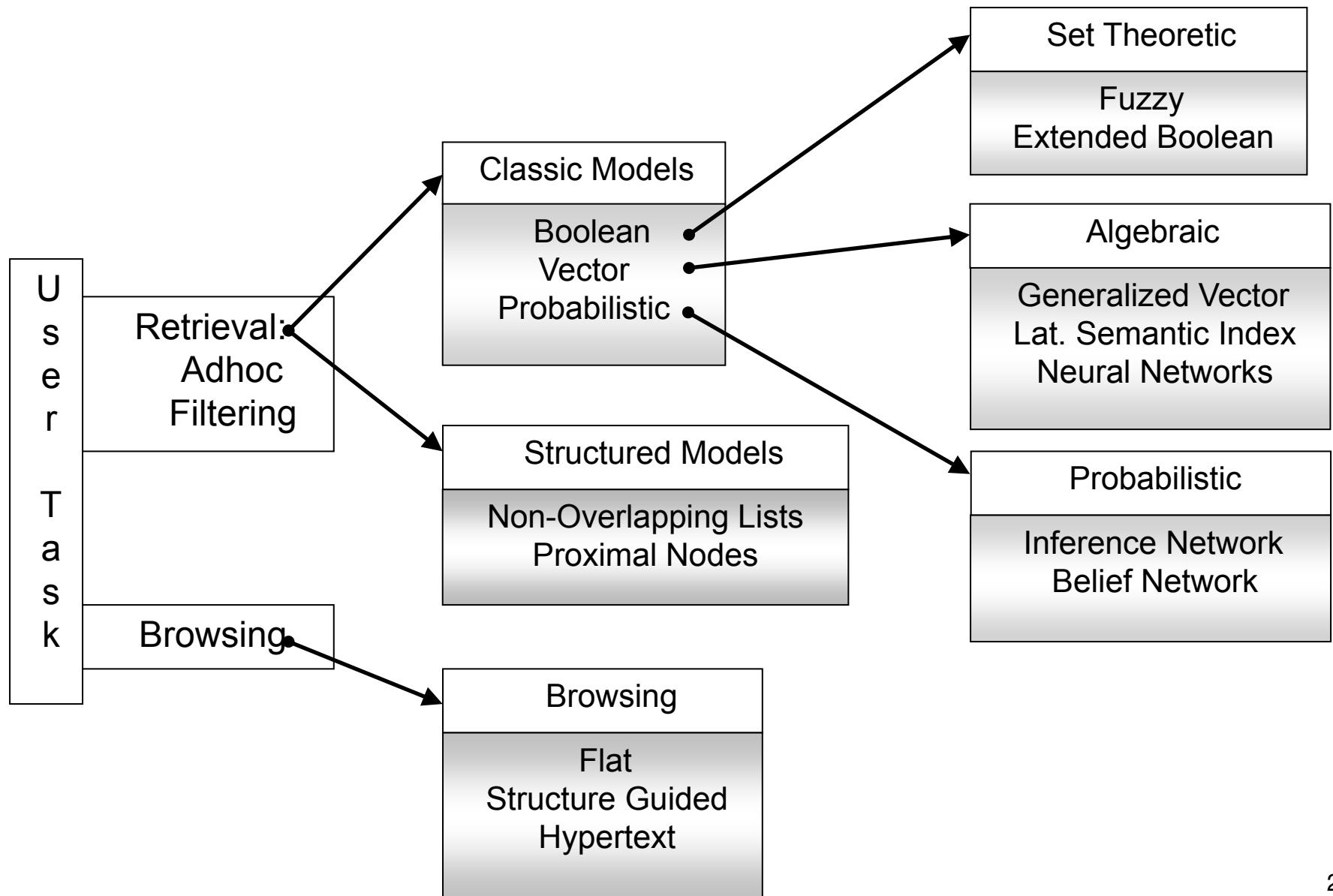


---

# **Chapter 02**

# **Modeling**

# IR Models



# Fuzzy logic

## Fuzzy Logic

“ทุกสิ่งบนโลกแท้จริงเป็นชิ้นเดียว มีได้มีเดียวซึ่งก็มีความแน่นอนแท้แน่น แต่มีหลายชิ้น หลายอย่าง หลายเหตุการณ์เกิดขึ้นอย่างไม่เที่ยงคง อาจเป็นสิ่งที่คุณแคร์และไม่แท้แน่น”

