Solutions to 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.

Answer. Consider a collection with the following normalized document vectors:

$$\begin{aligned} d_1 &= (0.6, 0, \ldots) \\ d_2 &= (0.5, 0, \ldots) \\ d_3 &= (0, 0.5, \ldots) \\ d_4 &= (0, 0.4, \ldots) \\ d_5 &= (0.4, 0.3, \ldots) \end{aligned}$$

Let $t_1 \ldots t_m$ be the set of terms in the collection and assume that the i^{th} component of each vector corresponds to term t_i. Let us apply the term-wise grouping technique with $k = 2$ for the query q = "t_1 t_2". Clearly, the preferred list for t_1 is $\{d_1, d_2\}$ and for t_2 is $\{d_3, d_4\}$. Hence, the candidate set for the query is $\{d_1, d_2, d_3, d_4\}$. Now, since t_1 and t_2 each occur in 3 documents, their IDFs are the same and the normalized query vector is therefore $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Computing cosines, we see that $cosine(q, d_1) = \frac{0.6}{\sqrt{2}}$, $cosine(q, d_2) = cosine(q, d_3) = \frac{0.5}{\sqrt{2}}$, $cosine(q, d_4) = \frac{0.4}{\sqrt{2}}$, and $cosine(q, d_5) = \frac{0.7}{\sqrt{2}}$. Therefore, d_5 is the top-ranked document for query q but is outside the candidate set specified by the term-wise grouping technique.

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.

Answer. There are of course many possible answers. One example (2-d case): assume the documents of the corpus are uniformly distributed on the first quadrant of the unit circle. If the chosen leaders are (0,1) and (1,0), then query points lying near (.707,.707) will lead to poor results.

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

Answer. In decreasing order of cosine similarity:

- Docs that have many rare words in common
- Docs that have only frequent words in common
- Docs that have no words in common

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.

Answer. The table below summarizes the relevant TF and IDF information for the entire collection. For all computations below, we shall use log base 2.

		TF			
Term	IDF	Doc 1	Doc 2	Doc 3	Doc 4
Digital	log 4	0	0	$\frac{1}{4}$	0
Filtering	log 4	0	0	0	$\frac{1}{4}$
Information	log 2	$\frac{1}{3}$	$\frac{1}{2}$	0	0
Retrieval	log 2	$\frac{1}{3}$	0	0	$\frac{1}{4}$
Speech	log 2	0	0	$\frac{1}{4}$	$\frac{1}{2}$
Storage	log 4	0	$\frac{1}{2}$	0	0
Synthesis	log 4	0	0	$\frac{1}{4}$	0
Systems	log 2	$\frac{1}{3}$	0	$\frac{1}{4}$	0

(i) Multiplying the IDF column with the TF column for Doc 1 yields the vector $(0, 0, \frac{1}{3}, \frac{1}{3}, 0, 0, 0, \frac{1}{3})$ which, upon normalization results in $(0, 0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, 0, 0, \frac{1}{\sqrt{3}})$.

(ii) Treating the query "Speech Systems" as a document, the normalized query vector computes to $(0, 0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0, \frac{1}{\sqrt{2}})$. Therefore, $cosine(q, doc1) = \frac{1}{\sqrt{6}}$, $cosine(q, doc2) = 0$, $cosine(q, doc3) = \frac{1}{\sqrt{5}}$, and $cosine(q, doc4) = \sqrt{\frac{2}{9}}$. Hence, the documents in decreasing order of rank are: Doc 4, Doc 3, Doc 1, Doc 2.

(iii) (a) $\frac{1}{\sqrt{15}}$ (b) $\frac{1}{3}\sqrt{\frac{2}{5}}$

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) ?

Answer: The first s columns of P, and the first s rows of R^T (i.e., the first s columns of R) are the only ones relevant to the product.

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

Answer: The first s columns of P span the same vector space as the n columns of A_s. The first s columns of P thus form an orthonormal basis for the document vectors represented by the columns of A_s.

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

Answer: A little thought will show that the column j of A_s is given by the product of P, Q_s, and the j'th column of R^T. Each column of $Q_s R^T$ represents a document vector, in an s-dimensional vector space with coordinates in terms of the basis given by the first s columns of P.

(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?

Answer: Take cosine between the s-dimensional vectors given by the i'th and j'th columns of $Q_s R^T$, instead of between the corresponding m-dimensional columns of A_s,

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

Answer:

$$A_s^T A_s = (PQ_s R^T)^T (PQ_s R^T) = RQ_s P^T PQ_s R^T = RQ_s Q_s R^T = (RQ_s)(Q_s R^T) = B^T B$$

Note: you will not be responsible for knowing all the linear algebra used in this problem, but you should understand the "big picture".