Birkbeck
(University of London)

MSc Examination for Internal Students

Dept of Computer Science and Information Systems

Information Retrieval and Organisation
(COIY064H7: 15 credits)

Date of Examination: Thursday 14th June 2012
Duration of Paper: 2:30pm – 4:30pm (2 hours)

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)

You want to compress the dictionary terms stored in the leaf pages of a B-tree. Apply front coding to the following list of terms:

- Jalaun
- Jalna
- Jalpaiguri
- Jamaica
- Jamal
- Jamb

2. (7 marks)

You are using a Boolean IR system with an inverted file index containing the following positional information:

cold, 2:	$< 1,1 :< 6 >;\ 4,1 :< 4 >>$
days, 1:	$< 3,1 :< 2 >>$
eat, 1:	$< 6,1 :< 1 >>$
hot, 2:	$< 1,1 :< 3 >;\ 4,1 :< 8 >>$
in, 3:	$< 2,1 :< 3 >;\ 4,2 :< 1,5 >>$
lot, 1:	$< 6,1 :< 3 >>$
nine, 1:	$< 3,1 :< 1 >>$
old, 1:	$< 3,1 :< 3 >>$
pease, 5:	$< 1,2 :< 1,4 >;\ 2,1 :< 1 >;\ 5,2 :< 1,3 >>$
porridge, 5:	$< 1,2 :< 2,5 >;\ 2,1 :< 2 >;\ 5,2 :< 2,4 >>$
pot, 3:	$< 2,1 :< 5 >;\ 4,2 :< 3,7 >>$
the, 4:	$< 2,1 :< 4 >;\ 4,2 :< 2,6 >;\ 6,1 :< 2 >>$

(a) Which of the terms have multiple occurrences within the same document? (3 marks)

(b) Which document IDs would the following query return?
"`NOT` (porridge `/1` hot)" (4 marks)

3. (8 marks)

You are testing the implementation of a minimum editing distance algorithm measuring the Levenshtein distance. The algorithm outputs the following matrix, which contains errors. Identify the cells of the matrix that contain wrong entries and provide the correct entries for these cells.

	“”	l	e	t
“”	0	1	2	4
c	1	1	2	3
a	2	2	2	3
t	3	2	2	1

4. (5 marks)

For a statistical analysis you want to count the number of occurrences of all terms appearing in a document collection, i.e., the output will be a list of all terms together with their frequencies:

a: 1,383,134
abbey: 387
aberdeen: 54
able: 1,987
...

As the document collection is very large, you want to parallelise this computation by using the Map-Reduce approach. Briefly describe the processing steps.

5. (12 marks)

You want to use cluster pruning to decrease the costs for computing the top-K documents.

(a) In a first step you have to compute the distances between document vectors. Using the cosine measure, determine the distance between each pair of the following vectors. Assume that the components of these vectors already include the TF-IDF weighting.

$$d_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 1 \end{pmatrix} \quad d_2 = \begin{pmatrix} 6 \\ 0 \\ 0 \\ 6 \\ 3 \end{pmatrix} \quad d_3 = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

(6 marks)

(b) Given a (different) set of nine documents divide up the document collection into three clusters. Documents d_1, d_2, and d_3 have been chosen as *leaders* of these clusters. Use the following distance matrix to assign the remaining documents to these leaders.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
d_1		0.2	0.4	0.2	0.6	0.8	0.4	0.1	0.2
d_2			0.1	0.5	0.1	0.7	0.6	0.3	0.7
d_3				0.4	0.3	0.5	0.2	0.4	0.4
d_4					0.8	0.2	0.4	0.4	0.3
d_5						0.5	0.8	0.3	0.4
d_6							0.2	0.9	0.3
d_7								0.6	0.7
d_8									0.2
d_9									

(6 marks)

6. (8 marks)

An IR system stores scientific papers and you want to build an index structure for the titles of the papers. This index should be able to answer phrase queries on these titles efficiently.

(a) How could you use the concept of a permuterm index to achieve this?

(6 marks)

(b) What would be the major disadvantage of such an index?

(2 marks)

7. (4 marks)

During the evaluation of two web search engines, A and B, experiments are run to determine the overlap of the indexed web pages. These experiments reveal that 25% of A's pages are also indexed by B and 50% of B's pages are also indexed by A. What is the ratio between the index sizes of A and B?

