LBSC 796 /INFM 718R: Homework 2 Answers

Part 1: Boolean Retrieval

1. Term frequency matrix

Term	Doc 1	Doc 2	Doc 3	Doc 4
cornell	1	0	0	0
for	0	1	0	0
implementation	1	0	1	0
information	0	1	1	1
introduction	0	0	0	1
manual	0	1	0	0
modern	0	0	0	1
of	1	0	1	0
retrieval	0	1	1	1
smart	1	1	1	0
system	1	1	1	0
the	2	1	1	0
to	0	0	0	1
user's	0	1	0	0

2. New matrix

Term	Doc 1	Doc 2	Doc 3	Doc 4
cornell	1	0	0	0
for	0	1	0	0
implementation	1	0	1	0
information	0	1	1	1
introduction	0	0	0	1
manual	0	1	0	0
modern	0	0	0	1
of	1	0	1	0
retrieval	0	1	1	1
smart	1	1	1	0
system	1	1	1	0
the	1	1	1	0
to	0	0	0	1
user's	0	1	0	0

3. Docs having "information" and "retrieval"

Term	Doc 1	Doc 2	Doc 3	Doc 4
information	0	1	1	1
retrieval	0	1	1	1
Information AND retrieval	0	1	1	1

Answer: Doc 2, Doc 3, and Doc 4.

4.
A. (smart AND implementation) OR (introduction AND retrieval)

step 1: (smart AND implementation)

Term	Doc 1	Doc 2	Doc 3	Doc 4
smart	1	1	1	0
implementation	1	0	1	0
smart AND implementation	1	0	1	0

step 2: (introduction AND retrieval)

Term	Doc 1	Doc 2	Doc 3	Doc 4
introduction	0	0	0	1
retrieval	0	1	1	1
introduction AND retrieval	0	0	0	1

step 3: (smart AND implementation) OR (introduction AND retrieval)

Term	Doc 1	Doc 2	Doc 3	Doc 4
smart AND implementation	1	0	1	0
introduction AND retrieval	0	0	0	1
(smart AND implementation) OR (introduction AND retrieval)	1	0	1	1

Answer: Doc 1, Doc 3 and Doc 4.

B. (Cornell OR SMART) AND (Implementation)

step 1: (Cornell OR SMART)

Term	Doc 1	Doc 2	Doc 3	Doc 4
cornell	1	0	0	0
smart	1	1	1	0
cornell OR smart	1	1	1	0

step 2: (Cornell OR SMART) AND (Implementation)

Term	Doc 1	Doc 2	Doc 3	Doc 4
cornell OR smart	1	1	1	0
implementation	1	0	1	0
(cornell OR smart) AND (implementation)	1	0	1	0

Answer: Doc1 and Doc3.

C. (information NOT retrieval)

Term	Doc 1	Doc 2	Doc 3	Doc 4
information	0	1	1	1
retrieval	0	1	1	1
information NOT retrieval	0	0	0	0

Answer: None.

5. Perform the same computation for the query: (information XOR system)

Term	Doc 1	Doc 2	Doc 3	Doc 4
information	0	1	1	1
system	1	1	1	0
information XOR system	1	0	0	1

Answer: Doc 1 and Doc 4.

Part 2: Vector Space Retrieval

1. Build the *w* matrix

	TF			***IDF***	***W_{ij}***		
	1	2	3		1	2	3
t1			5	log(3/1)=0.477			2.385
t2	4	1	3	log(3/3)=0.000			
t3	5		4	log(3/2)=0.176	0.880		0.704
t4	6	3	3	log(3/3)=0.000			
t5		1		log(3/1)=0.477		0.477	
t6	3		7	log(3/2)=0.176	0.528		1.232
t7		6	1	log(3/2)=0.176		1.056	0.176
t8	2			log(3/1)=0.477	0.954		

2. Build the w' matrix

	W_{ij}				***W'_{ij}***		
	1	2	3		1	2	3
t1			2.385				0.858
t2							
t3	0.880		0.704		0.628		0.253
t4				→			
t5		0.477				0.412	
t6	0.528		1.232		0.377		0.443
t7		1.056	0.176			0.911	0.063
t8	0.954				0.681		
length	**1.402**	**1.159**	**2.781**				

3. Compute rank order using vector space method for UNWEIGHTED query t2 t7

		W'_{ij}		
	query	1	2	3
t1				0.858
t2	1			
t3		0.628		0.253
t4				
t5			0.412	
t6		0.377		0.443
t7	1		0.911	0.063
t8		0.681		
Similarity score		**0**	**0.911**	**0.063**

Rank order: Doc 2, Doc 3, Doc 1

4. Compute rank order for WEIGHTED query t2(7) t7

		W'_{ij}		
	query	1	2	3
t1				0.858
t2	7			
t3		0.628		0.253
t4				
t5			0.412	
t6		0.377		0.443
t7	1		0.911	0.063
t8		0.681		
Similarity score		**0**	**0.911**	**0.063**

Rank order: Doc 2, Doc 3, Doc 1

5. *Doc 4* in the class example is missing here; the change has an overall impact on the IDF values, which leads to differences in W_{ij}, W'_{ij}, and hence document similarity scores.