# เซตแบบตันฉบับ (Crisp Set)

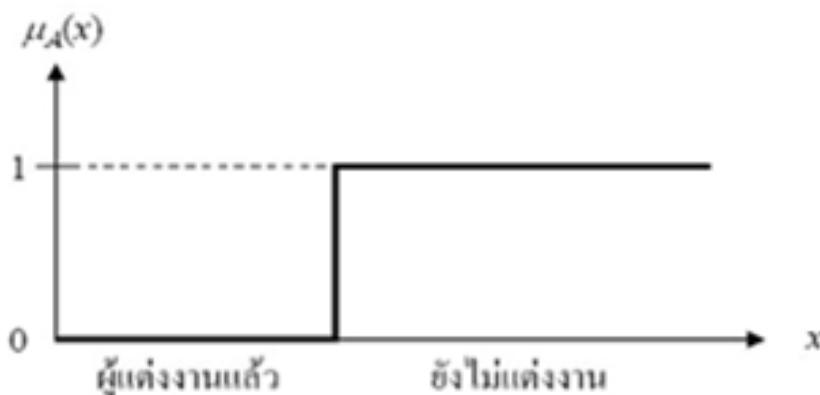
- กำหนดค่าความเป็นสมาชิกตามแนวคิดเลขฐานสอง
- เซตที่มีค่าความเป็นสมาชิกเป็น 0 หรือ 1 เท่านั้น
- ขอบเขตของเซตที่ตัดขาดจากกันแบบทันทีทันใด ไม่มีความต่อเนื่อง

$$\mu_A(x) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

# เซตแบบตื้นฉบับ (ต่อ)



ตัวอย่าง Crisp Set (Classical Set)



ตัวอย่างการแสดงค่าความเป็นสมาชิกของผู้ที่ยังไม่ได้เด็งงาน

# Fuzzy Logic

---

“ทุกสิ่งบนโลกแห่งความเป็นจริง มีได้มีเสพะสิ่งที่มีความแน่นอนเท่านั้น แต่มีหลายสิ่ง หลายอย่างหลายเหตุการณ์เกิดขึ้นอย่างไม่เที่ยงตรง อาจเป็นสิ่งที่คุณไม่รู้และไม่แน่นอน”

# Fuzzy Set

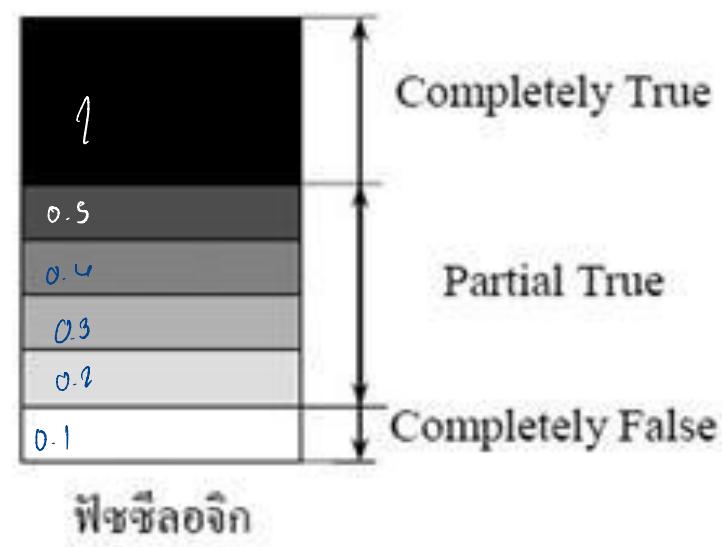
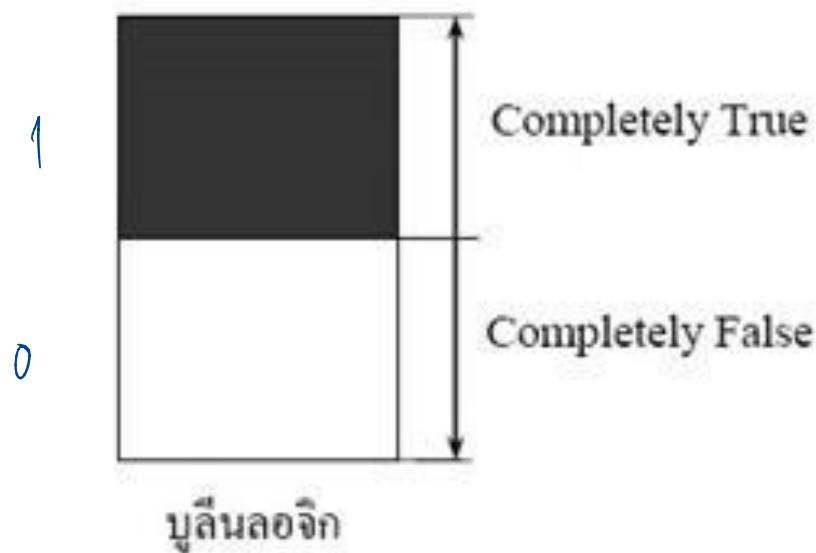
---

