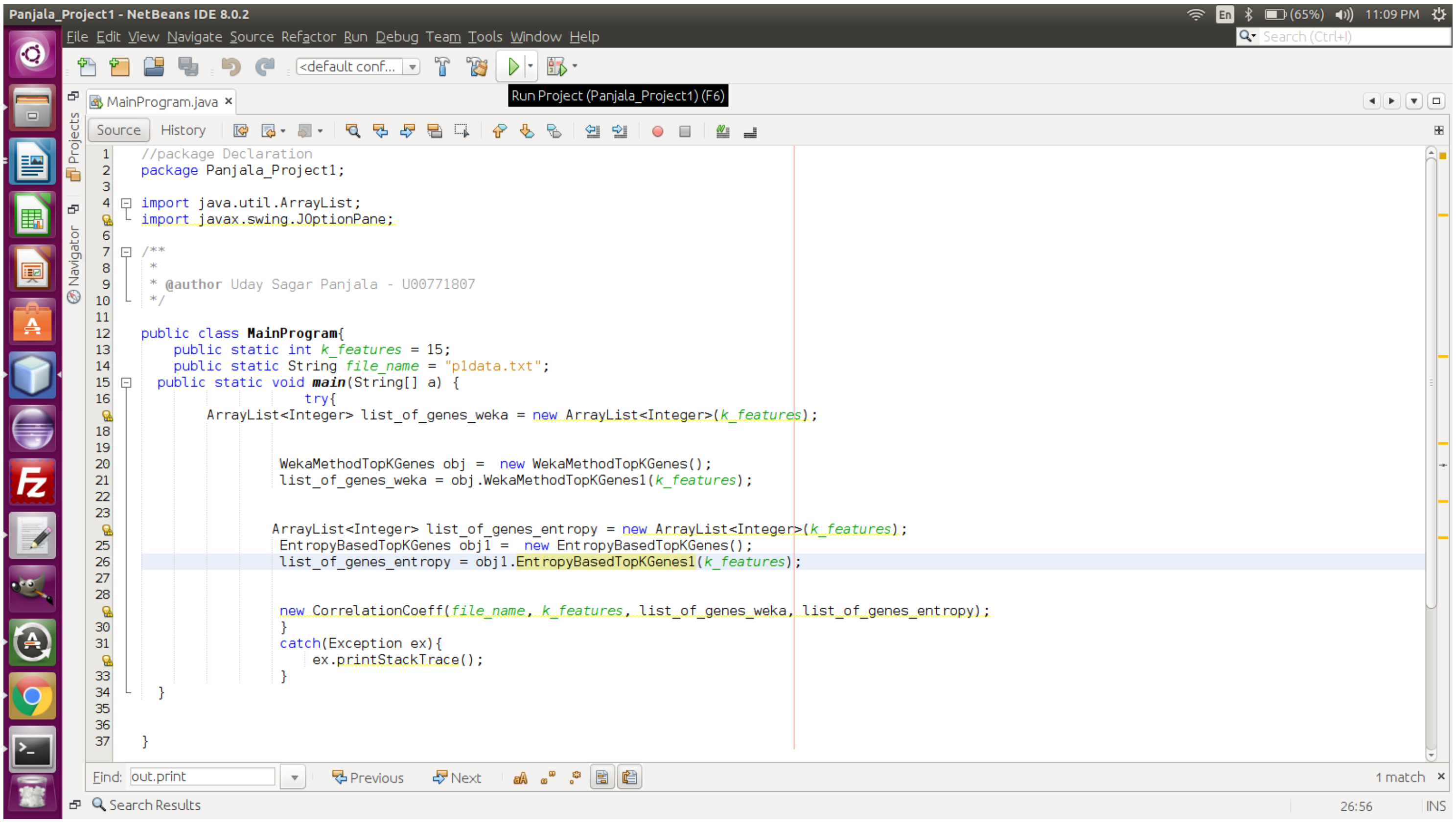
Data Mining – Project 1 – Summer 2016

Name: Uday Sagar Panjala
UID: U00771807
Course No. 7720-01
Project 1

Instructions:

