Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 4
April 18, 2011
Due date: May 2, 2011; Monday, by 5:00 pm

Notes: Handwritten answers are not acceptable. Answer the questions in the order given here.

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.

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

2. Using the ACM digital library find two articles by G. Salton (possibly with his co-workers) where he discusses the use of term discrimination values for term indexing. What does he suggest for the use TDV to improve recall and precision?

3. Create a corpus of at least 15 words (i.e., a collection of 15 words) and find the stems of two words of this corpus by using the successor variety method. Make sure that you have a reasonable corpus so that it will lead you to a reasonable direction in finding stems of these two words.

What do you think about the effectiveness of the successor variety method in general? What are the possible requirements of its success?

4. Consider the study by our research group on Turkish text retrieval (JASIST, 2008). Briefly describe the stemming methods used in this study. The study indicates that a simple stemming approach such as simple word truncation works as good as more complicated methods during information retrieval. What is the reason for this?

According to the same study what is the impact of the use of stopwords during indexing on information retrieval effectiveness? Explain why does it have such an effect (no effect)?

5. TDT (topic detection and tracking) questions.

a. What are the five main research topics studied by the TDT research initiative?

b. Consider the study by our research group on new event detection (NED) and topic tracking (TT) on Turkish news (JASIST, 2010). On which study the experimental results were more successful (new event detection or topic tracking)? Can you explain why?

c. What is the effect of the use of stopword list and stemming on effectiveness (any effect)? Are the effects of wordstopping and stemming the same as those of information retrieval? What can be the reasons of differences/no-differences, please explain briefly.

6. Find an example of using Monte Carlo method in performance evaluation. Make sure that it is different from the ones we studied in the classroom. Your examples may or may not be related to information retrieval activities.

7. Consider the Nuray & Can IPM article (Automatic ranking of information retrieval systems using data fusion, 2006).

a. Explain the method used for automatic ranking in that article.

b. Suggest your own method of automatic ranking: please try to be creative.

c. Consider four different information retrieval systems (A, B, C, D) ranking documents a,… f.
A= (a, c, b, d)
B= (b, c, a, e)
C= (c, a, 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.

8. Consider the SIGIR 2007 Athena Lecturer Award presentation of Karen Sparck Jones "Natural language and the information layer" (you may also want to watch its movie version from ACM digital library) and "Seven ages of information retrieval" of Michael Lesk. How do these two papers complement each other? What can you say about the future of information retrieval systems based on these two papers and your observations etc. Please write a short essay on these issues.

9. Consider the original paper Brin and Page paper (The anatomy of a large-scale hypertextual web search engine) and explain the major data structures used in this algorithm.

In your opinion which of the data structures is most important and what happens if we delete that part from the algorithm?

10. Consider the Carpineto, Osinski, Romano, and Weiss paper "A survey of web clustering engines." What is meant by cluster labeling, what is its purpose?

If you are asked to implement an algorithm for this purpose which of the algorithms described in the paper would you choose? Does it need any modifications if so how would you change it? Explain why.