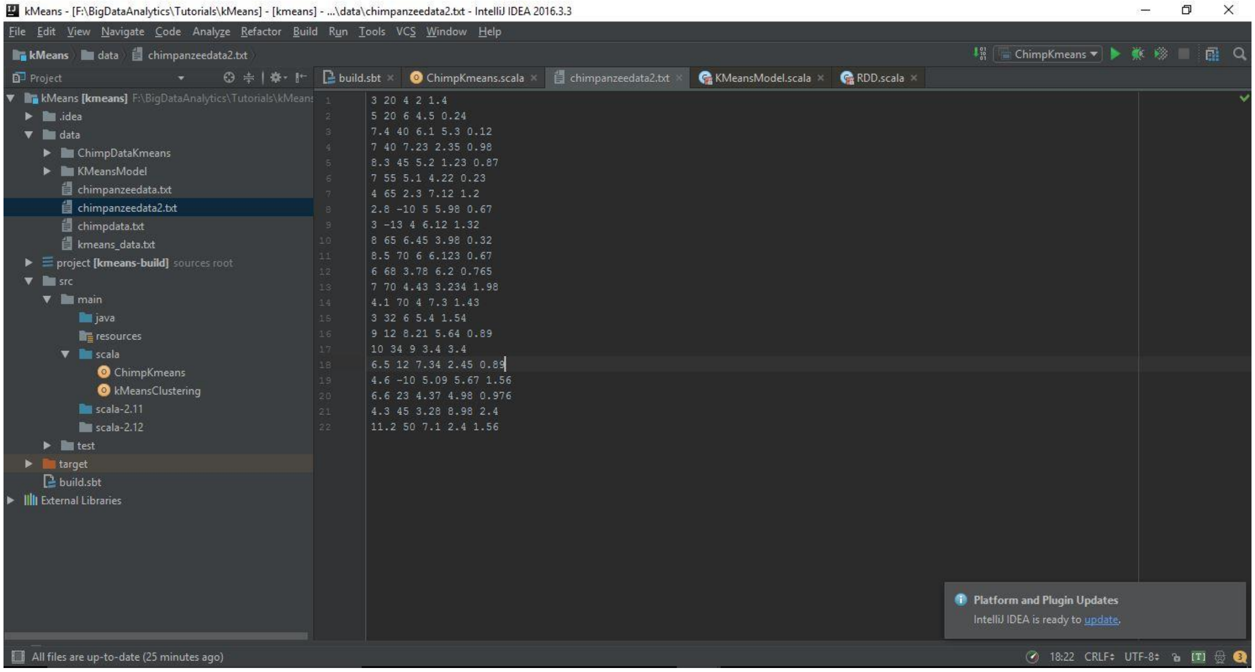
```
[3.0,20.0,4.0,2.0,1.4]
[7.0,70.0,4.43,3.234,1.98]
[5.0,20.0,6.0,4.5,0.24]
[4.1,70.0,4.0,7.3,1.43]
[3.0,32.0,6.0,5.4,1.54]
[7.4,40.0,6.1,5.3,0.12]
[7.0,40.0,7.23,2.35,0.98]
[9.0,12.0,8.21,5.64,0.89]
[8.3,45.0,5.2,1.23,0.87]
[10.0,34.0,9.0,3.4,3.4]
[7.0,55.0,5.1,4.22,0.23]
[6.5,12.0,7.34,2.45,0.89]
[4.0,65.0,2.3,7.12,1.2]
[4.6,-10.0,5.09,5.67,1.56]
[2.8,-10.0,5.0,5.98,0.67]
[6.6,23.0,4.37,4.98,0.976]
[3.0,-13.0,4.0,6.12,1.32]
[4.3,45.0,3.28,8.98,2.4]
[11.2,50.0,7.1,2.4,1.56]
[8.0,65.0,6.45,3.98,0.32]
[8.5,70.0,6.0,6.123,0.67]
[6.0,68.0,3.78,6.2,0.765]
17/02/09 00:25:01 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
17/02/09 00:25:01 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Within Set Sum of Squared Errors = 2309.227367499988
Clustering on training data:
([3.0,20.0,4.0,2.0,1.4],2)
([5.0,20.0,6.0,4.5,0.24],2)
([7.4,40.0,6.1,5.3,0.12],0)
([7.0,40.0,7.23,2.35,0.98],0)
([8.3,45.0,5.2,1.23,0.87],0)
([7.0,55.0,5.1,4.22,0.23],0)
([4.0,65.0,2.3,7.12,1.2],1)
([2.8,-10.0,5.0,5.98,0.67],2)
([3.0,-13.0,4.0,6.12,1.32],2)
([8.0,65.0,6.45,3.98,0.32],1)
```

Platform and Plugin Updates
IntelliJ IDEA is ready to update.

