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: Thursday, 22th May 2014
Duration of paper: 2:30pm – 4:30pm (2 hours)

RUBRIC

1. *This paper contains* 12 *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)**

Build the *positional* inverted file index for the following document collection. Please do not use any preprocessing on the tokens and do not compress the postings lists.

docID	docText
1	hickory dickory dock
2	the mouse ran up the clock
3	the clock struck one
4	the mouse ran down
5	hickory dickory dock

2. **(5 marks)**

Assume that you are using k-grams for doing wildcard searches. The search term entered by a user is `*tone` . What Boolean queries on a 2-gram index and a 3-gram index would be generated for this search term respectively?

3. **(10 marks)**

Fill the following matrix to compute the Levenshtein edit distance between two strings ‘kitten’and ‘sitting’.

	‘’	s	i	t	t	i	n	g
‘’								
k								
i								
t								
t								
e								
n								

4. **(5 marks)**

How would the following dictionary entries be stored using front coding?

> abandon, abandoned, abandoning, abandonment, accommodate, accommodation, accompanied

5. **(5 marks)**

Decode the following binary sequence encoded in γ-code:
1110001110111111101001

6. **(10 marks)**

Compute the tf-idf weights for the terms `car`, `auto`, `insurance`, and `best` for each document, given the term frequency (tf) and document frequency (df) information in the following table:

term	df	doc1-tf	doc2-tf	doc3-tf
`car`	200	1	100	10
`auto`	20	1	10	1
`insurance`	2000	100	10	1
`best`	20,000	100	1000	10

There are a total of 200,000 documents. Assume that the logarithmic variants are used for both the tf and idf values (and the logarithm to base 10, i.e., $\log_{10}$, is used).

7. **(5 marks)**

Build the suffix tree for the string *xabxa*.

8. **(10 marks)**

The following table shows how two human judges rated the relevance of a set of documents to a particular information need (0 = nonrelevant, 1 = relevant).

docID	1	2	3	4	5	6	7	8	9	10	11	12
Judge1	0	0	1	1	1	1	1	1	0	0	0	1
Judge2	0	0	1	1	0	0	0	0	1	1	1	0

Let us assume that you have developed an IR system that for this query returns the set of documents $\{4, 5, 6, 7, 8\}$.

(a) Calculate the precision, recall, and F_1 of your system if a document is considered relevant as long as either judge thinks that it is relevant.

(b) Calculate the precision, recall, and F_1 of your system if a document is considered relevant only if the two judges agree that it is relevant.

9. **(5 marks)**

The vendor of an IR system claims that their system outputs the following result for a TREC query. Is this a believable result? Briefly explain your answer.

Ranking	Recall	Precision
1. d_8	10%	80%
2. d_{32}	30%	70%
3. d_{98}	40%	60%
4. d_{124}	30%	50%
5. d_9	40%	40%
6. d_{78}	40%	30%
7. d_{73}	40%	20%

10. **(10 marks)**

Consider the following document collection that consists of 4 documents.

d_1: computer computer
d_2: computer science master degree advanced computer computer computer
d_3: information technology
d_4: computer science information technology

Suppose the query q is 'computer science'. 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.

11. **(10 marks)**

Consider the following collection of documents that belong to two classes: UK and US.

	docID	docText	class
TRAINING	d_1	London, England.	UK
	d_2	York, England.	UK
	d_3	New England.	US
	d_4	Newark, New Jersey.	US
TEST	d_5	York. New York.	?

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

12. **(15 marks)**

Suppose that the pair-wise similarity information of a document collection $\{d_1, d_2, d_3, d_4, d_5, d_6\}$ is given by the following table.

d_1						
d_2	0.7					
d_3	0.4	0.8				
d_4	0.2	0.1	0.5			
d_5	0.0	0.1	0.0	0.6		
d_6	0.1	0.4	0.2	0.3	0.4	
	d_1	d_2	d_3	d_4	d_5	d_6

(a) If it is known that $\{d_1, d_2, d_3\}$ are *Facts* and $\{d_4, d_5\}$ are *Myths*, how will document d_6 be classified by the kNN algorithm with $k = 3$, $k = 4$, and $k = 5$ respectively? Please use the *standard* kNN algorithm but not its weighted variant. (5 marks)

(b) Use the single-link HAC algorithm to cluster these documents, and draw the generated dendrogram. (5 marks)

(c) Use the complete-link HAC algorithm to cluster these documents, and draw the generated dendrogram. (5 marks)

