An Example of Text Clustering with HAC

Dell Zhang
16/03/2012

Document Collection
d_1, d_2, d_3, d_4, d_5 .

Similarity Matrix

d_1					
d_2	0.6				
d_3	0.5	0.7			
d_4	0.4	0.6	0.6		
d_5	0.8	0.4	0.5	0.9	
	d_1	d_2	d_3	d_4	d_5

// Single-Link HAC

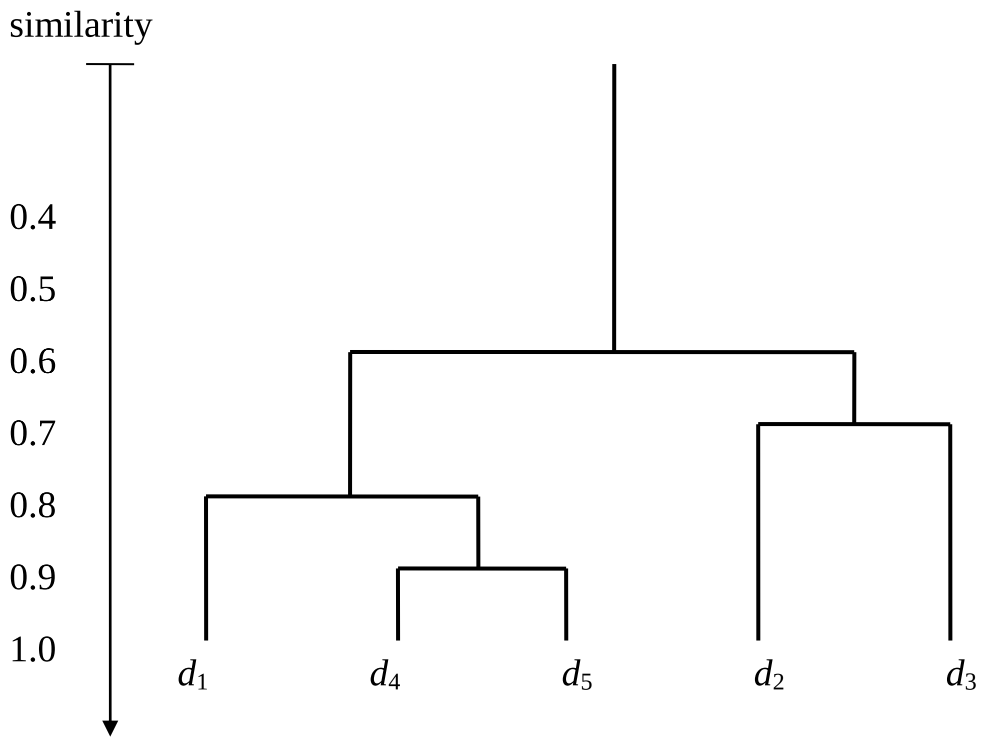
Step 1:
$\{d_1\}$ $\{d_2\}$ $\{d_3\}$ $\{d_4\}$ $\{d_5\}$

Step 2: maximum similarity = $sim(\{d_4\}, \{d_5\}) = 0.9$
$\{d_1\}$ $\{d_2\}$ $\{d_3\}$ $\{d_4, d_5\}$

Step 3: maximum similarity = $sim(\{d_1\}, \{d_4, d_5\}) = sim(d_1, d_5) = 0.8$
$\{d_1, d_4, d_5\}$ $\{d_2\}$ $\{d_3\}$

Step 4: maximum similarity = $sim(\{d_2\}, \{d_3\}) = 0.7$
$\{d_1, d_4, d_5\}$ $\{d_2, d_3\}$

Step 5: maximum similarity = $sim(\{d_1, d_4, d_5\}, \{d_2, d_3\}) = 0.6$
$\{d_1, d_4, d_5, d_2, d_3\}$



// Complete-Link HAC

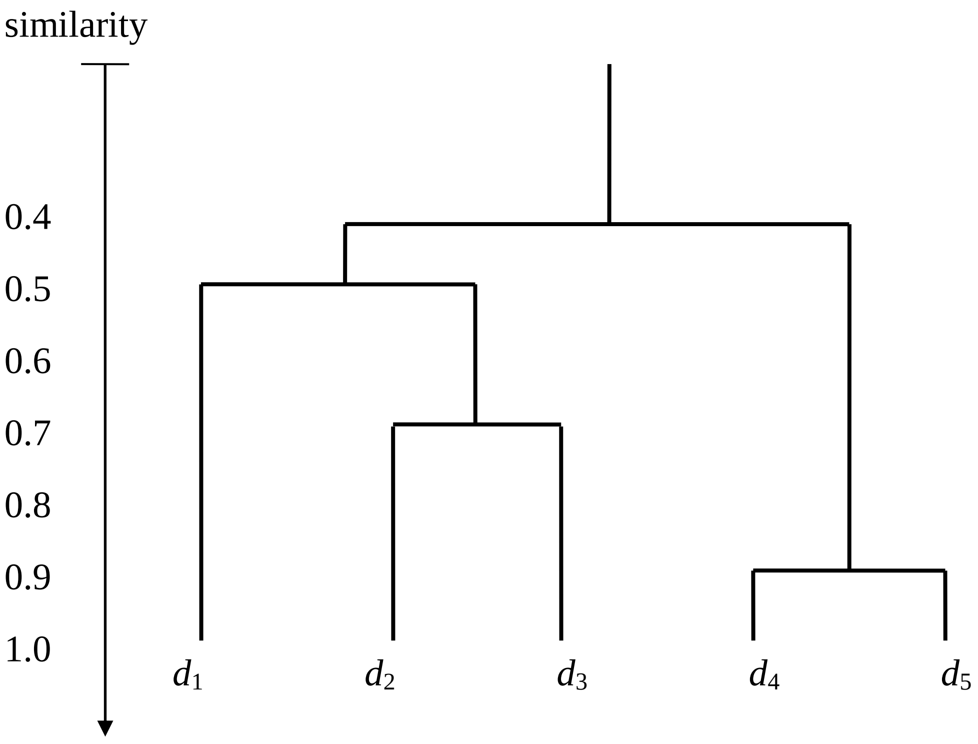
Step 1:
$\{d_1\}$ $\{d_2\}$ $\{d_3\}$ $\{d_4\}$ $\{d_5\}$

Step 2: maximum similarity = $sim(\{d_4\}, \{d_5\}) = 0.9$
$\{d_1\}$ $\{d_2\}$ $\{d_3\}$ $\{d_4, d_5\}$

Step 3: maximum similarity = $sim(\{d_2\}, \{d_3\}) = 0.7$
$\{d_1\}$ $\{d_2, d_3\}$ $\{d_4, d_5\}$

Step 4: maximum similarity = $sim(\{d_1\}, \{d_2, d_3\}) = sim(d_1, d_3) = 0.5$
$\{d_1, d_2, d_3\}$ $\{d_4, d_5\}$

Step 5: maximum similarity = $sim(\{d_1, d_2, d_3\}, \{d_4, d_5\}) = 0.4$
$\{d_1, d_2, d_3, d_4, d_5\}$


similarity
0.4
0.5
0.6
0.7
0.8
0.9
1.0
d_1
d_2
d_3
d_4
d_5