All files are up-to-date (3 minutes ago)


kMeans - [F:\BigDataAnalytics\Tutorials\kMeans] - [kmeans] - ...\src\main\scala\ChimpKmeans.scala - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
kMeans src main scala ChimpKmeans.scala
ChimpKmeans
Run ChimpKmeans
Platform and Plugin Updates
IntelliJ IDEA is ready to update.
All files are up-to-date (4 minutes ago)
79:1 CRLF UTF-8

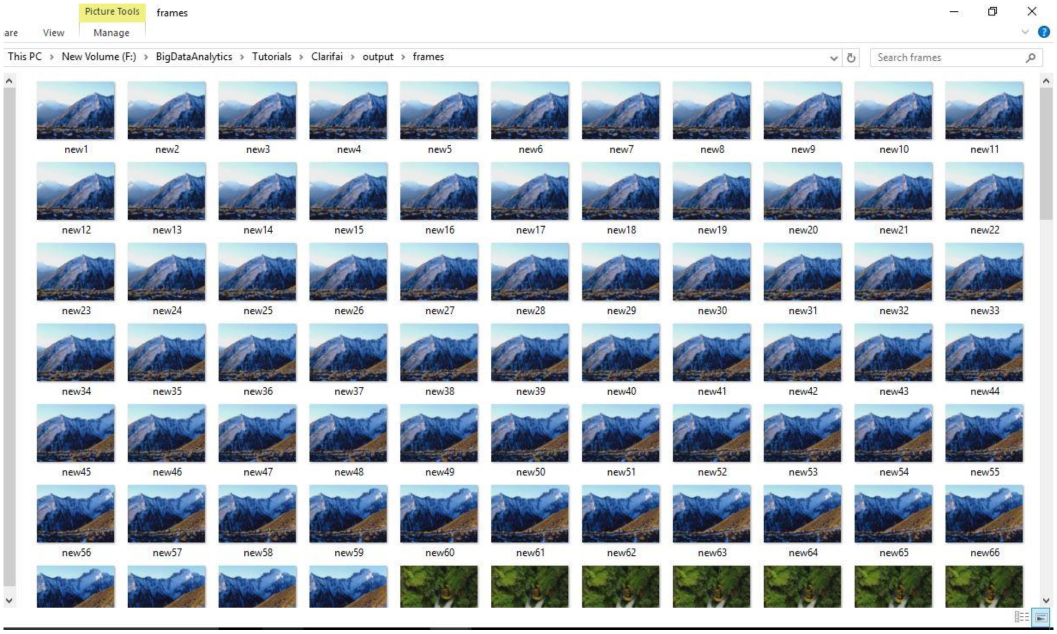
```
[8.0,65.0,6.45,3.98,0.32]
[8.5,70.0,6.0,6.123,0.67]
[6.0,68.0,3.78,6.2,0.765]
17/02/09 00:25:01 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
17/02/09 00:25:01 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Within Set Sum of Squared Errors = 2309.227367499988
Clustering on training data:
([3.0,20.0,4.0,2.0,1.4],2)
([5.0,20.0,6.0,4.5,0.24],2)
([7.4,40.0,6.1,5.3,0.12],0)
([7.0,40.0,7.23,2.35,0.98],0)
([8.3,45.0,5.2,1.23,0.87],0)
([7.0,55.0,5.1,4.22,0.23],0)
([4.0,65.0,2.3,7.12,1.2],1)
([2.8,-10.0,5.0,5.98,0.67],2)
([3.0,-13.0,4.0,6.12,1.32],2)
([8.0,65.0,6.45,3.98,0.32],1)
([8.5,70.0,6.0,6.123,0.67],1)
([6.0,68.0,3.78,6.2,0.765],1)
([7.0,70.0,4.43,3.234,1.98],1)
([4.1,70.0,4.0,7.3,1.43],1)
([3.0,32.0,6.0,5.4,1.54],0)
([9.0,12.0,8.21,5.64,0.89],2)
([10.0,34.0,9.0,3.4,3.4],0)
([6.5,12.0,7.34,2.45,0.89],2)
([4.6,-10.0,5.09,5.67,1.56],2)
([6.6,23.0,4.37,4.98,0.976],2)
([4.3,45.0,3.28,8.98,2.4],0)
([11.2,50.0,7.1,2.4,1.56],0)
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.

Process finished with exit code 0
```

