ข้อ 2. สมมติในระบบมีเอกสารทั้งหมด 10 เอกสารดังนี้ (bird, cat, dog, tiger คือ Keyword โดย Keyword เหล่านี้ไม่มีความสัมพันธ์กัน)

D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,tiger,cat,dog}
D3: {dog,bird,bird}
D4: {cat,tiger}
D5: {tiger,tiger,dog,tiger,cat}
D6: {bird,cat,bird,cat,tiger,tiger,bird}
D7: {bird,tiger,cat,dog}
D8: {dog,cat,bird}
D9: {cat,dog,tiger}
D10: {tiger,tiger,tiger}

ผู้ใช้ส่งคำเรียกค้น "**I don't love cat and dog. But I love tiger as well as bird.**" เข้าไปในระบบ จงตอบคำถาม

2.1 เพื่อให้ได้คำตอบในคำถาม 2.2 ผู้ใช้สามารถเลือกใช้โมเดลใดได้บ้างเพราะอะไร (เลือกได้เฉพาะตัวเลือกที่ให้มา)

A) BM25 Model
B) Fuzzy Model
C) Extend Boolean Model
D) Vector Model
E) Probabilistic Model
F) Generalized Vector Model

2.2 ให้นักศึกษาแสดงวิธีคำนวณหา **Ranking** ของเอกสารทุกเอกสารในระบบ ตามที่ผู้ใช้ต้องการ

2.3 หากระบบกำหนดให้เอกสารที่ 9 ตรงประเด็นมากกว่าเอกสารที่ 2 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย(**35 คะแนน**)

Answer

2.1 ใช้ Extend Boolean Model เนื่องจากลักษณะของ Query เป็นแบบ Boolean และโจทย์กำหนดให้ Keyword ไม่สัมพันธ์กัน

A) BM25 Model ใช้ได้แต่ไม่ครอบคลุมการใช้ Boolean Expression (And Or Not)

B) Fuzzy Model ใช้ไม่ได้ เพราะโจทย์กำหนดให้ Keyword ไม่มีความสัมพันธ์กัน

C) Extend Boolean Model

D) Vector Model ใช้ได้แต่ไม่ครอบคลุมการใช้ Boolean Expression (And Or Not)

E) Probabilistic Model ใช้ได้แต่ไม่ครอบคลุมการใช้ Boolean Expression (And Or Not)

F) Generalized Vector Model ใช้ไม่ได้ เพราะโจทย์กำหนดให้ Keyword ไม่มีความสัมพันธ์กัน

Query = “I don’t love cat and dog. But I love tiger as well as bird”

ไม่รักแมวและสุนัข แสดงว่าไม่ต้องการเอกสารที่มีแมวหรือมีสุนัข

รักเสือพอๆกับรักนก แสดงว่าต้องการเอกสารที่มีนกหรือมีเสือ

Query = (Bird OR Tiger) AND NOT(Cat OR Dog)

ขั้นตอนที่ 1

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,tiger,cat,dog}
D3: {dog,bird,bird}
D4: {cat,tiger}
D5: {tiger,tiger,dog,tiger,cat}
D6: {bird,cat,bird,cat,tiger,tiger,bird}
D7: {bird,tiger,cat,dog}
D8: {dog,cat,bird}
D9: {cat,dog,tiger}
D10: {tiger,tiger,tiger}

	Bird	Cat	Dog	Tiger	Max
Doc1	3	2	2	0	3
Doc2	0	2	1	1	2
Doc3	2	0	1	0	2
Doc4	0	1	0	1	1
Doc5	0	1	1	3	3
Doc6	3	2	0	2	3
Doc7	1	1	1	1	1
Doc8	1	1	1	0	1
Doc9	0	1	1	1	1
Doc10	0	0	0	3	3
n	5	8	7	7	

