Boolean Retrieval (Chapter 1)

Exercise 1/1

Recommend a query processing strategy for (*tangerine* OR *trees*) AND (*marmalade* OR *skies*) AND (*kaleidoscope* OR *eyes*) with respect to the following postings list sizes:

eyes 213 312

kaleidoscope 87 009

marmalade 107 913

skies 271 658

tangerine 46 653

trees 316 812

We use a database trick where we filter out the results with the clause of the shortest intermediate result first. Operations OR is understood as addition and AND as multiplication. Compose the equations:

$$\textit{tangerine} \text{ OR } \textit{trees} = 46\,653 + 316\,812 = 363\,465$$

$$\textit{marmalade} \text{ OR } \textit{skies} = 107\,913 + 271\,658 = 379\,571$$

$$\textit{kaleidoscope} \text{ OR } \textit{eyes} = 87\,009 + 213\,312 = 300\,321$$

After sorting these with respect to sizes and we get the ordering

$$\textit{kaleidoscope} \text{ OR } \textit{eyes} < \textit{tangerine} \text{ OR } \textit{trees} < \textit{marmalade} \text{ OR } \textit{skies}$$

we see that the query is best processed in the following sequence:

1. $a =$ *kaleidoscope* OR *eyes*
2. $b =$ *tangerine* OR *trees*
3. $c =$ *marmalade* OR *skies*
4. $d = a$ AND b
5. $e = d$ AND c

Exercise 1/2

What is the best order for processing the query *ostrich* AND *hippo* AND *giraffe* if we know that the number of occurrences of the animals are 100, 500, 300, respectively?

(*ostrich* AND *giraffe*) AND *hippo*

Exercise 1/3

Create an inverted index composed of the following collection of documents:

Doc 1: new home sales top forecasts

Doc 2: home sales rise in July

Doc 3: increase in home sales in July

Doc 4: July new home sales rise

Very easy procedure. Start with an empty table. If the term already appears in the table as a key, add the document ID only. Otherwise, take each term of a document and add it as a key to the table with the ID of the document. This way we get the inverted index represented in the following table.

forecasts	1			
home	1	2	3	4
in	2	3		
increase	3			
July	2	3	4	
new	1	4		
rise	2	4		
sales	1	2	3	4
top	1			

Table 1: Inverted index

Exercise 1/4

Create an inverted index composed of the following collection of documents:

Doc 1: hippo ostrich ostrich giraffe

Doc 2: lion frog giraffe hippo

Doc 3: ostrich frog bat giraffe lion frog

---

bat	3		
frog	2	3	
giraffe	1	2	3
hippo	1	2	
lion	2	3	
ostrich	1	3	

Table 2: Inverted index