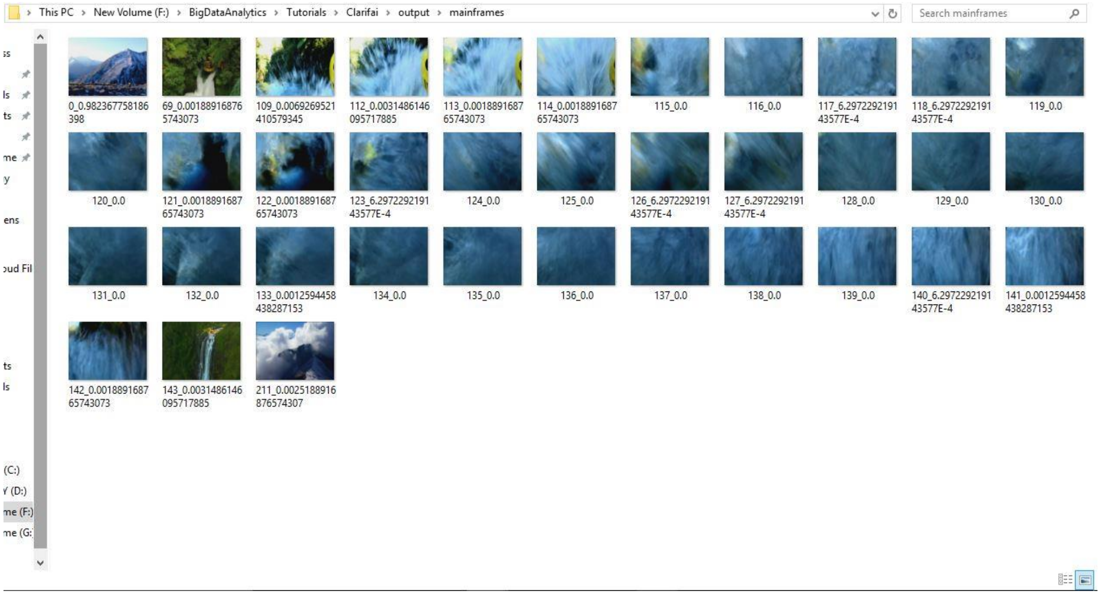

kMeans - [F:\BigDataAnalytics\Tutorials\kMeans] - [kmeans] - ...\data\chimpanzeedata2.txt - IntelliJ IDEA 2016.3.3
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
kMeans data chimpanzeedata2.txt
ChimpKmeans
Project
build.sbt
ChimpKmeans.scala
chimpanzeedata2.txt
KMeansModel.scala
RDD.scala
kMeans [kmeans] F:\BigDataAnalytics\Tutorials\kMeans
.idea
data
ChimpDataKmeans
KMeansModel
chimpanzeedata.txt
chimpanzeedata2.txt
chimpdata.txt
kmeans_data.txt
project [kmeans-build] sources root
src
main
java
resources
scala
ChimpKmeans
kMeansClustering
scala-2.11
scala-2.12
test
target
build.sbt
External Libraries
Platform and Plugin Updates
IntelliJ IDEA is ready to update.
All files are up-to-date (25 minutes ago)
18:22 CRLF UTF-8

