Name: Vaibhav Agrawal
Roll Number: 15CE10057
Assignment 3: Clustering
Machine Learning (CS60050)

AAAI needs your help !!!

The Association for the Advancement of Artificial Intelligence (AAAI) organizes a conference of Artificial Intelligence, which is one of the most prestigious conferences relevant to the field of AI. Every year many researchers across the world submit to the conference and after a rigorous process of review and evaluation, only a selected set of papers get accepted. 150 such papers got accepted in AAAI this year. The submissions were found to span over several different domains of computer science such as: Machine Learning, Optimization, Knowledge-Based systems, Robotics, Natural Language Processing, etc.

You can find a dataset containing all the accepted submissions **here** . In this dataset you will find the following relevant attributes of each paper.

- **Title:** Free Text; Title of the paper
- **Keywords:** Free Text: author-generated keywords
- **Topics:** Categorical; author-selected, low-level keywords from conference-provided list
- **High-level Domains:** Categorical; author-selected, high-level keywords from conference-provided list. There are 9 distinct high-level domains in the dataset.
- **Abstract:** Free Text: abstract of the paper

AAAI wants an automated unsupervised script to group these documents into different clusters, so that all papers having similar high-level domains will be grouped together. For example: Let's assume there is a paper by Harish on a novel Clustering Algorithms, and there is another paper by Surya on a novel Classification Algorithm. Topics of the two papers might be different such as: {Clustering, Unsupervised, Machine Learning} and {Classification, Supervised, Machine Learning}, however both come under the same high-level domain , i.e., Machine Learning.

To complete this task there has to be a notion of similarity among different papers. For this assignment, the simple **jaccard coefficient of two sets of topics** is considered as the notion of similarity between the two papers. For example, if the set of topics for a paper is H = {Clustering, Unsupervised, Machine Learning} and if the set of topics for another paper is S = {Classification, Supervised, Machine Learning}, then the Jaccard Coefficient between the two papers is $JC_{HS} = JC_{SH} = \frac{|H \cap S|}{|H \cup S|} = \frac{1}{5} = 0.2$.

1. Implement a **bottom-up hierarchical clustering algorithm** considering the aforementioned notion of similarity, to find **9 (nine)** clusters using both the (i) **complete linkage** and (ii) **single linkage** strategies. State the clusters identified in your report.

2. With the same notion of similarity, design a graph where each node is a research paper, and an edge is to be drawn between two nodes if their topics are very similar (you can decide some threshold for this purpose, based on the quality of the clusters - see later). Now apply the **Girvan-Newman clustering algorithm** on this graph to find **9 (nine)** clusters. State the clusters identified in your report. For this part, you can use any graph library (e.g., the Networkx package in Python, or igraph library in C) for creating / updating the graph, and evaluating the centralities. However, direct use of any implementation of Girvan-Newman clustering algorithm will be penalised.
3. Consider the 'gold standard' clustering to be based on the high-level domains associated with the papers, i.e., all papers of a particular high-level domain (according to the dataset) constitute one cluster.
   Now you have three sets of clusters (having 9 clusters each) of the given set of research papers, one identified by the hierarchical clustering method with complete linkage, the second identified by the hierarchical clustering method with single linkage, and the third identified by the graph-based method. The final step is to evaluate the quality of these three clusterings. To this end, implement the **Normalized Mutual Information (NMI)** metric mentioned in this document to evaluate the quality of the clusters that you have got. Report the NMI values of the three clusterings.

Note that the clustering algorithms that you will implement in Parts 1 and 2 should use only the Topics (low-level keywords) in the data; they should NOT use the High-level domains. The High-level domains should be used only in Part 3, for evaluation of the clusterings.

