Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
April 15, 2011
Due date: Friday, April 27, 2012; noon time: leave a hardcopy in my mailbox.

Notes: Handwritten answers are not acceptable.

1. Consider the following document by term D matrix. Calculate the TDV of t1 and t2 using the space density and cover coefficient concepts. In both cases use the approximate methods. For the space density method use the similarity measure of your own choice.

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

2. Consider the following binary string:
0100110100110110111

a. Draw the PAT tree for the first 9 sistrings (show some of the steps so that it can be followed). Also obtain the corresponding PAT array.

b. In a PAT tree environment how can we efficiently implement a simple pattern search? Briefly explain.

c. In a PAT tree environment how can we efficient implement a search such as <P1> n <P2> where P1 and P2 indicate patterns for two sistrings and n indicates the maximum distance between the beginning positions of them.

3. Consider the paper " Automatic ranking of information retrieval systems using data fusion" by Nuray and Can article (*Information Processing and Management*, 2006).

a. What is the intuition behind bias? Does it make sense? Explain your answer and suggest a possible improvement in that method.

b. Consider four different information retrieval systems (A, B, C, D) ranking documents a,… f.
A= (c, a, b, d)
B= (b, c, a, e)
C= (a, c, b, f)
D= (a, b = c) // b and c are assigned the same rank!

Perform data fusion by using the reciprocal rank, Borda count, and Condorcet methods. Please show your steps.

4. Consider the paper "A vector space model for automatic indexing" by Salton, Wong, and Yang (*Comm. of the ACM*, Nov. 1975).

According this study what type of words (in terms of their document frequency) are identified as good, bad, and indifferent discriminators? Do the results match the intuition behind the idf component of tf.idf? Please explain.

What do they mean right-to-left and left-to-right transformation? What do they try to achieve by doing these operations?

5. Consider the paper "Information retrieval on Turkish texts" by our research group - Bilkent IR group- (published in *Journal of the American Society for Information Science and Technology* -JASIST-, 2008). Which of the matching function(s) defined in the paper would you prefer in a dynamic environment that involves document additions and deletions? Explain why.

6. Study the paper "Document ranking and the vector-space model" by Dik Lun Lee et al. (*IEEE Software*, 1997). Briefly explain the query feedback method that you like most and discuss its applicability in the web environment.

7. How can we combine traditional term weighting approaches (as defined by Salton and Buckley in their 1988 *Information Processing and Management* article) with modern page ranking approaches such as Google's PageRank algorithm?

8. What is meant by search engine optimization (SEO)? How can we modify the Google's PageRank algorithm to avoid the effects of SEO?