CS-E4650 - Assignment 2

Arttu Häkkinen - 596077

October 24, 2021

Task 2

a)

Pairwise Jaccard's distances:

	t1	t2	t3	t4	t5	t6	t7	t8
t1	0.00	0.33	0.89	0.75	0.75	0.57	0.89	0.89
t2	0.33	0.00	1.00	0.57	0.57	0.57	0.75	0.75
t3	0.89	1.00	0.00	0.89	1.00	0.89	0.75	0.75
t4	0.75	0.57	0.89	0.00	0.33	0.75	0.75	0.57
t5	0.75	0.57	1.00	0.33	0.00	0.75	0.75	0.57
t6	0.57	0.57	0.89	0.75	0.75	0.00	0.89	0.89
t7	0.89	0.75	0.75	0.75	0.75	0.89	0.00	0.33
t8	0.89	0.75	0.75	0.57	0.57	0.89	0.33	0.00

b)

We iterate distance matrix by merging the transactions with smallest Jaccard's Distance. After merging, the new value of the combined transactions will be the largest of the original distance matrix entries regarding the merged transactions, E.g. if t_1 and t_2 are merged into t_{12}, and $d_J(t_1, t_3) = 1$ and $d_J(t_2, t_3) = .89$, the new matrix entry would be $d_J(t_{12}, t_3) = 1$.

b) Complete linkage metric

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
t_1	0							
t_2	**,33**	0						
t_3	,89	1	0					
t_4	,75	,57	,89	0				
t_5	,75	,57	1	,33	0			
t_6	,57	,57	,89	,75	,75	0		
t_7	,89	,75	,75	,75	,75	,89	0	
t_8	,89	,75	,75	,57	,57	,89	,33	0

→ Merge t_1 and t_2, since their d_j is smallest.

	t_{12}	t_3	t_4	t_5	t_6	t_7	t_8
t_{12}	0						
t_3	1	0					
t_4	,75	,89	0				
t_5	,75	1	**,33**	0			
t_6	,57	,89	,75	,75	0		
t_7	,89	,75	,75	,75	,89	0	
t_8	,89	,75	,57	,57	,89	,33	0

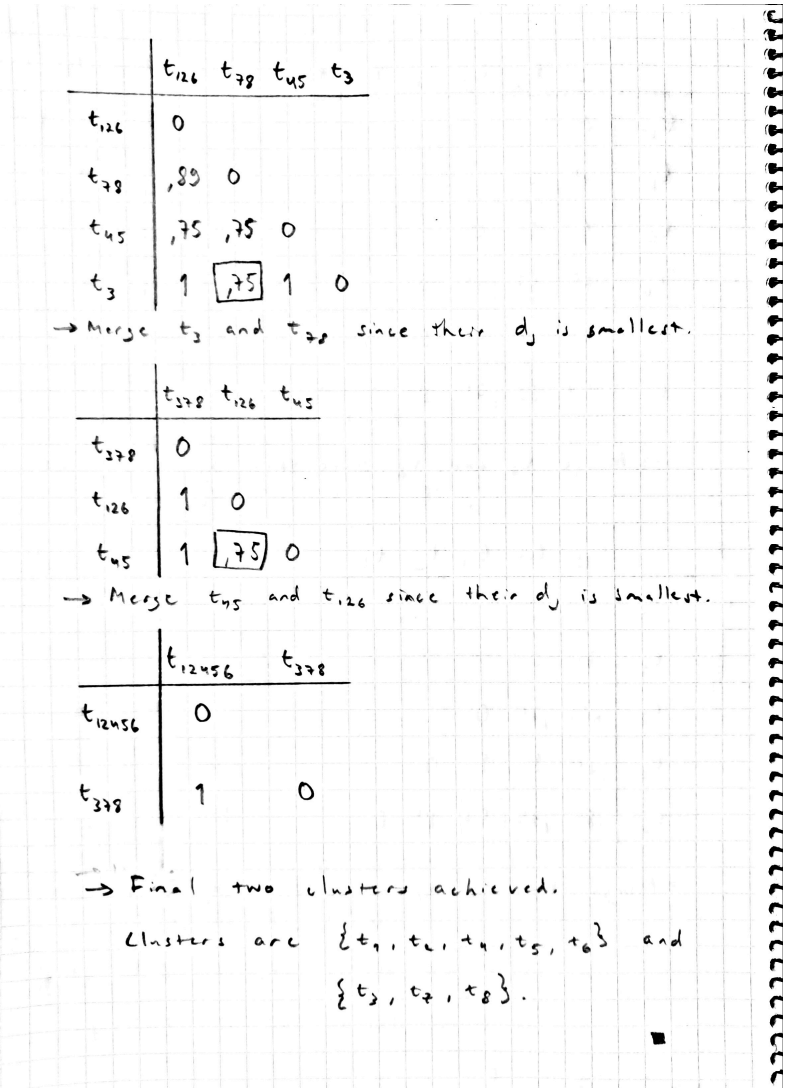
→ Merge t_4 and t_5, since their d_j is smallest.

