Problem Set 09

1. Cluster (5p)

We have the following (fictional) language similarity matrix:

	Czech	Polish	Russian	English	Danish	Swedish
Czech		0.85	0.70	0.30	0.25	0.20
Polish	0.85		0.40	0.25	0.70	0.80
Russian	0.70	0.40		0.30	0.10	0.20
English	0.30	0.25	0.30		0.75	0.80
Danish	0.25	0.70	0.10	0.75		0.95
Swedish	0.20	0.80	0.20	0.80	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.

2. Votes (5p)

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.

Deadline:

December 5, 2016 at 8:00 AM