

# IR Coursework (1)

Sven Helmer  
SCSIS, Birkbeck, University of London

Spring 2012

This is part 1 of the coursework (worth 12 marks) covering the topics of Sven's lectures. The total coursework, together with Dell's part, will be worth 20 marks. The deadline of this coursework is 24.02.2012 (final cut-off deadline 09.03.2012).

1. (3 marks)

Draw the inverted file index for the following documents (do not use any preprocessing on the tokens and do not compress the postings lists):

DocID	document content
1	hickory dickory dock
2	the mouse ran up the clock
3	the clock struck one
4	the mouse ran down
5	hickory dickory dock

(a) using only sorted DocIDs

(b) adding positional information to the posting entries as well.

2. (1 marks)

Assume you are using k-grams for doing wildcard searches. The search term entered by a user is \*tone\*.

(a) What Boolean query on a 2-gram index would be generated for this search term?

(b) What Boolean query on a 3-gram index would be generated for this search term?

3. (2 marks)

How would the following dictionary entries be stored using front coding?

abandon, abandoned, abandoning, abandonment, accommodate, accommodation, accompanied

4. (2 marks)

Decode the following binary sequence encoded in Gamma code:  
111000111011111101001

5. (2 marks)

Compute the tf-idf weights for the terms **car**, **auto**, **insurance**, **best**, for each document, given the tf and df values in the following table:

term	df	doc1-tf	doc2-tf	doc3-tf
car	200	1	100	10
auto	20	1	10	1
insurance	2000	100	10	1
best	20,000	100	1000	10

There are a total of 200,000 documents. Assume that the logarithmic variants are used for the tf and idf values (assume that  $\log_{10}$ , a logarithm to the base 10, is used).

6. (2 marks)

The vendor of an IR system claims that their system outputs the following result for a TREC query. Is this a believable result? Briefly explain your answer.

Ranking	Recall	Precision
1. $d_8$	10%	80%
2. $d_{32}$	30%	70%
3. $d_{98}$	40%	60%
4. $d_{124}$	30%	50%
5. $d_9$	40%	40%
6. $d_{78}$	40%	30%
7. $d_{73}$	40%	20%