



4/15/2023

# CS-351-Artificial Intelligence



2020244 - Mohsin Zia  
2020474 - Syed Irtaza Haider

INSTRUCTOR: PROFESSOR ZAHID HALIM

---

## Clustering Techniques!

---

### HOW TO RUN?

1. Open the folder in vsCode.
2. Go to main.py.
3. Click on run python file.

### EXPLANATION:

#### ➤ UTILITY LIBRARIES USED:

```
import numpy as np
import matplotlib.pyplot as plt
from numpy import ndarray
import pandas as pd
import math
import networkx as nx
import os
import shutil
```

#### ➤ TASK 01:

##### ✓ FUNCTIONS FOR READ/WRITE FROM/TO A TEXT FILE:

- Function read\_data reads data from a text file and returns a tuple containing (numberOfRows, numberOfColumns, Data).
- Function write\_data writes data from matrix to a text file including (numberOfRows, numberOfColumns, Data).

```
> def read_data(file_name: str, seperator=None) → tuple: ...
...
> def write_data(file_name: str, arrayType: str, data: list) → ndarray
```

✓ **FUNCTIONS FOR COMPUTING PEARSON'S CORRELATION COEFFICIENT:**

- Function `correlationCoefficient` computes and returns the correlation value between two rows using Pearson's correlation coefficient formula.
- Function `pearsonCorrelationCoefficientMatrix` computes and returns the correlation matrix of NxN dimension.

```
def correlationCoefficient(dataX: list, dataY: list) → float: ...

def pearsonCorrelationCoefficientMatrix(rows: int, data: ndarray) → ndarray: ...
```

✓ **FUNCTION FOR COMPUTING DISCRETIZED MATRIX:**

- Function `discretize` computes and returns the discretize matrix from the correlation matrix data. If the value in the correlation matrix is  $\geq$  mean of it's respective column, then it's replaced with 0 else it is kept 1. Here doing vice versa of what we should've made us closer to sample output.

```
def discretize(totalRows: int, corr_matrix: list) → ndarray: ...
```

✓ **FUNCTION FOR COMPUTING PIXEL MATRIX FOR COLOR CODED IMAGE:**

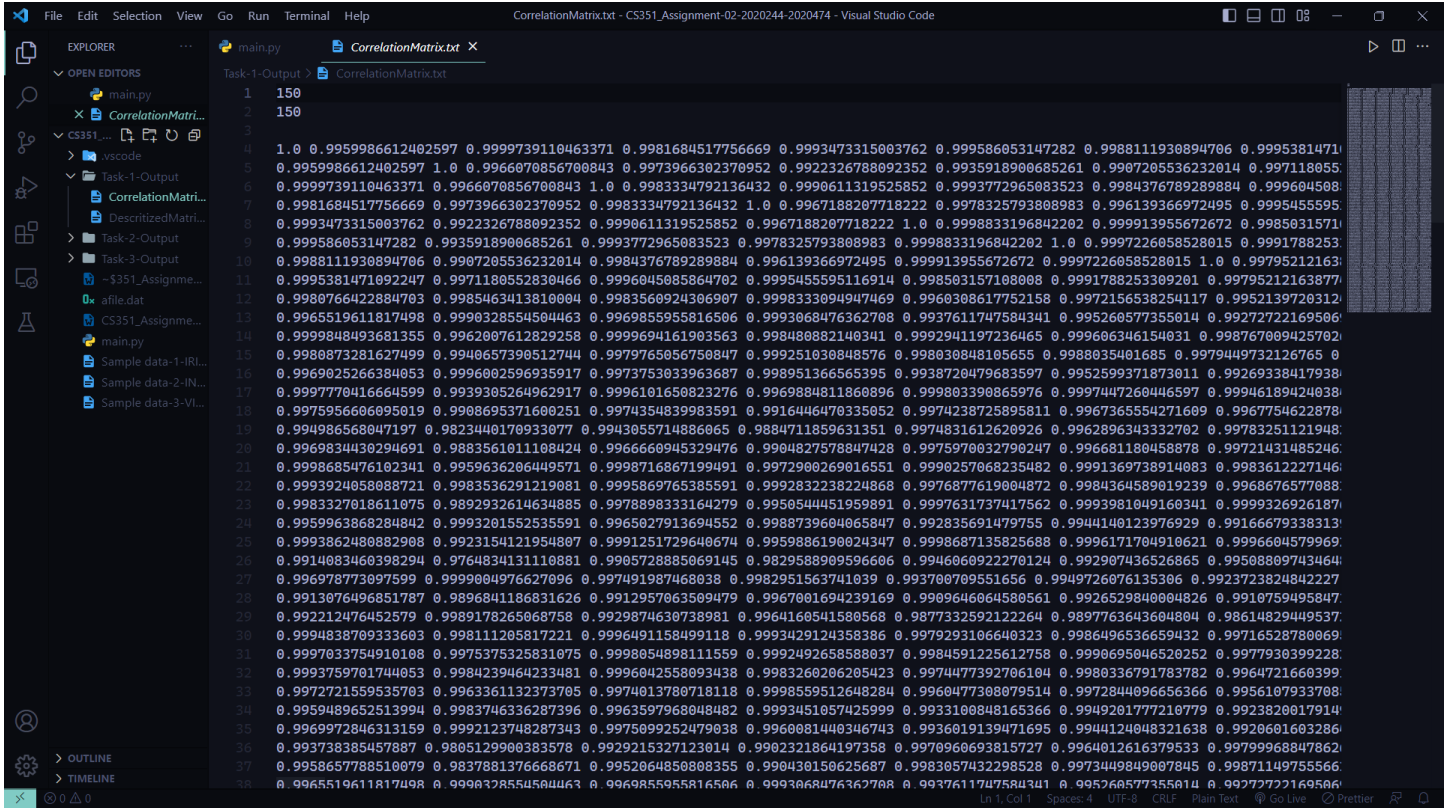
- Function `colourCodedImage` computes and returns the Pixel Matrix of the correlation matrix data. It is calculated by dividing each element by maximum value of their respective column and then multiplying each value with 255. Then it's placed in 3-D Matrix [R, G, B]. We compute the value of G here as we are required to print sample image in green shade and then place it like [0      G      0].

```
def colourCodedImage(totalRows: int, corr_matrix: list) → ndarray: ...
```

