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: Friday, 14th June 2019
Duration of paper: 1:30 pm – 3:30 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: `The Annotated Alice`
d_2: `More Annotated Alice`
d_3: `The Annotated Alice: The Definitive Edition`
d_4: `The Annotated Alice: 150th Anniversary Deluxe Edition`

2. **(10 marks)**

Find out whether `drive` is closer to `divers` or to `brief` according to minimum edit distances. The allowed operations include insertion (with cost 1), deletion (with cost 1), and substitution (with cost **2**). 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 `metoo` in a permuterm index? (2 marks)

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

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

(d) What is the biggest value that can be encoded by a 9-bit γ-code? (3 marks)

4. **(10 marks)**

The following ranked list shows the relevance judgements of the twenty 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	$\surd$	$\surd$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\times$
position	11	12	13	14	15	16	17	18	19	20
relevance	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\surd$

There are five more relevant document in the collection, but the system missed them.

Calculate the following performance measures of the system with respect to this query:

(a) Precision P, (2 marks)

(b) Recall R, (2 marks)

(c) F_1 measure, (2 marks)

(d) Precision-Recall Break-Even Point ($PRBEP$), (2 marks)

(e) Mean Average Precision (MAP) for this query. (2 marks)

5. **(10 marks)**

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

- d_1: `Which university is the best university?`
- d_2: `Birkbeck College is the best university!`

Suppose the query q is '`best university`'. 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 $\lambda = 0.6$) and the entire collection.

6. **(10 marks)**

The following document collection consists of book reviews that are either positive (P) or negative (N).

	docID	docText	class
TRAINING	d_1	easy read	P
	d_2	easy read funny	P
	d_3	hard read	N
	d_4	hard read dull	N
	d_5	hard hard dull	N
	d_6	hard funny	N
TEST	d_7	very good read hard read	?

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

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> cause all of me </s>
<s> loves all of you </s>
<s> give your all to me </s>
<s> i will give my all to you </s>
```

(a) What are the probability estimations $P(\texttt{your}|\texttt{all to})$ and $P(\texttt{you}|\texttt{all to})$ in a trigram language model (without smoothing)? (2 marks)

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

8. **(10 marks)**

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

	black	white	red
comedy	4	1	1
action	0	0	2
horror	1	2	0

(a) Is `comedy` more similar to `action` or `horror`, according to the cosine similarity (based on the raw counts in the table)? (5 marks)

(b) Is `horror` more similar to `black` or `white`, according to the Positive Pointwise-Mutual-Information (PPMI)? (5 marks)

9. **(10 marks)**

Give a brief answer to each of the following questions.

(a) What is the *distributional hypothesis* in linguistics? (2 marks)

(b) Why do dense word vectors usually work better than sparse word vectors in NLP tasks? (3 marks)

(c) Use *skip-gram with negative sampling (SGNS)* to generate training examples for the target word t = "eggs" in the following sentence, with the context window length $L = \pm 2$ and the ratio of negative examples to positive examples $k = 1.5$.
"A full English breakfast typically includes bacon, sausages, eggs and a beverage such as coffee or tea." (5 marks)

10. **(10 marks)**

Give a brief answer to each of the following questions.

(a) Consider a neural unit with just one input $x = -1$.
What will be its output if the activation function is sigmoid ?
What will be its output if the activation function is ReLU?
Tip: $e \approx 2.72$. (2 marks)

(b) What are the advantages and disadvantages of neural language models in comparison with traditional n-gram language models? (4 marks)

(c) Label the following sentence using the IOB encoding for named entity recognition with two entity types: ORG (organisation) and LOC (location).
"Blue Prism has offices in London, Paris, and New York etc." (4 marks)

