Information Retrieval 21-22

Exercises week 1

With solutions

Exercise 1

Rewrite these Boolean queries in disjunctive normal form:

a) NOT (NOT Caesar OR roman) AND Brutus
b) (Euro OR NOT Dollar) AND NOT Yen

Solution Exercise 1

a) Caesar AND (NOT roman) AND Brutus

b) (Euro AND NOT Yen) OR (NOT Dollar AND NOT Yen)

Exercise 2

Consider these documents:

Doc 1 new recipe for cookies

Doc 2 cookie dough recipe

Doc 3 new chocolate dough

Doc 4 recipe for chocolate cookies

a) Draw the term-document incidence matrix for this document collection

b) Draw the inverted index representation for this collection.

Solution Exercise 2a

	DOC1	DOC2	DOC3	DOC4
new	1	0	1	0
recipe	1	1	0	1
for	1	0	0	1
cookies	1	0	0	1
cookie	0	1	0	0
dough	0	1	1	0
chocolate	0	0	1	1

Solution Exercise 2b

new	1,3
recipe	1,2,4
for	1,4
cookies	1,4
cookie	2
dough	2,3
chocolate	3,4

Exercise 3

For the document collection shown in Exercise 2, what are the returned results for these queries:

a) cookies OR cookie
b) recipe AND NOT (chocolate OR dough)

Solution Exercise 3

a) 1,2 and 4

b) 1

Exercise 4

For the queries below, can we still run through the intersection time O*(x+y),* where *x* and *y* are the lengths of the posting lists for Brutus and Caesar? If not, what can we achieve?

a) Brutus AND NOT Caesar
b) Brutus OR NOT Caesar

Solution Exercise 4

a) Time is O(x+y). Instead of collecting documents that occur in both postings lists, collect those that occur in the first one and not in the second.

b) Time is O(N) (where N is the total number of documents in the collection) assuming we need to return a complete list of all documents satisfying the query. This is because the length of the results list is only bounded by N, not by the length of the postings lists.

Exercise 5

Recommend a query processing order for the query below, given the postings list sizes in the table:

(rocket OR launch) AND (sky OR universe) AND (sun OR orbit)

Term	Postings size
rocket	81300
launch	109999
sky	43123
universe	48491
sun	62687
orbit	577513

Solution Exercise 5

- rocket OR launch: 81300 + 109999 = 191299
- sky OR universe: 43123 + 48491 = 91614
- sun OR orbit: 62687 + 577513 = 640200
- Process in increasing order of their frequency (sum)

Term	Postings size
rocket	81300
launch	109999
sky	43123
universe	48491
sun	62687
orbit	577513