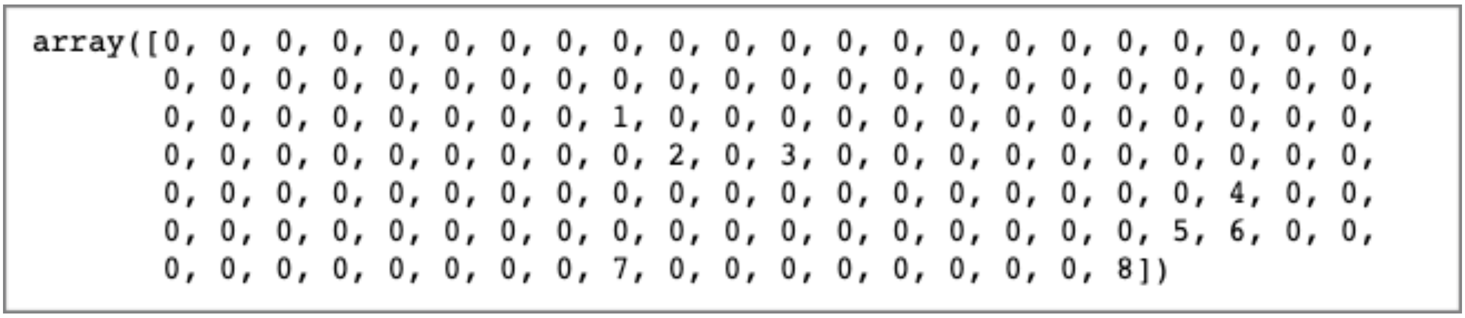
Report:

1. For each of the three clustering methods -- final learned clusters and the number of entities in each of them

Ans 1.) "Clusters" matrix is an array of size 150. Each index corresponds to a paper in the AAAI. 9 clusters are identified, each clusters is given by index ranging from 0 to 8. For example: In the array above, first paper (i.e. clusters[0]) belongs to first cluster given by index = 0.

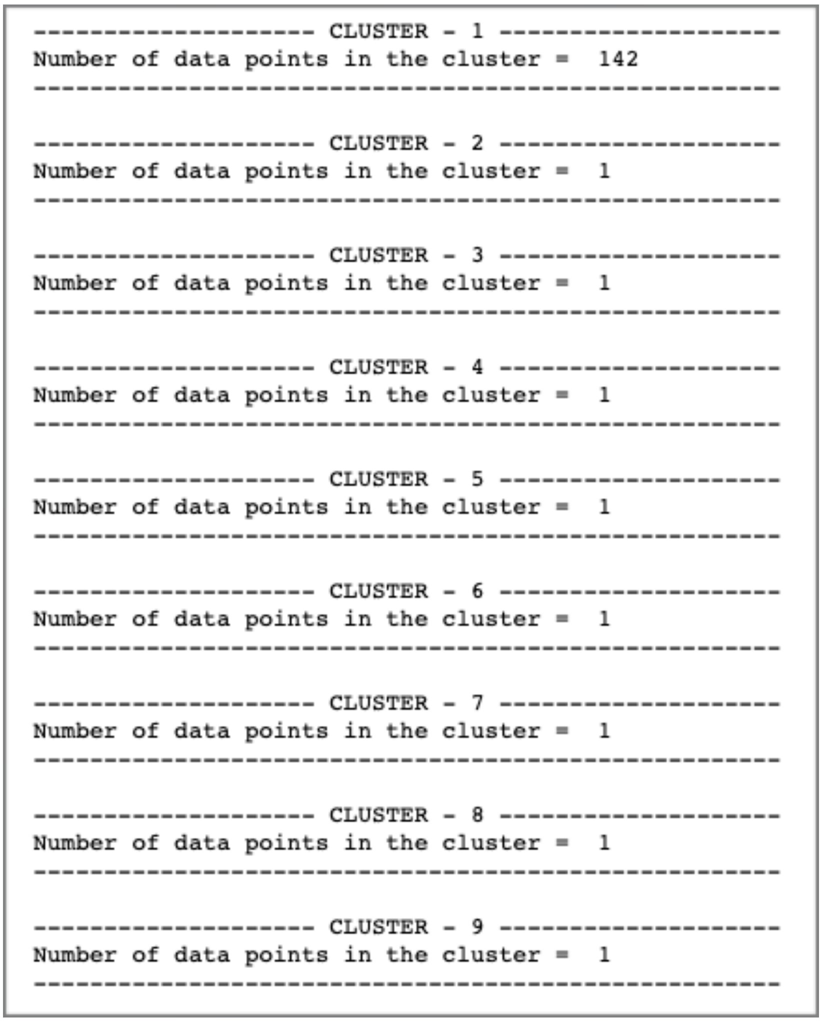
- ***Single Linkage Bottom-up hierarchical clustering algorithm*** **:**

