# Birkbeck
## (University of London)

**MSc Examination for Internal Students**

**Department of Computer Science and Information Systems**

**Information Retrieval and Organisation (COIY064H7)**
**Credit Value: 15**

**Date of Examination: xxxday x xxx 2011**
**Duration of Paper: xx:00 - xx:00 (2 hours)**

*RUBRIC*

1. *This paper contains* xx *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. (5 marks)

(a) Briefly explain the difference between Tokenisation and Normalisation. (4 marks)

(b) What is the most common algorithm used for stemming English? (1 mark)

2. (6 marks)

Assume an IR system needs an inverted file index that maps 2-grams to the terms that contain the 2-grams for tolerant retrieval. Start with an empty index and insert the terms *the*, *their*, *theme*, and *they* into this index. What will the final index look like?

3. (4 marks)

10,000,000 numbers are sorted in ascending order using a replacement selection sorting algorithm in the first step of external sorting. The main memory buffer can hold 1,000,000 numbers.

(a) How many blocks would be produced in the output if the numbers in the input are already sorted in ascending order? (2 marks)

(b) How many blocks would be produced in the output if the numbers in the input are sorted in descending order? (2 marks)

4. (3 marks)

A system supports wildcard queries with a permuterm index. What steps would have to be taken to process the following query: ***A*B*** ? Assume that only prefix queries can be answered efficiently by the permuterm index.

5. (4 marks)

During an evaluation of an IR system a query returns the following answer set containing ten documents.

| Ranking | Recall | Precision |
|---|---|---|
| 1. $d_{25}$ | 0% | 0% |
| 2. $d_3$ | 20% | 50% |
| 3. $d_{70}$ | 20% | 33% |
| 4. $d_{20}$ | 20% | 25% |
| 5. $d_8$ | 40% | 40% |
| 6. $d_{39}$ | 40% | 33% |
| 7. $d_{14}$ | 60% | 43% |
| 8. $d_{52}$ | 60% | 38% |
| 9. $d_{63}$ | 60% | 33% |
| 10. $d_{16}$ | 60% | 30% |

Which of the documents in this list are relevant? How many relevant documents are not returned by the system?

6. (5 marks)

An IR system uses a Gamma code to encode postings lists. The following bitstring is returned by an inverted file index as a gap encoded list of DocIDs. Is this a sequence properly encoded in Gamma code? If yes, write down the decoded list of DocIDs. If no, briefly explain why this bitstring is not a correct Gamma code.

11100011111010111100011111100100

7. (10 marks)

There is a variant of the Levenshtein distance called Damerau-Levenshtein distance, which allows one more edit operation (in addition to insert, delete, and replace): a transposition of two adjacent characters. For example, to turn 'lit*er*' into 'lit*re*' would require one transposition operation, switching the positions of the letters 'e' and 'r'. Comparing this to the computation of the minimum cost of a cell in the Levenshtein distance algorithm, what modifications do you have to make to this calculation to implement a Damerau-Levenshtein distance?

8. (8 marks)

Given the vectors for the query $q$ and the two documents $d_1$ and $d_2$ as follows

$$\vec{q} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \quad \vec{d_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \quad \vec{d_2} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

we see that the user is interested in documents containing the terms $t_1$, $t_4$, and $t_6$.

(a) Which of the above documents are in the answer set of the Boolean query $t_1$ `AND` $t_4$ `AND` $t_6$? (2 marks)

(b) Which of the above documents are in the answer set of the Boolean query $t_1$ `OR` $t_4$ `OR` $t_6$? (2 marks)

(c) Which of the above documents has a higher ranking if you apply the cosine measure? Do not apply any TF-IDF weighting, but do apply normalisation. (4 marks)

9. (10 marks)

An IR system uses Rocchio's algorithm for relevance feedback. After a first iteration, the documents $d_1$ and $d_2$ are marked as relevant by the user, while the documents $d_3$, $d_4$, and $d_5$ are marked as irrelevant. The vectors for the query $q_0$ and the documents are given below:

$$\vec{q_0} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \quad \vec{d_1} = \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix} \quad \vec{d_2} = \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} \quad \vec{d_3} = \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} \quad \vec{d_4} = \begin{pmatrix} 0 \\ 4 \\ 5 \end{pmatrix} \quad \vec{d_5} = \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}$$

As a reminder, the formula used in Rocchio's algorithm is:

$$\vec{q_m} = \alpha\vec{q_0} + \beta\frac{1}{|D_r|}\sum_{\vec{d_j}\in D_r}\vec{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d_j}\in D_{nr}}\vec{d_j}$$

(a) Assume that the parameters used in this first iteration are $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.5$. What is the modified query vector used for the second round? (6 marks)

(b) Does it make sense to adjust the parameters $\alpha$, $\beta$, and $\gamma$ in between iterations? If yes, briefly explain how the parameters would be modified. If no, briefly explain your reasoning. (4 marks)

10. (5 marks)

Briefly explain how the language modelling approach to IR can capture the intuition that terms with high *tf-idf* scores in the vector space model should influence the ranking more than other terms.

11. (10 marks)

Suppose the document collection consists of the following two documents.
$d_1$: "We know which university is the best university."
$d_2$: "Georgia Tech is the best university in Georgia."
Suppose the query $q$ is "best university".
Show how the above documents should be ranked using MLE unigram language models from the documents and the collection, mixed with $\lambda = 0.6$ as the weight for document models.

12. (10 marks)

Consider the following collection of documents that belong to two classes:
amphibian (A) and non-amphibian (B).

| | docID | docText | class |
|---|---|---|---|
| TRAINING | $d_1$ | red frogs eat red flies | A |
| | $d_2$ | green frogs eat flies | A |
| | $d_3$ | green lizards eat all flies | B |
| TEST | $d_4$ | red green green | ? |

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

13. (5 marks)

Compare the speed of two learning algorithms for text classification:
Naive Bayes (NB) and $k$-Nearest-Neighbours (kNN).
Which is usually faster with respect to their training time? Why?
Which is usually faster in classifying a new, previously unseen document? Why?

14. (5 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 A B C C |
| cluster 2 | A B C C |
| cluster 3 | A B B C |

What is the purity of the above clustering?

15. (5 marks)

Given a set of five points $\{1, 2, 3, 4, 5\}$ on an axis, what clusters will 2-means clustering produce if 4 is chosen as the initial seed for one cluster and 5 is chosen as the initial seed for the other cluster?
Use Euclidean distance. For example, the distance between points 2 and 5 is 3.

16. (5 marks)

Given a set of five points $\{0, 2, 3, 6, 9\}$ on an axis, perform single-link Hierarchical Agglomerative Clustering (HAC) and draw the generated dendrogram.
Use Euclidean distance. For example, the distance between points 2 and 5 is 3.