## Example

เซตของอายุคน อาจแบ่งเป็น วัยทารก วัยเด็ก วัยรุ่น วัยกลางคนและวัยชรา แต่ละช่วงอายุคนไม่สามารถระบุได้แน่ชัดว่าวัยทารกับวัยเด็กแยกจากกันแน่ชัดช่วงใด วัยทารกอาจถูกตีความว่าเป็นอายุระหว่าง 0 ถึง 1 ปี บางคนอาจตีความว่าวัยทารกอยู่ในช่วงอายุ 0 ถึง 2 ปี เชตของเหตุการณ์ที่ไม่แน่นอนเช่นนี้เรียกว่า “ฟูซซีเซต” (Fuzzy Set)

# Fuzzy Logic

- ฟิชชีล็อกิกมีลักษณะที่พิเศษกว่าตรรกศาสตร์แบบจริงเท็จ (Boolean logic) โดยมีการเพิ่มแนวคิด **ความจริงบางส่วน (partial true)** เข้ามา ซึ่งจะมีค่าความจริงอยู่ในช่วงระหว่างจริง (completely true) กับเท็จ (completely false)



# Fuzzy set

- ถูกนำเสนอในรูปแบบของค่าของเขตที่คลุมเครือ
- ยอมให้มีค่าความเป็นสมาชิกของเซตมีได้มากกว่า 2 ค่า
  - ความเป็นสมาชิกจะมีค่าระหว่าง  $0 - 1$  **[0,1]** ลักษณะ เช่น  $[1,1)$   
โดย  $0$  หมายถึง ไม่มีความเป็นสมาชิกเลย  
 $1$  หมายถึง ความเป็นสมาชิกโดยสมบูรณ์  
ระหว่าง  $0 - 1$  หมายถึง ความเป็นสมาชิกแค่บางส่วน  
ดังนั้น ความเป็นสมาชิกของพื้นที่จะมีความต่อเนื่อง

# Fuzzy set

ฟิชชี่เซต A ของเอกภพ U ถูกกำหนดโดยฟังก์ชันความเป็นสมาชิก

Ex. ทุกคน

ทุกคน = 1  
คนเล็กทุกคน = 0.1 หมายความว่าคนที่มากกับทุกคน แต่ไม่ใช่ทุกคน

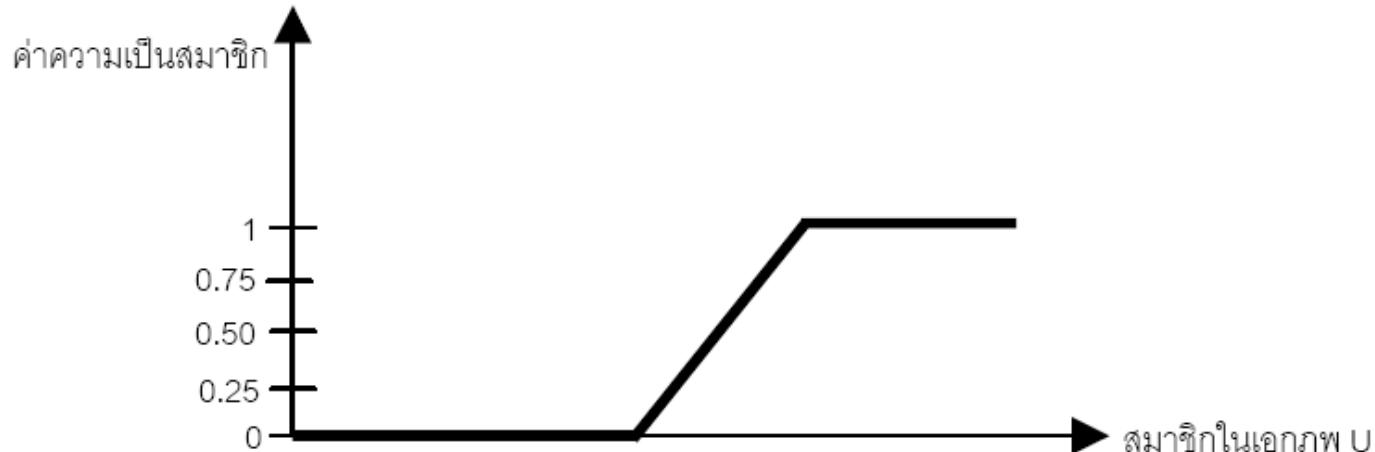
$$\mu_A: U \rightarrow [0,1]$$

กำหนดสมาชิกแต่ละตัว n ของเอกภพ คือ

$$\mu_A(u)$$

โดยมีค่าอยู่ในช่วง

[0,1]



กราฟแสดงความเป็นสมาชิก

Fuzzy ก็คือ keyword ที่ไม่แน่นอนในเอกสารใดๆ นิยามสัมภ์ที่นักเขียน

## Example

Query = “**cat and dog**”

