Machine Learning: Homework 8

Laurent HAYEZ

December 4, 2016

Exercise 1. *We have the following (fictional) language similarity matrix:*

	Czech	*Polish*	*Russian*	*English*	*Danish*	*Swedish*
Czech		*0.85*	*0.7*	*0.3*	*0.25*	*0.2*
Polish	*0.85*		*0.4*	*0.25*	*0.7*	*0.8*
Russian	*0.7*	*0.4*		*0.3*	*0.1*	*0.2*
English	*0.3*	*0.25*	*0.3*		*0.75*	*0.8*
Danish	*0.25*	*0.7*	*0.1*	*0.75*		*0.95*
Swedish	*0.2*	*0.8*	*0.2*	*0.8*	*0.95*	

Do clustering (by hand) using two different techniques, once a complete link agglomerative clustering, and once a single link agglomerative clustering. Illustrate the intermediate steps and draw the final dendrograms.

Solution. 1. We start with the complete link method.

	Czech	Polish	Russian	English	Danish	Swedish
Czech		0.85	0.7	0.3	0.25	0.2
Polish	0.85		0.4	0.25	0.7	0.8
Russian	0.7	0.4		0.3	0.1	0.2
English	0.3	0.25	0.3		0.75	0.8
Danish	0.25	0.7	0.1	0.75		0.95
Swedish	0.2	0.8	0.2	0.8	0.95	

We construct the next table according to the following rule

$$\text{sim}(c_i, c_j) = \min_{x \in c_i,\ y \in c_j} \{\text{sim}(x, y)\}.$$

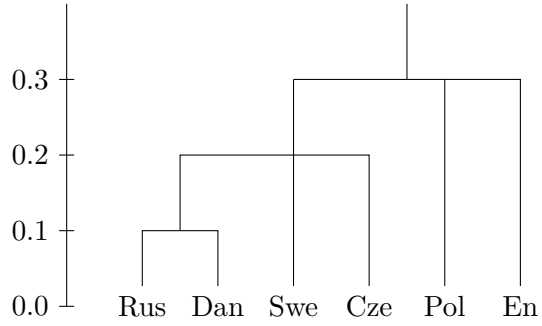
	Czech	Polish	Russian, Danish	English	Swedish
Czech		0.85	0.25	0.3	0.2
Polish	0.85		0.4	0.25	0.8
Russian, Danish	0.25	0.4		0.3	0.2
English	0.3	0.25	0.3		0.8
Swedish	0.2	0.8	0.2	0.8	

	Czech	Polish	Russian, Danish, Swedish	English
Czech		0.85	0.2	0.3
Polish	0.85		0.4	0.25
Russian, Danish, Swedish	0.2	0.4		0.3
English	0.3	0.25	0.3	

	Polish	Rus, Dan, Swe, Cze	English
Polish		0.4	0.25
Rus, Dan, Swe, Cze	0.4		0.3
English	0.25	0.3	

	Pol, En	Rus, Dan, Swe, Cze
Pol, En		0.3
Rus, Dan, Swe, Cze	0.3	

The resulting dendrogram is the following.



2. We do the same but with the single link.

	Czech	Polish	Russian	English	Danish	Swedish
Czech		0.85	0.7	0.3	0.25	0.2
Polish	0.85		0.4	0.25	0.7	0.8
Russian	0.7	0.4		0.3	0.1	0.2
English	0.3	0.25	0.3		0.75	0.8
Danish	0.25	0.7	0.1	0.75		0.95
Swedish	0.2	0.8	0.2	0.8	0.95	

We construct the next table according to the following rule

$$\text{sim}(c_i, c_j) = \max_{x \in c_i,\ y \in c_j} \{\text{sim}(x, y)\}.$$

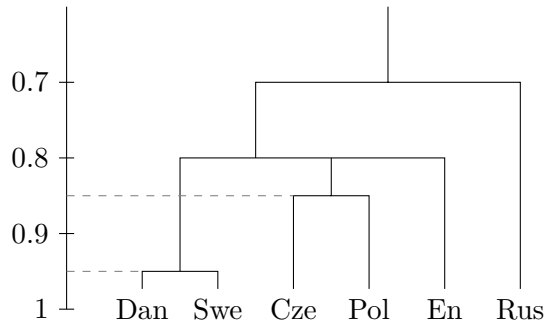
	Czech	Polish	Russian	English	Dan, Swe
Czech		0.85	0.7	0.3	0.25
Polish	0.85		0.4	0.25	0.8
Russian	0.7	0.4		0.3	0.2
English	0.3	0.25	0.3		0.8
Dan, Swe	0.8	0.8	0.2	0.8	

	Cze, Pol	Russian	English	Dan, Swe
Cze, Pol		0.7	0.3	0.8
Russian	0.7		0.3	0.2
English	0.3	0.3		0.8
Dan, Swe	0.8	0.2	0.8	

	Cze, Pol, Dan, Swe	Russian	English
Cze, Pol, Dan, Swe		0.7	0.8
Russian	0.7		0.3
English	0.8	0.3	

	Cze, Pol, Dan, Swe, En	Russian
Cze, Pol, Dan, Swe, En		0.7
Russian	0.7	

The resulting dendrogram is the following.



□

Exercise 2. *The dataset Canton.txt shows the votes per canton (as percentage of yes) for all federal votes during the last three years (including topic, date, and vote number). Transform it to arff and then visualize the cluster tree using WEKA. Cluster all the cantons together with the average link hierarchical clustering applying a Manhattan distance without normalization. We want the distance between clusters to be interpreted as branch length and make sure leafs have the names of the cantons.*

Submit the arff file, visualization, and WEKA output.

Solution. Command used in WEKA:

```
weka.clusterers.HierarchicalClusterer -N 1 -L AVERAGE -P -B -A "weka.core.ManhattanDistance -R first-last"
```

□