Problem Set 2

Problem 1. Give an example that illustrates how the *term-wise grouping* technique (Slide 14, lecture 4) may fail to accurately retrieve the "top k" documents for a query.

Problem 2. Give an example that illustrates how the *sampling and pre-grouping* technique (Slide 16, lecture 4) may fail to accurately retrieve the "top k" documents for a query.

Problem 3. Work out the exercise described on slide 38 of lecture 3.

Problem 4. This problem deals with the vector space model (using TF-IDF weights) as applied to the following collection of 4 documents:

Doc 1 : Information Retrieval Systems
Doc 2 : Information Storage
Doc 3 : Digital Speech Synthesis Systems
Doc 4 : Speech Filtering, Speech Retrieval

(i) Compute all non-zero entries in the normalized vector for Doc 1.

(ii) Rank all the documents in the collection for the query "Speech Systems"?

(iii) Compute the cosine similarities between (a) docs 1 and 2 (b) docs 3 and 4.

Problem 5. This problem will try to give more insight into LSI. Consider the equation from Lecture 4, Slide 33:

$$A_s = PQ_sR^T$$

(i) What are the only useful columns of P, and the useful rows of R^T (equivalently, the useful columns of R) (Consider PQ_s and Q_sR^T in turn) ?

(ii) What is special about the vector space spanned by these columns of P?

(iii) The columns of P are orthonormal. Using this, and the answer to part (ii), what does each column of Q_sR^T represent?

(iv) Inner products (and thus cosine) are invariant under a change of orthonormal bases. Knowing this, and the answers to the above, what is an efficient way to calculate the cosine similarity between two documents d_i and d_j?

(v) Let $B = Q_sR^T$. Note that B is an $s \times n$ matrix, with $s << m$. Show that the document-document similarity matrix $A_s^TA_s$ is also given by B^TB.