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 tiger dog cat"** เข้าไปในระบบ มีเอกสารจำนวน 7 เอกสารถูกส่งออกมาคือ **D5,D7,D2,D6,D9,D1,D10** หากผู้เรียกค้นอนุมานว่าเอกสารที่ตอบออกมานี้ ถ้ามี Keyword ตามต้องการอย่างน้อย 2 ใน 3 ของที่ป้อนเข้าไปถือว่าตรงประเด็น จงตอบคำถาม

2.1 เพื่อคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ผู้เรียกค้น**สามารถเลือกใช้โมเดลใดได้บ้าง 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 9 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 tiger dog cat"** เข้าไปในระบบ มีเอกสารจำนวน 7 เอกสารถูกส่งออกมาคือ **D5,D7,D2,D6,D9,D1,D10** หากผู้เรียกค้นอนุมานว่าเอกสารที่ตอบออกมานี้ ถ้ามี Keyword ตามต้องการอย่างน้อย 2 ใน 3 ของที่ป้อนเข้าไปถือว่าตรงประเด็น จงตอบคำถาม

2.1 เพื่อคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ผู้เรียกค้น**สามารถเลือกใช้โมเดลใดได้บ้าง 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 9 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย(35 **คะแนน**)

Answer

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

BM25 1

Query = `tiger dog tiger 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 tiger 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

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

$N = 10$ $\quad R = 5$

$n_{Bird} = 5$ $\quad r_{Bird} = 3$

$n_{Cat} = 8$ $\quad r_{Cat} = 5$

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

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

R R R R R

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

BM25 2

Query = `tiger dog tiger dog cat`

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

$$idf_{bird} = log \frac{(3 + 0.5)/(5 - 3 + 0.5)}{(5 - 3 + 0.5)/(10 - 5 - 5 + 3 + 0.5)} = 0.292$$

$$idf_{cat} = log \frac{(5 + 0.5)/(5 - 5 + 0.5)}{(8 - 5 + 0.5)/(10 - 8 - 5 + 5 + 0.5)} = 0.895$$

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

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

	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

$N = 10$ $R = 5$

$n_{Bird} = 5$ $r_{Bird} = 3$

$n_{Cat} = 8$ $r_{Cat} = 5$

$n_{Dog} = 6$ $r_{Dog} = 4$

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

$\boldsymbol{Avdl} = 4.1$

BM25 2

Query = tiger dog tiger 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	0.292
Cat	0.895
Dog	0.623
Tiger	0.623

	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

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

$$sim\ (d_1, q) = 0.292 * \frac{(2.25)3}{1.25\left((1 - 0.75) + 0.75 * \frac{7}{4.1}\right) + 3} * \frac{201 * 0}{200 + 0} + 0.895 * \frac{(2.25)2}{1.25\left((1 - 0.75) + 0.75 * \frac{7}{4.1}\right) + 2} * \frac{201 * 1}{200 + 1}$$

$$+0.623 * \frac{(2.25)2}{1.25\left((1 - 0.75) + 0.75 * \frac{7}{4.1}\right) + 2} * \frac{201 * 2}{200 + 2} + 0.623 * \frac{(2.25)0}{1.25\left((1 - 0.75) + 0.75 * \frac{7}{4.1}\right) + 0} * \frac{201 * 2}{200 + 2}$$

$$= 2.456$$

BM25 2

Query = tiger dog tiger dog cat

	Sim
Doc1	2.456
Doc2	1.363
Doc3	0.958
Doc4	1.649
Doc5	3.354
Doc6	2.456
Doc7	2.607
Doc8	1.649
Doc9	2.607
Doc10	1.704

Rank →

	Sim
Doc5	3.354
Doc7	2.607
Doc9	2.607
Doc1	2.456
Doc6	2.456
Doc10	1.704
Doc4	1.649
Doc8	1.649
Doc2	1.363
Doc3	0.958

	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
q	*0*	*1*	*2*	*2*

สรุป

Query = tiger dog tiger dog cat

	Sim
Doc5	3.354
Doc7	2.607
Doc9	2.607
Doc1	2.456
Doc6	2.456
Doc10	1.704
Doc4	1.649
Doc8	1.649
Doc2	1.363
Doc3	0.958

	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
q	*0*	*1*	*2*	*2*

2.3 หากระบบกำหนดให้เอกสารที่ 3 ตรงประเด็นมากกว่าเอกสารที่ 9 โมเดลที่เลือกมาให้คำตอบถูกต้องหรือไม่ ถ้าผิดต้องแก้ไขอย่างไรจงอธิบาย

จากการคำนวณเอกสาร 3 ตรงประเด็นน้อยกว่าเอกสาร 9 ทั้งนี้เนื่องจากการเรียกค้นต้องการ Cat Dog Tiger แต่เอกสาร 3 มี Bird Dog แต่เอกสาร 9 มี Cat Dog Tiger ดังนั้นในความเป็นจริง เอกสาร 9 จึงต้องประเด็นมากกว่า ซึ่งตามที่โจทย์กำหนดมาจึงคลาดเคลื่อนกับความเป็นจริง หากต้องการให้เอกสาร 3 ตรงประเด็นมากกว่า จะต้องเปลี่ยนคำเรียกค้นให้มี Bird จึงจะเป็นจริงตามโจทย์

ข้อ 3. เอกสารหนึ่งมีข้อความดังนี้

```
Love is a variety of different feelings, states, and attitudes
that ranges from interpersonal affection ("I love my mother")
to pleasure ("I loved that meal"). It can refer to an emotion
of a strong attraction and personal attachment.It can also be a
virtue representing human kindness, compassion, and affection—
"the unselfish loyal and benevolent concern for the good of
another". (I love you my student)
```