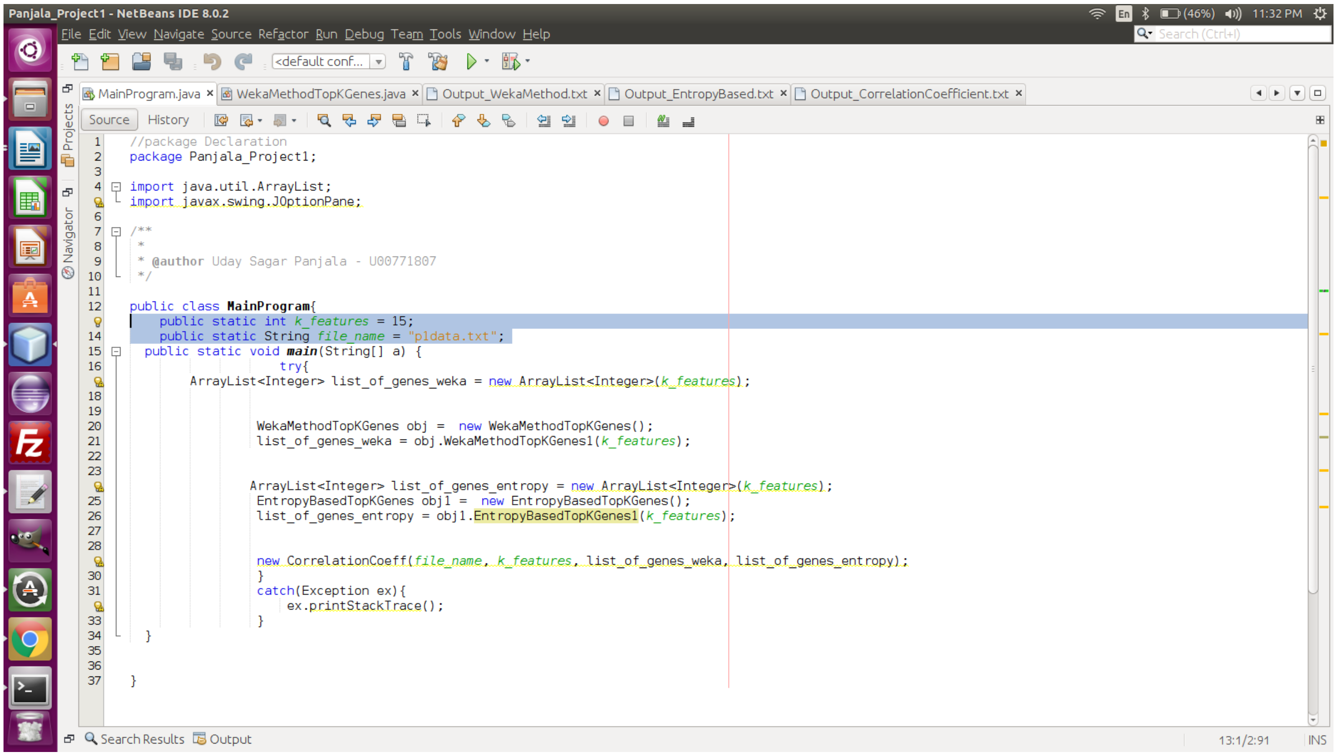
1. Please open the project via **NetBeans.** The project name is Panjala_Project1 (lastName_Project1).

2. The main file is named as MainProgram.java and could be run by clicking on the "run" button as shown in the screenshot below.


Panjala_Project1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Team Tools Window Help
Search (Ctrl+I)
<default conf...
Run Project (Panjala_Project1) (F6)
MainProgram.java
Source
History
Projects
Navigator
Find: out.print
Previous
Next
1 match
Search Results
26:56
INS

```java
//package Declaration
package Panjala_Project1;

import java.util.ArrayList;
import javax.swing.JOptionPane;

/**
 *
 * @author Uday Sagar Panjala - U00771807
 */

public class MainProgram{
    public static int k_features = 15;
    public static String file_name = "p1data.txt";
  public static void main(String[] a) {
                try{
        ArrayList<Integer> list_of_genes_weka = new ArrayList<Integer>(k_features);


            WekaMethodTopKGenes obj =  new WekaMethodTopKGenes();
            list_of_genes_weka = obj.WekaMethodTopKGenes1(k_features);


           ArrayList<Integer> list_of_genes_entropy = new ArrayList<Integer>(k_features);
            EntropyBasedTopKGenes obj1 =  new EntropyBasedTopKGenes();
            list_of_genes_entropy = obj1.EntropyBasedTopKGenes1(k_features);


            new CorrelationCoeff(file_name, k_features, list_of_genes_weka, list_of_genes_entropy);
            }
            catch(Exception ex){
                ex.printStackTrace();
            }
  }


}
```

