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, 25th May 2017
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. *This paper is not prior-disclosed.*
4. *The use of non-programmable electronic calculators is permitted.*

1. **(10 marks)**

Write regular expressions for the following language. By "word", we mean an alphabetic string separated from other words by white-space, any relevant punctuation, line breaks, and so forth.

(a) The set of all words ending with the letter `s`. (2 marks)

(b) The set of all strings that represent prices in the UK, i.e., the pound sign followed by a real number which has optionally two digits after the decimal point. For example, £`1500` and £`19.99` . (3 marks)

(c) The set of all strings that have both the word `Birkbeck` and the word `College` in them (but **not** words like `Colleges` which merely contain the word `College` as a part). (2 marks)

(d) The set of all strings that start at the beginning of the line with a word and finish at the end of the line with the same word. (3 marks)

2. **(10 marks)**

Find out whether `drive` is closer to `brief` or to `divers` according to minimum edit distances. The allowed operations include insertion (with cost 1) and deletion (with cost 1), but not substitution. Show your work using the edit distance grids.

3. **(10 marks)**

Give a brief answer to each of the following questions.

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

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

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

(d) What is the biggest value that can be encoded by a 7-bit γ-code? (3 marks)

4. **(10 marks)**

Consider a fictitious document collection that contains the following two documents.

- d_1: The European Union Act 2011 prevents additional powers being passed to Brussels without a referendum.
- d_2: EU will not ban Channel 5 perfumes over allergy findings.

Suppose the query q is 'European Union'. Show how the above documents should be ranked for q, using a unigram language model with Jelinek-Mercer smoothing that mixes the distributions estimated from the specific document (weight $\lambda = 0.4$) and the entire collection.

5. **(10 marks)**

The following document collection consists of (preprocessed and simplified) movie reviews that are either positive (P) or negative (N).

	docID	docText	class
TRAINING	d_1	the most fun film of the summer	P
	d_2	very powerful	P
	d_3	just plain boring	N
	d_4	entirely predictable and lacks energy	N
	d_5	no surprises and very few laughs	N
TEST	d_6	there is just no fun	?

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

6. **(10 marks)**

Give a brief answer to each of the following questions.

(a) How does the language modelling approach to IR capture the intuition that terms with high *tf-idf* scores in the vector space model should influence the ranking more than other terms? (4 marks)

(b) What are the assumptions that make the Naïve Bayes algorithm computationally tractable? How can the Naïve Bayes algorithm be effective when it relies on such oversimplifying assumptions? (6 marks)

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		
	truth: *positive*	truth: *negative*
system: *positive*	40	40
system: *negative*	10	3600
class 2		
	truth: *positive*	truth: *negative*
system: *positive*	90	90
system: *negative*	60	7200

(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`	`duck`	`wine`
`London`	6	1	1
`Paris`	0	0	2
`Beijing`	1	2	0

(a) Is `London` closer to `Paris` or `Beijing`, according to the cosine similarity (based on the raw counts in the table)? (5 marks)

(b) Is `Beijing` closer to `fish` or `duck`, 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., $p(w_k|w_j)$? (3 marks)

(c) Which two clusters would be merged at each step of Brown clustering? (3 marks)

