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 $d_J(t_{12},t_3)=1$.

-			tz			t.	t,	t ₃
+,	0				3			
t _z	- -	0						
	[33]							
+,	,95							
t,	1,42	, 2,	, 85	0				
t ₅	'32	f2,	1	,33	0			
t,	123	,57	, 85	,75	,75	0		
4	ده,	,75	,75	, 3 5	,75	,85	0	
+3	1,85	,75	,75	,57	,57	,85	, 33	0
→	Merge	t,	and	t _e	, sin	cc +	heir	dj is smallest.
	t12	f3	ŧ"	t ₅	t,	tą	t,	
t12	0							
t3	1	0						
t "	,75	,85	0					
ts	1,75		[,33]	0				
ŧ,	1		,75		0			
t,			, 7 5			0		
t 3	-		,57				0	
D	1.37			H				

tys 0 to 2,75 0 to 3,75 ,57 ,80 0 to 3,75 ,57 ,80 0 to 3,57 ,80 ,75 ,80 0 to 4,57 ,75 0 to 2,89 ,75 0 to 2,89 ,75 0 to 3,85 ,75 ,57 ,80 0 >Mers c t,2 and to since their do is smalle		tus	t,2	t,	t.	t	tg		1				
t ₃ 1 1 0 t ₆ ,75 ,57 ,80 0 t ₇ ,75 ,80 ,75 ,80 0 t ₈ ,57 ,80 ,75 ,80 [,33] 0 > Marge t ₇ and t ₈ , since their d ₂ is smalle t ₇ t ₇ t ₇ t ₇ t ₈ t ₈ t ₇ 0 t ₁₂ ,89 ,75 0 t ₁₂ ,89 ,75 0 t ₁₃ ,75 1 1 0 t ₆ ,89 ,75 [,57] ,89 0	tus	0											
t ₆ ,75 ,57 ,80 0 t ₇ ,75 ,80 ,75 ,80 0 t ₈ ,57 ,80 ,75 ,80 ,33 0 -> Merse t ₇ and t ₈ , since their d ₂ is smalle t ₇₈ t ₇₅ t ₇₂ t ₃ t ₆ t ₇₈ 0 t ₇₂ ,75 0 t ₇₂ ,75 1 1 0 t ₆ ,89 ,75 ,57 ,89 0	ŧ12	75	0		S S B S								
to 1,75,89,75,85 0 to 1,57,85,75,85,33 0 -> Marge to and to since their do is smalle to 1,88 to 1,2 to 1,2 to 1,2 to 1,2 to 1,3 to 1,4 to 1	f3	1	1	0									
t ₈ ,57 ,85 ,75 ,85 ,33 0 -> Merse t ₂ and t ₈ , since their d ₃ is smalle t ₂ t ₃ t ₄ t ₅ t ₇ t ₇ t ₈ t ₆ t ₁₂ ,39 ,75 0 t ₁₂ ,39 ,75 0 t ₁₃ ,75 1 1 0 t ₆ ,89 ,75 ,57 ,89 0	t,	₹,	,57	,80	•	A 180 a 100 a		-					-
-=> Merge to and to, since their do is smalled to the top of the	t _a	,75	,85	, 7 5	.85	٥		A					
-= Merge to and to, since their do is smalled too to the to	t ₈	,57	,85	, 75	,85	,33	0	H	- +				
t ₁₂ ,85 ,75 0 t ₁₂ ,89 ,75 0 t ₃ ,75 1 1 0 t ₆ ,89 ,75 ,57 ,89 0	-> 1	Merse	. t _a	• • •	l t	s, sie	cc	the	ir d	١	is	٦~	. 11e
t ₁₂ ,75 0 t ₁₂ ,85 ,75 0 t ₂ ,75 1 1 0 t ₆ ,85 ,75 ,57 ,80 0		t 78	t ₅₅	t,2	t ₃	te							
t ₁₂ ,89 ,75 0 t ₃ ,75 1 1 0 t ₆ ,89 ,75 ,57 ,89 0	tąş	0				2.		1 - 0					
t ₃ ,75 1 1 0 t ₆ ,85 ,75 ,57 ,85 0	tys	, 7 5	0								1		
e, 18, 17, 2f, 88, 3+	t12	,89	,75	0						1			, ,
	+3	,75	- 1	1	0								
-> Merge tiz and to since their do is smalle	ŧ,	,89	,75	[,57]	,85	0							14
	→Me	٠ ٠	+, ₂	and	ŧ,	۱۰ د	ce	the	in +	لم	زا	Jan .	. 11e
			1)	

t,26	t ₁₂₆ t ₇₈ t ₄₅ t ₃
	,82 0
+ + + + + + + + + + + + + + + + + + + +	0 2f, 2f,
	1 75 1 0
> Merse	to and too since their do is smallest.
	t278 t,26 tus
+278	٥
t,26	1 0
t 45	1 (75) 0
- Mer	se tys and tize since their old is smallest.
	t,2456 t378
tizuse	
t338	1 0
-> F:	nal two clusters achieved.
۷١.	asters are Eta, to, ta, ts, to and
	{ t ₃ , t ₄ , 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.

 $\mathbf{c})$

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

e) Sin	sle	lin le	٠, د	net	~:•									
					t ₅	ti	t,	t		,				
ŧ,	0								7					
+a	33	0												
£3	,85	1	0											
ŧ,	,75	,57	,89	0										
+ ₅	25.	,57	1	,33	0									
te	,57	,57	,85	, 75	.75	0								
t	,85	,75	,75	,35	.25	.83	0				1			
+8	.85	26,	25,	,57	,57	,85	.33	0						
	> ^	vers.	e t.	, an	d to			y 2		-			-	
	t,2	t3	tn	f2	+6	t,	t _g		-			33		
t,2	0													
٤,	,85	0												
٤,	+2,	,85	0											
ts	123	1	133	0										
+6	,57	,85	,75	,35	0		4							
t,	'32	,35	,95	:35	,85	0								
+ E	,75	,75	,57	, 53	,85	,33	0							
	-> Me		L	1	ts.									-

	tus	t,2	£ 3	46	t,	ts	
+42							
	,5 7						
43	,89	,85	0				
	,75			0			
ŧ,	25,	,75	2 €,	,85	0		
ŧε	, 2 3	,75	135	,85	1,33	0	
_	> Mei	د	t,	م م ما	t g	•	
	t = 1	÷45	6,2	÷ 3	ŧ,		
t,,	1						
tus	F2,	0					
t,2	,75	,57	0				
t ,	,75	.85	,82	0			
ŧ,	,85	26,	,53	,85	0		
	LO A1	, 10	Inm.	n P	Air.	0,57 dis	va lue

tyses 0 tyses 0 tyses 50 tyses 50		tisso t,2 t3 tc
t ₁₂ ,57 0 t ₂ ,75 ,83 0 t ₆ ,75 ,53 ,83 0 -> Again, con/A pick and combination to be mirged with 0,57 value. (hoose t ₁₂ and t ₆ . t ₁₂ t ₁₂ 0 t ₁₂ t ₁₃ 0 t ₂ ,83 ,75 0 -> Pick t ₁₂ and t ₄ ; se t ₁₂ t ₁₃ 0 -> Final two clusters are lt, t ₂ , t ₄ , t ₅ , t ₆ , t ₇ .	t4578	
ty 175 ,88 0 th 175 [57] ,88 0 Again, con/A pick any continetion to be merged with 0,57 value. (hoose ty and to. tussy [57] 0 tussy [57] 0 -> Pick tye and tyese tusses ty tusses () -> Pick tye and tyese tusses () -> Pick tye and tyese tusses () -> Pick tye and tyese there () -> Pick tyes	+12	,57
to ,75 [57],85 0 Again, con/A pick any combination to be marged with 0,57 value. Choose tiz and to. tize tusse to tize 0 tusse 0 tyse 1,57 0 tyse 1,57 0 tyses 0 tiansers to 1 tiansers to 1 Tiansers to 1 Final two clasters are lity, text, to, to, to,	£3	75, 25, 25,
→ Again, could pick any combination to be marged with 0,57 value. Choose tiz and to. tize tusso to tize 0 -> Pick tize and tyse tauses to tizes o tizes o -> Pick tize and tyse tizes o tizes o -> Pick tize and tyse and tyse tizes o -> Final two clusters are lite, to, to, to, to, to, to, to, to, to, to	te	75, [2], 26,
$\begin{array}{cccccccccccccccccccccccccccccccccccc$		with 0,57 value.
ty 150 , 75 0 ty 150		5126 fuzzb fz
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	t,26	0
toursers tourse	t4578	.53
+> Pick t, 26 and tyrse toursess to toursess 0 toursess 0 -> Final two clasters are lite, to, to, to, to, to,	€3	0 24, 22,
thusers t_3 $t_{1243635}$ t_3 t_75 $t_{1243635}$ t_{1	-5	Pick 6,26 and 64528
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$		
tizers 0 t_3 t_3 t_3 t_4 t_5 t_6 t_7 t_8		
t_3 , t_3 0 . Step and t_4 , t_5 , t_6 , t_7 . And t_8 .	£1242838	
\rightarrow Final two clasters are $\{t_1, t_2, t_4, t_5, t_6, t_7, \dots \}$	+3	٠ 0 26,
	->	final two clusters are Ety, tz, ta, ts, ta, ta, t
{ + ₃ },		\{\epsilon\} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \

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