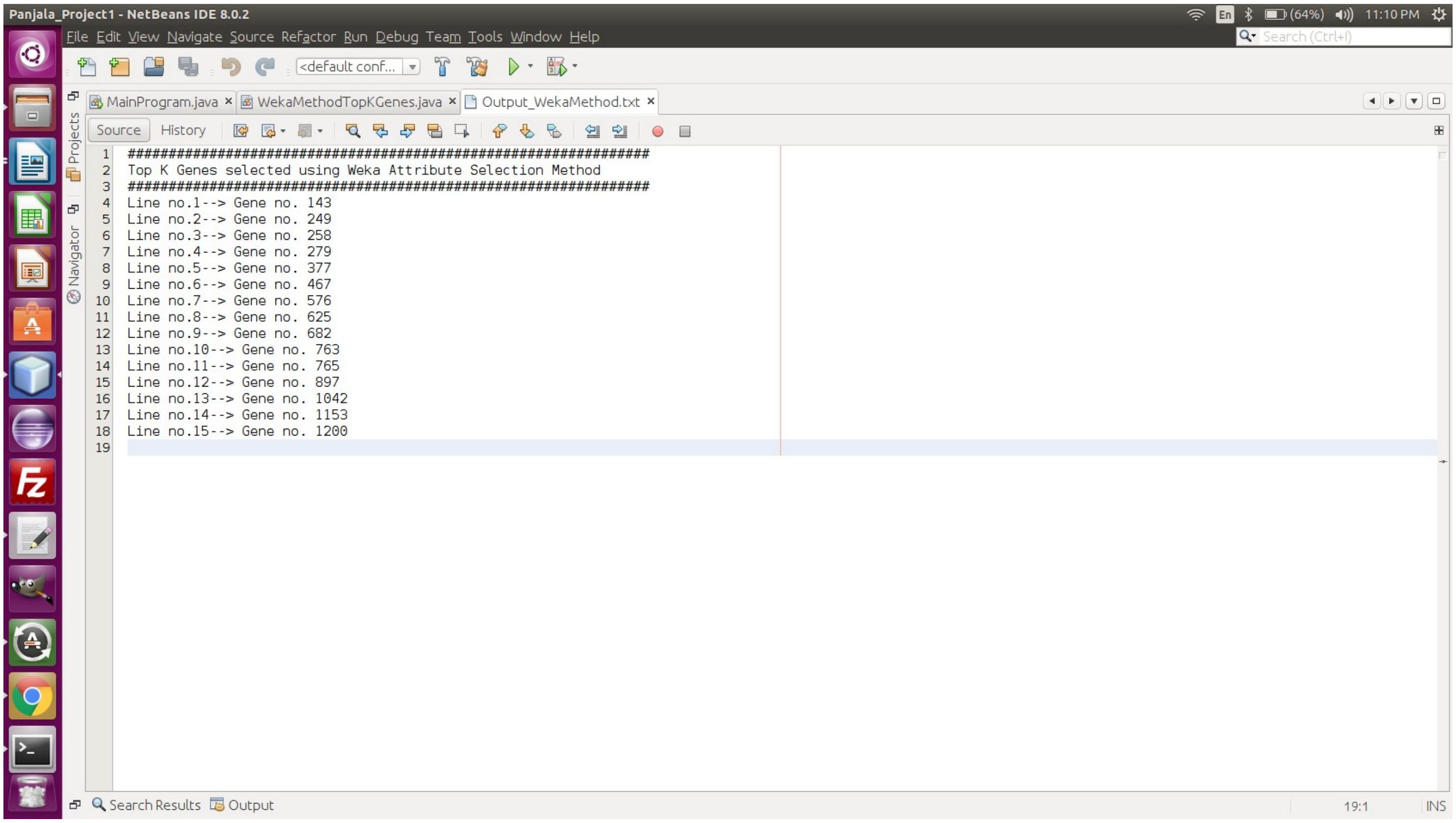
The value of the top K features and the input dataset file can be changed in the above file as highlighted in figure below



```java
//package Declaration
package Panjala_Project1;

import java.util.ArrayList;
import javax.swing.JOptionPane;

/**
 *
 * @author Uday Sagar Panjala - U00771807
 */

public class MainProgram{
    public static int k_features = 15;
    public static String file_name = "pldata.txt";
  public static void main(String[] a) {
                    try{
        ArrayList<Integer> list_of_genes_weka = new ArrayList<Integer>(k_features);


                 WekaMethodTopKGenes obj =  new WekaMethodTopKGenes();
                 list_of_genes_weka = obj.WekaMethodTopKGenes1(k_features);


                ArrayList<Integer> list_of_genes_entropy = new ArrayList<Integer>(k_features);
                 EntropyBasedTopKGenes obj1 =  new EntropyBasedTopKGenes();
                 list_of_genes_entropy = obj1.EntropyBasedTopKGenes1(k_features);


                 new CorrelationCoeff(file_name, k_features, list_of_genes_weka, list_of_genes_entropy);
                 }
                 catch(Exception ex){
                     ex.printStackTrace();
                 }
  }


}
```