clusters_single =



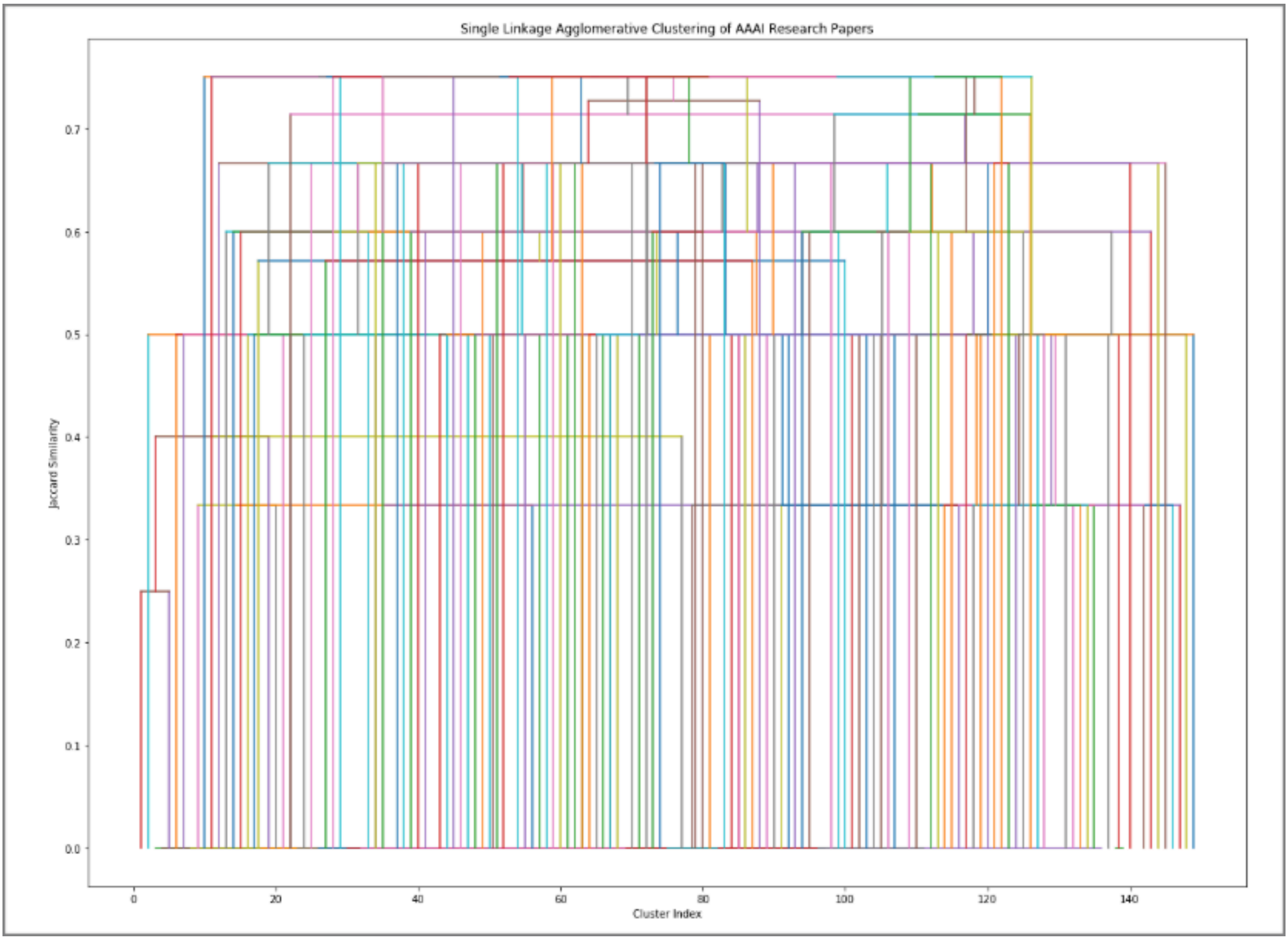
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       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, 5, 6, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 8])
```

Number of Entities:



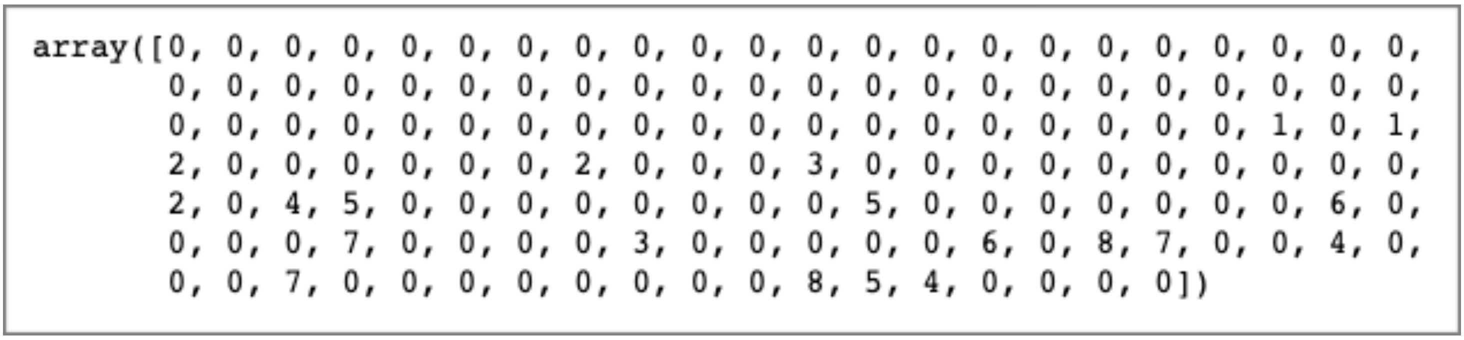
```
-------------------- CLUSTER - 1 --------------------
Number of data points in the cluster =  142
-----------------------------------------------------