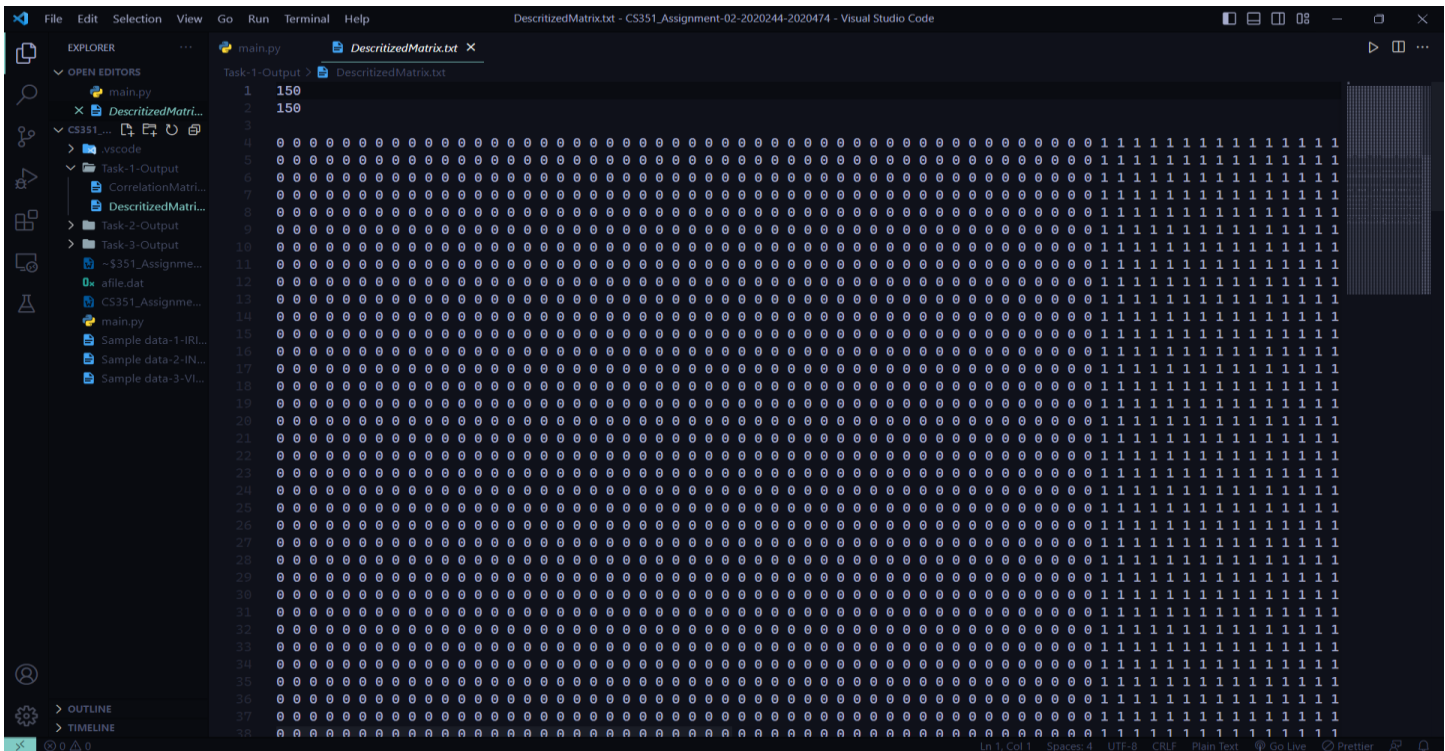
✓ **USING MATPLOTLIB FOR VISUALIZATION:**

```
fig, axes = plt.subplots(nrows=1, ncols=2)
axes[0].set_title('Discretized Matrix')
axes[0].imshow(discretizedMatrix, cmap='gray')
axes[1].set_title('Color Coded Correlation Matrix')
axes[1].imshow(pixelMatrix.astype(int))
plt.show()
```