$d_1 = \{dog, cat, bird, zebra, zoo\}$  เกี่ยวกับ สวนสัตว์

$d_2 = \{cat, kitty, fish\}$  เกี่ยวกับ แมว

$d_3 = \{dog, puppy, house, robber\}$  ความเมือง

$d_4 = \{ant, sugar\}$  เกี่ยวกับ - หัวใจ

Boolean =  $d_1$

Fuzzy =  $d_1, d_2, d_3$

# Fuzzy set

ตัวดำเนินการหลัก ๆ ของฟูซซี่เซต

Complement = NOT

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

Intersection = AND

$$\mu_{A \cap B} = \mu_A(x) \wedge \mu_B(x) = \min(\mu_A(x), \mu_B(x))$$

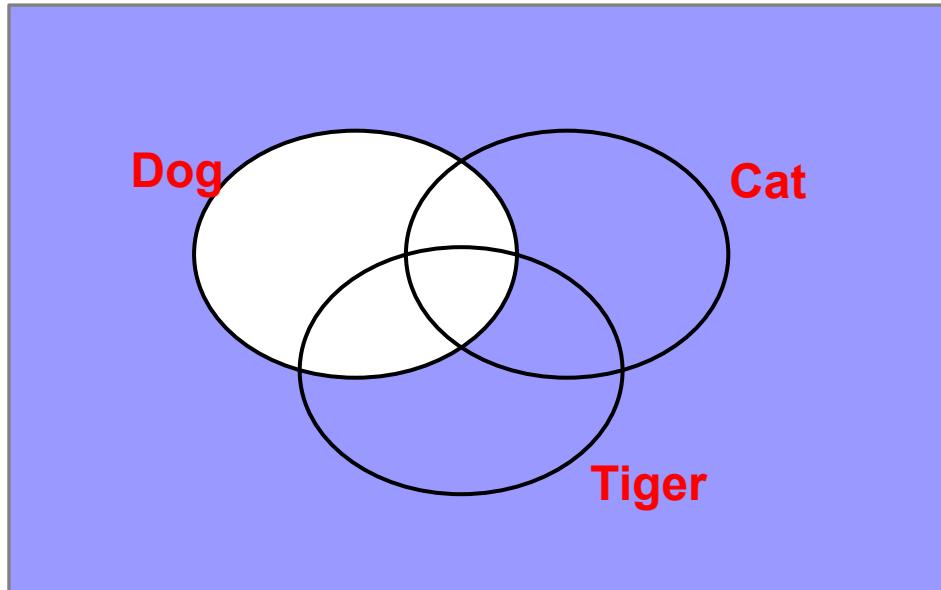
Union = OR

$$\mu_{A \cup B} = \mu_A(x) \vee \mu_B(x) = \max(\mu_A(x), \mu_B(x))$$

