Tutorial 1

1. There are a number of ways to compare two objects (vectors) **x** and **y** that consist of *n* binary attributes. The comparison of two such vectors leads to the following four quantities:

   f_{00}= the number of attributes with a value of 0 in both **x** and **y**.
   f_{01}= the number of attributes with a value of 0 in **x** and 1 in **y**.
   f_{10}= the number of attributes with a value of 1 in **x** and 0 in **y**.
   f_{11}= the number of attributes with a value of 1 in both **x** and **y**.

   Based on these quantities, we can define the following two measures:
   Simple Matching Coefficient (SMC):

   $$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

   Jaccard coefficient

   $$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

   a. Calculate the value of the Simple Matching Coefficient and the Jaccard coefficient for the two vectors x=(1,0,0,0,0,1,1,0) and y=(0,0,1,0,1,0,1,0).
   b. What is the main difference between these two measures?

2. The cosine similarity for two vectors **x** and **y** with continuous attributes is defined as follows:

   $$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

   where · indicates the dot product between two vectors, $\sum_{k=1}^{n} x_k y_k$ (x_k and y_k are the *k*-th attributes of **x** and **y** respectively), and ||**x**|| is the length of vector **x**,

   $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x}\cdot\mathbf{x}}$

   a. Calculate the value of the cosine similarity for the two vectors **x** = (3,5,0,1,0,1) and **y** = (2,6,0,2,3,0).

b. If two vectors have a cosine similarity of 1, are they identical?
c. What is the geometric interpretation of the cosine similarity?

3. We consider the problem of document data analysis. Let t_{ij} be the frequency of the i-th word (term) in the j-th document and m be the number of documents. Consider the variable transformation defined by

$$t'_{ij} = t_{ij} \log \frac{m}{n_i}$$

where n_i is the number of documents in which the i-th term appears.

a. What is the effect of this transformation if a term occurs in one document? In every document?
b. What is the purpose of this transformation?