-------------------- CLUSTER - 2 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 3 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 4 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 5 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 6 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 7 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 8 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------

-------------------- CLUSTER - 9 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------
```

Dendogram:



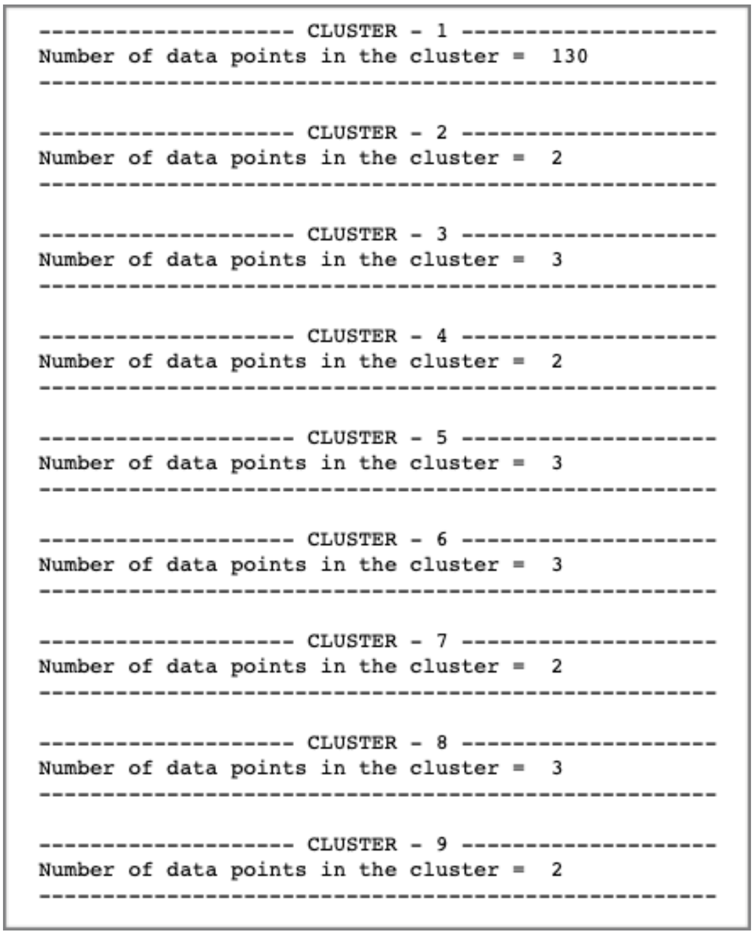
- ## *Complete Linkage Bottom-up hierarchical clustering algorithm*

clusters_complete =



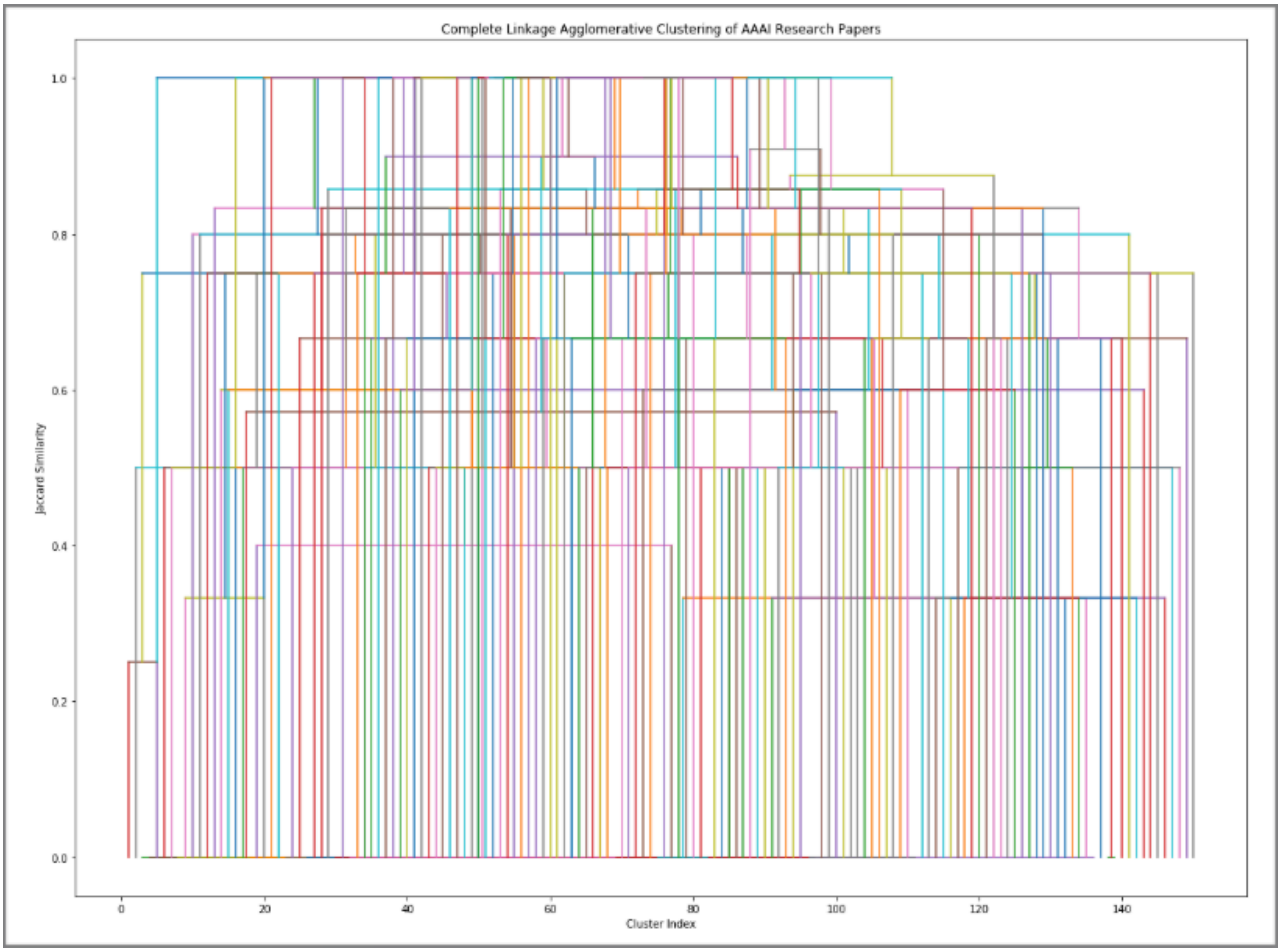
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       2, 0, 4, 5, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 6, 0,
       0, 0, 0, 7, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 6, 0, 8, 7, 0, 0, 4, 0,
       0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 8, 5, 4, 0, 0, 0, 0])
```

Number of Entities:



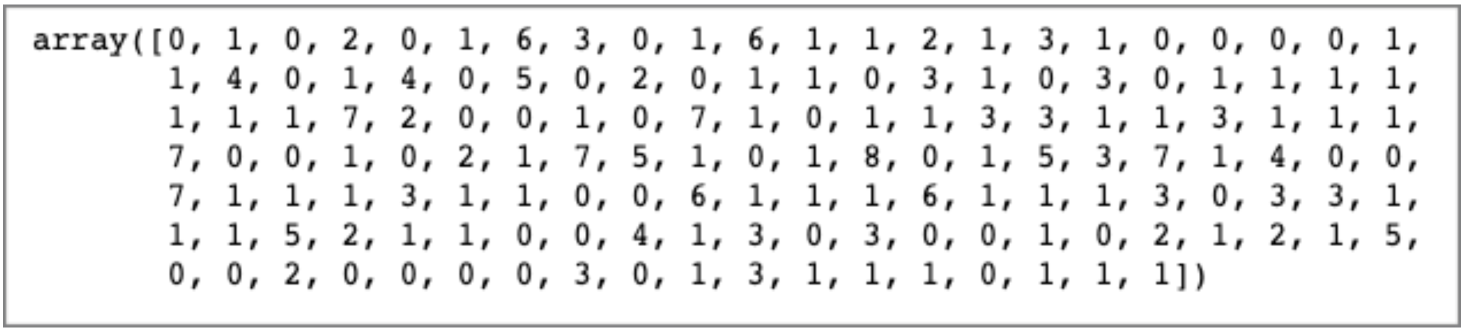
```
-------------------- CLUSTER - 1 --------------------
Number of data points in the cluster =  130
-----------------------------------------------------

-------------------- CLUSTER - 2 --------------------
Number of data points in the cluster =  2
-----------------------------------------------------

-------------------- CLUSTER - 3 --------------------
Number of data points in the cluster =  3
-----------------------------------------------------

-------------------- CLUSTER - 4 --------------------
Number of data points in the cluster =  2
-----------------------------------------------------

-------------------- CLUSTER - 5 --------------------
Number of data points in the cluster =  3
-----------------------------------------------------

-------------------- CLUSTER - 6 --------------------
Number of data points in the cluster =  3
-----------------------------------------------------

-------------------- CLUSTER - 7 --------------------
Number of data points in the cluster =  2
-----------------------------------------------------

-------------------- CLUSTER - 8 --------------------
Number of data points in the cluster =  3
-----------------------------------------------------

-------------------- CLUSTER - 9 --------------------
Number of data points in the cluster =  2
-----------------------------------------------------
```

Dendogram:



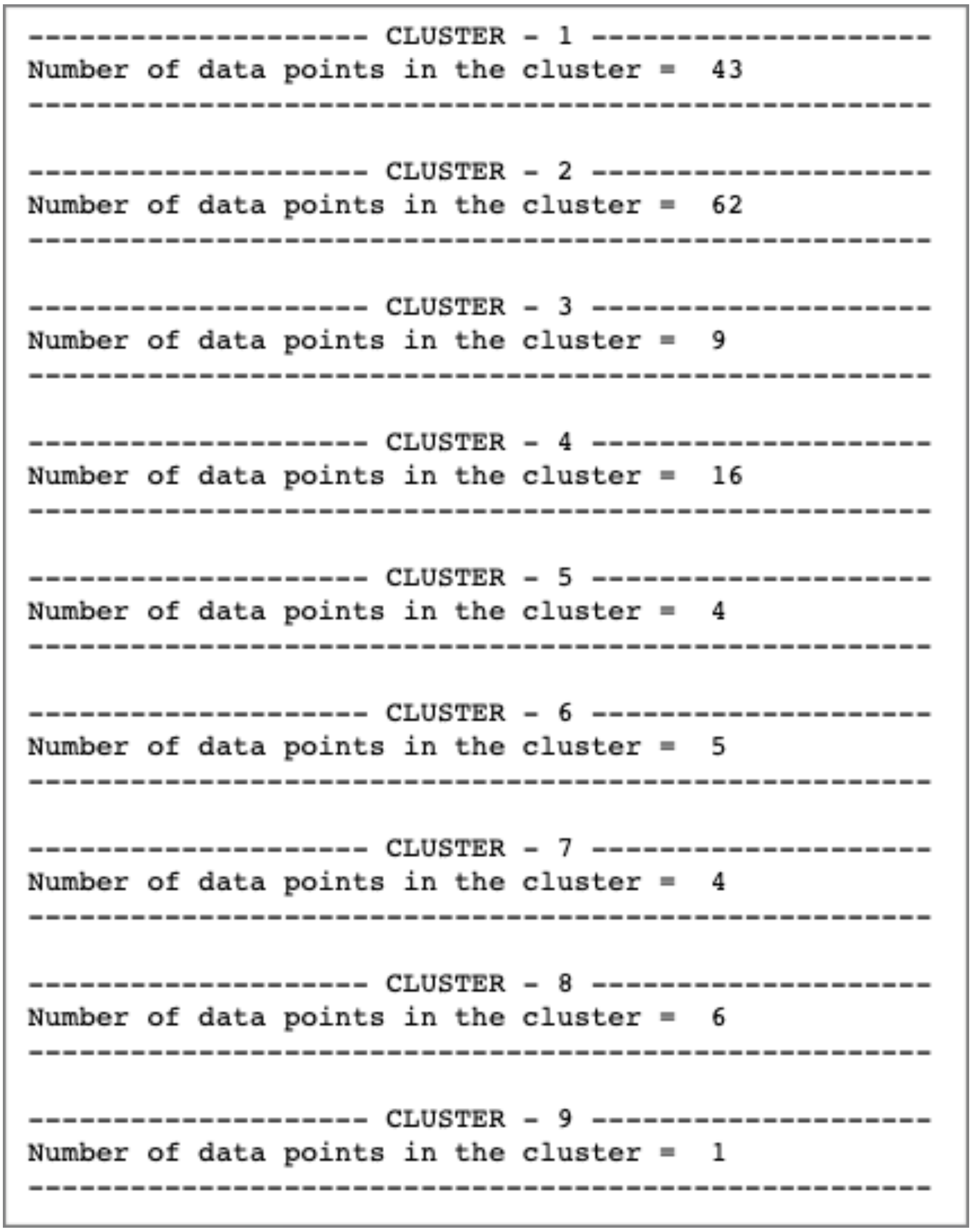
- ***Girvan-Newman clustering algorithm***

