



มหาวิทยาลัยรามคำแหง

ข้อสอบวิชา COS4351 ระบบค้นคืนสารสนเทศ ชุดที่ 1

ชื่อ.....รหัส.....

คำสั่ง ข้อสอบมีทั้งหมด 14 ข้อ จำนวน 3 หน้า รวม 100 คะแนน

1. (15 คะแนน) กำหนดเอกสารมีค่าสำคัญดังนี้
DOC1 = { AA , BB , CC , DD , EE}
DOC2 = { BB, DD , EE , XX}
จงหาความคล้ายคลึงกันของเอกสารทั้งสองโดยวิธี
1.1 Simple matching
1.2 Dice's Coefficient
1.3 Jaccard's Coefficient
1.4 Overlap Coefficient
1.5 จงสร้างโครโมโซมต้นกำเนิดตามเอกสารที่กำหนดโดยวิธีจีเนติกอัลกอริทึม
2. (10 คะแนน) จงอธิบายถึงความแตกต่างระหว่าง Database และ IR ในประเด็นที่สำคัญ
3. (5 คะแนน) จงอธิบายถึงความสำคัญของกระบวนการกระทำซ้ำ(Feedback)ของระบบIR
4. (10 คะแนน) กำหนดความสามารถของระบบ IR ระบบ1 และระบบ 2 ดังนี้

ความสามารถ	ระบบ 1	ระบบ 2
HIT	2000	4000
WASTED	4000	2000
MISSED	6000	8000
PASSED	8000	6000

จงหา Recall ,Precision, Fallout ของสองระบบนี้เป็นอย่างไร ระบบใดดีกว่าจงให้เหตุผลพร้อมอธิบายให้เข้าใจ

5. (5 คะแนน) สมมติว่า เอกสาร D1, D2, D3, D4 ,D5 มีคีย์เวิร์ดดังนี้

D1 = (K1 ,K2, K3 , K4 , K5)

D2 = (K3, K4, K5, K6)

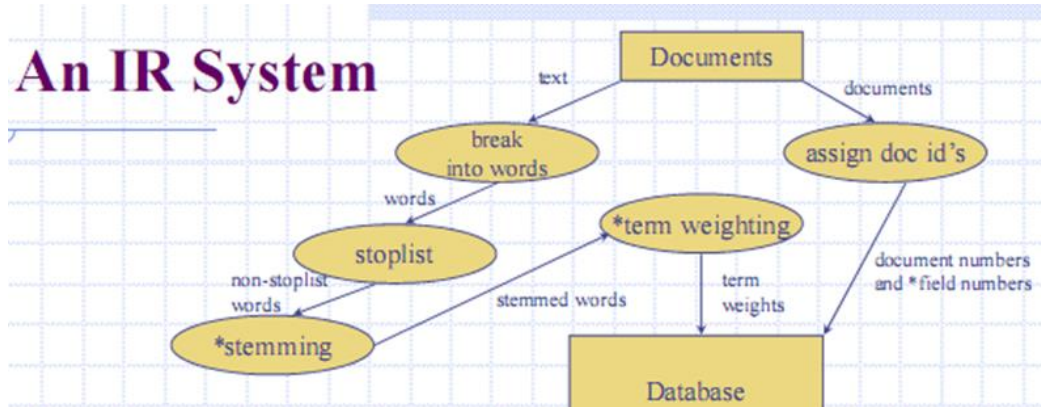
D3 = (K2, K3, K4, K5)

D4 = (K1, K2, K3, K4, K6)

D5 = {K3 , K5, K6}

เมื่อนำมาเก็บในโครงสร้างของ Inverted file จะเก็บข้อมูลในลักษณะใด

6. (5 คะแนน)ท่านคิดว่าการวิเคราะห์ข้อความ (Text Analysis) ในการพัฒนาระบบ IR เพื่อมีวัตถุประสงค์อะไร จงอธิบายให้เข้าใจ
7. (5 คะแนน)Index ในระบบIRคืออะไร ใช้ประโยชน์อย่างไร อธิบายให้เข้าใจ
8. (5 คะแนน)ปัญหาในการวิเคราะห์ข้อความมีอะไรบ้าง มีวิธีการแก้ไขอย่างไร
9. (5 คะแนน)การทำงานของเสิร์จเอนจิน(Search Engine)โดยทั่วไปประกอบด้วย 3 ส่วนหลักอะไรบ้าง จงอธิบายพอเข้าใจ
10. (10 คะแนน)พิจารณารูปของระบบ IR ต่อไปนี้