	t_{45}	t_{12}	t_3	t_6	t_7	t_8
t_{45}	0					
t_{12}	,75	0				
t_3	1	1	0			
t_6	,75	,57	,89	0		
t_7	,75	,89	,75	,89	0	
t_8	,57	,89	,75	,89	**,33**	0

→ Merge t_7 and t_8, since their d_j is smallest.

	t_{78}	t_{45}	t_{12}	t_3	t_6
t_{78}	0				
t_{45}	,75	0			
t_{12}	,89	,75	0		
t_3	,75	1	1	0	
t_6	,89	,75	**,57**	,89	0

→ Merge t_{12} and t_6 since their d_j is smallest.



	t_{126}	t_{78}	t_{45}	t_3
t_{126}	0			
t_{78}	,89	0		
t_{45}	,75	,75	0	
t_3	1	,75	1	0

→ Merge t_3 and t_{78} since their d_j is smallest.

	t_{378}	t_{126}	t_{45}
t_{378}	0		
t_{126}	1	0	
t_{45}	1	,75	0

→ Merge t_{45} and t_{126} since their d_j is smallest.

	t_{12456}	t_{378}
t_{12456}	0	
t_{378}	1	0

→ Final two clusters achieved.

Clusters are $\{t_1, t_2, t_4, t_5, t_6\}$ and $\{t_3, t_7, t_8\}$.

■

So, it is proved that that it is possible to yield two different clusterings: t_{12456} and t_{378}. Also we can prove that another clustering of two clusters would be reached if we return in the above tables to the table where we still have four clusters left: $t_{126}, t_{78}, t_{45}, t_3$. If we were to choose to merge t_{45} and t_{78} instead, we would reach different two cluster end result.

c)

Same as in part b, but now instead of updating with the largest pairwise distance, update with the smallest pairwise distance.

c) Single linkage metric

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
t_1	0							
t_2	**,33**	0						
t_3	,89	1	0					
t_4	,75	,57	,89	0				
t_5	,75	,57	1	,33	0			
t_6	,57	,57	,89	,75	,75	0		
t_7	,89	,75	,75	,75	,75	,89	0	
t_8	,89	,75	,75	,57	,57	,89	,33	0

→ Merge t_1 and t_2

	t_{12}	t_3	t_4	t_5	t_6	t_7	t_8
t_{12}	0						
t_3	,89	0					
t_4	,57	,89	0				
t_5	,57	1	**,33**	0			
t_6	,57	,89	,75	,75	0		
t_7	,75	,75	,75	,75	,89	0	
t_8	,75	,75	,57	,57	,89	,33	0