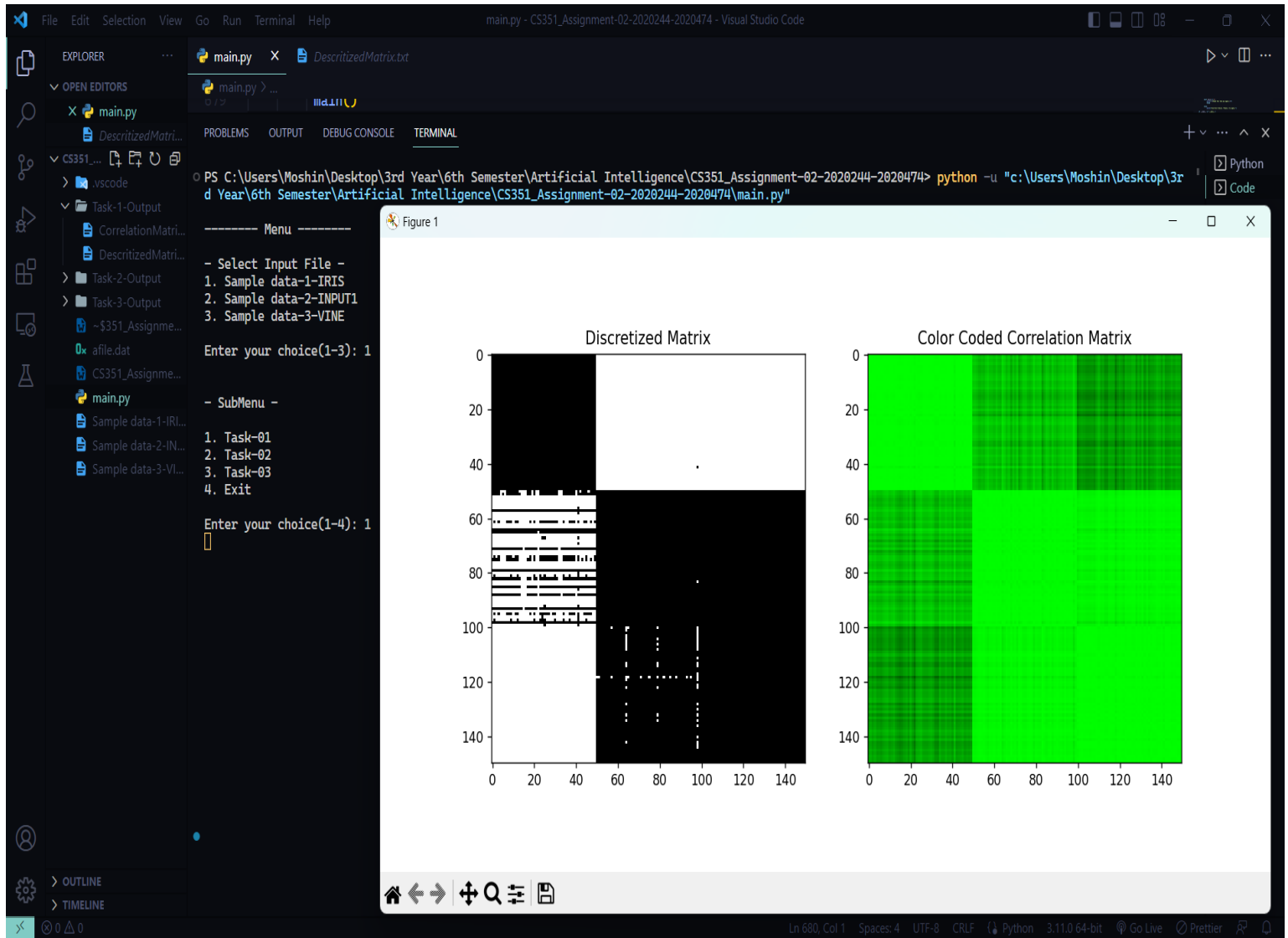
- *Correlation Matrix:*



- *Discretized Matrix:*



- *Bitmap Visualization:*



## ➤ TASK 02:

## ✓ FUNCTION FOR PERMUTATION/SHUFFLING:

- Function `permuteMatrix` provides constant shuffle over a period. It shuffles the rows of `inputData`. A random seed (50) is used to ensure that results are reproducible.

```
> def permuteMatrix(data: list) → list: ...
```

## ✓ FUNCTION FOR CALCULATING SIGNATURE VALUE AND SORTING VIA SIGNATURE:

- Function `signatureCalculation` computes and returns sorted input data via their signature values. We passed the permuted data as input to this function, but I guess it really doesn't matter here if we pass the non-permuted data, in the end it's getting sorted via its signature values.

```
> def signatureCalculation(number_of_rows: int, number_of_cols: int, shuffledData: list) → list: ...
```

## ✓ USING MATPLOTLIB FOR VISUALIZATION:

```
# Plotting Discretized Matrix before and after permutation.
fig, axes = plt.subplots(nrows=1, ncols=2)
axes[0].set_title('Before Permutation')
axes[0].imshow(discretizedMatrixBeforePermutation, cmap='gray')
axes[1].set_title('After Permutation')
axes[1].imshow(discretizedMatrixAfterPermutation, cmap='gray')
plt.show()
```

```
# Plotting Pixel Matrix after SignatureWiseSorting And Rearrangement.
plt.suptitle('After Signature-Wise-Sorting')
plt.imshow(pixelMatrix.astype(int))
plt.show()
```

✓ **OUTPUT SCREENSHOTS:**

- *Input after permutation:*

```

File Edit Selection View Go Run Terminal Help
EXPLORER
  OPEN EDITORS
    main.py
    InputAfterShuffli...
  CS351_...
    .vscode
    Task-1-Output
    Task-2-Output
      CorrelationMatri...
      DescritizedMatri...
      InputAfterShuffli...
      InputSortedViaS...
      SignatureCorrel...
    Task-3-Output
    ~$351_Assignme...
    ~WRL0003.tmp
    afile.dat
    Assignment-2.pdf
    CS351_Assignme...
    main.py
    Sample data-1-IRI...
    Sample data-2-IN...
    Sample data-3-VI...
  Task-2-Output > InputAfterShuffling.txt
    1 150
    2 4
    3
    4 5.6 3.0 4.1 1.3
    5 6.3 2.5 4.9 1.5
    6 5.4 3.4 1.7 0.2
    7 5.4 3.9 1.3 0.4
    8 6.5 3.0 5.2 2.0
    9 6.7 3.1 5.6 2.4
    10 5.7 2.5 5.0 2.0
    11 5.1 3.3 1.7 0.5
    12 4.8 3.0 1.4 0.1
    13 6.2 2.2 4.5 1.5
    14 5.1 3.4 1.5 0.2
    15 7.4 2.8 6.1 1.9
    16 4.9 3.1 1.5 0.1
    17 6.8 3.0 5.5 2.1
    18 5.7 2.8 4.5 1.3
    19 5.0 3.0 1.6 0.2
    20 5.8 2.7 3.9 1.2
    21 5.3 3.7 1.5 0.2
    22 5.5 2.4 3.7 1.0
  
