ข้อ 2. สมมติเอกสารในระบบมีทั้งหมด 10 เอกสาร (bird, cat, dog, tiger คือ **Keyword** ซึ่งไม่มีความสัมพันธ์กัน)

D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,cat,cat,cat}
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}

เมื่อส่งคำเรียกค้น**” tiger dog cat dog cat”** เข้าไปในระบบ มีเอกสารจำนวน 8 เอกสารถูกส่งออกมาคือ **D5,D7,D2,D6,D9,D8,D10,D1** หากผู้เรียกค้นอนุมานว่าเอกสารที่ตอบออกมานี้ ถ้ามี Keyword ตามต้องการอย่างน้อย 2 ใน 3 ของที่ป้อนเข้าไปถือว่าตรงประเด็น จงตอบคำถาม

2.1 เพื่อคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ผู้เรียกค้นสามารถเลือกใช้โมเดลใดได้บ้าง เพราะอะไร (ตอบให้ครบทุกโมเดลที่เป็นไปได้ โดยเลือกจากโมเดลที่ให้มาเท่านั้น)

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

2.2 จากข้อ 2.1 ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ตามโมเดลที่ผู้เรียกค้นเลือก 1 โมเดล

2.3 หากระบบกำหนดให้เอกสารที่ 3 ตรงประเด็นมากกว่าเอกสารที่ 6 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย(35 คะแนน)

ข้อ 2. สมมติเอกสารในระบบมีทั้งหมด 10 เอกสาร (bird, cat, dog, tiger คือ **Keyword** ซึ่งไม่มีความสัมพันธ์กัน)

D1: {bird,cat,bird,cat,dog,dog,bird}
D2: {cat,cat,cat,cat}
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}

เมื่อส่งคำเรียกค้น **" tiger dog cat dog cat"** เข้าไปในระบบ มีเอกสารจำนวน 8 เอกสารถูกส่งออกมาคือ **D5,D7,D2,D6,D9,D8,D10,D1** หากผู้เรียกค้นอนุมานว่าเอกสารที่ตอบออกมานี้ ถ้ามี Keyword ตามต้องการอย่างน้อย 2 ใน 3 ของที่ป้อนเข้าไปถือว่าตรงประเด็น จงตอบคำถาม

2.1 เพื่อคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ผู้เรียกค้นสามารถเลือกใช้โมเดลใดได้บ้าง เพราะอะไร (ตอบให้ครบทุกโมเดลที่เป็นไปได้ โดยเลือกจากโมเดลที่ให้มาเท่านั้น)

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

2.2 จากข้อ 2.1 ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ตามโมเดลที่ผู้เรียกค้นเลือก 1 โมเดล

2.3 หากระบบกำหนดให้เอกสารที่ 3 ตรงประเด็นมากกว่าเอกสารที่ 6 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย(30 คะแนน)

Answer

2.1 เลือกใช้ BM25 Model เนื่องจากลักษณะของ Query เป็น keyword แยกกัน ไม่มี Expression นอกจากนี้โจทย์กำหนดให้ Keyword ไม่สัมพันธ์กัน และได้กำหนดเอกสารที่ตรงประเด็น

BM25 1

Query = *tiger dog cat dog cat*

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}
```

$$sim\ (d_j, q) = \sum_{i \epsilon q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1) f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl}\right) + f_i} \cdot \frac{(k_2 + 1) qf_i}{k_2 + qf_i}$$

d_j - เอกสารที่ j

R – จำนวนเอกสารที่ตรงประเด็น

N – จำนวนเอกสารทั้งหมด

r_i – จำนวนเอกสารที่ตรงประเด็นที่มี keyword i

n_i – จำนวนเอกสารทั้งหมดที่มี keyword i

f_i – ความถี่ของ keyword i ในเอกสาร j

dl – จำนวนคำของเอกสาร j

avdl – จำนวนคำเฉลี่ยของทุกเอกสาร

qf_i – ความถี่ของ keyword i ใน query

b – ค่าคงที่โดยตาม TREC จะใช้ค่า 0.75 ($0.5 < b < 0.8$)

k_1 – ค่าคงที่โดยตาม TREC จะใช้ค่า 1.25 ($1.2 < k_1 < 2$)

k_2 – ค่าคงที่โดยปกติจะอยู่ในช่วง 0 - 1000

BM25 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}
```

