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: Monday, 6th June 2016
Duration of paper: 2:30pm – 4:30pm (2 hours)

RUBRIC

1. *This paper contains* ten *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.*

©Birkbeck College 2016

1. **(10 marks)**

Give a brief answer (in one sentence) to each of the following questions.

(a) What is the *tf-idf* weighting scheme? (2 marks)

(b) What is the *idf* of a term that occurs in half of the documents in the collection? (2 marks)

(c) What are the most widely known algorithms for English *stemming* and *phonetic correction* respectively? (2 marks)

(d) What roles does *smoothing* play in the language modelling approach to IR? (4 marks)

2. **(10 marks)**

(a) Show in detail how the Levenshtein edit distance between two words "`dell`" and "`adele`" could be computed. (8 marks)

(b) What are the corresponding edits that change the former to the latter? (2 marks)

3. **(10 marks)**

(a) What are the entries for the search term "`brexit`" in a 2-gram index? (2 marks)

(b) What are the entries for the search term "`brexit`" in a permuterm index? (3 marks)

(c) What is the biggest value that can be encoded by a one-byte VB code? (2 marks)

(d) Decode the following *gap* sequence in γ-code to the original postings list of document IDs: `11100101100011010`. (3 marks)

4. **(10 marks)**

(a) Is it true that if two documents are identical their Jaccard coefficient must be 1? Justify your answer. (2 marks)

(b) Is it true that if two documents' Jaccard coefficient is 1 they must be identical? Justify your answer. (2 marks)

(c) What are the possible estimations of Jaccard coefficient that you can make using the MinHash algorithm with 4 random permutations, ? (2 marks)

(d) How does the number of random permutations affect the effectiveness and efficiency of MinHash for near duplicate detection? (2 marks)

(e) Would MinHash still work if we use maximum instead of minimum in the algorithm? Justify your answer. (2 marks)

5. **(10 marks)**

(a) Build the generalised suffix tree for the following two documents.
Document 1: `Mr`
Document 2: `Mrs` (8 marks)

(b) What is the main advantage of suffix arrays in comparison to suffix trees? (2 marks)

6. **(10 marks)**

The following ranked list shows the relevance judgements of the twenty search results that an IR system retrieved for a given query ($\surd$ = relevant, × = nonrelevant).

position	1	2	3	4	5	6	7	8	9	10
relevance	$\surd$	$\surd$	×	$\surd$	×	×	×	$\surd$	×	×
position	11	12	13	14	15	16	17	18	19	20
relevance	×	×	×	×	×	×	×	×	×	×

There is one more relevant document in the collection, but the system missed it.

Calculate the following performance measures of this system with respect to this query:

(a) Precision P, (2 marks)

(b) Recall R, (2 marks)

(c) F_1 measure, (2 marks)

(d) $PRBEP$, (2 marks)

(e) Average Precision AP (as in MAP). (2 marks)

7. **(10 marks)**

Give a brief answer (in one sentence) to each of the following questions.

(a) What is the *probability ranking principle*? (2 marks)

(b) What is the *cluster hypothesis*? (2 marks)

(c) What is the optimal number of clusters for *cluster pruning* in terms of efficiency? (2 marks)

(d) Can the *purity* score of clustering ever be 0? If so, provide an example. If not, explain why. (2 marks)

(e) When will *microaveraging* and *macroaveraging* give the same results? (2 marks)

8. **(10 marks)**

The following document collection consists of book reviews that are either positive (P) or negative (N).

	docID	docText	class
TRAINING	d_1	easy read	P
	d_2	easy read funny	P
	d_3	hard read	N
	d_4	hard read dull	N
	d_5	hard hard dull	N
	d_6	hard funny	N
TEST	d_7	read read hard	?

Train a classifier for sentiment analysis using the (multinomial) Naive Bayes algorithm (with Laplace smoothing), and then apply it to predict the class of the test document.

9. **(10 marks)**

Consider the following small collection of six documents represented as data points in a two-dimensional vector space:
$d_1 = (0, 2),\ d_2 = (0, 3),\ d_3 = (3, 2),\ d_4 = (3, 4),\ d_5 = (5, 5),\ d_6 = (6, 6).$

(a) Which document, d_4 or d_6, is more similar to document d_2, according to the cosine similarity? (5 marks)

(b) Which pairs of documents have cosine similarity 1 between themselves? (2 marks)

(c) Is it true that if a pair of documents have cosine similarity 1 between themselves, they will always have the same cosine similarity with any query vector q? Why? (3 marks)

10. **(10 marks)**

Consider the following small collection of six documents represented as data points in a two-dimensional vector space:
$d_1 = (0, 2),\ d_2 = (0, 3),\ d_3 = (3, 2),\ d_4 = (3, 4),\ d_5 = (5, 5),\ d_6 = (6, 6).$
Compute the clustering of those data points (based on the *Euclidean distance*) using the following two HAC algorithms, and show the clustering results as dendrograms.

(a) The single-link HAC algorithm. (5 marks)

(b) The complete-link HAC algorithm. (5 marks)