Assignment 3: Clustering

-By Vaibhav Poddar
16CS10051

NOTE: Numbers in [ ] are the document Ids with starting index 0.

PART1

Cluster learned by Hierarchical Clustering method with Single Linkage are:

Cluster Number	***Size of the Cluster***	***Documents in the Cluster***
Cluster1	1	[52]
Cluster2	1	[75]
Cluster3	1	[77]
Cluster4	1	[107]
Cluster5	1	[128]
Cluster6	45	[0, 1, 2, 4, 8, 10, 17, 18, 19, 20, 24, 28, 29, 31, 37, 39, 49, 50, 55, 67, 68, 70, 74, 76, 79, 81, 86, 87, 95, 96, 106, 112, 116, 117, 121, 123, 126, 131, 132, 133, 135, 136, 137, 138, 146]
Cluster7	64	[3, 5, 6, 9, 11, 14, 16, 21, 22, 25, 27, 30, 32, 33, 36, 40, 41, 42, 43, 44, 45, 46, 51, 54, 56, 57, 60, 61, 63, 64, 65, 69, 72, 78, 80, 84, 89, 90, 91, 93, 94, 97, 98, 99, 100, 101, 102, 103, 104, 109, 110, 111, 114, 115, 119, 130, 140, 141, 143, 144, 145, 147, 148, 149]
Cluster8	22	[7, 12, 13, 15, 23, 26, 35, 47, 48, 53, 59, 62, 66, 73, 83, 85, 88, 92, 113, 118, 127, 134]
Cluster9	14	[34, 38, 58, 71, 82, 105, 108, 120, 122, 124, 125, 129, 139, 142]

Cluster learned by Hierarchical Clustering method with Complete Linkage are:

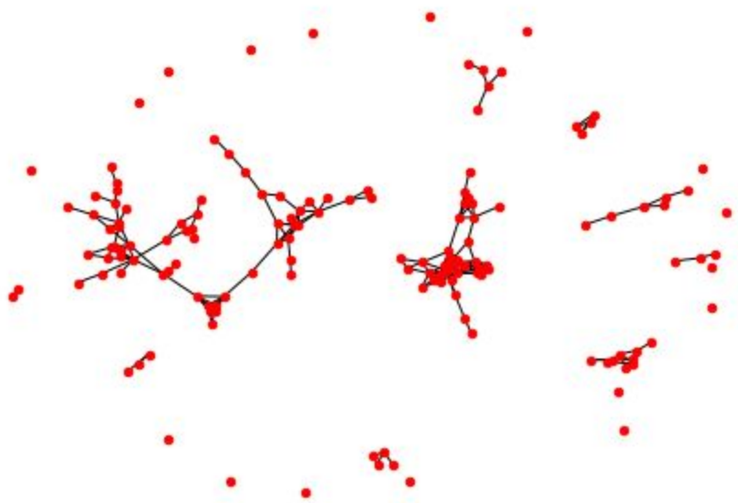
Cluster Number	***Size of the Cluster***	***Documents in the Cluster***
Cluster1	11	[46, 57, 60, 62, 78, 84, 90, 92, 104, 130, 145]
Cluster2	11	[6, 7, 11, 15, 32, 35, 59, 94, 97, 101, 143]
Cluster3	14	[24, 25, 37, 39, 41, 50, 70, 89, 91, 100, 111, 116, 121, 144]
Cluster4	12	[0, 4, 9, 10, 18, 27, 44, 47, 53, 76, 83, 140]
Cluster5	16	[12, 23, 34, 61, 63, 64, 65, 85, 107, 115, 118, 119, 139, 141, 147, 148]
Cluster6	18	[1, 8, 19, 22, 36, 45, 52, 69, 72, 79, 106, 109, 110, 117, 128, 133, 137, 138]
Cluster7	16	[5, 28, 33, 56, 66, 73, 74, 75, 77, 81, 88, 93, 112, 114, 131, 149]
Cluster8	23	[3, 13, 14, 26, 30, 38, 40, 42, 48, 54, 58, 71, 82, 99, 103, 105, 113, 120, 122, 127, 129, 134, 142]
Cluster9	29	[2, 16, 17, 20, 21, 29, 31, 43, 49, 51, 55, 67, 68, 80, 86, 87, 95, 96, 98, 102, 108, 123, 124, 125, 126, 132, 135, 136, 146]

PART2

Cluster learned by Girvan-Newman clustering algorithm are:(Threshold=0.279)

Cluster Number	***Size of the Cluster***	***Documents in the Cluster***
Cluster1	53	[130, 5, 6, 11, 141, 14, 143, 16, 145, 144, 147, 148, 21, 22, 25, 32, 33, 36, 40, 41, 42, 45, 46, 51, 54, 56, 57, 60, 61, 63, 64, 65, 69, 72, 78, 80, 84, 89, 90, 91, 93, 94, 97, 100, 101, 103, 104, 109, 110, 111, 114, 115, 119]
Cluster2	37	[0, 1, 2, 4, 133, 132, 135, 8, 136, 137, 138, 17, 18, 19, 20, 146, 24, 29, 31, 37, 39, 49, 50, 55, 67, 68, 70, 76, 79, 86, 95, 96, 106, 116, 117, 123, 126]
Cluster3	7	[66, 73, 13, 47, 48, 83, 88]
Cluster4	6	[35, 7, 15, 59, 92, 62]
Cluster5	6	[38, 105, 142, 120, 122, 125]
Cluster6	5	[12, 85, 118, 23, 26]
Cluster7	4	[34, 139, 58, 82]
Cluster8	4	[112, 81, 74, 131]
Cluster9	3	[99, 3, 30]

Graph when threshold = 0.279



PART3

Method	***NMI value***
Hierarchical Clustering method with Single Linkage	0.5066227644244067
Hierarchical Clustering method with Complete Linkage	0.3872525324557641
Girvan-Newman clustering algorithm	0.5453413235246195

Finding the best Threshold for Girvan-Newman clustering algorithm

Threshold: 0.999	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.969	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.939	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.909	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.879	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.849	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.819	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.789	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.759	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.729	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.699	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.669	Best_Threshold: 0.999	Max_NMI: 0.17657887391666793
Threshold: 0.639	Best_Threshold: 0.639	Max_NMI: 0.19288852050387326
Threshold: 0.609	Best_Threshold: 0.639	Max_NMI: 0.19288852050387326
Threshold: 0.579	Best_Threshold: 0.579	Max_NMI: 0.20575705950689016
Threshold: 0.549	Best_Threshold: 0.579	Max_NMI: 0.20575705950689016
Threshold: 0.519	Best_Threshold: 0.579	Max_NMI: 0.20575705950689016
Threshold: 0.489	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.459	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.429	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.399	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.369	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.339	Best_Threshold: 0.489	Max_NMI: 0.43461632368966047
Threshold: 0.309	Best_Threshold: 0.309	Max_NMI: 0.5272084210000898
Threshold: 0.279	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.249	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.219	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.189	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.159	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.129	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.099	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.069	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.039	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***
Threshold: 0.009	***Best_Threshold: 0.279***	***Max_NMI: 0.5453413235246195***

- Best Threshold for Girvan-Newman clustering algorithm is in the range **[0.009 - 0.279]**
- Therefore for the **PartB** I chose Threshold to be **0.279,** where if the Jaccard Coefficient between two paper is greater or equal to 0.279 only then, there will be an edge between those 2 documents.

-By Vaibhav Poddar
16CS10051

********** THE END **********