Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 1
February 21, 2014
Due date: March 10, 2014; Monday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable.

1. Consider the following search results for two queries Q1 and Q2 (the documents are ranked in the given order, the relevant documents are shown in bold).

Q1: D1, **D2**, D3, **D4**, D5, **D6**, D7, D8, D9, D10.

Q2: **D1**, D2, **D3**, D4, **D5**, D6, D7, D8, **D9**, and D10.

For Q1 and Q2 the total number of relevant documents are, respectively, 15 and 4 (in Q1 12 of the relevant documents are not retrieved).

a. Using the TREC interpolation rule, in a table give the precision value for the 11 standard recall levels 0.0, 0.1, 0.2, … 1.0. Please also draw the corresponding recall-precision graph as shown in the first figure of TREC-6 Appendix A (its link is available on the course web site).

Hint. "Interpolated" means that, for example, precision at recall 0.10 (i.e., after 10% of rel docs for a query have been retrieved) is taken to be MAXIMUM of precision at all recall points >= 0.10. Values are averaged over all queries (for each of the 11 recall levels). These values are used for Recall-Precision graphs. (This paragraph is taken from: http://ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm.)

Please do this for each query separately and obtain one table for both queries using the average of two values at each recall point.

b. Find R-Precision (TREC-6 Appendix A for definition) for Query1 and Query2.

c. Find MAP for these queries. If it is not possible to calculate explain why.

2. Consider document-based partitioning and term-based partitioning approaches as define in the Zobel-Moffat paper "Inverted files for text search engines" (*ACM Computing Surveys*, 2006). Please also consider the following document by term binary D matrix for m= 6 documents (rows), n= 6 terms (columns).
Describe a two ways of indexing across a cluster of machines.

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

Under which condition you would prefer one approach over the other one? Please briefly explain.

3. In this part again consider the Zobel-Moffat paper.

a. Understand the skipping concept as applied to the inverted index construction.

Assume that we have the following posting list for term a: <1, 5> <4, 1> <9, 3> <10, 4> <12, 4> <17, 4> <18, 3>, <22, 2> <24, 4> <33, 4> <38, 5> <43, 5> <55, 3><64, 2> <68, 4> <72, 5> <75, 1> <88, 2>.. The posting list indicates that term-a appears in d1 five times and in d3 twice, etc.

Assume that we have the following posting list for term-b: <12, 3> <40, 2> <66, 1>.

Consider the following conjunctive Boolean query: term-a **and** term-b. If no skipping is used how many comparisons do you have to find the intersection of these two lists?

Introduce a skip structure, draw the corresponding figure then give the number of comparisons involved to process the same query.

b. Give a posting list of term-a (above it is given in standard sorted by document number order) in the following forms: 1), a) ordered by $f_{d,t}$, b) ordered by frequency information in prefix form.

Which method would you prefer and why?

4. What is meant by Cranfield approach to testing in information retrieval?

5. Is pooling a reliable approach for the construction of test collections? Find the paper by Zobel regarding this issue, it may help.

What is bpref? Is it anyway related to the pooling concept?

6. Please study the paper "Data stream clustering," by Silva et al. *ACM Computing Survey*, 2013.

a. Describe the similarities and differences of the data stream and traditional information retrieval environments.
b. What is the concept of time window, what are the advantages of using time window?
c. What is meant by abstraction?
d. Choose a data structure defined in the data abstraction part and explain its components?
e. In data stream clustering how do we use the standard clustering algorithms such as k-means?

7. Study the paper "A survey of Web clustering search engines" by Carpineto et al. *ACM Computing Survey*, 2009. Find two of the web search engines defined in the paper and compare their user interface, performance in general, by conduction a simple set experiments. Please also define your experiments. If the engines mentioned in the paper do not exist find other similar search engines on the web. Note that you are not writing a research you are trying to convince people with IR knowledge. Try to be more convincing by providing some quantitative observations.

8. On data forms used in clustering:

a. What is nominal data? Give two examples.
b. What is the type of data (D matrix) we use in document clustering (nominal etc.)?
c. Can we use nominal data in clustering algorithm we have in IR? If your answer is no how can we convert nominal data into a meaningful form so that they can be used for clustering? Do some research on this on the Web. State your references.