Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

Natural Language Processing and Information Retrieval (COIY064H7)

========== ONLINE ==========

CREDIT VALUE: 15 credits

Date of examination: Tuesday, 2nd June 2020
Duration of paper: 2:00 pm – 5:00 pm (3 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.*

1. **(10 marks)**

Assume that we would like to develop a search engine for the following small document collection.

d_1: `First they came for the socialists.`
d_2: `Then they came for the trade unionists.`
`Then they came for the Jews.`
d_3: `Then they came for me.`

(a) Build the *positional* inverted index. Do not use any preprocessing (except tokenisation and case-folding) or index compression techniques. The document frequency and term frequency information should be included in the index. (8 marks)

(b) What are the TF-IDF weights for the terms "they" and "the" in the vector for d_2? The logorithm should be used in the computation of IDF (with base 2) but not TF. *Tip*: $\log_2(3) \approx 1.6$. (2 marks)

2. **(10 marks)**

Compute the edit distance between two words '`brexit`' and '`megxit`'. The allowed operations include insertion (with cost 1) and deletion (with cost 1), but not substitution. Show your work using the edit distance grid.

3. **(10 marks)**

Give a brief answer to each of the following questions.

(a) How would the following dictionary entries be stored using front coding? `computation`, `compute`, `computer`, `computing`, `data` (2 marks)

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

(c) What is the smallest value that would require two bytes to be encoded using the VB code? (2 marks)

(d) What is the original postings list of document IDs for the *gap* sequence in γ-code `11100011101111110100 10` ? (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	$\surd$	$\surd$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\surd$
position	11	12	13	14	15	16	17	18	19	20
relevance	$\times$	$\times$	$\times$	$\times$	$\surd$	$\times$	$\times$	$\times$	$\times$	$\times$

There are four more relevant documents 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 three documents.

- d_1: `Which university is the best?`
- d_2: `Birkbeck College is the best!`
- d_3: `We are proud of Birkbeck.`

Suppose the query q is '`best university`'. Show how those documents should be ranked with respect to 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.

6. **(10 marks)**

Each document in the following collection consists of the key words extracted from the description of a mobile app. All mobile apps are categorised into two classes according to whether they are appropriate for kids (Y) or not (N).

	docID	docText	class
TRAINING	d_1	learn novel	Y
	d_2	learn music video	Y
	d_3	adult video	N
	d_4	adult video novel	N
	d_5	learn video	Y
	d_6	learn music novel	Y
TEST	d_7	learn video teach video	?

Show how the (multinomial) Naive Bayes algorithm (with Laplace smoothing) can be used to train a classifier and 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> I am not </s>
<s> the guy </s>
<s> who told the guy </s>
<s> that you are the guy </s>
<s> who gave the guy the black eye </s>
```

(a) What are the probability estimations $P(\texttt{<s>}|\texttt{the guy})$ and $P(\texttt{</s>}|\texttt{the guy})$ 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> I am the guy </s> (8 marks)

8. **(10 marks)**

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

	pepper	ginger	onion	garlic
pork	1	2	0	2
beef	0	0	2	2
fish	4	1	0	1

(a) Is pork more similar to beef or fish, according to the cosine similarity (based on the raw counts in the table)? (5 marks)

(b) Is pork more similar to pepper or ginger, according to the Positive Pointwise-Mutual-Information (PPMI)? (5 marks)

9. **(10 marks)**

Give a brief answer to each of the following questions.

(a) Assume that for a given term-document matrix $\mathbf{C}$ whose (i, j) entry is the TF-IDF weight of the i-th term in the j-th document, the Singular Value Decomposition (SVD) is given by $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Let $\mathbf{U}_k$ denote the reduced matrix formed by retaining only the first k columns of $\mathbf{U}$. What does the matrix $\mathbf{U}_k\mathbf{U}_k^T$ represent? (2 marks)

(b) Label the following sentence using the IOB encoding for named entity recognition with the following entity types: PER (Person), ORG (organisation), and LOC (Location).
"Zoubin Ghahramani holds joint appointments at University College London and the Alan Turing Institute." (2 marks)

(c) In *one-of classification*, the classifier would assign exactly one class to each document, and every document would be assigned exactly one class. Show that in this setup, (i) the total number of false positive decisions equals the total number of false negative decisions, and (ii) the micro-averaged F_1 is always identical to the accuracy, i.e., the fraction of documents that are classified correctly. (6 marks)

10. **(10 marks)**

Give a brief answer to each of the following questions.

(a) Given an input vector $\mathbf{z} = (-1, -1, -1, -1)$, what would be the output of the softmax function? (2 marks)

(b) Prove that the sigmoid function is a special case of the softmax function for binary classification. (2 marks)

(c) Prove that Naive Bayes and Logistic Regression for binary classification are both linear classifiers, i.e., they both use a linear function as the decision boundary between the positive class and the negative class. (6 marks)