```

- *Input after sorting via signature:*

```

File Edit Selection View Go Run Terminal Help
EXPLORER
  OPEN EDITO...
    main.py
    InputSortedViaSi...
  CS351_...
    .vscode
    Task-1-Output
    Task-2-Output
      CorrelationMatri...
      DescritizedMatri...
      InputAfterShuffli...
      InputSortedViaS...
      SignatureCorrel...
    Task-3-Output
    ~$351_Assignme...
    ~WRL0003.tmp
    afile.dat
    Assignment-2.pdf
    CS351_Assignme...
    main.py
    Sample data-1-IRI...
    Sample data-2-IN...
  Task-2-Output > InputSortedViaSignature.txt
    1 7.7 3.8 6.7 2.2
    2 7.9 3.8 6.4 2.0
    3 7.7 2.6 6.9 2.3
    4 7.2 3.6 6.1 2.5
    5 7.6 3.0 6.6 2.1
    6 7.7 2.8 6.7 2.0
    7 7.7 3.0 6.1 2.3
    8 7.3 2.9 6.3 1.8
    9 6.8 3.2 5.9 2.3
    10 7.2 3.2 6.0 1.8
    11 6.7 3.3 5.7 2.5
    12 7.4 2.8 6.1 1.9
    13 6.9 3.2 5.7 2.3
    14 6.3 3.3 6.0 2.5
    15 7.1 3.0 5.9 2.1
    16 6.7 3.1 5.6 2.4
    17 6.7 3.3 5.7 2.1
    18 6.3 3.4 5.6 2.4
    19 7.2 3.0 5.8 1.6
    20 6.5 3.0 5.8 2.2
  