3. The main program triggers all the three tasks.

4. The first task is to calculate the top K features based on weka. A call to weka is made internally to the program and can be seen in “**WekaMethodTopKGenes.java**”. The output is saved in the file, “Output_WekaMethod.txt”. It is shown below.



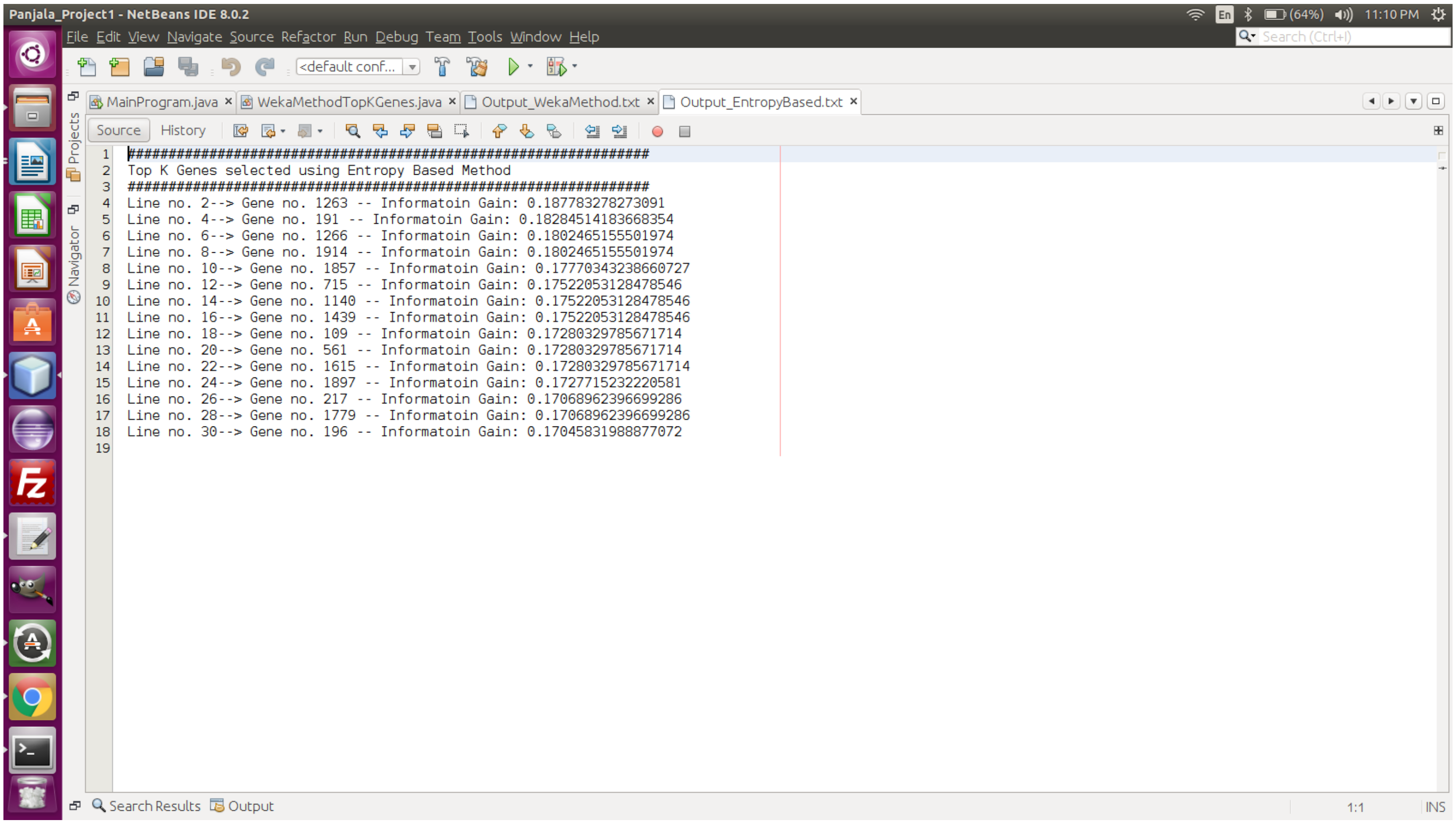
```
################################################################
Top K Genes selected using Weka Attribute Selection Method
################################################################
Line no.1--> Gene no. 143
Line no.2--> Gene no. 249
Line no.3--> Gene no. 258
Line no.4--> Gene no. 279
Line no.5--> Gene no. 377
Line no.6--> Gene no. 467
Line no.7--> Gene no. 576
Line no.8--> Gene no. 625
Line no.9--> Gene no. 682
Line no.10--> Gene no. 763
Line no.11--> Gene no. 765
Line no.12--> Gene no. 897
Line no.13--> Gene no. 1042
Line no.14--> Gene no. 1153
Line no.15--> Gene no. 1200
```