ขั้นตอนที่ 1

Only Doc1

	Bird	Cat	Dog	Tiger	Max
Doc1	3	2	2	0	3
Doc2	0	2	1	1	2
Doc3	2	0	1	0	2
Doc4	0	1	0	1	1
Doc5	0	1	1	3	3
Doc6	3	2	0	2	3
Doc7	1	1	1	1	1
Doc8	1	1	1	0	1
Doc9	0	1	1	1	1
Doc10	0	0	0	3	3
n	5	8	7	7	

$$tf_{bird} = \frac{3}{3} = 1.000$$
$$tf_{cat} = \frac{2}{3} = 0.667$$
$$tf_{dog} = \frac{2}{3} = 0.667$$
$$tf_{tiger} = \frac{0}{3} = 0.000$$

$$idf_{bird} = \log(\frac{10}{5}) = 0.301$$
$$idf_{cat} = \log(\frac{10}{8}) = 0.097$$
$$idf_{dog} = \log(\frac{10}{7}) = 0.155$$
$$idf_{tiger} = \log(\frac{10}{7}) = 0.155$$

$$idf_{norm,\,bird} = \frac{0.301}{0.301} = 1.000$$
$$idf_{norm,\,cat} = \frac{0.097}{0.301} = 0.322$$
$$idf_{norm,\,dog} = \frac{0.155}{0.301} = 0.515$$
$$idf_{norm,\,tiger} = \frac{0.155}{0.301} = 0.515$$

$$\mathbf{w_{bird} = 1.000 * 1.000 = 1.000}$$
$$\mathbf{w_{cat} = 0.667 * 0.322 = 0.215}$$
$$\mathbf{w_{dog} = 0.667 * 0.515 = 0.343}$$
$$\mathbf{w_{tiger} = 0.000 * 0.515 = 0.000}$$

ขั้นตอนที่ 1

น้ำหนักของแต่ละ Keyword ในแต่ละเอกสาร

	Bird	Cat	Dog	Tiger
Doc1	1.000	0.215	0.343	0.000
Doc2	0.000	0.322	0.257	0.257
Doc3	1.000	0.000	0.257	0.000
Doc4	0.000	0.322	0.000	0.515
Doc5	0.000	0.107	0.172	0.515
Doc6	1.000	0.215	0.000	0.343
Doc7	1.000	0.322	0.515	0.515
Doc8	1.000	0.322	0.515	0.000
Doc9	0.000	0.322	0.515	0.515
Doc10	0.000	0.000	0.000	0.515

ขั้นตอนที่ 2

Query = "I don't love cat and dog. But I love tiger as well as bird"

Query = (Bird OR Tiger) AND NOT(Cat OR Dog)

	Bird	Cat	Dog	Tiger
Doc1	1.000	0.215	0.343	0.000
Doc2	0.000	0.322	0.257	0.257
Doc3	1.000	0.000	0.257	0.000
Doc4	0.000	0.322	0.000	0.515
Doc5	0.000	0.107	0.172	0.515
Doc6	1.000	0.215	0.000	0.343
Doc7	1.000	0.322	0.515	0.515
Doc8	1.000	0.322	0.515	0.000
Doc9	0.000	0.322	0.515	0.515
Doc10	0.000	0.000	0.000	0.515

$$sim(q_{or}, d_j) = \sqrt{\frac{W_{1,j}^2 + W_{2,j}^2}{2}}$$

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}{2}}$$

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{\left(1 - \sqrt{\frac{W_{bird,j}^2 + W_{tiger,j}^2}{2}}\right)^2 + \left(1 - \left(1 - \sqrt{\frac{W_{cat,j}^2 + W_{dog,j}^2}{2}}\right)\right)^2}{2}}$$

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{\left(1 - \sqrt{\frac{W_{bird,j}^2 + W_{tiger,j}^2}{2}}\right)^2 + \left(\sqrt{\frac{W_{cat,j}^2 + W_{dog,j}^2}{2}}\right)^2}{2}}$$

ขั้นตอนที่ 2

Query = "I don't love cat and dog. But I love tiger as well as bird"

Query = (Bird OR Tiger) AND NOT(Cat OR Dog)

