Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

Information Retrieval and Organisation (COIY064H7)

CREDIT VALUE: 15 credits

Date of examination: Wednesday, 12th June 2013
Duration of paper: 10:00am – 12:00noon (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. **(10 marks)**

Shown below is a portion of a positional index in the format:
`term`: $doc_1 : \langle position_1, position_2, \ldots \rangle$; $doc_2 : \langle position_1, position_2, \ldots \rangle$; etc.

`angels`: $2 : \langle 36, 174, 252, 651 \rangle$; $4 : \langle 12, 22, 102, 432 \rangle$; $7 : \langle 17 \rangle$;
`fools`: $2 : \langle 1, 17, 74, 222 \rangle$; $4 : \langle 8, 78, 108, 458 \rangle$; $7 : \langle 3, 13, 23, 193 \rangle$;
`fear`: $2 : \langle 87, 704, 722, 901 \rangle$; $4 : \langle 13, 43, 113, 433 \rangle$; $7 : \langle 18, 328, 528 \rangle$;
`in`: $2 : \langle 3, 37, 76, 444, 851 \rangle$; $4 : \langle 10, 20, 110, 470, 500 \rangle$; $7 : \langle 5, 15, 25, 195 \rangle$;
`rush`: $2 : \langle 2, 66, 194, 321, 702 \rangle$; $4 : \langle 9, 69, 149, 429, 569 \rangle$; $7 : \langle 4, 14, 404 \rangle$;
`to`: $2 : \langle 47, 86, 234, 999 \rangle$; $4 : \langle 14, 24, 774, 944 \rangle$; $7 : \langle 199, 319, 599, 709 \rangle$;
`tread`: $2 : \langle 57, 94, 333 \rangle$; $4 : \langle 15, 35, 155 \rangle$; $7 : \langle 20, 320 \rangle$;
`where`: $2 : \langle 67, 124, 393, 1001 \rangle$; $4 : \langle 11, 41, 101, 421, 431 \rangle$; $7 : \langle 16, 36, 736 \rangle$;

Consider any expression within a pair of double quotation marks as a phrase query, for this question. Which document(s), if any, match each of the following queries at which positions?

(a) "fools rush in" (5 marks)

(b) "fools rush in" AND "angels fear to tread". (5 marks)

2. **(5 marks)**

If you wanted to search for `s*ng` in a permuterm wildcard index, what key(s) would you do the lookup on?

3. **(10 marks)**

Fill the following Levenshtein matrix to compute the edit distance between two strings 'obama' (input) and 'romney' (output).

	''	r	o	m	n	e	y
''							
o							
b							
a							
m							
a							

4. **(5 marks)**

Compute variable byte codes and γ-codes for the postings list 776, 801, 1101, 312513. Use gaps instead of docIDs for all but the first entry.

(a) Give the solution for variable byte codes as a sequence of 8-bit blocks. (3 marks)

(b) Give the solution for the γ-codes of the postings list as a sequence of 4 pairs of bit strings, where the first bit string of each pair corresponds to a length and the second to an offset. (2 marks)

5. **(5 marks)**

Consider the following sequence of γ-coded gaps:
01111000111011111101011110101110111.

(a) What is the sequence of gaps? (3 marks)

(b) What is the sequence of postings? Assume that the first entry is the docID of the first document, and so on. (2 marks)

6. **(5 marks)**

Compute the Jaccard coefficient between the query 'digital phones' and the document 'digital phones and video phones and other phones'. Treat 'and' and 'other' as stop words.

7. **(15 marks)**

The following `R`'s and `N`'s represent relevant (R) and nonrelevant (N) documents respectively in a ranked list of 10 documents retrieved in response to a query from a collection of 10,000 documents. These 10 documents are the complete result set of the system. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 4 relevant documents. Assume that there are 8 relevant documents in total in the collection.

`R R N R N  N N R N N`

(a) What is the precision? (3 marks)

(b) What is the recall? (3 marks)

(c) What is the F_1 measure? (3 marks)

(d) What is the precision/recall break-even point? (3 marks)

(e) What is the MAP for this single query? (3 marks)

8. **(5 marks)**

Why do we usually have to face a trade-off between precision and recall?

9. **(10 marks)**

Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d_1 and d_2. She judges d_1, with the content "CDs software cheap CDs" relevant and d_2 with content "cheap thrills DVDs" nonrelevant. Assume that we use direct term frequency (with no scaling and no document frequency), and do not length-normalize vectors. Using the Rocchio relevance feedback (with parameters $\alpha = 1$, $\beta = 0.8$, $\gamma = 0.2$), what would the revised query vector be after relevance feedback? Keep in mind that negative term weights are treated in a special way.

10. **(10 marks)**

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

- d_1: The European Union Act 2011 prevents additional powers being passed to Brussels without a referendum.
- d_2: EU will not ban Channel 5 perfumes over allergy findings.

Suppose the query q is 'European Union'. Show how the above documents should be ranked for q, using a unigram language model with Jelinek-Mercer smoothing that mixes the distributions estimated from the specific document (weight $\lambda = 0.4$) and the entire collection (weight $1 - \lambda = 0.6$).

11. **(10 marks)**

Consider the following collection of documents that belong to two classes: China (C) and Japan (J).

	docID	docText	class
TRAINING	d_1	Kyoto Tokyo Taiwan	J
	d_2	Japan Kyoto	J
	d_3	Taipei Taiwan	C
	d_4	Macao Taiwan Beijing	C
TEST	d_5	Taiwan Taiwan Kyoto	?

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

12. **(5 marks)**

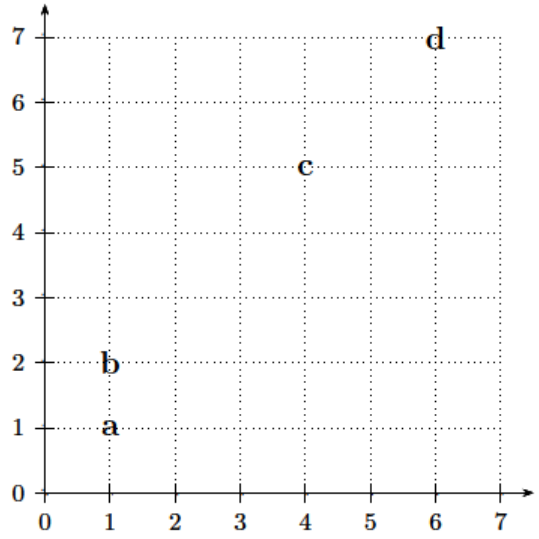
Please answer the following questions about k-means clustering.

(a) Why are documents that do not use the same term for the concept *car* likely to end up in the same cluster in k-means clustering? (2 marks)

(b) Two of the possible termination conditions for k-means clustering are: (1) assignment does not change, (2) centroids do not change. Do these two conditions imply each other or not? Why? (3 marks)

13. **(5 marks)**

Perform a 2-means clustering (based on Euclidean distance) to convergence for the points below.



Start with the two seeds **a** and **b**. For each iteration give (1) the coordinates of the centroids, and (2) the assignments of points to centroids.

