Marking ID:	1990
Year:	2016
Course Code:	CS5100
Course Tutor:	Dr Volodya Vovk
Assignment No.:	3
Degree Title:	MSc Data Science and Analytics

Declaration of Academic Integrity:
This is to remind students that submitting this work acknowledges that the assignment is entirely their own work and has not been submitted previously for another course or programme at RHUL or any other institution.

Summary

I created a table called distancetable which displayed pairwise distance between all objects in nci dataset. I then classified the 64 objects into 32 clusters. After several steps of clustering, I obtained the root of the clustering which is the whole dataset.

With respect to the rule of inspecting the performance, I cut the tree in the second clustering which had 16 clusters because the number of the labels were 14 which was the closest number. I then used the sum of distinct labels in all 16 clusters as the criterion of examining performance (the smaller the better).

For performance of single linkage, the number was 41. I then clustered them with complete linkage whose result was 41 as well. The result of average linkage was equal to 38 which is much better. Finally, the performance of centroid linkage was 40. In this case, the result might show that the performance ranking is : average linkage > centroid linkage > single linkage ~ complete linkage.

Regarding k-means clustering, the result is equal to 30, which is significantly smaller than the result of hierarchical agglomerative clustering. Thus, it shows that k-means clustering creates better performance than hierarchical agglomerative clustering does.