	Bird	Cat	Dog	Tiger
Doc1	1.000	0.215	0.343	0.000
Doc2	0.000	0.322	0.257	0.257
Doc3	1.000	0.000	0.257	0.000
Doc4	0.000	0.322	0.000	0.515
Doc5	0.000	0.107	0.172	0.515
Doc6	1.000	0.215	0.000	0.343
Doc7	1.000	0.322	0.515	0.515
Doc8	1.000	0.322	0.515	0.000
Doc9	0.000	0.322	0.515	0.515
Doc10	0.000	0.000	0.000	0.515

$$sim(q_{and}, d_1) = 1 - \sqrt{\frac{\left(1 - \sqrt{\frac{W_{bird,1}^2 + W_{tiger,1}^2}{2}}\right)^2 + \left(\sqrt{\frac{W_{cat,1}^2 + W_{dog,1}^2}{2}}\right)^2}{2}}$$

$$sim(q_{and}, d_1) = 1 - \sqrt{\frac{\left(1 - \sqrt{\frac{1.000^2 + 0.000^2}{2}}\right)^2 + \left(\sqrt{\frac{0.215^2 + 0.343^2}{2}}\right)^2}{2}}$$

$$sim(q_{and}, d_1) = 0.710$$

ขั้นตอนที่ 3

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,tiger,cat,dog}
D3: {dog,bird,bird}
D4: {cat,tiger}
D5: {tiger,tiger,dog,tiger,cat}
D6: {bird,cat,bird,cat,tiger,tiger,bird}
D7: {bird,tiger,cat,dog}
D8: {dog,cat,bird}
D9: {cat,dog,tiger}
D10: {tiger,tiger,tiger}

Query = (Bird OR Tiger) AND NOT(Cat OR Dog)

	Sim
Doc1	0.710
Doc2	0.386
Doc3	0.756
Doc4	0.522
Doc5	0.539
Doc6	0.792
Doc7	0.664
Doc8	0.633
Doc9	0.457
Doc10	0.550

Ranking	Sim
Doc6	0.792
Doc3	0.756
Doc1	0.710
Doc7	0.664
Doc8	0.633
Doc10	0.550
Doc5	0.539
Doc4	0.522
Doc9	0.457
Doc2	0.386

Rank → Doc6,Doc3,Doc1,Doc7,Doc8,Doc10,Doc5,Doc4,Doc9,Doc2

ขั้นตอนที่ 3

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

```
D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,tiger,cat,dog}
D3: {dog,bird,bird}
D4: {cat,tiger}
D5: {tiger,tiger,dog,tiger,cat}
D6: {bird,cat,bird,cat,tiger,tiger,bird}
D7: {bird,tiger,cat,dog}
D8: {dog,cat,bird}
D9: {cat,dog,tiger}
D10: {tiger,tiger,tiger}
```

Query = (Bird OR Tiger) AND NOT(Cat OR Dog)

Ranking	Sim
Doc6	0.792
Doc3	0.756
Doc1	0.710
Doc7	0.664
Doc8	0.633
Doc10	0.550
Doc5	0.539
Doc4	0.522
Doc9	0.457
Doc2	0.386

Rank → Doc6,Doc3,Doc1,Doc7,Doc8,Doc10,Doc5,Doc4,Doc9,Doc2

2.3 หากระบบกำหนดให้เอกสารที่ 9 ตรงประเด็นมากกว่าเอกสารที่ 2 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย

จากผลลัพธ์ที่ได้เอกสาร 9 ตรงประเด็นมากกว่าเอกสารที่ 2 จึงเป็นไปตามที่โจทย์กำหนด ทั้งนี้ความต้องการของผู้เรียกค้นต้องการเอกสารที่มี Bird หรือ Tiger แต่ไม่ต้องการเอกสารที่ Cat หรือ มี Dog จะเห็นว่า เอกสารที่ 2 ปรากฏ Cat สองครั้ง และ Dog หนึ่งครั้ง ซึ่งมากกว่าเอกสารที่ 9 ที่ปรากฏ Cat และ Dog เพียงหนึ่งครั้ง ดังนั้นเอกสารที่ 9 จึงตรงประเด็นมากกว่า

ข้อ 3. เมื่อใช้งานคำเรียกค้น **Q = 7Dog-3Cat+Bird** ระบบส่งผลลัพธ์ออกมาคือเอกสาร 5 เอกสารดังนี้

D1: "Dog is a animal like to fight cat that same tiger. Cats eat fish. Dogs eat ham"

D2: "Birds fly over tiger. A Dog run to catch a Bird. A cat smile beside the river"

D3: "A tiger is crying for a lost bird. A cat is being hungry while the dog eat all fishs. That cat complain to its friend"

D4: "All birds sing the song. A dog is flighting the cat. The cat crying and go to complain with mother cat"

D5: "My dog wife is being stomach, A cat tell her boy friend because its want to have son child cat"