5. The second task is to find the top K genes present in the 62 cases each having around 2000 genes with the class given. The information gain is calculated for each gene based on two split methods. In the first split method, a split is made in the half which prepares two bins. A second split is made in the left bin. For the second split method, the second split is made in right bin.

6. The entropy of all the genes is calculated which is used to get the split information which is inturn used to calculate the information gains of each split using the formula as shown below.

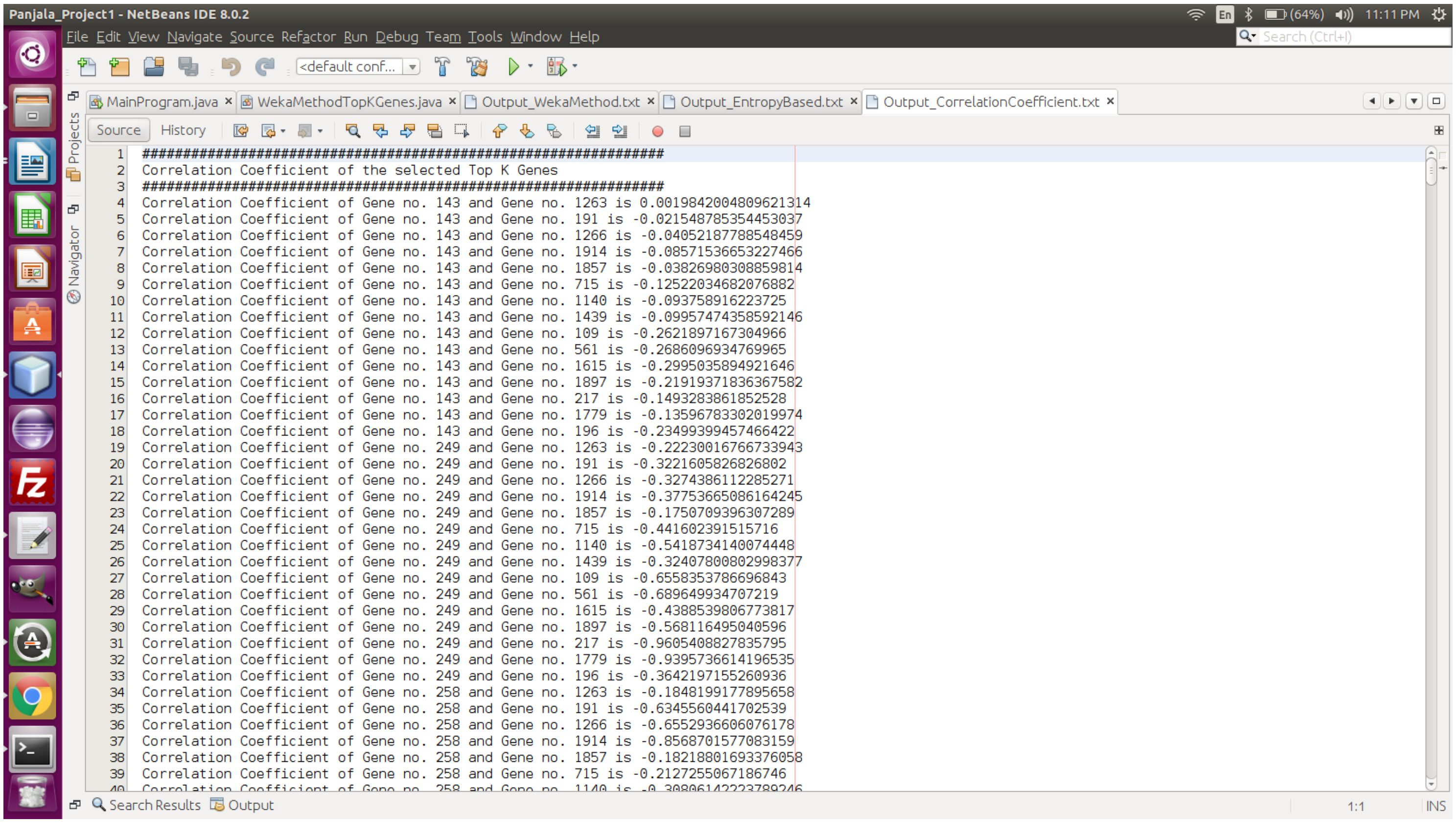
Information Gain = Entropy – Information Split

7. The entropy based top K genes are outputted in “Output_EntropyBased.txt” file. It is shown below.



```
################################################################
Top K Genes selected using Entropy Based Method
################################################################
Line no. 2--> Gene no. 1263 -- Informatoin Gain: 0.187783278273091
