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, 25th May 2021
Duration of paper: 10:00 am – 12:00 pm (2 hours)

RUBRIC

1. *This paper contains* seven *questions for a total of* $15 \times 6 + 10 = 100$ *marks.*
2. *Students should attempt to answer* ***all*** *of them.*
3. *This paper is not prior-disclosed.*

1. **(15 marks)**

Suppose that we have the following query q and document collection $\{d_1, d_2, d_3\}$:

q:	`future of online education`
d_1:	`Birkbeck is shaping the future of online education.`
d_2:	`Birkbeck is great in online education.`
d_3:	`Online education is the future of education.`

The text preprocessing steps include tokenization, case-folding, and stop-word-removal, but not lemmatization/stemming. Here the set of stopwords are {`in`, `is`, `of`, `the`}.

(a) Consider the the instantiation of the vector space model where documents and queries are represented as raw *term frequency* (TF) vectors.
What is the cosine similarity between each document and the query? (12 marks)

(b) What are the *inverse document frequency* (IDF) weights for the query terms using the logarithm with base 10? (3 marks)

2. **(15 marks)**

Suppose that you have a query $q =$ '`online courses`' and you are considering the following three documents $\{d_1, d_2, d_3\}$ within a large document collection.

d_1:	`Birkbeck provides excellent online courses`
d_2:	`Birkbeck provides affordable evening education`
d_3:	`online education is affordable`

In addition, you are given the probabilities of some words in the unigram collection model.

term	prob.	term	prob.	term	prob.
`affordable`	0.001	`birkbeck`	0.025	`courses`	0.125
`education`	0.075	`evening`	0.005	`excellent`	0.002
`is`	0.300	`online`	0.250	`provides`	0.003

How should you rank those three documents with respect to q using the unigram *query likelihood model* with *Jelinek-Mercer smoothing* (where the unigram document model weight $\lambda = 0.6$)?

3. **(15 marks)**

Suppose a query has a total of 4 relevant documents in a collection of 100 documents. System A and System B have each retrieved 10 documents, and the relevance status of the ranked lists is shown below.

$$\text{System A: } [-\ +\ -\ -\ +\ -\ -\ -\ -\ -]$$

$$\text{System B: } [+\ +\ -\ -\ +\ -\ -\ -\ -\ -]$$

where the leftmost entry corresponds to the highest ranked document, and the rightmost entry corresponds to the lowest ranked document. A "$+$" indicates a relevant document and a "$-$" corresponds to a non-relevant one. For example, the top ranked document retrieved by System A is non-relevant, whereas the top ranked document retrieved by B is relevant.

(a) What is the F_1 measure of each system? (5 marks)

(b) What is the *PRBEP* of each system? (5 marks)

(c) What is the *average-precision* (as in *MAP*) of each systems? (5 marks)

4. **(15 marks)**

You are given the following corpus of one sentence to train a trigram language model (with Laplace smoothing). Include the special sentence *start* and *end* symbols `<s>` and `</s>` in your counts just like any other token.

```
This is the rat that ate the malt that lay in the house
that Jack built.
```

What is the *perplexity* of this smoothed model on the following test sentence?

```
This is the house that Jack built.
```

5. **(15 marks)**

Consider the following corpus consisting of only two sentences.

```
Born to sweet delight.
Born to endless night.
```

Let us define the context of a word as the two words to the left and the two words to the right from the target word, occurred within the same sentence (if there are any).

(a) Compute the *first-order co-occurrence* vectors for the words "`delight`" and "`night`" using Positive Pointwise Mutual Information (PPMI). (12 marks)

(b) Compute the *second-order co-occurrence* between the words "`delight`" and "`night`" using cosine similarity based on their first-order co-occurrence vectors. (3 marks)

6. **(15 marks)**

(a) Prove that Logistic Regression with the *sigmoid* function is a special case of Multinomial Logistic Regression with the *softmax* function for binary classification. (5 marks)

(b) 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. (10 marks)

7. **(10 marks)**

Assume that documents are being classified into two categories, C_1 and C_2, such that a document can belong to more than one category. The table below shows the prediction of a classifier, denoted by "y" or "n", in addition to the true label (ground truth) represented by a "$+$" or "$-$", where a correct prediction is either $y(+)$ or $n(-)$.

	C_1	C_2
d_1	$y(+)$	$y(+)$
d_2	$n(-)$	$y(+)$
d_3	$n(+)$	$n(-)$
d_4	$y(-)$	$y(+)$
d_5	$n(+)$	$n(-)$

(a) What is the classification *accuracy* of this classifier? (2 marks)

(b) What is the *macro-averaged* F_1 measure of this classifier? (4 marks)

(c) What is the *micro-averaged* F_1 measure of this classifier? (4 marks)

