ASSIGNMENT 3

Q1 a) and b)

Each cluster is represented by an array
The array contains the index number which indicates the Title of the paper as read in the DataFrame named dataset

COMPLETE LINKAGE CLUSTER

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

NMI: 0.335782149799

SINGLE LINKAGE CLUSTER

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

NMI: 0.239357256535

GRAPH CLUSTERING

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

NMI: 0.571182388846

C)

For the thresholding 0.1 has been selected. So an edge would be connected between two nodes if the similarity (jaccard coefficient) between the two exceeds the value of 0.1.
0.1 has been selected because this gives a reasonably dense initial graph (we can not work with a sparse graph that already has more than 9 components) and the resulting NMI after making 9 clusters (or communities) is reasonably high too. Also checking for more thresholding values, we see that 0.1 is amongst the best ones.
NMI: 0.571182388846