3.1 เพื่อให้ได้คำตอบในข้อ 3.2 นักศึกษาควรใช้ Model ใดเพราะอะไร (เลือกได้เฉพาะตัวเลือกที่ให้มา)

A) Knuth-Morris-Pratt (KMP) **B)** Breadth-first search (BFS) **C)** Depth-first search (DFS) **D)** Boyer Moor

3.2 หากคำว่า ”love” **หากพบ** ”Love” **หรือ** ”love” **ถือว่าตรงประเด็น**
จงแสดงการค้นหาคำว่า ”love” **ในเอกสารด้านบนอย่างละเอียด โดยแสดงผลตำแหน่งที่ปรากฏคำที่เรียกค้นนี้**
(20 คะแนน)

Answer

3.1 เลือกใช้ Boyer Moor **หรือ** Knuth-Morris-Pratt (KMP) ได้ทั้งสองวิธี ซึ่งเป็นกระบวนการค้นหาข้อมูลตามลำดับ (Sequential Search)

3.2

1 11 22 32 42 54
Love is a variety of different feelings, states, and attitudes
68 80 94 108 116 128
that ranges from interpersonal affection ("I love my mother") to pleasure ("I
141 152
loved that meal").
166 178 191 198 213 222
It can refer to an emotion of a strong attraction and personal attachment.
250 257 270 276 286 302
It can also be a virtue representing human kindness, compassion, and affection—
317 327 337 348 364 372 384
"the unselfish loyal and benevolent concern for the good of another".(I love you
my student)

Boyer Moore

Value = length – index – 1

l = 4-0-1=3

o = 4-1-1 =2

v = 4-2-1 = 1

e = 4

Letter	l	o	v	e	*
Value	3	2	1	4	4

3.2

0 1

lovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelov
Love is a variety 2f different feelings, states, 3 and attitudes 4 5
elovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovel
that ranges from interpersonal affection ("I *love* my mother") to pleasure ("I
ove lovelovelovelo
*love*d that meal").
velovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelove
It can refer to an emotion of a strong attraction and personal attachment.
lovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelov
It can 6 also be 7 a virtue 8 representing 10 human kindness, compassion, and affection— 11
elovelovelovelove lovelovelove 9 lovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovelovel

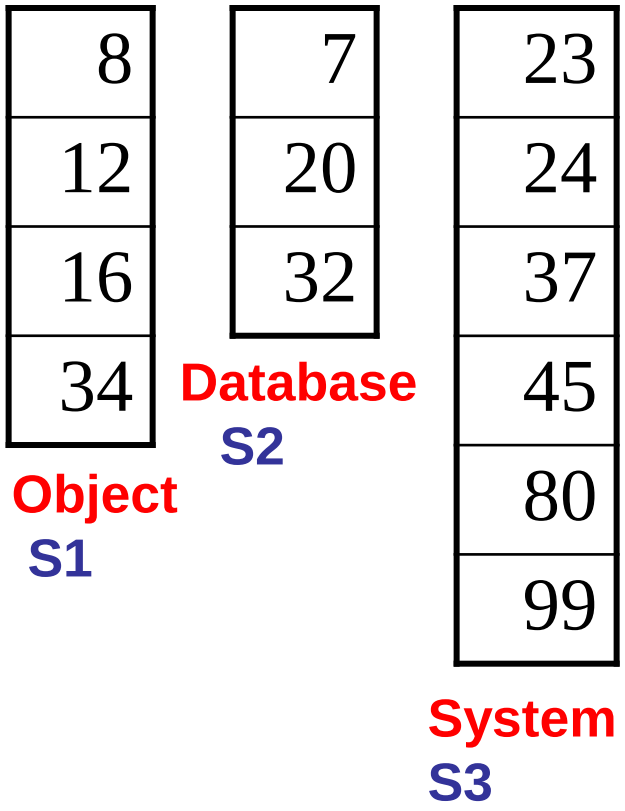
3.2

Item	Current Position	Match	Next Move Position	Move By	Found At Position
0	1	Y	5		1
1	25		29	e	
2	80		84	e	
3	108	Y	112	e	108
4	132		136	e	
5	140		141	v	
	141	Y	145	e	141
6	317		321	e	
7	325		327	o	
8	335		339	e	
9	339		341	o	
	341		345	e	
10	349		353	e	
11	382		403	l	
	385	Y	389	e	385

Letter	l	o	v	e	*
Value	3	2	1	4	4

ข้อ 4. Phrasal Search คือการค้นหาวลีในระบบ เช่น ค้นหาวลี Object Database System ซึ่งจะค้นหาเอกสารที่มีวลีดังกล่าวตอบสนองให้กับ User ให้นักศึกษาเขียนอัลกอริธึมของ Phrasal Search ที่ค้นหาข้อมูล โดยเอกสารที่ตรงประเด็น สามารถมี Keyword ที่ต้องการ อยู่ห่างกันสูงสุด 8 คำได้ พร้อมยกตัวอย่างประกอบด้วย **(15 คะแนน)**

Query : Object Database System



P-S+i → **P-8S+8i**

S=2 (Shortest Array)

Round 1

P=7

i=1→ P-8S+8i

→ 7-16+8 = -1 (1..6) (Not found in S1)

Round 2

P=20

i=1→ P-8S+8i

→ 20-16+8 = 12 (12..19) (Found in S1)

i=3→ P-8S+8i

→ 20-16+24 = 28 (21..28) (Found in S3)

Add this document to Answer Set