clusters_girvan =



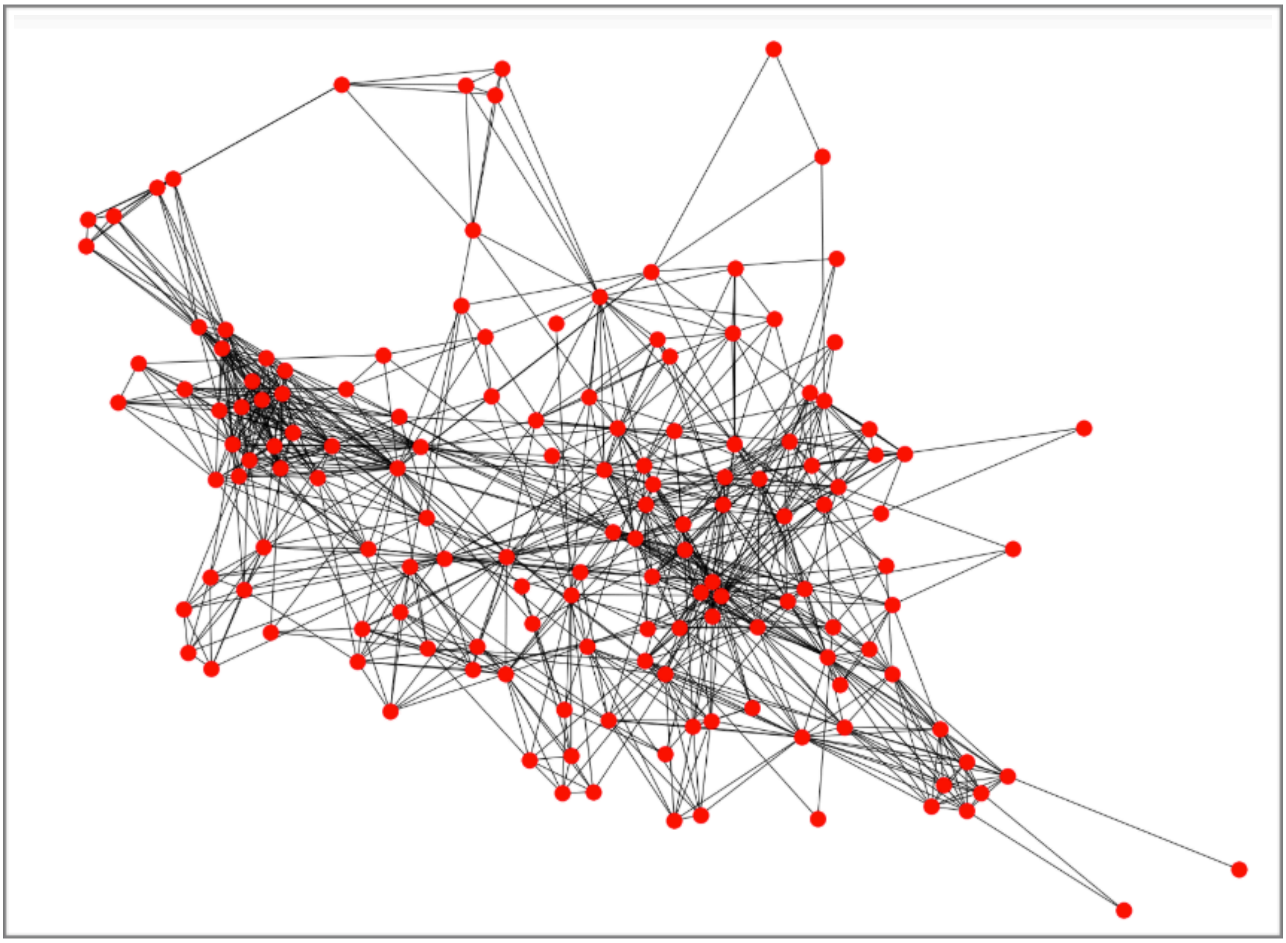
```
array([0, 1, 0, 2, 0, 1, 6, 3, 0, 1, 6, 1, 1, 2, 1, 3, 1, 0, 0, 0, 0, 1,
       1, 4, 0, 1, 4, 0, 5, 0, 2, 0, 1, 1, 0, 3, 1, 0, 3, 0, 1, 1, 1, 1,
       1, 1, 1, 7, 2, 0, 0, 1, 0, 7, 1, 0, 1, 1, 3, 3, 1, 1, 3, 1, 1, 1,
       7, 0, 0, 1, 0, 2, 1, 7, 5, 1, 0, 1, 8, 0, 1, 5, 3, 7, 1, 4, 0, 0,
       7, 1, 1, 1, 3, 1, 1, 0, 0, 6, 1, 1, 1, 6, 1, 1, 1, 3, 0, 3, 3, 1,
       1, 1, 5, 2, 1, 1, 0, 0, 4, 1, 3, 0, 3, 0, 0, 1, 0, 2, 1, 2, 1, 5,
       0, 0, 2, 0, 0, 0, 0, 3, 0, 1, 3, 1, 1, 1, 0, 1, 1, 1])
```