โดย **Dog,Cat,Tiger,Bird** คือ Keyword ของระบบ

หากต้องการปรับคำเรียกค้น (Query) ให้มีผลลัพธ์มีความตรงประเด็นมากยิ่งขึ้น โดยโจทย์กำหนดว่า**ห้ามใช้** Metric Cluster และ Scalar Cluster นักศึกษาควรใช้**โมเดลใดเพราะอะไร จงแสดงวิธีคำนวณ**ในโมเดลที่เลือกใช้ (23 คะแนน)

Answer

เลือกใช้ Associate Cluster เนื่องจากต้องการปรับเปลี่ยน Query ดังเดิม โดยใช้เอกสารที่ถูกส่งออกมาและพิจารณาเฉพาะความถี่ของ Keyword และโจทย์ไม่ต้องการให้ใช้วิธีอื่น

4.2

หาความสัมพันธ์ระหว่าง Keyword โดยพิจารณาความถี่ของ Keyword

	d_1	d_2	d_3	d_4	d_5
Bird	0	2	1	1	0
Cat	2	1	2	3	2
Dog	2	1	1	1	1
Tiger	1	1	1	0	0

$$C_{1,4} = (m_{1,1} * m^t_{1,4}) + (m_{1,2} * m^t_{2,4}) + (m_{1,3} * m^t_{3,4}) + (m_{1,4} * m^t_{4,4}) + (m_{1,5} * m^t_{5,4})$$

$$= 0 * 1 + 2 * 1 + 1 * 1 + 1 * 0 + 0 * 0$$

$$= 3$$

ความสัมพันธ์ระหว่าง Keyword

C	Bird	Cat	Dog	Tiger
Bird	6	7	4	3
Cat	7	22	12	5
Dog	4	12	8	4
Tiger	3	5	4	3

Normalize ความสัมพันธ์ระหว่างคีย์เวิร์ด

$$S_{u,v} = \frac{C_{u,v}}{C_{u,u} + Cv_{,v} - Cu_{,v}}$$

C	Bird	Cat	Dog	Tiger
Bird	6	7	4	3
Cat	7	22	12	5
Dog	4	12	8	4
Tiger	3	5	4	3

$$S_{1,3} = \frac{C_{1,3}}{C_{1,1} + C_{3,3} - C_{1,3}}$$

$$S_{1,3} = \frac{4}{6 + 8 - 4} = 0.40$$

Term Relation

S	Bird	Cat	Dog	Tiger
Bird	1.000	0.333	0.400	**0.500**
Cat	0.333	1.000	**0.667**	0.250
Dog	0.400	**0.667**	1.000	0.571
Tiger	**0.500**	0.250	0.571	1.000

1. {Bird,Tiger}
2. {Cat,Dog}
3. {Dog,Cat}
4. {Tiger,Bird}

S	Bird	Cat	Dog	Tiger
Bird	1.000	0.333	0.400	**0.500**
Cat	0.333	1.000	**0.667**	0.250
Dog	0.400	**0.667**	1.000	0.571
Tiger	**0.500**	0.250	0.571	1.000

Term Relation

1. *{Bird,Tiger}*
2. *{Cat,Dog}*
3. *{Dog,Cat}*
4. *{Tiger,Bird}*

Original Query

q = 7Dog-3Cat+Bird

New Query

q' = 7(Dog + 0.667Cat) - 3 (Cat + 0.667Dog) + (Bird+0.5Tiger)
= 4.332Dog – 0.669Cat + Bird + 0.5Tiger

ข้อ 4. สมมติบริษัทแห่งหนึ่งมีพนักงานทั้งหมด 10 คน (ชื่อ A,B,C,..,J) ซึ่งมีข้อมูลตามตารางด้านล่าง

Name	Address	Room#
A	Bangkok	101
B	Nonthaburi	201
C	Lopburi	202
D	Bangkok	102
E	Bangkok	103
F	Bangkok	104
G	Lopburi	203
H	Chiangmai	204
I	Nonthaburi	302
J	Bangkok	301

