Birkbeck
(University of London)

MSc Examination for Internal Students

School of Computer Science and Information Systems

Information Retrieval and Organisation (COIY064H7)
Credit Value: 15

Date of Examination: Thursday 4 June 2009
Duration of Paper: 10:00 - 12:00

RUBRIC

1. *This paper contains 13 questions for a total of 100 marks.*
2. *Students should attempt to answer* **all** *of them.*
3. *This paper is not prior-disclosed.*
4. *The use of non-programmable electronic calculators is permitted.*

1. (3 marks)

What is the IDF of a term that occurs in every document? Compare this with the use of stop word lists.

2. (3 marks)

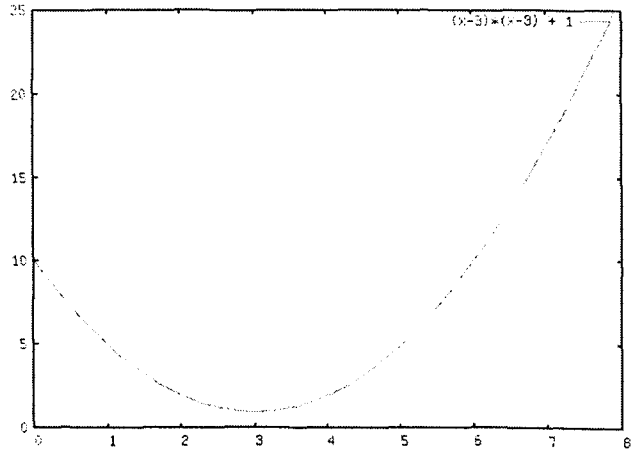
What is the *cluster hypothesis* in Information Retrieval?

3. (6 marks)

Consider the function

$$f(n) = (n-3)^2 + 1$$

which is plotted below.



(a) Would $f(n)$ be a valid function for term frequencies in a TF-IDF ranking? Briefly explain your answer. (3 marks)

(b) Would $f(n)$ be a valid function for inverse document frequencies in a TF-IDF ranking? Briefly explain your answer. (3 marks)

4. (8 marks)

For tolerant retrieval, the following inverted file with 3-grams has been constructed:

directory	postings lists
$$s	1,2,3,4
$se	2,3
a$$	3
e$$	1
ea$	3
ell	2,4
he$	1
hel	4
lls	2,4
ls$	2,4
s$$	2,4
sea	3
sel	2
she	1,4

Assume that a user searches for the misspelt word `solls`.

(a) Break down `solls` into its 3-grams. (3 marks)

(b) Assuming that at most one error is allowed, which documents have to be fetched to check if they contain words similar to `solls` and why? (5 marks)

5. (7 marks)

Consider a fictitious document collection that contains the following 2 documents.

d_1: The profit is down.
d_2: The profit decreases further.

Suppose the query q is 'profit down'. Show how the above documents should be ranked for q, using an unigram language model that mixes the distributions estimated from the specific document and the entire collection with equal weights.

6. (8 marks)

Suppose the pair-wise similarity table for a document collection $\{d_1, d_2, d_3, d_4, d_5, d_6\}$ is as follows.

	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1.0	0.8	0.5	0.4	0.4	0.4
d_2	0.8	1.0	0.4	0.5	0.5	0.5
d_3	0.5	0.4	1.0	0.8	0.8	0.7
d_4	0.4	0.5	0.8	1.0	0.9	0.5
d_5	0.4	0.5	0.8	0.9	1.0	0.7
d_6	0.4	0.5	0.7	0.5	0.7	1.0

Draw the dendrogram generated by the Single-Link Hierarchical Agglomerative Clustering (HAC) algorithm. Assume that some documents in this collection are about economics and the others are about finance. How should the dendrogram be cut to cluster the documents?

7. (10 marks)

Consider a fictitious document collection. The 6 documents in this collection are represented as 6 points in a two-dimensional vector space as follows.

$\vec{d_1}$: (1,1)	$\vec{d_2}$: (2,1)	$\vec{d_3}$: (4,1)
$\vec{d_4}$: (1,2)	$\vec{d_5}$: (2,2)	$\vec{d_6}$: (4,2)

Suppose that the distance between a pair of documents is measured by the Euclidean distance between their corresponding points. Show how the k-means algorithm (with $k = 2$) clusters these documents, using $\vec{d_2}$ and $\vec{d_3}$ as seeds first, and then using $\vec{d_2}$ and $\vec{d_5}$.

8. (16 marks)

Consider the following document collection consisting of the four documents

Doc 1: Mary had a little lamb
Doc 2: little lamb, little lamb
Doc 3: and everywhere that Mary went
Doc 4: Mary went, Mary went

having a vocabulary consisting of the following terms:

a
and
everywhere
had
lamb
little
Mary
that
went

(a) Assume that we use TF=$1 + \log_2(tf_{t,d})$ for computing the term frequency of term t in document d. ($\log_2$ is the binary logarithm, i.e. $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(4) = 2$, and so on.) If we only apply the term frequency TF to document vectors, what do the vectors for documents 1, 2, 3, and 4 look like? (8 marks)

(b) Assume that we use IDF=$\log_2(\frac{N}{df_t})$ for computing the inverse document frequency of a term t. If we only apply IDF to query vectors, what do the query vectors for the query "`lamb, little`" and the query "`everywhere`" look like? (4 marks)

(c) Given the query vector $\vec{q}$ and the document vector $\vec{d_5}$ defined as follows

$$\vec{q} = \begin{pmatrix} 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \end{pmatrix} \qquad \vec{d_5} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \\ 6 \\ 2 \\ 10 \\ 0 \\ 0 \end{pmatrix}$$

compute the cosine-similarity. Assume that TF-IDF has already been applied to the vectors, but that normalisation still has to be done. (4 marks)

9. (4 marks)

Give two reasons why relevance feedback has been little used in web search.

10. (8 marks)

When merging postings lists while processing conjunctive queries of the form t_1 AND t_2 AND ... AND t_n, the processing can be sped up by merging the postings lists of the terms in ascending order of their lengths.

(a) Could a similar technique be used to optimise the query processing of queries that have the form (t_1 OR t_2) AND (t_3 OR t_4) AND ... AND (t_{n-1} OR t_n)? If yes, how would you estimate the size of the result of (t_i OR t_{i+1})? If no, briefly explain your answer. (4 marks)

(b) Can a query of the form t_1 AND NOT t_2 be processed in linear time? If yes, explain under which assumptions this is possible. If no, briefly justify your answer. (4 marks)

11. (12 marks)

Consider the following collection of documents that belong to two classes: Healthy (H) and Unhealthy (U).

	docID	docText	class
TRAINING	d_1	sugar salt fruit	H
	d_2	sugar salt	U
	d_3	sugar	U
	d_4	sugar sugar salt salt	U
	d_5	sugar sugar salt fruit	H
TEST	d_6	fruit salt salt	?

Show how the Naive Bayes algorithm (with Laplace smoothing) can be used to train a classifier and predict the class of test document.

12. (5 marks)

What are the assumptions that make the Naive Bayes algorithm computationally tractable? How can the Naive Bayes algorithm be effective when it relies on such oversimplifying assumptions?

13. (10 marks)

(a) When compressing gaps in a postings list, we can use a trick to store gaps more efficiently, since there are no gaps of size 0. This can also be applied to VB-encoding. How can you do this and what is the largest gap you can encode in one byte in VB-code? (4 marks)

(b) This trick can be taken further in VB-encoding as high-order bytes also have to be larger than 0 (otherwise we would not use them). What is the largest gap you can encode in two bytes in VB-code? (6 marks)