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, 1st June 2015
Duration of paper: 10:00am – 12:00pm (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. *The use of non-programmable electronic calculators is permitted, but programmable electronic devices such as smart phones must be switched off.*
4. *This paper is not prior-disclosed.*

1. **(10 marks)**

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

(a) What are *stop words*? (2 marks)

(b) What is *stemming*? (2 marks)

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

(d) What is the difference between standard *relevance feedback* and *pseudo relevance feedback*? (4 marks)

2. **(10 marks)**

Fill the following matrix to compute the Levenshtein edit distance between two words '`Saturday`' and '`Sunday`'.
What are the corresponding edits that change the former to the latter?

	''	S	u	n	d	a	y
''							
S							
a							
t							
u							
r							
d							
a							
y							

3. **(5 marks)**

Suppose that we would like to build an index for tolerant retrieval.

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

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

4. **(15 marks)**

Suppose that we need to perform near-duplicate detection in a collection of three documents. Their sets of shingle fingerprints are as follows.
$D_1 : \{0, 1, 2\}$
$D_2 : \{1, 3, 4\}$
$D_3 : \{0, 2, 3\}$

(a) Compute their pairwise Jaccard coefficients. (3 marks)

(b) Estimate their pairwise Jaccard coefficients using MinHash with the following two hash functions.
$h_1(x) = (3x + 1) \mod 5$
$h_2(x) = (4x + 2) \mod 5$ (12 marks)

5. **(5 marks)**

Decode the following *gap* sequence in γ-code to the original postings list of document IDs:
`11011111000011101111`

6. **(5 marks)**

Build the suffix tree for the string `googol` .

7. **(10 marks)**

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

position	1	2	3	4	5	6	7	8	9	10
relevance	$\surd$	$\surd$	$\times$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\times$	$\surd$

There are two more relevant documents in the collection, but the system missed them.

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)

8. **(10 marks)**

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

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

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

(c) What is *cluster pruning*? (2 marks)

(d) What is the difference between *single-link* and *complete-link* hierarchical agglomerative clustering algorithms? (4 marks)

9. **(10 marks)**

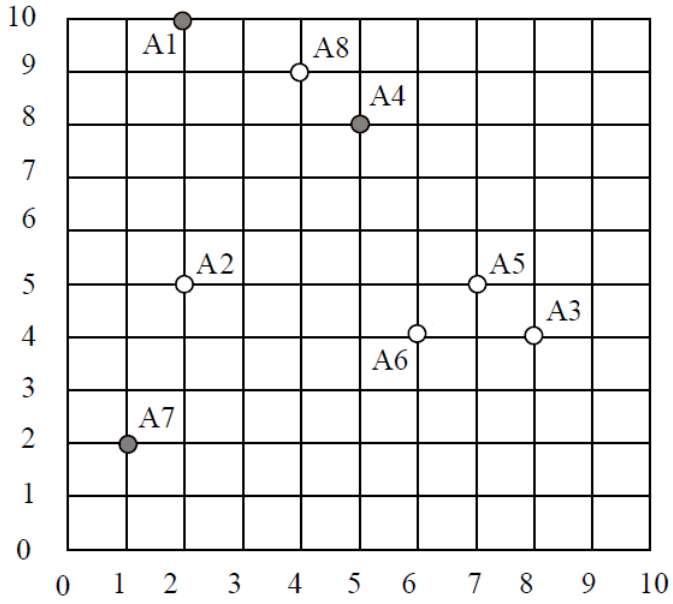
Each document in the following collection consists of the key words extracted from the description of a mobile app. All mobile apps are categorised into two classes according to whether they are appropriate for kids (Y) or not (N).

	docID	docText	class
TRAINING	d_1	child game	Y
	d_2	child video	Y
	d_3	child music game	Y
	d_4	child music video	Y
	d_5	adult video	N
	d_6	adult video game	N
TEST	d_7	child video video	?

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

10. **(20 marks)**

Consider the following small dataset of eight documents represented as data points in a vector space:
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).



(a) Use the k-means algorithm (based on *Euclidean distance*) to cluster those documents into three clusters, assuming that the seeds (initial cluster centres) are A1, A4 and A7. (10 marks)

(b) Use the standard k-nearest-neighbours algorithm (based on *cosine similarity*) to classify a new document A9=(0,5), with $k = 1$, $k = 3$, and $k = 5$ respectively. Suppose that the documents $\{$A1, A4, A7$\}$ are labelled as positive while the other documents $\{$A2, A3, A5, A6, A8$\}$ are labelled as negative. (10 marks)

