Tutorial 1: Retrieval Models
(SEMESTER II: 2001-2002)

CS4241 Multimedia Information Systems

1. We have learnt about the *Association Coefficients* for features that can take binary values i.e. they can have two states (0 and 1). For this case, we can define similarity measures such as Sokal & Mitchener:

$$d(X,Y) \;=\; \frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$$

   Now suppose that features can have *three* states i.e. they can assume three different values.

   (a) Explain how would you extend the notion of Association Coefficients for this case.

   (b) Then state the Sokal & Mitchener coefficient for the 3-state case.

2. Assume it is given that a document space consists of 4 unique terms. Let:

$$D_1 = (2,3,5,1) \qquad D_2 = (3,6,2,5)$$

$$D_3 = (1,2,5,5) \qquad Q = (0,1,1,0)$$

   If the total number of documents is 1000 and the document frequencies of the terms are $(30,50,100,50)$, compute:

   - Document-query similarities on the term-independence assumption.
   - Document-query similarities if we know that term 3 and term 4 belong to the same class (i.e. $t_3 \cdot t_4 = 1$).

   Use the vector-space model with tf-idf weights.

3. Explain how the 0.5 formula for probabilistic retrieval is obtained.

4. Query-based relevance feedback process is based on the following iterative procedure:

$$Q^{(i+1)} \;=\; Q^{(i)} \;+\; \alpha \sum_{D_i \in R} D_i \;-\; \beta \sum_{D_j \in NR} D_j$$

   Explain:

   - why most experiments show that the information contained in relevant documents is more valuable than those in the non-relevant documents for the feedback process?
   - why using only the highest-ranked non-relevant document has been found to be effective.