Query = *tiger dog cat dog cat*

	Bird	Cat	Dog	Tiger	Length
Doc1	*3*	*2*	*2*	*0*	7
~~*Doc2*~~	~~0~~	~~4~~	~~0~~	~~0~~	~~4~~
Doc3	**2**	**0**	**1**	**0**	**3**
Doc4	**0**	**1**	**0**	**1**	**2**
Doc5	*0*	*1*	*1*	*3*	5
Doc6	*3*	*2*	*0*	*2*	7
Doc7	*1*	*1*	*1*	*1*	4
Doc8	*1*	*1*	*1*	*0*	*3*
Doc9	*0*	*1*	*1*	*1*	3
~~*Doc10*~~	~~0~~	~~0~~	~~0~~	~~3~~	~~3~~

$$Avdl = \frac{41}{10} = 4.1$$

$N = 10$ $\quad R = 6$

$n_{Bird} = 5$ $\quad r_{Bird} = 4$

$n_{Cat} = 8$ $\quad r_{Cat} = 6$

$n_{Dog} = 6$ $\quad r_{Dog} = 5$

$n_{Tiger} = 6$ $\quad r_{Tiger} = 4$

R R R R R R

เอกสารที่ส่งออกมา **D5,D7,D2,D6,D9,D8,D10,D1**

BM25 2

Query = *tiger dog cat dog cat*

	Bird	Cat	Dog	Tiger
Doc1	*3*	*2*	*2*	*0*
~~*Doc2*~~	~~*0*~~	~~*4*~~	~~*0*~~	~~*0*~~
Doc3	**2**	**0**	**1**	**0**
Doc4	**0**	**1**	**0**	**1**
Doc5	*0*	*1*	*1*	*3*
Doc6	*3*	*2*	*0*	*2*
Doc7	*1*	*1*	*1*	*1*
Doc8	*1*	*1*	*1*	*0*
Doc9	*0*	*1*	*1*	*1*
~~*Doc10*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~*3*~~

$$idf_i = log\frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}$$

$$idf_{bird} = log\frac{(4 + 0.5)/(6 - 4 + 0.5)}{(5 - 4 + 0.5)/(10 - 5 - 6 + 4 + 0.5)} = 0.623$$

$$idf_{cat} = log\frac{(6 + 0.5)/(6 - 6 + 0.5)}{(8 - 6 + 0.5)/(10 - 8 - 6 + 6 + 0.5)} = 1.114$$

$$idf_{dog} = log\frac{(5 + 0.5)/(6 - 5 + 0.5)}{(6 - 5 + 0.5)/(10 - 6 - 6 + 5 + 0.5)} = 0.932$$

$$idf_{tiger} = log\frac{(4 + 0.5)/(6 - 4 + 0.5)}{(6 - 4 + 0.5)/(10 - 6 - 6 + 4 + 0.5)} = 0.255$$

$N = 10$ $\quad R = 6$

$n_{Bird} = 5$ $\quad r_{Bird} = 4$

$n_{Cat} = 8$ $\quad r_{Cat} = 6$

$n_{Dog} = 6$ $\quad r_{Dog} = 5$

$n_{Tiger} = 6$ $\quad r_{Tiger} = 4$

$\boldsymbol{Avdl} = 4.1$

BM25 2

Query = *tiger dog cat dog cat*

d_j - เอกสารที่ j
R – จำนวนเอกสารที่ตรงประเด็น
N – จำนวนเอกสารทั้งหมด
r_i – จำนวนเอกสารที่ตรงประเด็นที่มี keyword i
n_i – จำนวนเอกสารทั้งหมดที่มี keyword i
f_i – ความถี่ของ keyword i ในเอกสาร j
dl – จำนวนคำของเอกสาร j
avdl – จำนวนคำเฉลี่ยของทุกเอกสาร
qf_i – ความถี่ของ keyword i ใน query
b – ค่าคงที่โดยตาม TREC จะใช้ค่า 0.75 ($0.5 < b < 0.8$)
k_1 – ค่าคงที่โดยตาม TREC จะใช้ค่า 1.25 ($1.2 < k_1 < 2$)
k_2 – ค่าคงที่โดยปกติจะอยู่ในช่วง 0 - 1000

	idf
Bird	0.623
Cat	1.114
Dog	0.932
Tiger	0.255

	Bird	Cat	Dog	Tiger
Doc1	*3*	*2*	*2*	*0*
~~*Doc2*~~	~~*0*~~	~~*4*~~	~~*0*~~	~~*0*~~
Doc3	**2**	**0**	**1**	**0**
Doc4	**0**	**1**	**0**	**1**
Doc5	*0*	*1*	*1*	*3*
Doc6	*3*	*2*	*0*	*2*
Doc7	*1*	*1*	*1*	*1*
Doc8	*1*	*1*	*1*	*0*
Doc9	*0*	*1*	*1*	*1*
~~*Doc10*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~*3*~~

