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, 11th June 2018
Duration of paper: 10:00 am – 12:00 pm (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.*

1. **(10 marks)**

Build the *positional* inverted index for the following document collection. Do not use any token preprocessing or index compression technique. The document frequency and term frequency information should be included in the index.

d_1: `Everybody look`
d_2: `to their left`
d_3: `Everybody look`
d_4: `to their right`
d_5: `It is not about`
d_6: `the money money money`

2. **(10 marks)**

Write regular expressions to extract the following set of strings from a given document. By "word", we mean an alphabetic string separated from other words by white-space, any relevant punctuation, line breaks, and so forth.

(a) All lower-case words ending with the suffix `ing` . (2 marks)

(b) All strings with two consecutive repeated words (e.g., "`The The`" and "`bug bug`" but not "`The bug`" or "`The big bug`"). (3 marks)

(c) All strings that start with the word `you` and finish with the word `me` or the other way around. (2 marks)

(d) All lines that start at the beginning of the line with a word and finish at the end of the line with an integer. (3 marks)

3. **(10 marks)**

Compute the Levenshtein edit distance between two words '`sitting`' and '`kitten`' (with insertion cost 1, deletion cost 1, and substitution cost 1).
Show your work using the edit distance grid, and also represent the final result as an alignment between those two words indicating the editing operations required to convert the former to the latter.

4. **(10 marks)**

Give a brief answer to each of the following questions.

(a) What are the entries for the search term `apple` in a permuterm index? (2 marks)

(b) What is the Jaccard coefficient between `mary` and `may` using bigrams? (3 marks)

(c) What is the VB code for the document ID 130? (2 marks)

(d) What is the original postings list of document IDs for the *gap* sequence in γ-code `11100011101111101001` ? (3 marks)

5. **(10 marks)**

Consider the following small document collection.

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

Show how the first two documents (d_1 and d_2) should be ranked with respect to the query $q =$ "computer science", using an unigram language model that mixes the distributions estimated from the specific document and the entire collection with equal weights.

6. **(10 marks)**

Given the following movie reviews each labelled with a genre action (A) or comedy (C)

d_1: `fast, furious, shoot` (A)
d_2: `fun, laugh, love, love` (C)
d_3: `laugh, fly, fast, fun, fun` (C)
d_4: `furious, shoot, shoot, fun` (A)
d_5: `fly, fast, shoot, love` (A)

train a movie genre classifier using the standard (multinomial) Naïve Bayes algorithm (with Laplace smoothing) first, and then apply it to predict the most likely genre for a new movie review

d_6: `furious, fun, fantastic, fun`

7. **(10 marks)**

We are given the following corpus where the special symbols `<s>` and `</s>` represent the beginning and the end of sentence respectively.

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

(a) What is the probability of the following sentence computed using a bigram language model (with Laplace smoothing)?
`<s> I like ham </s>` (8 marks)

(b) What are the probability estimations $P(\texttt{Sam}|\texttt{I am})$ and $P(\texttt{am}|\texttt{Sam I})$ in a trigram language model (without smoothing)? (2 marks)

8. **(10 marks)**

The following table shows the performance of a multi-class multi-label text classifier on a document collection with two classes, where "truth" represents the real class labels of documents (i.e., the gold standard) and "system" represents the output of the classifier.

class 1		
	system: *positive*	system: *negative*
truth: *positive*	400	100
truth: *negative*	400	360
class 2		
	system: *positive*	system: *negative*
truth: *positive*	900	600
truth: *negative*	900	720

(a) Compute the macroaveraged precision and recall. (4 marks)

(b) Compute the microaveraged F_1 measure. (6 marks)

9. **(10 marks)**

The following table shows a simplified term-term co-occurrence matrix.

	`fish`	`chips`	`fries`
`Japanese`	6	1	1
`English`	1	2	0
`French`	0	0	2

(a) Is `Japanese` closer to `French` or `English`, according to the cosine similarity (based on the raw counts in the table)? (5 marks)

(b) Is `English` closer to `fish` or `chips`, according to the Positive Pointwise-Mutual-Information (PPMI)? (5 marks)

10. **(10 marks)**

Give a brief answer to each of the following questions.

(a) What is the difference between the two models for generating word embeddings: skip-gram and CBOW? (4 marks)

(b) How does the skip-gram model compute the probability of word w_k given word w_j, i.e., $p(w_k|w_j)$? (2 marks)

(c) Let C denote a term-document matrix where the (i, j) entry corresponds to the tf-idf weight for the i-th term in the j-th document. What does the (i, j) entry in the matrix C^TC represent? What does the (i, j) entry in the matrix CC^T represent? (4 marks)