```

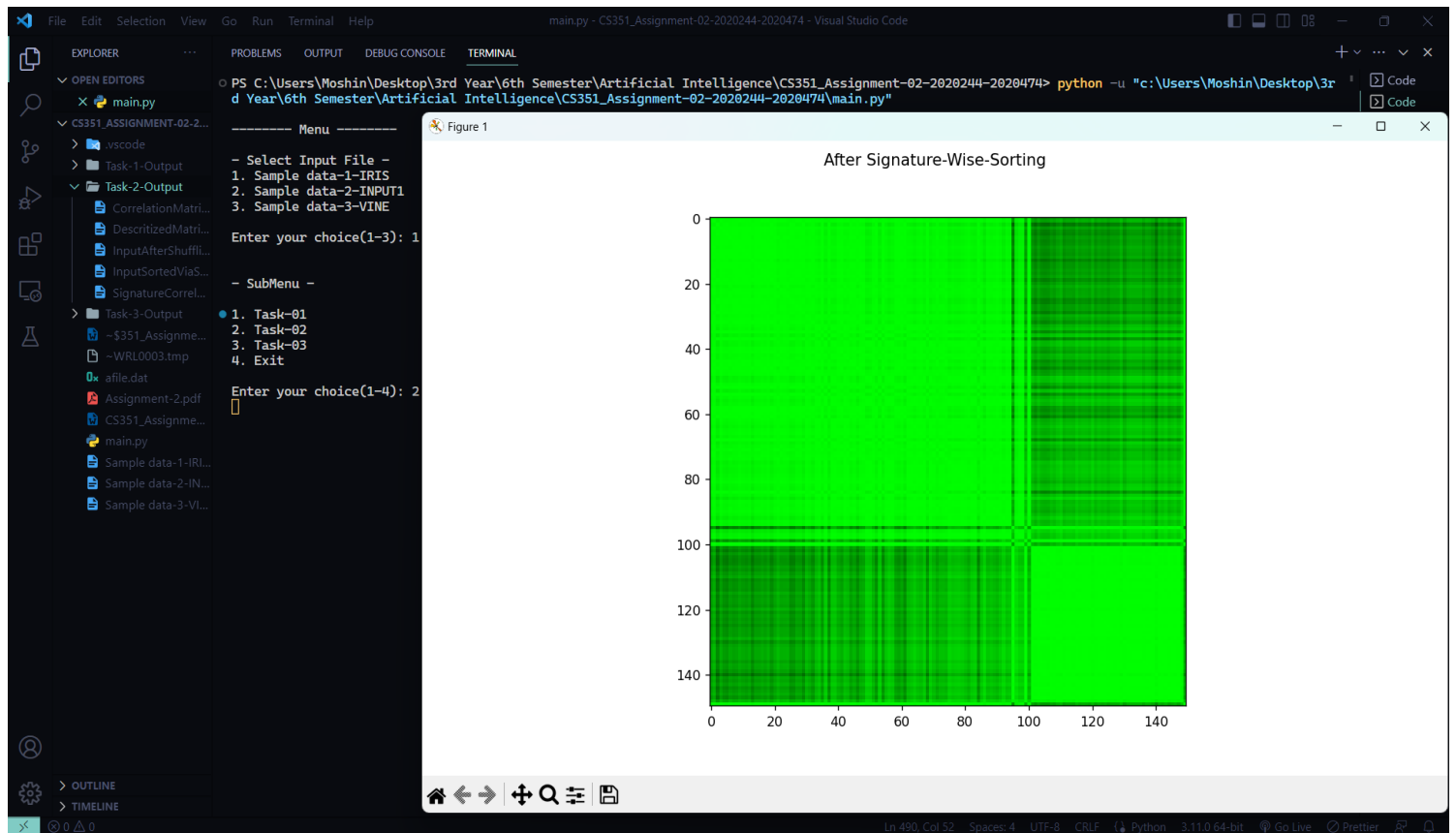
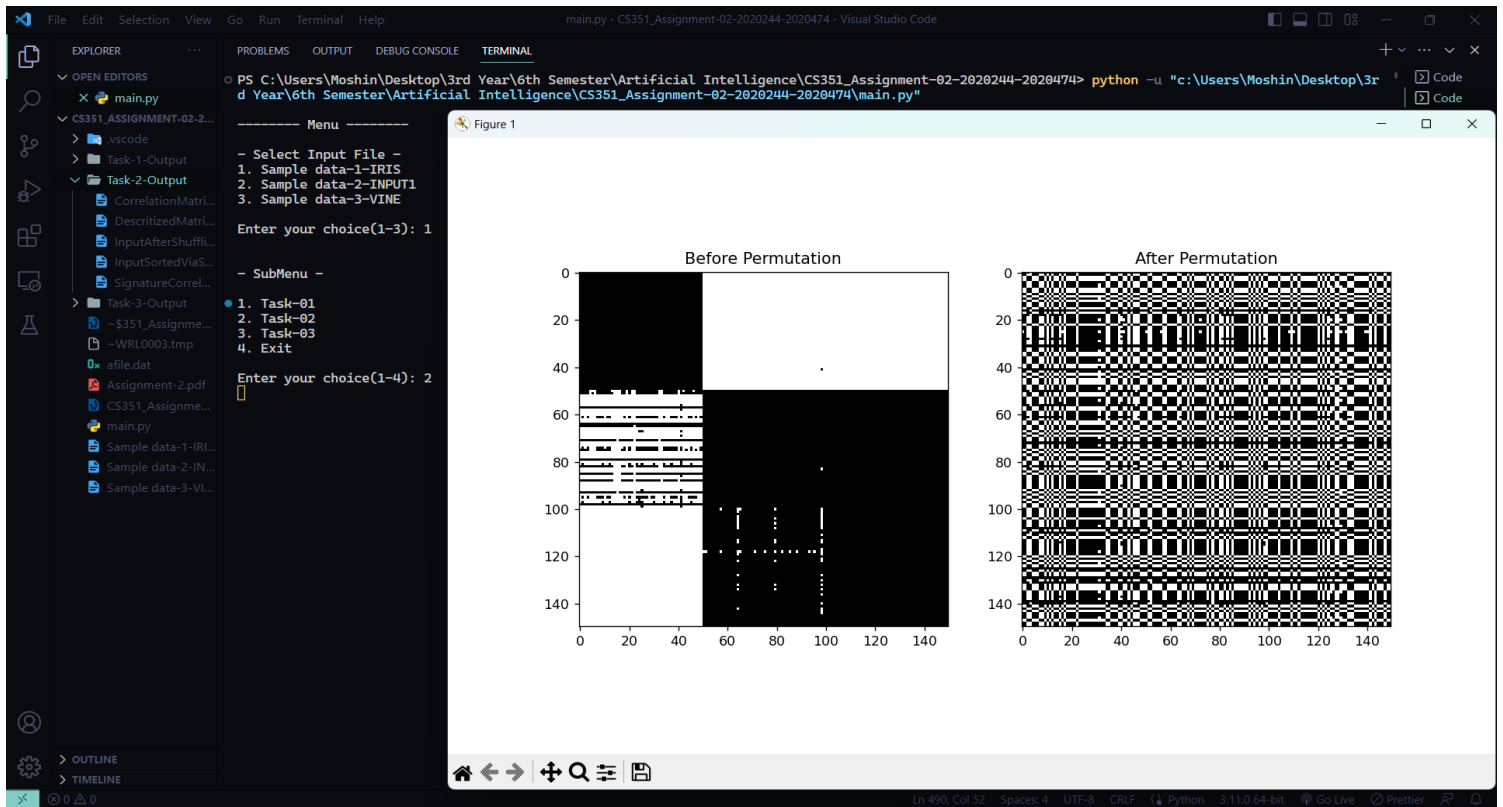