Number of Entities:



```
-------------------- CLUSTER - 1 --------------------
Number of data points in the cluster =  43
-----------------------------------------------------

-------------------- CLUSTER - 2 --------------------
Number of data points in the cluster =  62
-----------------------------------------------------

-------------------- CLUSTER - 3 --------------------
Number of data points in the cluster =  9
-----------------------------------------------------

-------------------- CLUSTER - 4 --------------------
Number of data points in the cluster =  16
-----------------------------------------------------

-------------------- CLUSTER - 5 --------------------
Number of data points in the cluster =  4
-----------------------------------------------------

-------------------- CLUSTER - 6 --------------------
Number of data points in the cluster =  5
-----------------------------------------------------

-------------------- CLUSTER - 7 --------------------
Number of data points in the cluster =  4
-----------------------------------------------------

-------------------- CLUSTER - 8 --------------------
Number of data points in the cluster =  6
-----------------------------------------------------

-------------------- CLUSTER - 9 --------------------
Number of data points in the cluster =  1
-----------------------------------------------------
```

Graph Network:



2. NMI values of the three clusterings you obtained

Ans 2.)

Clustering Algorithm	Normalized Mutual Information (NMI)
Single Linkage Bottom-up hierarchical	-1.332268
Complete Linkage Bottom-up hierarchical	-0.775029
Girvan-Newman clustering	0.320487

3. What are the thresholds that you are considering while doing the assignments have to be clearly mentioned. Note that you can decide upon what sort of thresholding you want to do based on the goodness of your clusters.

Ans 3.) The Jaccard Similarity is given by $JC_{HS} = JC_{SH} = \frac{|H \cap S|}{|H \cup S|}$. The distance between each label or paper is given by Jaccard Distance = (1 - Jaccard Similarity). Thus more the Jaccard Distance less would be the similarity between two entities. I have selected 1 as the threshold. Since, Jaccard distance of 1 implies that the similarity is zero between two entities, thus we start with a graph with maximum connectivity, with our assumed threshold.