Information Retrivel: Assignment 1

Shifei Chen

---

1 Exercises

Exercise 1.1

$$
\begin{aligned}
new &\rightarrow 1, 4 \\
home &\rightarrow 1, 2, 3, 4 \\
sales &\rightarrow 1, 2, 3, 4 \\
top &\rightarrow 1 \\
forecasts &\rightarrow 1 \\
rise &\rightarrow 2, 4 \\
in &\rightarrow 2, 3 \\
july &\rightarrow 2, 3, 4 \\
increase &\rightarrow 3
\end{aligned}
$$

Exercise 1.2 a

Terms	Doc 1	Doc 2	Doc 3	Doc 3
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
schizophrenia	1	1	1	1
new	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

Exercise 1.2 b

$$
\begin{aligned}
breakthrough &\rightarrow 1 \\
drug &\rightarrow 1, 2 \\
for &\rightarrow 1, 3, 4 \\
schizophrenia &\rightarrow 1, 2, 3, 4 \\
new &\rightarrow 2, 3, 4 \\
approach &\rightarrow 3 \\
treatment &\rightarrow 3 \\
of &\rightarrow 3 \\
hopes &\rightarrow 4 \\
patients &\rightarrow 4
\end{aligned}
$$

Exercise 1.3 a

schizophrenia AND drug = 1111 AND 1100 = 1100

So this query will return Doc 1 and Doc 2

Exercise 1.3 b

for AND NOT(drug OR approach)
= 1011 AND NOT (1100 OR 0010)
= 1011 AND NOT 1110
= 1011 AND 0001
= 0001

So this query will return Doc 4

2 Questions

2.1 Use the term-document incidence matrix in Figure 1.1 to return the documents related to the query "(Brutus OR Caesar) AND NOT(Antony OR Cleopatra)"

(Brutus OR Caesar) AND NOT(Antony OR Cleopatra)
= (110100 OR 110111) AND NOT (110001 OR 100000)
= 110111 AND NOT 110001
= 110111 AND 001110
= 000110

So this query will return Hamlet and Othello

2.2 Can you find a general way to process arbitrary Boolean queries as the query in previous exercise?

I think the general process for boolean queries should be like

1. Build the term-document incidence matrix for this particular document collection
2. Replace all of the terms in the query with their binary representations
3. Calculate the binary result by executing binary operations, such as AND, OR or NOT
4. Look up in the term-document incidence matrix to figure out which document is in the final binary result

It might also be a good idea to convert the query into a postfix expression with the Shunting Yard Algorithm[1]. For example the query in the previous question "(Brutus OR Caesar) AND NOT(Antony OR Cleopatra)" would be turned into "Brutus Caesar OR Antony Cleopatra OR NOT AND".

[1] http://www.oxfordmathcenter.com/drupal7/node/628