- Correlation Matrix after sorting via signature.

```
File Edit Selection View Go Run Terminal Help SignatureCorrelationMatrix.txt - CS351_Assignment-02-202044-2020474 - Visual Studio Code
EXPLORER
main.py SignatureCorrelationMatrix.txt
OPEN EDITORS
main.py
SignatureCorrel...
CS351_...
vscode
Task-1-Output
Task-2-Output
CorrelationMatri...
DescritizedMatri...
InputAfterShuffl...
InputSortedViaS...
SignatureCorrel...
Task-3-Output
-$351_Assignme...
-WRL0003.tmp
afile.dat
Assignment-2.pdf
CS351_Assignme...
main.py
Sample data-1-IR...
Sample data-2-IN...
Sample data-3-VI...
OUTLINE
TIMELINE
1 150
2 150
3
4 1.0 0.9969265282801588 0.9754748788193148 0.9976735539203875 0.992454270928333 0.9895090508874873 0.9849478745073577
5 0.9969265282801588 1.0 0.9635838640213761 0.9967403201417416 0.9855578498965926 0.9815300758785935 0.9853486146826014
6 0.9754748788193148 0.9635838640213761 1.0 0.981900973821415 0.9948758002578576 0.9969109983727068 0.987360079208383
7 0.9976735539203875 0.9967403201417416 0.981900973821415 1.0 0.995830906374389 0.9936156026299201 0.9942990176873734
8 0.992454270928333 0.9855578498965926 0.9948758002578576 0.995830906374389 1.0 0.9997430108575116 0.9941562239158039
9 0.9895090508874873 0.9815300758785935 0.9969109983727068 0.9936156026299201 0.9997430108575116 1.0 0.9935953551967754
10 0.9849478745073577 0.9853486146826014 0.987360079208383 0.9942990176873734 0.9941562239158039 0.9935953551967754
11 0.9969265282801588 0.990174246276099 0.991033436939984 0.9978396749886277 0.9994298438377482 0.9984322272361296 0.993
12 0.99616092228038466 0.9916180508181198 0.98995200865922 0.9986707093363719 0.9991761448313977 0.9980134105880651 0.99
13 0.998969324544828 0.9979537569340188 0.9784784934259774 0.9996747304571576 0.9943061488107112 0.9916519858790411 0.99
14 0.9946910235589236 0.9926660785338124 0.9880471518285147 0.9991142984173286 0.9980057088970975 0.9966928204047779 0.99
15 0.9921946990523245 0.98928663514408374 0.9912874784826377 0.9977011258948318 0.9986799401680176 0.9979463598872328 0.99
16 0.9940269997277056 0.9938914896840415 0.9847617160663854 0.9991399440574812 0.9962217529975328 0.9945842080429466 0.99
17 0.9912476548921286 0.9786838745689347 0.9904717159004716 0.9886335977782496 0.9955378890851222 0.9952264556997722 0.99
18 0.99370037516999487 0.9914882346168784 0.989146719290506 0.9986531588096706 0.9982033995845564 0.9970972719402965 0.99
19 0.9913525231956634 0.9897780788829625 0.9893977015370282 0.9976748590050771 0.9975115341118936 0.996507868377131 0.99
20 0.999155830887885 0.9978212082984812 0.9978596570327789 0.9995970265743797 0.9944153173467996 0.9917740463883987 0.99
21 0.9993523628817766 0.9950533760990793 0.982588380781241 0.9985736781940172 0.9961410126085493 0.99397255420746 0.98
22 0.9974747824559683 0.9982803909906299 0.9769932366047659 0.9996823772219638 0.9932317523597589 0.9905839485469884 0.99
23 0.9943078038965893 0.9867134580335148 0.9933613689719064 0.996006006489486 0.9996661931824827 0.9991522668011459 0.99
24 0.9926565308945218 0.9935917859393915 0.983231875131057 0.9985915485967843 0.9949999175119389 0.9933023113741177 0.99
25 0.9956246842747881 0.9850713795552598 0.9622154746850845 0.9871423907354275 0.97747443634408485 0.9750580653181324 0.99
26 0.997401597511447 0.99815641135604633 0.975437537175306 0.998718407211012 0.99267747489646293 0.9896902133933598 0.987
27 0.9950510392637024 0.9962832951630802 0.9800659357896997 0.9994090704175747 0.994169539634228 0.9919382833731574 0.99
28 0.9831408302532382 0.9879883239629544 0.976440680632507 0.9931377413865061 0.9879581505981719 0.9863581283083201 0.99
29 0.9901198062203477 0.9832453449221153 0.9961718211056861 0.9946983321721008 0.998155880533661 0.9989990794365533 0.99
30 0.9924317514364079 0.9858072463775824 0.9947709920299058 0.996006006489486 0.9999926190366516 0.9997164408648389 0.99
31 0.9998971988201654 0.9976068836948101 0.9761794970445951 0.9984857454897693 0.9930170442526836 0.9901142792868682 0.99
32 0.9991295367966591 0.997620574851733 0.98019632014549302 0.9996473285052515 0.9951940582231704 0.9927125326235762 0.99
33 0.9900180887866926 0.9826654936308796 0.9964532863996134 0.9943280794844118 0.998269643075252 0.999696938258557 0.99
34 0.9916317745454512 0.9979679811203384 0.9627932300257599 0.9955919381840696 0.98367730939096877 0.97993692393040125 0.99
```



- *Bitmap Visualization:*



➤ **TASK 03:**✓ **FUNCTION FOR SETUP OF PERMUTATED CORRELATION MATRIX:**

- Function setPermutatedMatrix takes permutated data set's correlation matrix and based on a threshold, it updates that matrix. It iterates through each value in the matrix and if the value is less than the threshold it makes it 0, else it is **not** updated.

```
> def setPermutatedMatrix(number_of_rows: int, number_of_cols: int, shuffledData: list) → list: ...
```

✓ **FUNCTIONS FOR COMPUTING NODE WEIGHTS OF THE UPDATED MATRIX WITH RESPECT TO THRESHOLD:**

- Function getNodeWeights computes and returns the nodeWeights matrix by calculating node weights of each row. It is calculated as the sum of all columns in a respective row.

```
> def getNodeWeights(number_of_rows: int, number_of_cols: int, weightedMatrix: list) → list: ...
```

✓ **FUNCTION FOR MAKING CLUSTER OF THE ROW IN CORRELATION MATRIX WITH MAXIMUM NODE WEIGHTS:**

- Function getIndexOfMaxWeight computes and returns the index of the node with maximum weightage. Once it is visited, matrix of node weights at that particular index is placed with -1 so that that node is excluded from our to-cluster list.
- Function makeCluster makes the graph of element with maximum node weight, it connects the edges to the elements in graphList that have survived the threshold test and then it connects to them and forms a cluster. Adding nodes and edges in between is done in this function.

```
> def getIndexOfMaxWeight(number_of_rows: int, nodeWeights: list) → int: ...
```

```
> def makeClusters(index: int, graphList: list): ...
```