Line no. 4--> Gene no. 191 -- Informatoin Gain: 0.18284514183668354
Line no. 6--> Gene no. 1266 -- Informatoin Gain: 0.1802465155501974
Line no. 8--> Gene no. 1914 -- Informatoin Gain: 0.1802465155501974
Line no. 10--> Gene no. 1857 -- Informatoin Gain: 0.17770343238660727
Line no. 12--> Gene no. 715 -- Informatoin Gain: 0.17522053128478546
Line no. 14--> Gene no. 1140 -- Informatoin Gain: 0.17522053128478546
Line no. 16--> Gene no. 1439 -- Informatoin Gain: 0.17522053128478546
Line no. 18--> Gene no. 109 -- Informatoin Gain: 0.17280329785671714
Line no. 20--> Gene no. 561 -- Informatoin Gain: 0.17280329785671714
Line no. 22--> Gene no. 1615 -- Informatoin Gain: 0.17280329785671714
Line no. 24--> Gene no. 1897 -- Informatoin Gain: 0.1727715232220581
Line no. 26--> Gene no. 217 -- Informatoin Gain: 0.17068962396699286
Line no. 28--> Gene no. 1779 -- Informatoin Gain: 0.17068962396699286
Line no. 30--> Gene no. 196 -- Informatoin Gain: 0.17045831988877072
```