A เป็นโปรแกรมเมอร์มือหนึ่งของบริษัทและ A เป็นพนักงานที่กว้างขวางมีเพื่อนสนิทหลายคน คือ **B,D,E,I,J** ในเรื่องงาน A ได้สร้างระบบเรียกค้นขึ้น ซึ่งภายในประกอบด้วย 2 อัลกอริธึม ต่อมา A ต้องการทราบว่าอัลกอริธึมใดเหมาะสมกว่ากันเมื่อเรียกค้นพนักงานที่มีที่อยู่ใน Bangkok ซึ่งให้ผลลัพธ์ดังนี้

อัลกอริธึมที่ 1 ให้ผลลัพธ์ตามลำดับคือ **A,D,B,E,I,F,H**

อัลกอริธึมที่ 2 ให้ผลลัพธ์ตามลำดับคือ **J,F,A,C,D,E,B**

จากผลลัพธ์ที่ได้นี้ให้นักศึกษาคำนวณหาว่าอัลกอริธึมใดเหมาะสมกว่ากัน ด้วยตัวเลขเท่าใดบ้าง (ทุกวิธีที่สามารถประเมินได้) **(17 คะแนน)**

Answer

ข้อมูลที่ตรงประเด็นทั้งหมด 5 ข้อมูลคือ

A,D,E,F,J

1	A	D	B	E	I	F	H	AVG
Precision	1.00	1.00	0.67	0.75	0.60	0.67	0.57	**0.85**
Recall	0.20	0.40	0.40	0.60	0.60	0.80	0.80	-
F	0.33	0.57	0.50	0.67	0.60	0.73	0.67	**0.57**
E(β=2)	0.24	0.45	0.43	0.63	0.60	0.77	0.74	**0.52**

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

$$Coverage = \frac{R_k}{U}$$

$$Novelty = \frac{R_u}{R_u + R_k}$$

$R_k =$ *Relevant Docs known to the user which were retrieved*

$R_u =$ *Relevant Docs previously unknown to the user which were retrieved*

$U =$ *Relevant Docs known to the user*

A สนิทกับ B,D,E,I,J จึงทราบว่า A,D,E,J ตรงประเด็น

$$Coverage = \frac{R_k}{U} = \frac{\text{รู้ว่าตรงประเด็นแล้วออกมา}(A,D,E)}{\text{รู้ว่าตรงประเด็น}} = \frac{3}{4} = 0.75$$

$$Novelty = \frac{R_u}{R_u + R_k} = \frac{\text{ไม่รู้ว่าตรงประเด็นแล้วออกมา}(F)}{\text{ไม่รู้ว่าตรงประเด็นแล้วออกมา}(F) + \text{รู้ว่าตรงประเด็นแล้วออกมา}(A,D,E)} = \frac{1}{1+3} = 0.25$$

2	J	F	A	C	D	E	B	AVG
Precision	1.00	1.00	1.00	0.75	0.80	0.83	0.71	**0.93**
Recall	0.20	0.40	0.60	0.60	0.80	1.00	1.00	-
F	0.33	0.57	0.75	0.67	0.80	0.91	0.83	**0.67**
E(β=2)	0.24	0.45	0.65	0.63	0.80	0.96	0.93	**0.62**

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

$$E = \frac{(1+\beta^2)PR}{\beta^2 P+R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

$$Coverage = \frac{R_k}{U}$$

$$Novelty = \frac{R_u}{R_u + R_k}$$

$R_k =$ *Relevant Docs known to the user which were retrieved*

$R_u =$ *Relevant Docs previously unknown to the user which were retrieved*

$U =$ *Relevant Docs known to the user*

A สนิทกับ **B,D,E,I,J** จึงทราบว่า **A,D,E,J** ตรงประเด็น

$$Coverage = \frac{R_k}{U} = \frac{\text{รู้ว่าตรงประเด็นแล้วออกมา}(A,D,E,J)}{\text{รู้ว่าตรงประเด็น}} = \frac{4}{4} = 1.00$$

$$Novelty = \frac{R_u}{R_u + R_k} = \frac{\text{ไม่รู้ว่าตรงประเด็นแล้วออกมา}(F)}{\text{ไม่รู้ว่าตรงประเด็นแล้วออกมา}(F) + \text{รู้ว่าตรงประเด็นแล้วออกมา}(A,D,E,J)} = \frac{1}{1+4} = 0.20$$

อัลกอริธึมที่2 มีความตรงประเด็นมากกว่าอัลกอริธึมที่ 1 ด้วยเครื่องมือวัดทุกเครื่องมือ