✓ **FUNCTION FOR VISUALIZATION OF WEIGHTED GRAPH IN FORM OF CLUSTERS:**

- Function visualizeWeightedGraph takes input a weighted graph and index with maximum weight, it then used matplotlib to create a figure. Further it uses networkx library to draw the nodes and edges between source and destination. It can also add weights in between edges but it doesn't look good when we are visualizing it. It saves the figure in png format.

```
> def visualizeWeightedGraph(G, index: int): ...
```

✓ **CREATING A DIRECTORY FOR OUR CLUSTER IMAGES:**

```

605     # removing if the directory is already present
606     shutil.rmtree('./Task-3-Output/GraphClusterVisualization',
607                  ignore_errors=True)
608
609     # creating a directory for saving png file of clusters
610     os.mkdir('./Task-3-Output/GraphClusterVisualization')

```

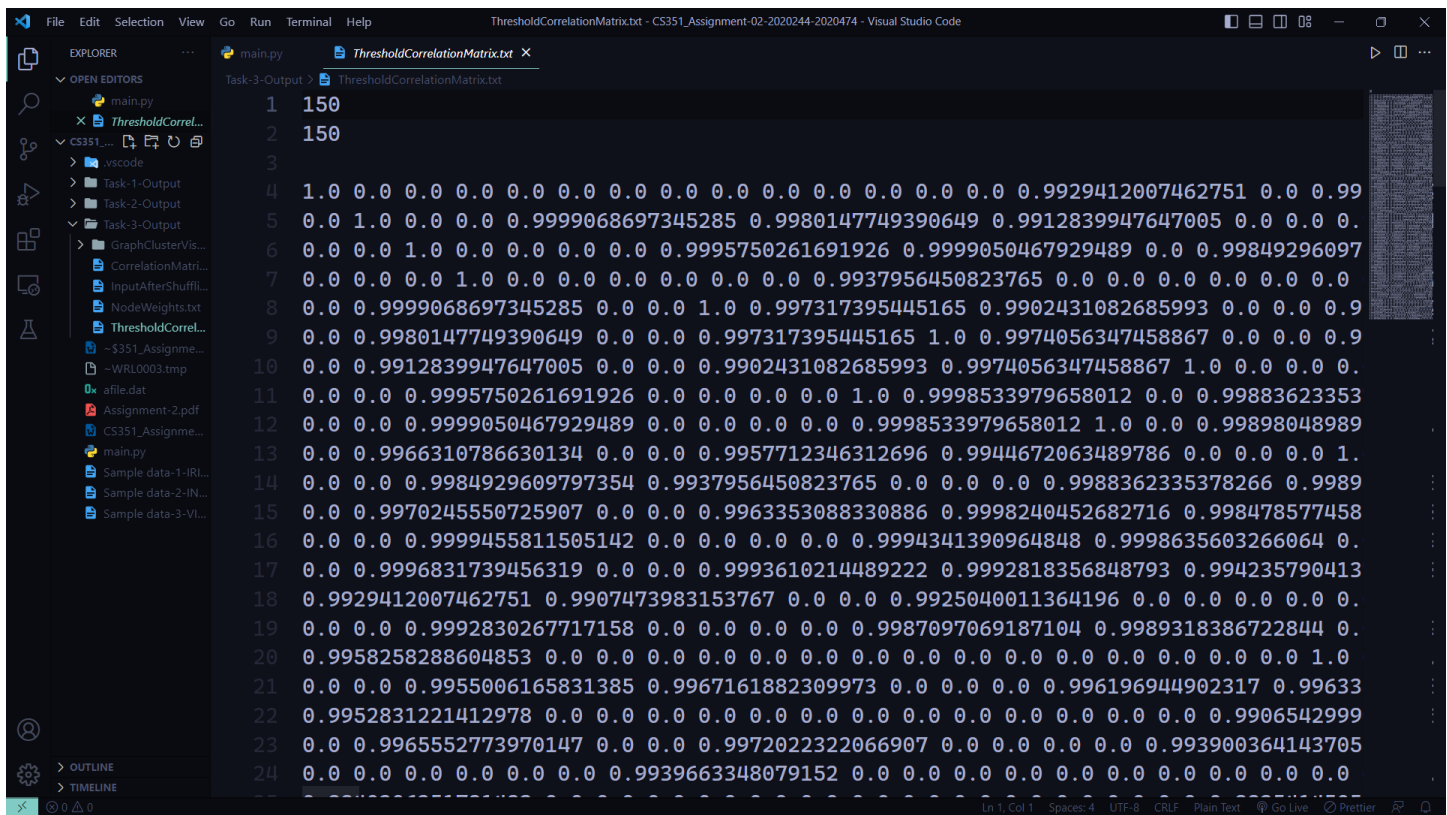
✓ **OUTPUT SCREENSHOTS:**

- *Node Weights Matrix:*
  - *The threshold is set to 0.99.*

The screenshot shows the Visual Studio Code interface with the 'NodeWeights.txt' file open. The file contains a list of 24 numerical values, each preceded by a line number from 1 to 24. The Explorer sidebar on the left shows the project structure, including folders for 'Task-1-Output', 'Task-2-Output', and 'Task-3-Output', with 'Task-3-Output' expanded to show subfolders like 'GraphClusterVis...' and files like 'NodeWeights.txt'.

