## **Information Processing and Retrieval**

# Lab 5: Organizing document collections - Pen-and-paper exercises - Solutions

### 2.1 Performing clustering:

 $D_1$ 

	$d_1$	$d_2$	$d_3$	$d_4$
$d_1$	0	8	6	4
$d_2$	8	0	2	6
$d_3$	6	2	0	6
$d_4$	4	6	6	0

 $D_2$ 

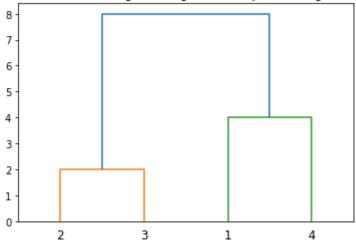
	$d_1$	$(d_2, d_3)$	$d_4$
$d_1$	0	8	4
$(d_2,d_3)$	8	0	6
$d_4$	4	6	0

 $D_3$ 

	$(d_1, d_4)$	$(d_2,d_3)$
$(d_1, d_4)$	0	8
$(d_2,d_3)$	8	0

Dendogram obtained with the complete (maximum) link criterion.

Hierarchical Clustering Dendrogram - Complete linkage criteria



Number of points in node (or index of point if no parenthesis).

#### 2.2.1 External scores:

Purity( $\Omega$ ,C) = 0.75

$$RI = \frac{TP + TN}{TP + FP + FN + TN} = \frac{3}{6} = 0.5$$

#### 2.2.1 Internal scores:

- $s(d_1) = 0.43$
- $s(d_2) = 0.71$
- $s(d_3) = 0.67$
- $s(d_4) = 0.33$
- $s(+) = avg(d_1, d_4) = 0.38$
- $s(-) = avg(d_2, d_3) = 0.69$
- s(o) = avg(s(+),s(-)) = 0.54