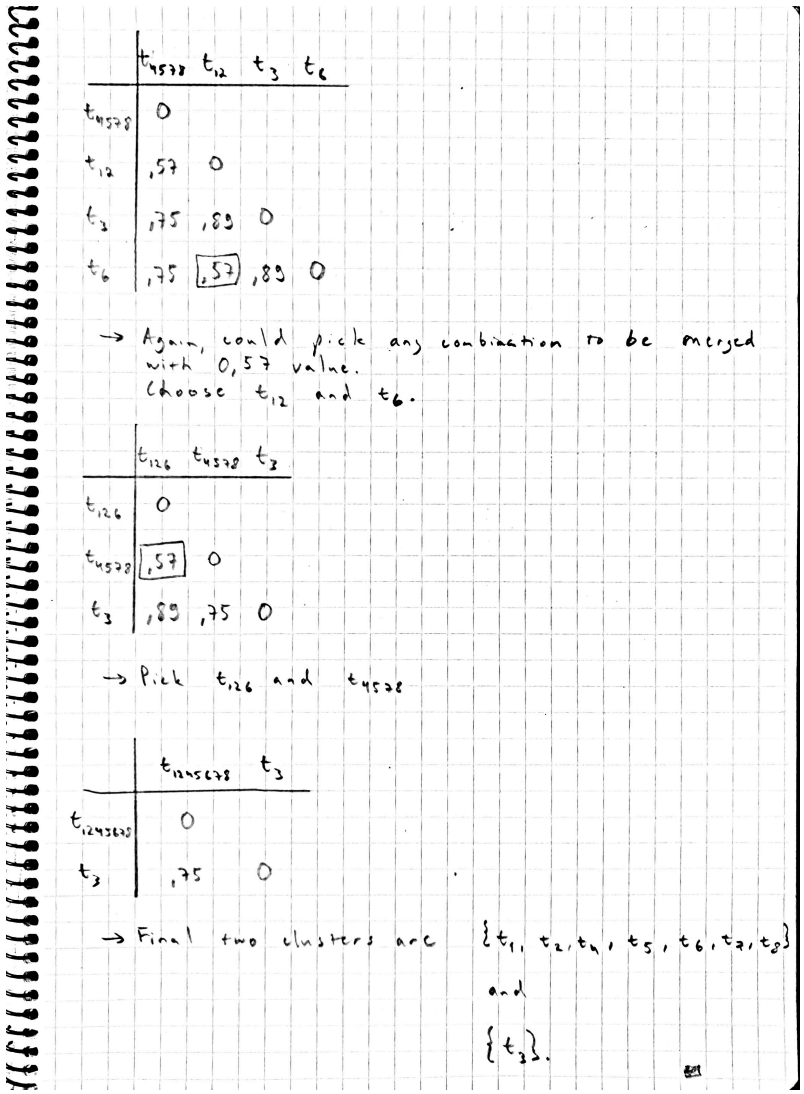
→ Merge t_4 and t_5.

	t_{45}	t_{12}	t_3	t_6	t_7	t_8
t_{45}	0					
t_{12}	,57	0				
t_3	,89	,89	0			
t_6	,75	,57	,89	0		
t_7	,75	,75	,75	,89	0	
t_8	,57	,75	,75	,89	,33	0

→ Merge t_7 and t_8.

	t_{78}	t_{45}	t_{12}	t_3	t_6
t_{78}	0				
t_{45}	,57	0			
t_{12}	,75	,57	0		
t_3	,75	,89	,89	0	
t_6	,89	,75	,57	,89	0

→ Could merge any 0,57 distance value row, column pair.
Pick t_{45} and t_{78} for merging.



The order of the data doesn't matter. With single linkage metric clustering, t_3 would always form a cluster on its own, as long as the clustering is proceeded until there are N=2 clusters. This is because the smallest pairwise distances of t_3 equal to .75, and since there are in total 9 pairwise distances in the matrix M that are smaller than .75, the clustering would never have other N=2 clusterings than $t_{1245678}$ and t_3 - no matter the order we choose to merge the equal distances in the distance matrix M.

References

[1] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html