Line Number	Node Weight
1	37.829184
2	63.696564
3	42.855684
4	37.832880
5	64.699600
6	47.828390
7	37.848264
8	43.863631
9	43.869552
10	30.835539
11	47.899688
12	47.835339
13	42.873655
14	55.777938
15	53.716542
16	33.847737
17	38.888478
18	48.887303
19	44.859111
20	60.725785
21	23.868011
22	49.829191
23	39.832470
24	56.718753

- *Threshold Correlation Matrix:*



- *Directory of png's overview/Clusters saved:*



- *Most dense and least dense clusters formed:*

```

PS C:\Users\Moshin\Desktop\3rd Year\6th Semester\Artificial Intelligence\CS351_Assignment-02-2020244-2020474> python -u "c:\Users\Moshin\Desktop\3rd Year\6th Semester\Artificial Intelligence\CS351_Assignment-02-2020244-2020474\main.py"

----- Menu -----
- Select Input File -
1. Sample data-1-IRIS
2. Sample data-2-INPUT1
3. Sample data-3-VINE
Enter your choice(1-3): 1

- SubMenu -
1. Task-01
2. Task-02
3. Task-03
4. Exit
Enter your choice(1-4): 3
Enter the threshold: 0.99

Most dense cluster: 112

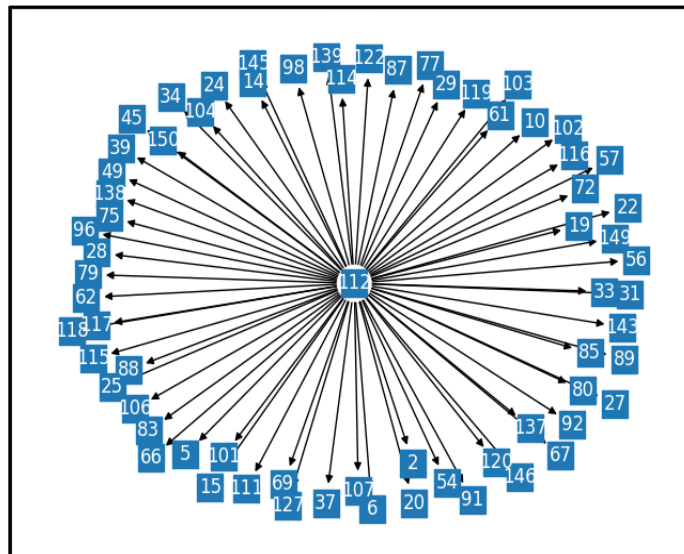
Least dense cluster: 41

- SubMenu -
1. Task-01
2. Task-02
3. Task-03
4. Exit
Enter your choice(1-4): 4

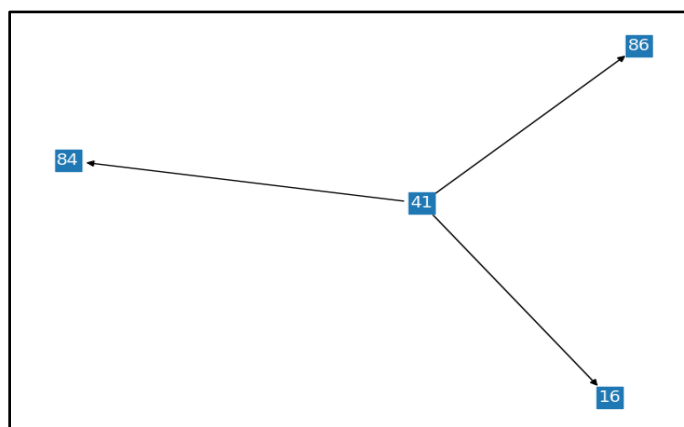
Thank You! See you again.
PS C:\Users\Moshin\Desktop\3rd Year\6th Semester\Artificial Intelligence\CS351_Assignment-02-2020244-2020474>

```

- *Most dense cluster:*



- *Least dense cluster:*



➤ **COMPARISON OF TASK 2 AND 3:**

✓ **TASK-2:**

- The primary goal of task 2 is to demonstrate the dissimilarity in bitmap by carrying out permutations on the data and presenting the bitmap of the discretized iris data before and after the permutation process. Subsequently, the permuted data undergoes sorting based on its signature values, and a color-coded image of the correlation data is generated and displayed.

✓ **TASK-3:**

- The primary objective of task 3 is to identify clusters within the correlation matrix subsequent to the permutation of data. To accomplish this, a threshold value is established, which in our case is 0.99, and the correlation matrix is updated accordingly. Following this, we pinpoint the node with the highest weight and construct a cluster including its surviving neighbors from the threshold process. This process is repeated iteratively until clusters are generated for each row in our graph list. Ultimately, this technique enables us to effectively determine the clusters present within the correlation matrix.

➤ **WORK DISTRIBUTION AMONG GROUP MEMBERS:**

Throughout the completion of this assignment, it is evident that we have made substantial contributions towards its successful completion. From the initiation of the project to the final submission, we both have actively participated in the task allocation, research, and implementation phases. Additionally, the collaborative nature of the project enabled both of us to offer valuable insights and suggestions, leading to the development of a comprehensive and well-structured final submission. We both have played an integral role throughout the assignment, and combined efforts have resulted in a successful outcome.

➤ **INSIGHTS:**

During the completion of this assignment, we have gained several valuable insights. Firstly, we discovered that the Pearson correlation matrix, although correlation is always symmetric process, didn't produce symmetric bitmap for discretized data. While initially considering that we may have made an error in our formula, using the predefined `pd.DataFrame.corr()` function produced results that were consistent with our output. Despite this, doubts remained about the accuracy of the sample image produced by us. It should also be noted that discretized data is calculated by taking the mean of the column and updating each value in that column. I think that it's the reason why the bitmap is not looking symmetric. Another observation made was that sorting via signature sorts both the input data and shuffled data to the same matrix, rendering the shuffling process is unnecessary here. Finally, we recognized the immense power of the Python language. Tasks that would have taken a significant amount of time in other programming languages such as C++, such as bitmap visualization (I first tried visualization using QT after completing these tasks in C++ and it was becoming too complex 😞), were easily accomplished using Matplotlib in just two lines of code in Python.

---

THE END.

---