$$sim\ (d_j, q) = \sum_{i \epsilon q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1) f_i}{k_1\left((1-b) + b \cdot \frac{dl}{avdl}\right) + f_i} \cdot \frac{(k_2 + 1) qf_i}{k_2 + qf_i}$$

$$sim\ (d_1, q) = 0.623 * \frac{(2.25)3}{1.25\left((1-0.75) + 0.75 * \frac{7}{4.1}\right) + 3} * \frac{201 * 0}{200 + 0} + 1.114 * \frac{(2.25)2}{1.25\left((1-0.75) + 0.75 * \frac{7}{4.1}\right) + 2} * \frac{201 * \mathbf{2}}{200 + \mathbf{2}}$$

$$+0.932 * \frac{(2.25)2}{1.25\left((1-0.75) + 0.75 * \frac{7}{4.1}\right) + 2} * \frac{201 * \mathbf{2}}{200 + \mathbf{2}} + 0.255 * \frac{(2.25)0}{1.25\left((1-0.75) + 0.75 * \frac{7}{4.1}\right) + 0} * \frac{201 * \mathbf{1}}{200 + \mathbf{1}}$$

$$= \mathbf{4.683}$$

BM25 2

Query = *tiger dog cat dog cat*

	Sim
Doc1	4.683
Doc2	3.374
Doc3	1.433
Doc4	1.909
Doc5	3.496
Doc6	2.843
Doc7	3.342
Doc8	3.145
Doc9	3.342
Doc10	0.351

Rank →

	Sim
Doc1	4.683
Doc5	3.496
Doc2	3.374
Doc7	3.342
Doc9	3.342
Doc8	3.145
Doc6	2.843
Doc4	1.909
Doc3	1.433
Doc10	0.351

	Bird	Cat	Dog	Tiger
Doc1	*3*	*2*	*2*	*0*
~~Doc2~~	~~0~~	~~4~~	~~0~~	~~0~~
Doc3	2	0	1	0
Doc4	0	1	0	1
Doc5	*0*	*1*	*1*	*3*
Doc6	*3*	*2*	*0*	*2*
Doc7	*1*	*1*	*1*	*1*
Doc8	*1*	*1*	*1*	*0*
Doc9	*0*	*1*	*1*	*1*
~~Doc10~~	~~0~~	~~0~~	~~0~~	~~3~~

สรุป

Query = *tiger dog cat dog cat*

	Sim
Doc1	4.683
Doc5	3.496
Doc2	3.374
Doc7	3.342
Doc9	3.342
Doc8	3.145
Doc6	2.843
Doc4	1.909
Doc3	1.433
Doc10	0.351

	Bird	Cat	Dog	Tiger
Doc1	*3*	*2*	*2*	*0*
~~*Doc2*~~	~~*0*~~	~~*4*~~	~~*0*~~	~~*0*~~
Doc3	2	0	1	0
Doc4	0	1	0	1
Doc5	*0*	*1*	*1*	*3*
Doc6	*3*	*2*	*0*	*2*
Doc7	*1*	*1*	*1*	*1*
Doc8	*1*	*1*	*1*	*0*
Doc9	*0*	*1*	*1*	*1*
~~*Doc10*~~	~~*0*~~	~~*0*~~	~~*0*~~	~~*3*~~

2.3 หากระบบกำหนดให้เอกสารที่ 3 ตรงประเด็นมากกว่าเอกสารที่ 6 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย

จากการคำนวณเอกสาร **3** ตรงประเด็นน้อยกว่าเอกสาร **6** ทั้งนี้เนื่องจากการเรียกค้นต้องการ **Cat Dog Tiger** แต่เอกสาร **3** มี **Bird Dog** แต่เอกสาร **6** มี **Bird Cat Dog Tiger** ดังนั้นในความเป็นจริงเอกสาร **6** จึงต้องประเด็นมากกว่า ซึ่งตามที่โจทย์กำหนดมาจึงคลาดเคลื่อนกับความเป็นจริง หากต้องการให้เอกสาร **3** ตรงประเด็นมากกว่า จะต้องเปลี่ยนคำเรียกค้นให้มี **Bird** จึงจะเป็นจริงตามโจทย์