

Birkbeck
(University of London)

MSc Examination for Internal Students

School of Computer Science and Information Systems

Information Retrieval and Organisation (COIY064H7)

Mock Exam

1. (4 marks)
 An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

2. (6 marks)
 A document collection is indexed using the following inverted file (the entries in the postings lists are document IDs):

⋮	
Farnham	2,3,9,11,17,18,21
Faversham	3,12,17
Folkestone	15,16
Godalming	1,2,3,5,9,11,15,18,21,35
Hailsham	1,2,15,16,18
Haslemere	3,5,18,22,25
Hastings	11,15,16,17
⋮	

The following (Boolean) query is processed on the IR system: $\text{Farnham} \wedge \text{Godalming} \wedge \text{Haslemere}$. Which documents are returned in the answer set? Provide all intermediate results (assume that the system optimises the execution of the query).

3. (6 marks)
 Compute the similarity of the two terms *keyboard* and *fingerboard* by using 2-grams and the Jaccard coefficient.

4. (2 marks)
 Consider a positional index of the following shape:
 love: 1: <12>; 2: <23,32,43>; 3:<53>
 hell: 1: <25>; 2:<34,40>; 5:<38>

where the numbers before a colon (:) are document IDs, the numbers in the bracketed lists are positions.
 Is it possible that there is a document containing the phrase “love is hell”?

5. (6 marks)
 Assume you index the terms *Mars*, *landed* and *rover* in the following document: “After a successful landing on Mars, the Mars rover Opportunity landed on a Mars plain in the Meridiani section of Mars. The ship landed at an excellent landing spot.”
 Assume that there are 1000 documents in the complete collection: 100 contain *Mars*, 1 contains *landed*, and 10 contain *rover*. Using these, compute the tf-idf weights for the vector components of *Mars*, *landed*, and *rover* for this document.

Use the following formulas for computing the TF- and IDF-values:
 $\text{TF} = 1 + \log_{10}(tf_{t,d})$ and $\text{IDF} = \log_{10}(\frac{N}{df_t})$.

Use the following approximations for the logarithms:
 $\log_{10}(2) \approx 0.3$, $\log_{10}(3) \approx 0.5$, $\log_{10}(4) \approx 0.6$,

6. (10 marks)
 Consider a fictitious document collection that contains the following 4 documents.

d_1 : click go the shears boys click click click
d_2 : click click
d_3 : metal here
d_4 : metal shears click here

Suppose the query q is 'click shears'. Show how the above documents should be ranked for q , using an unigram language model that mixes the distributions estimated from the specific document and the entire collection with equal weights.

7. (16 marks)

Consider the following collection of documents that belong to two classes: American (A) and British (B).

	docID	docContent	class
TRAINING	d_1	New York	A
	d_2	New York, US	A
	d_3	UK. London, UK.	B
	d_4	London, UK	B
TEST	d_5	New York. New York. London. London. London.	?

Show how the Naive Bayes algorithm (with Laplace smoothing) can be used to train a classifier and predict the class of test document. Explain why the classification result is consistent or inconsistent with our intuition.

8. (10 marks)

Consider the following collection of documents that belong to two classes: World-News (W) and Business-News (B).

	docID	docText	class
TRAINING	d_1	Iraq election	W
	d_2	French executive injured	W
	d_3	Chief executive smiles	B
	d_4	Krispy Kreme executive resigns	B
TEST	d_5	Executive suite	?

Show how the 3NN algorithm predicts the class of test document, using raw term frequency, no IDF, and cosine similarity. Would the same result be guaranteed using the 1NN algorithm? Why or why not?