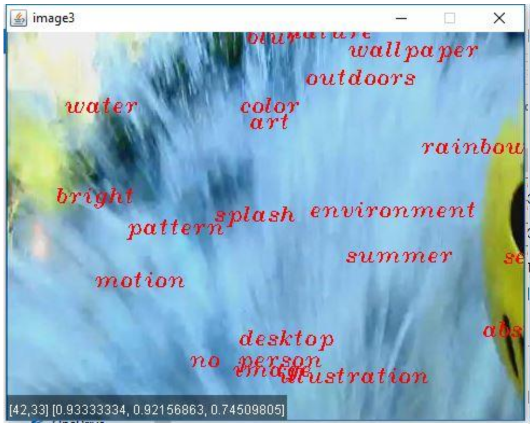
```
3 20 4 2 1.4
5 20 6 4.5 0.24
7.4 40 6.1 5.3 0.12
7 40 7.23 2.35 0.98
8.3 45 5.2 1.23 0.87
7 55 5.1 4.22 0.23
4 65 2.3 7.12 1.2
2.8 -10 5 5.98 0.67
3 -13 4 6.12 1.32
8 65 6.45 3.98 0.32
8.5 70 6 6.123 0.67
6 68 3.78 6.2 0.765
7 70 4.43 3.234 1.98
4.1 70 4 7.3 1.43
3 32 6 5.4 1.54
9 12 8.21 5.64 0.89
10 34 9 3.4 3.4
6.5 12 7.34 2.45 0.89
4.6 -10 5.09 5.67 1.56
6.6 23 4.37 4.98 0.976
4.3 45 3.28 8.98 2.4
11.2 50 7.1 2.4 1.56
```

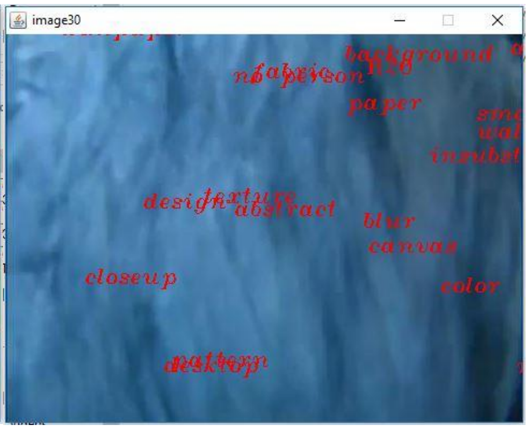
Video Annotation:

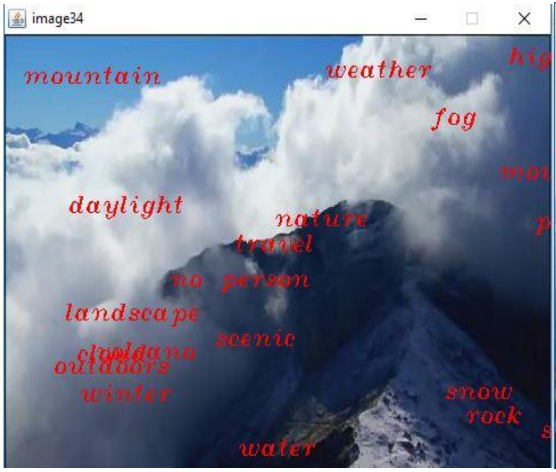
A sample video sample4.mkv is taken as input and key frame detection and video annotation codes are run on the input video. There are 270 frames, 35 main frames generated from a video of 10 seconds length. The video output files are taken into folders **frames** and **main frames**. The below are the outputs generated from the key frame detection and video annotation:

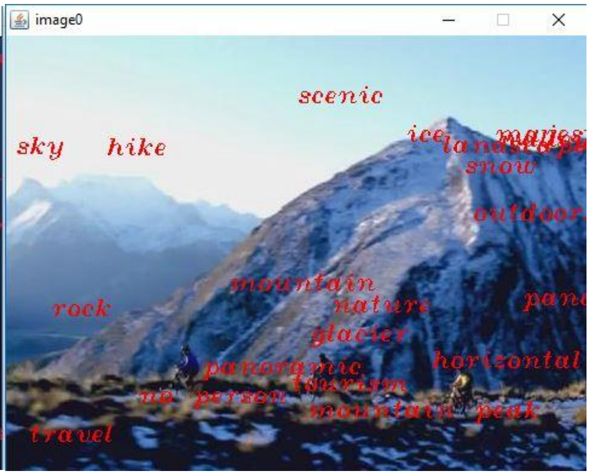












Summary:

The video contains a total number of 270 frames out of which 35 frames are the main/key frames. The video annotation is based on these key/main frames. The video is a mixture of several outdoor sceneries and the total length of the video is 10 seconds. The video starts with main frame **image 0** as seen above contains mountain, sky, nature, hike etc as described above. The video is about exploring the nature. The example summary for **image 0** – The image contains video of hike travelling on mountain top with a mountain on the background and the video is horizontally filmed, it is scenic and made as a travel video with a panoramic view. Similarly, all the key frames can be summarized with the annotations extracted from each frame.