8. The correlation coefficients are calculated between the top K Genes selected using weka and top K Genes using entropy based information gain. We get a total of K * K combinations. The output is saved in “Output_CorrelationCoefficient.txt” and is shown below.



```
####################################################################
Correlation Coefficient of the selected Top K Genes
####################################################################
Correlation Coefficient of Gene no. 143 and Gene no. 1263 is 0.0019842004809621314
Correlation Coefficient of Gene no. 143 and Gene no. 191 is -0.021548785354453037
Correlation Coefficient of Gene no. 143 and Gene no. 1266 is -0.04052187788548459
Correlation Coefficient of Gene no. 143 and Gene no. 1914 is -0.08571536653227466
Correlation Coefficient of Gene no. 143 and Gene no. 1857 is -0.03826980308859814
Correlation Coefficient of Gene no. 143 and Gene no. 715 is -0.12522034682076882
Correlation Coefficient of Gene no. 143 and Gene no. 1140 is -0.093758916223725
Correlation Coefficient of Gene no. 143 and Gene no. 1439 is -0.09957474358592146
Correlation Coefficient of Gene no. 143 and Gene no. 109 is -0.2621897167304966
Correlation Coefficient of Gene no. 143 and Gene no. 561 is -0.2686096934769965
Correlation Coefficient of Gene no. 143 and Gene no. 1615 is -0.2995035894921646
Correlation Coefficient of Gene no. 143 and Gene no. 1897 is -0.21919371836367582
Correlation Coefficient of Gene no. 143 and Gene no. 217 is -0.1493283861852528
Correlation Coefficient of Gene no. 143 and Gene no. 1779 is -0.13596783302019974
Correlation Coefficient of Gene no. 143 and Gene no. 196 is -0.23499399457466422
Correlation Coefficient of Gene no. 249 and Gene no. 1263 is -0.22230016766733943
Correlation Coefficient of Gene no. 249 and Gene no. 191 is -0.3221605826826802
Correlation Coefficient of Gene no. 249 and Gene no. 1266 is -0.3274386112285271
Correlation Coefficient of Gene no. 249 and Gene no. 1914 is -0.37753665086164245
Correlation Coefficient of Gene no. 249 and Gene no. 1857 is -0.1750709396307289
Correlation Coefficient of Gene no. 249 and Gene no. 715 is -0.441602391515716
Correlation Coefficient of Gene no. 249 and Gene no. 1140 is -0.5418734140074448
Correlation Coefficient of Gene no. 249 and Gene no. 1439 is -0.32407800802998377
Correlation Coefficient of Gene no. 249 and Gene no. 109 is -0.6558353786696843
Correlation Coefficient of Gene no. 249 and Gene no. 561 is -0.689649934707219
Correlation Coefficient of Gene no. 249 and Gene no. 1615 is -0.4388539806773817
Correlation Coefficient of Gene no. 249 and Gene no. 1897 is -0.568116495040596
Correlation Coefficient of Gene no. 249 and Gene no. 217 is -0.9605408827835795
Correlation Coefficient of Gene no. 249 and Gene no. 1779 is -0.9395736614196535
Correlation Coefficient of Gene no. 249 and Gene no. 196 is -0.3642197155260936
Correlation Coefficient of Gene no. 258 and Gene no. 1263 is -0.1848199177895658
Correlation Coefficient of Gene no. 258 and Gene no. 191 is -0.6345560441702539
Correlation Coefficient of Gene no. 258 and Gene no. 1266 is -0.6552936606076178
Correlation Coefficient of Gene no. 258 and Gene no. 1914 is -0.8568701577083159
Correlation Coefficient of Gene no. 258 and Gene no. 1857 is -0.18218801693376058
Correlation Coefficient of Gene no. 258 and Gene no. 715 is -0.2127255067186746
Correlation Coefficient of Gene no. 258 and Gene no. 1140 is -0.30806142223789246
```

9. The correlation coefficients values here are mostly negative which indicates that both variables move in the opposite direction. If one variable increases, the other variable decreases with the same magnitude, and vice versa.

10. For example, suppose a study is conducted to assess the relationship between outside temperature and heating bills. The study concludes that there is a negative correlation between the prices of heating bills and the outdoor temperature. The correlation coefficient is calculated to be -0.96. This strong negative correlation signifies that as the temperature decreases outside, the prices of heating bills increase and vice versa.

Citations:

http://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp