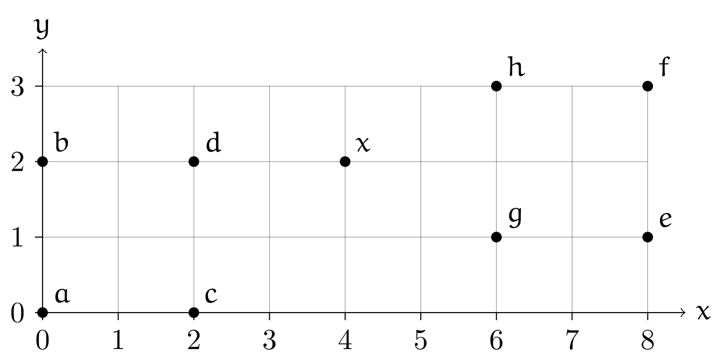
Cvičení 10

Pojmy potřebné pro zvládnutí tohoto cvičení: vzdálenosti, podobnosti a nepodobnosti, aglomerativní hierarchické shlukování, k-means, k-medoids, matice vzdáleností (nepodobností), dendrogram.

Příklad 1: Shlukování - rekapitulace.

a) Jaké shlukovací algoritmy a na jaká data můžete použít? Uveďte základní myšlenky algoritmů.
b) Pro které algoritmy je vhodné vytvořit nejprve matici vzdáleností?
c) Jakými způsoby můžete reprezentovat výsledky shlukovacího algoritmu?

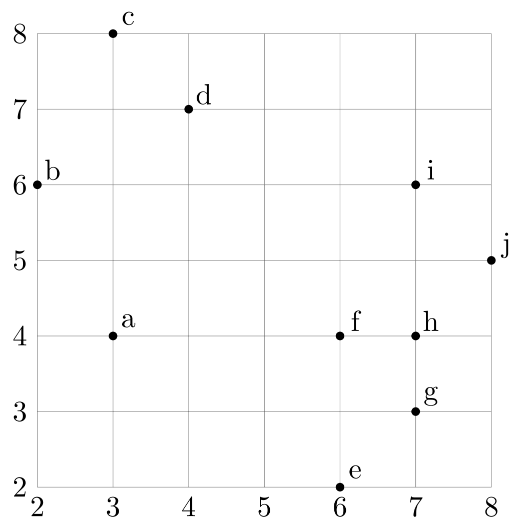
Příklad 2: Použijte hierarchické shlukovací metody (SL,CL,AL,CeL) pro nalezení dendrogramů v níže uvedených datech s využitím manhattenovské (euklidovské a čebyševovské) vzdálenosti. Z vytvořených dendrogramů odhadněte nejvhodnější počet shluků. Zamyslete se, jak a které dendrogramy se změní, pokud prohodíte pojmenování bodů ”x”a ”d”.



Příklad 3: Vypočtěte vzdálenosti mezi danými shluky C_i pro manhattanskou, čebyševovskou a Ružičkovu metriku nad danými daty. Metody pro výpočet mezishlukové vzdálenosti jsou: nejbližší sousedé (single linkage), nejvzdálenější sousedé (complete linkage), průměrná vzdálenost mezi shluky (average linkage) a vzdálenost centroidů (centroid linkage).

připomínám, že Ružičkova vzdálenost je založena na Růžičkově podobnosti:
$d_R(x, y) = \frac{\sum_{i=1}^{n} max(x_i,y_i) - \sum_{i=1}^{n} min(x_i,y_i)}{\sum_{i=1}^{n} max(x_i,y_i)}$.

$C_1 = \{c, d\}$,
$C_2 = \{f, g, h\}$,
$C_3 = \{i, j\}$,
$C_4 = \{a, b, c, d\}$



$d_1^{SL}(C_1, C_2) =$	$d_R^{SL}(C_1, C_2) =$	$d_\infty^{SL}(C_1, C_2) =$
$d_1^{CL}(C_1, C_2) =$	$d_R^{CL}(C_1, C_2) =$	$d_\infty^{CL}(C_1, C_2) =$
$d_1^{AL}(C_1, C_2) =$	$d_R^{AL}(C_1, C_2) =$	$d_\infty^{AL}(C_1, C_2) =$
$d_1^{Ce}(C_1, C_2) =$	$d_R^{Ce}(C_1, C_2) =$	$d_\infty^{Ce}(C_1, C_2) =$

Příklad 4: Vytvořte matice vzdáleností pomocí manhattanské a čebyševovské metriky nad daty z příkladu 3. Pro které shlukovací algoritmy jsou tyto matice použitelné?

Pomocí jednotlivých metod aglomerativního hierarchického shlukování vytvořte dendrogramy reprezentující hierarchii shluků.

Příklad 5: Vytvořte matici vzdáleností pro vektorová data z tabulky 1 pomocí manhattanské a čebyševovy metriky a vytvořte matici vzdáleností (Jaccard) pro data z tabulky 2, kde symbol x reprezentuje situaci, že prvek a_jV patří do množiny popisující objekt o_i.

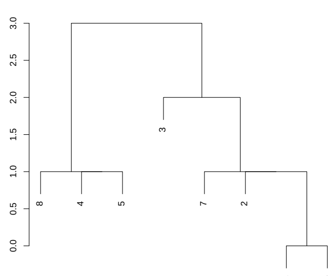
$U \setminus A$	a_1	a_2	a_3
o_1	0	1	2
o_2	1	1	3
o_3	0	0	1
o_4	1	3	2
o_5	0	3	2
o_6	0	1	2
o_7	1	1	2
o_8	1	3	3

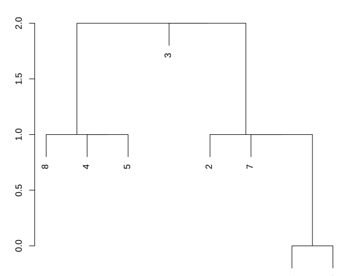
Tabulka 1: Aprox. pr.

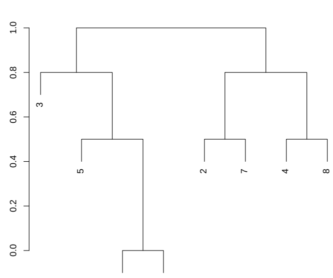
$U \setminus f(o_i, a_j)$	$a_1 0$	$a_1 1$	$a_2 0$	$a_2 1$	$a_2 3$	$a_3 1$	$a_3 2$	$a_3 3$
o_1	x			x			x	
o_2		x		x				x
o_3	x		x			x		
o_4		x			x		x	
o_5	x				x		x	
o_6	x			x			x	
o_7		x		x			x	
o_8		x			x			x

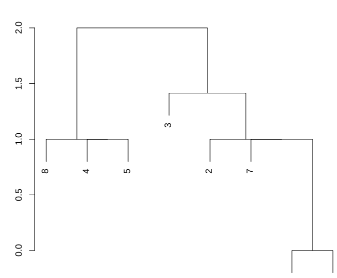
Tabulka 2: Aproximační prostor po one-hot-encoding

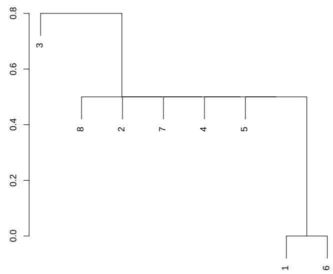
Pro vytvořené (níže uvedené) dendrogramy určete, jaká byla použita metoda pro výpočet vzdáleností mezi prvky a jaká byla použita metoda pro výpočet vzdálenosti mezi shluky. Na jaké úrovni řezu ve vytvořených dendrogramech získáte právě 3 shluky? Je zde nějaké odlehlé pozorování (outlier)?

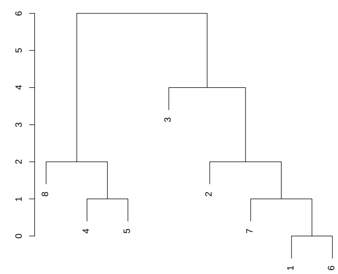




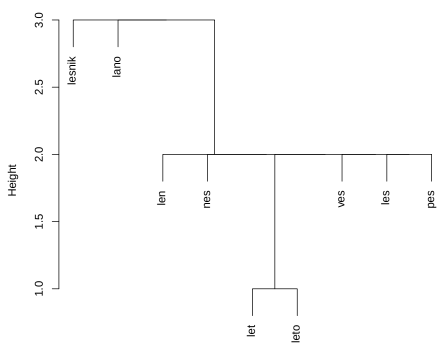


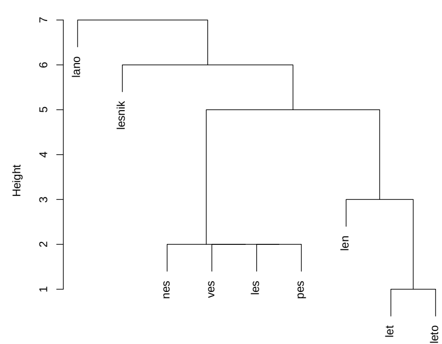


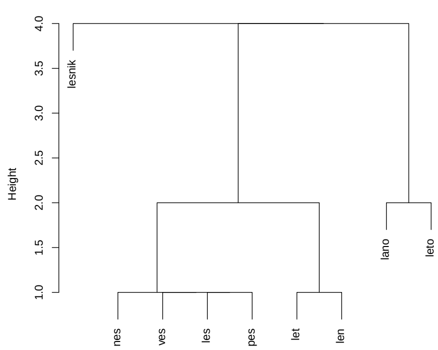


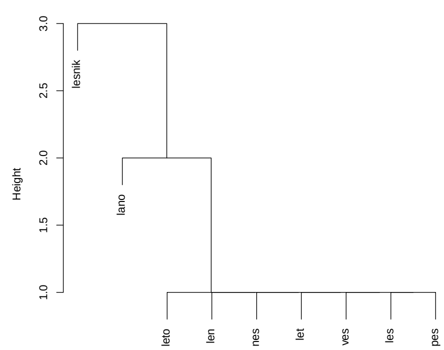


Příklad 6: Vytvořte matice vzdáleností pomocí editační (Levenshteinovy) a LCS metriky nad množinou slov {les, pes, ves, let, nes, lano, len, lesník, léto}. Pro vytvořené (níže uvedené) dendrogramy určete, jaká byla použita metoda pro výpočet vzdáleností mezi prvky a jaká byla použita metoda pro výpočet vzdálenosti mezi shluky. Na jaké úrovni řezu ve vytvořených dendrogramech získáte právě 3 shluky?

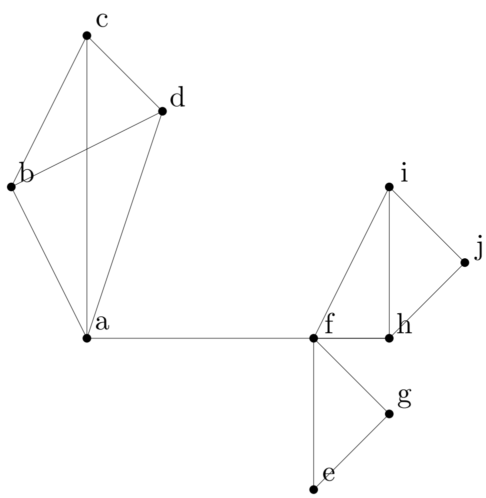








Příklad 7:
Pro daný graf vytvořte matici vzdáleností - použita metrika je délka nejkratší neorientované cesty v daném grafu. Pomocí hierarchického shlukování nalezněte shluky (komunity) v daném grafu. Porovnejte metody single linkage a complete linkage. Jaké další metody můžete použít pro hierarchické shlukování na grafech? Můžete stejný postup použít i v případě orientovaného grafu?



Příklad 8: Pro data z příkladu 3 vytvořte shluky pomocí algoritmu k-means (k-průměrů) a PAM (k-medoidů) pro počet shluků k=2 a 3. Jako počáteční reprezentanty shluků pro k=2 zvolte prvky $c_1 = g$ a $c_2 = h$ a pro k=3 zvolte prvky $r_1 = g$, $r_2 = h$ a $r_3 = e$. Určete nové reprezentanty shluků po první a druhé iteraci obou algoritmů pro k=2 a 3.

Příklad 9: Použijte metody k-means, k-medoid a DBSCAN pro nalezení shluků v níže uvedených datech. Pro metody založené na reprezentantech uvažujte k=2 a dvě různé varianty pro počáteční centroidy ($c_1 = b, c_2 = f$ a $c'_1 = c, c'_2 = h$), pro metody založenou na hustotě uvažujte $\epsilon = 2$ a $minPt = 3$. Porovnejte získané výsledky a rozhodněte, které ze shluků nalezených k-means je "lepší" a proč. Porovnejte lepší výsledek z k-means s výsledkem z DBSCANu a zamyslete se, proč to tak dopadlo.