# Fuzzy Logic

---

## Complement

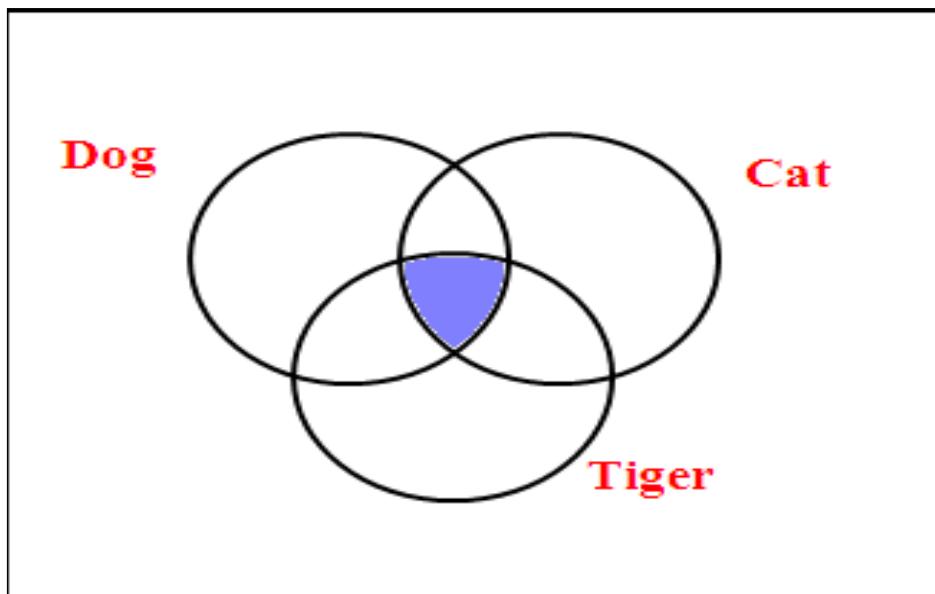


$$\mu_{\overline{Dog}}(x) = (1 - \mu_{Dog}(x))$$

# Fuzzy Logic

## Intersection

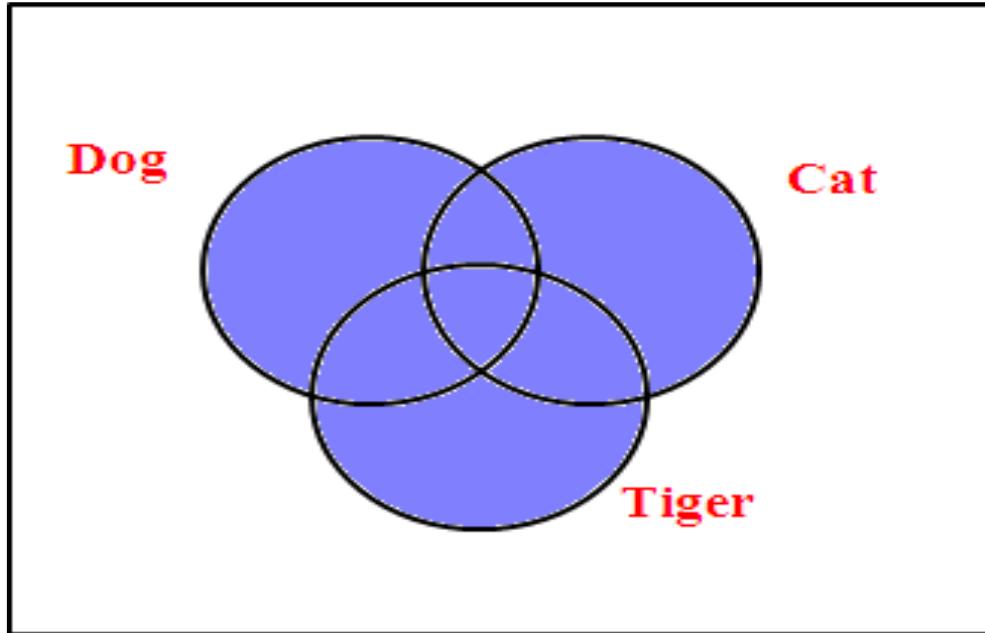
$\cap$  = अनुभव



$$\mu_{Dog \cap Cat \cap Tiger}(x) = \mu_{Dog}(x) \cdot \mu_{Cat}(x) \cdot \mu_{Tiger}(x)$$

# Fuzzy Logic

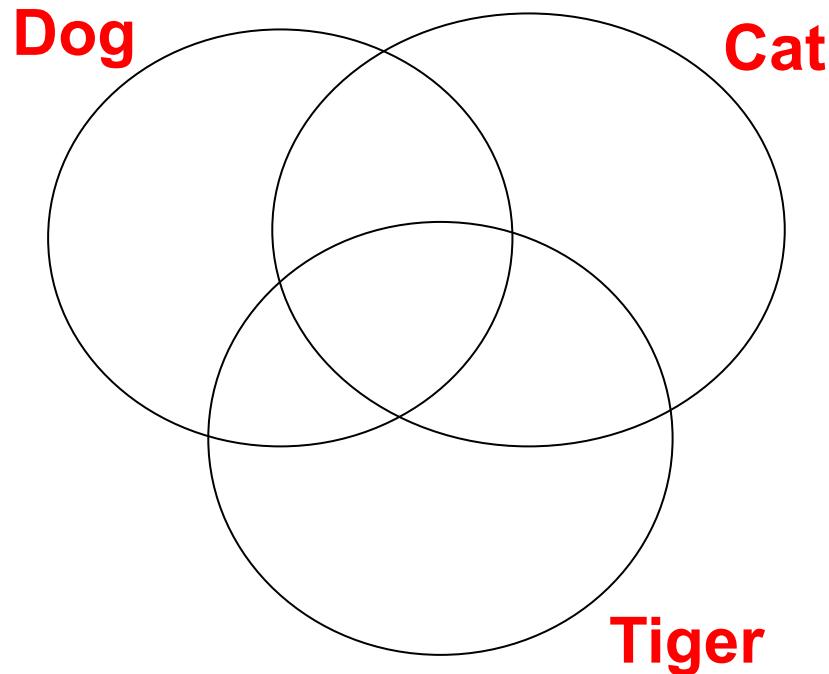
## Union



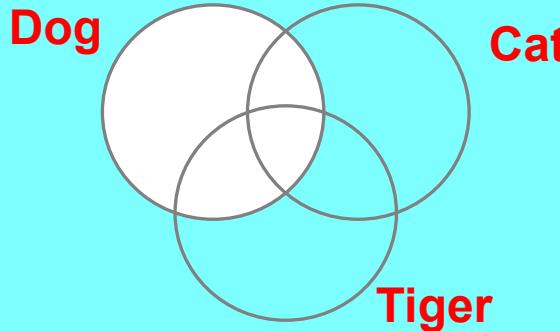
$$\mu_{Dog \cup Cat \cup Tiger}(x) = 1 - (1 - \mu_{Dog}(x)).(1 - \mu_{Cat}(x)).(1 - \mu_{Tiger}(x))$$

