T-61.5060 Algorithmic methods in data mining

Exercises November 15, 2007

1. Let U be the set of attributes, and $U_{AB} = U \setminus \{A, B\}$ Consider the probe distance between A and B defined as
$$d_P(A, B) = \sum_{C \in U_{AB}} |Pr(C|A = 1) - Pr(C|B = 1)|.$$
Construct a small example dataset showing that d_P is not simply a function of the Jaccard distance of A and B.

2. Study how you would implement a similarity measure based on the ideas
"Two words are similar, if they occur in similar sentences." and
"Two sentences are similar, if they contain similar words."
Assume sentences are rows and words are columns, and that the order of the rows is ignored. (I.e., the data is a 0-1 matrix.)

3. Go through the proof of "clustering aggregation is a subcase of correlation clustering".