8. (8 marks)

Assume you are using the Jaccard coefficient to compute the query document score in an IR system. The Jaccard coefficient $J(Q, D)$ is defined as $\frac{Q \cap D}{Q \cup D}$, where Q is the set of terms in the query and D is the set of terms in document D.

Furthermore, assume that a document is relevant if it contains *all* of the query terms, otherwise it is considered irrelevant. Let $Q = \{\text{information}, \text{retrieval}\}$. Is it possible to find two documents, D_1 and D_2, such that D_1 is relevant, D_2 is irrelevant, and $J(Q, D_1) < J(Q, D_2)$? If yes, give an example for D_1 and D_2. If no, briefly explain your answer.

9. (10 marks)

Assume a situation where every document in the test collection has been assigned exactly one class, and that a classifier also assigns exactly one class to each document. This setup is called one-of classification.

(a) In one-of classification, is the total number of false positive decisions always equal to the total number of false negative decisions?
Please briefly explain your answer.

(2 marks)

(b) In one-of classification, is microaveraged precision always equal to microaveraged recall?
Please briefly explain your answer.

(2 marks)

(c) In one-of classification, is microaveraged F_1 always equal to accuracy?
Please briefly explain your answer.

(4 marks)

(d) In one-of classification, is microaveraged F_1 always equal to macroaveraged F_1?
Please briefly explain your answer.

(2 marks)

10. (10 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.

11. (10 marks)

The following table shows the result of flat clustering on a document collection, where each letter "A", "B" or "C" represents a document in the true class "A", "B" or "C" respectively.

cluster 1	A A B C
cluster 2	A B C C
cluster 3	A B B C

(a) What is the purity of the above clustering?

(4 marks)

(b) What is the Rand Index (RI) of the above clustering?

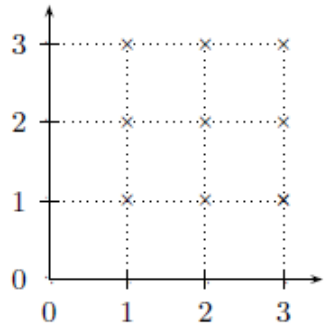
(4 marks)

(c) If we replace each document d in this collection with two identical copies of d from the same class, which clustering performance measure (purity or RI) will stay the same?

(2 marks)

12. (5 marks)

Consider the following 9 points in a two-dimensional vector space representing a collection of 9 documents:
$d_1 = (1, 1)$; $d_2 = (2, 1)$; $d_3 = (3, 1)$;
$d_4 = (1, 2)$; $d_5 = (2, 2)$; $d_6 = (3, 2)$;
$d_7 = (1, 3)$; $d_8 = (2, 3)$; $d_9 = (3, 3)$.



Suppose that the dissimilarity between each pair of documents is measured by Euclidean distance (straight-line distance).

(a) How will the 2-means clustering algorithm group those points, starting from the seeds d_1 and d_2?

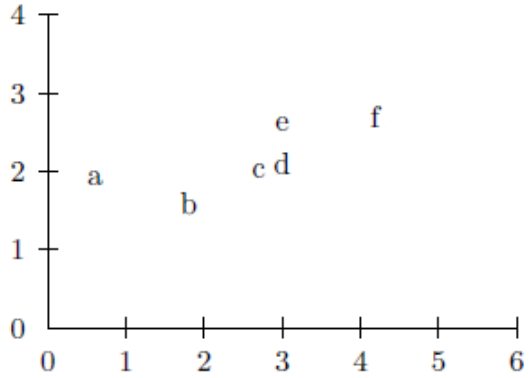
(2 marks)

(b) How will the 3-means clustering algorithm group those points, starting from the seeds d_3, d_6 and d_9?

(3 marks)

13. (10 marks)

Consider the following 6 points in a two-dimensional vector space representing a collection of 6 documents:
a = (0.6, 1.9), b = (1.8, 1.6), c = (2.7, 2.0), d = (3.0, 2.1), e = (3.0, 2.6), f = (4.2, 2.7).



Suppose that the similarity of two points (x_1, y_1) and (x_2, y_2) is defined as $-[(x_1 - x_2)^2 + (y_1 - y_2)^2]$.

(a) Compute single-link clustering of those points and depict the result as a dendrogram.

(5 marks)

(b) Compute complete-link clustering of those points and depict the result as a dendrogram.

(5 marks)