# Example

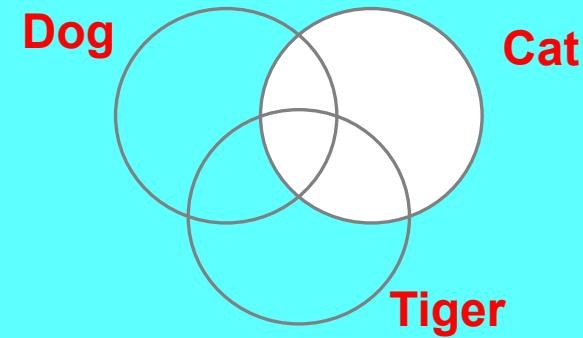
---



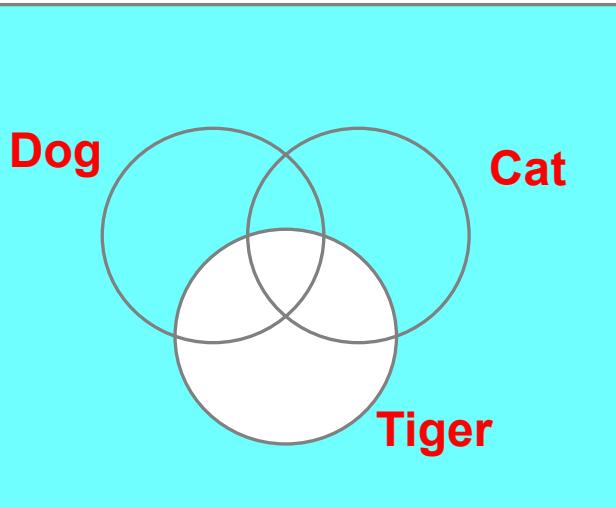
**Dog OR Cat OR Tiger**



$$1 - \mu_{Dog}(x)$$

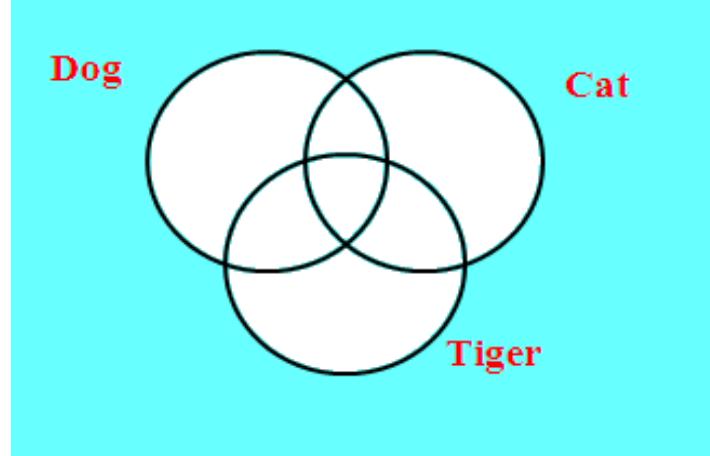


$$1 - \mu_{Cat}(x)$$



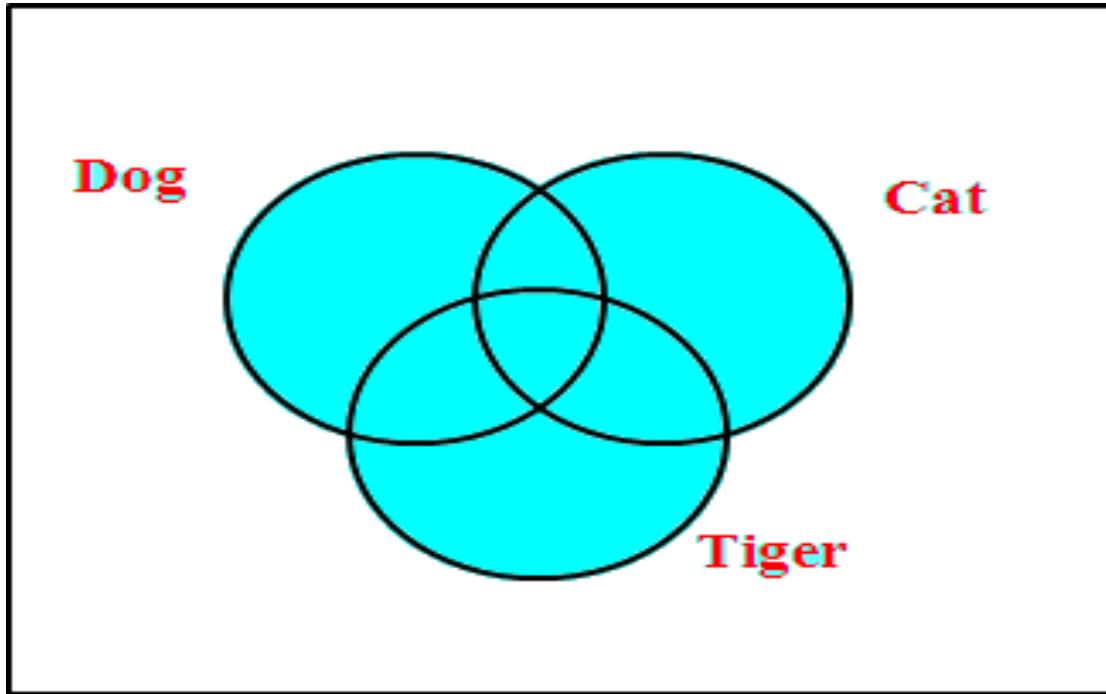