จงอธิบายถึงขั้นตอนการทำงานของระบบ IR ข้างต้นให้เข้าใจ

11. (5 คะแนน)กำหนดนิยามของคลัสเตอร์ในรูปแบบของกราฟ ที่ได้จากการวัดความคล้ายคลึงกันดังนี้ (object : 1,2,3,4,5,6)

Similarity matrix	1					
	2	.9				
	3	.6	.5			
	4	.8	.6	.6		
	5	.3	.9	.3	.8	
	6	.4	.5	.2	.4	.2
		1	2	3	4	5

ถ้ากำหนดค่า Threshold เท่ากับ 0.8 จงแสดงกราฟที่ได้จากการวัดความคล้ายคลึงนี้

12. (10 คะแนน)กำหนดให้ BA ,FI, MI ,NA , RM แทนเมืองในประเทศอิตาลี จงกระทำวิธีการของ Hierarchical clustering ในการหาระยะทางระหว่างเมืองต่างๆ โดยระยะทางมีหน่วยเป็นกิโลเมตร โดยใช้อัลกอริทึม single-linkage พร้อมวาดรูปต้นไม้

	BA	FI	MI	NA	RM
BA	0	200	300	50	400
FI	200	0	20	600	80
MI	300	20	0	90	100
NA	50	600	90	0	70
RM	400	80	100	70	0

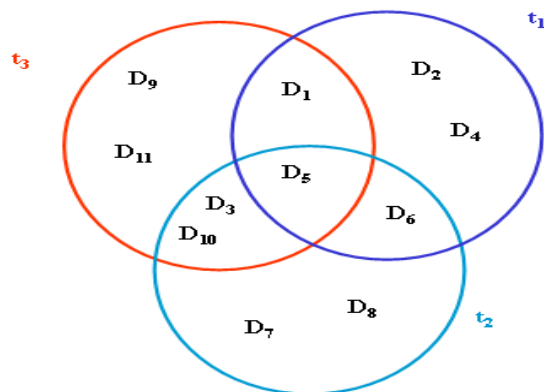
13. (5 คะแนน)กำหนดให้ $\{A, B, C\}$ เป็นเซตของสัญลักษณ์ของคีย์ R1, R2, R3, R4, R5, R6 เป็นเรคอร์ดที่เก็บบันทึกในโครงสร้างนี้ สมมติว่าคีย์แต่ละเรคอร์ดมีขนาดไม่เท่ากันดังนี้

R1	มีคีย์ว่า	ABC
R2	มีคีย์ว่า	AC
R3	มีคีย์ว่า	BBC
R4	มีคีย์ว่า	BCC
R5	มีคีย์ว่า	CA
R6	มีคีย์ว่า	CBA

จงสร้าง Double Chained Tree จากข้อมูลข้างต้น(วาดรูป)

14. (5 คะแนน)พิจารณารูปต่อไปนี้

docs	t1	t2	t3	RSV=Q.Di
D1	1	0	1	4
D2	1	0	0	1
D3	0	1	1	5
D4	1	0	0	1
D5	1	1	1	6
D6	1	1	0	3
D7	0	1	0	2
D8	0	1	0	2
D9	0	0	1	3
D10	0	1	1	5
D11	0	0	1	4
Q	1	2	3	
	q1	q2	q3	



Q is a query – also represented Boolean term combinations

โดย q1,q2,q3 คือ ข้อสอบถาม

t1,t2,t3 เป็น เทอมของคำสำคัญหรือดรรชนี

D1,D2,...D11 คือเอกสาร

จงอธิบายการดึงเอกสารเมื่อกำหนดค่า RSV มีค่าเท่ากับ 5