

Solutions for the Sample Exam

May 26, 2011

1 Exercise 1

Draw the term-document incidence matrix for this document collection. The answer could be found in the slides of Part 3.

	d1	d2	d3	d4
solution	1	0	0	0
found	1	1	0	0
for	1	0	1	1
laziness	1	1	1	1
old	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
of	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1

Draw the positional inverted index representation for this collections. The answer could be found in slides 35-36 of Part 3.

term	# documents containing the term	[Doc1, term-freq in Doc1: position1, position2 ...]
solution	1	[d1, 1: 1]
found	2	[d1, 1: 2], [d2, 1: 3]
for	3	[d1, 1: 3], [d3, 1: 3], [d4, 1: 3]
laziness	4	[d1, 1: 4], [d2, 1: 2], [d3, 1: 6], [d4, 1: 4]
old	3	[d2, 1: 1], [d3, 1: 1], [d4, 1: 1]
approach	1	[d3, 1: 2]
treatment	1	[d3, 1: 4]
of	1	[d3, 1: 5]
hopes	1	[d4, 1: 2]
patients	1	[d4, 1: 5]

2 Exercise 2

Recommend a query processing order for the query. First we approximate the OR operator with the sum of the frequencies and then execute the query from lowest frequency to highest. The execution order is: [(phones OR ears) AND (bush OR apricot)] AND (pudding OR brown)

3 Exercise 3

To execute OR query for two terms we need to return union of the doc ids in the posting lists of the two terms. As we have to visit all the postings, the skip pointers are not useful.

4 Exercise 4

Example document: "Open university is better than state university Israel"

5 Exercise 5

game\$, ame\$g, me\$ga, e\$gam, \$game

6 Exercise 6

The answer is in the slide 28-34 of Part 4.

7 Exercise 7

Relevant: “The term *open source* describes practices in production and development that promote access to the end product’s materials” score = $2/18 = 1/9$

Irrelevant: “We work in open space” score = $1/6$

8 Exercise 8

The relative scores of the documents is not affected. From properties of logarithms we know that $\log_a(x) = \log_a(b)\log_b(x)$. Hence assume that $idf_t = \log_a(N/df_t)$.

Now lets change its base to b . $\log_a(N/df_t) = \log_a(b) * \log_b(N/df_t) = c * \log_b(N/df_t)$. So changing the base from a to b changes the score by factor of $c = \log_a(b)$.

9 Exercise 9

Query:

term	tf	df	log tf	idf	tf-idf
film	1	100	$1+\log(1)=1$	$\log_{10}(10^4/10^2) = 2$	$1*2=2$
cameras	1	300	1	1.52	1.52
digital	0	1000	0	1	0

Document:

term	tf	log tf	normalized log tf
film	1	$1+\log(1)=1$	$1/\sqrt{1^2 + 1.3^2 + 1^2} = 0.52$
cameras	2	1.3	0.68
digital	1	1	0.52

The final similarity score is the dot product of the two vectors (query and document): $2*0.52+1.52*1.03+0*0.52=2.07$