$$1 - \mu_{Tiger}(x)$$

ສັນນະ AND ກັນ



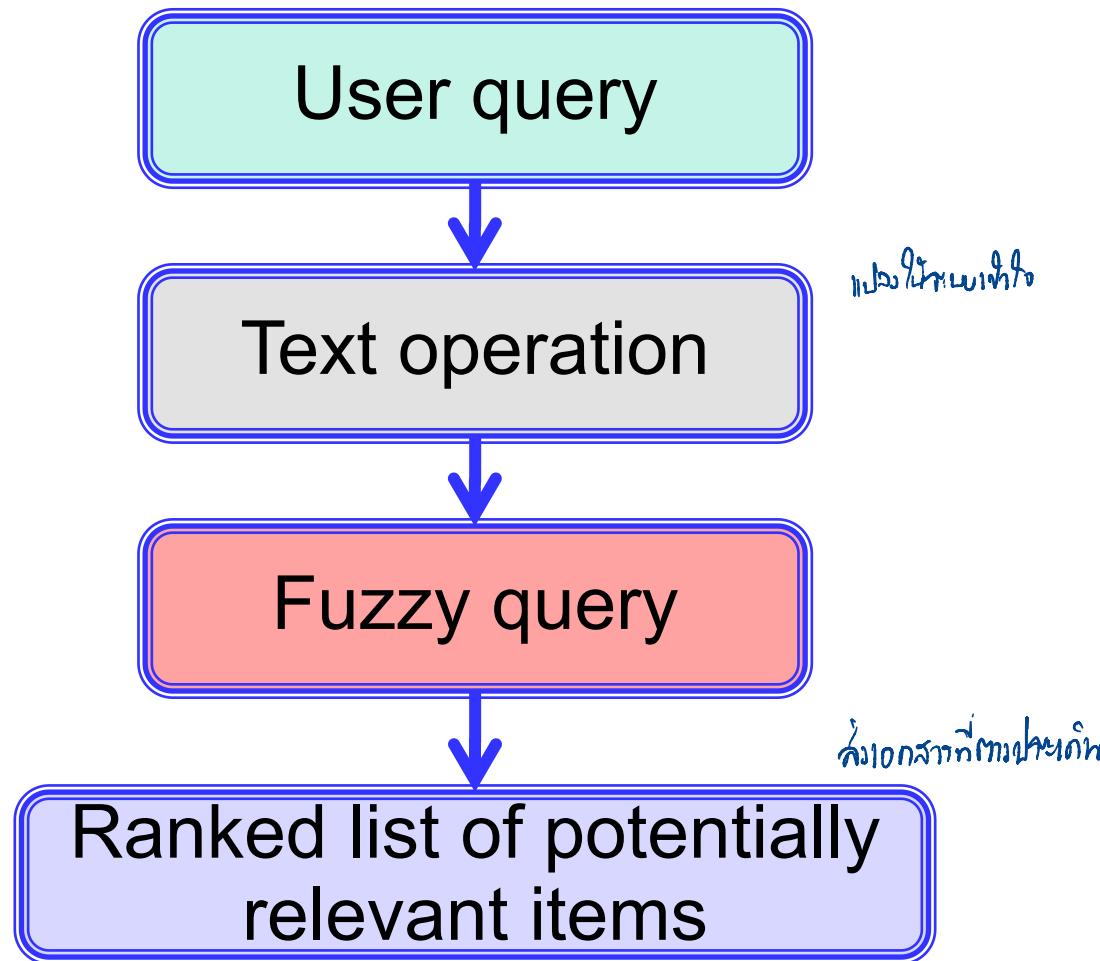
$$(1 - \mu_{Dog}(x)).(1 - \mu_{Cat}(x)).(1 - \mu_{Tiger}(x))$$

# Example



$$\mu_{Dog \cup Cat \cup Tiger}(x) = 1 - (1 - \mu_{Dog}(x)) \cdot (1 - \mu_{Cat}(x)) \cdot (1 - \mu_{Tiger}(x))$$

# Fuzzy Logic



# Index Term Relationship

$$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

ตัวเด่นๆ กัน

$c_{i,j}$  คือ ความสัมพันธ์ของคีย์เวิร์ด  $i$  กับ คีย์เวิร์ด  $j$

$n_{i,j}$  คือ จำนวนเอกสารที่มีทั้งคีย์เวิร์ด  $i$  และ คีย์เวิร์ด  $j$

$n_i$  คือ จำนวนเอกสารที่มีคีย์เวิร์ด  $i$

$n_j$  คือ จำนวนเอกสารที่มีคีย์เวิร์ด  $j$

# Example

ข้อ 2. สมมติในระบบมีเอกสาร 5 เอกสารดังนี้ (คำที่อยู่ในปีกภาคือ Keyword ซึ่ง Keyword มีความสัมพันธ์กัน)

ค่า ihm ก็คือ fuzzy model กับ GBSM

- D1 : {bird, cat, bird, cat, tiger, kitty, fish, fish}
- D2 : {cat, dog, tiger, zoo}
- D3 : {house, kitchen, bird, bird, cat, cat, kitty}
- D4 : {dog, rubber, house, dog}
- D5 : {tiger, forest, fish}

โดยมี Query : "ต้องการเอกสารที่มี cat หรือ kitty และไม่ต้องการเอกสารที่มี dog "

ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารแต่ละเอกสาร เพื่อเรียงลำดับเอกสารที่จะแสดงผลให้กับผู้เรียกค้น

# Example

## Answer

ความสัมพันธ์ระหว่าง Keyword

ចំណាំពិរុសន៍ ក្នុងបណ្តុះបណ្តាល

dog → dog X

- |         |  |
|---------|--|
| bird    | → cat, fish, house, kitchen, kitty, tiger            |
| cat     | → bird, dog, fish, house, kitchen, kitty, tiger, zoo |
| dog     | → cat, house, rubber, tiger, zoo                     |
| fish    | → bird, cat, forest, kitty, tiger                    |
| forest  | → fish, tiger  |
| house   | → bird , cat , dog, kitchen, kitty, rubber           |
| kitchen | → bird, cat, house, kitty                            |
| kitty   | → bird, cat, fish, house, kitchen, tiger             |
| rubber  | → dog, house   |
| tiger   | → bird, cat, dog, fish, forest, kitty, zoo           |
| zoo     | → cat, dog, tiger                                    |

# Example

## Answer

คำนวณความสัมพันธ์

Doc ถ้าฟอร์ด = ตามไปเก็บ 1

ถ้าล้วน = ลากผ่าน

\* ที่ keyword ที่ไม่ใช่ แต่ถ้าเก็บมาระยะหัก สุ่ม b and C b c b and C

$$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

bird,cat	= $2/(2+3-2) = 0.67$	cat,dog	= $1/(3+2-1) = 0.25$
bird,fish	= $1/(2+2-1) = 0.33$	cat,fish	= $1/(3+2-1) = 0.25$
bird,house	= $1/(2+2-1) = 0.33$	cat,house	= $1/(3+2-1) = 0.25$
bird,kitchen	= $1/(2+1-1) = 0.50$	cat,kitchen	= $1/(3+1-1) = 0.33$
bird,kitty	= $2/(2+2-2) = 1.00$	cat,kitty	= $2/(3+2-2) = 0.67$
bird,tiger	= $1/(2+3-1) = 0.25$	cat,tiger	= $2/(3+3-2) = 0.50$
		cat,zoo	= $1/(3+1-1) = 0.33$

# Example

---

$$\begin{array}{ll} \text{dog,house} & = 1/(2+2-1) = 0.33 \\ \text{dog,rubber} & = 1/(2+1-1) = 0.50 \\ \text{dog,tiger} & = 1/(2+3-1) = 0.25 \\ \text{dog,zoo} & = 1/(2+1-1) = 0.50 \end{array}$$

$$\begin{array}{ll} \text{fish,forest} & = 1/(2+1-1) = 0.50 \\ \text{fish,kitty} & = 1/(2+2-1) = 0.33 \\ \text{fish,tiger} & = 2/(2+3-2) = 0.67 \\ \text{forest,tiger} & = 1/(1+3-1) = 0.33 \end{array}$$

$$\begin{array}{ll} \text{house,kitchen} & = 1/(2+1-1) = 0.50 \\ \text{house,kitty} & = 1/(2+2-1) = 0.33 \\ \text{house,rubber} & = 1/(2+1-1) = 0.50 \end{array}$$

$$\text{kitchen,kitty} = 1/(1+2-1) = 0.50$$

$$\text{kitty,tiger} = 1/(2+3-1) = 0.25$$

$$\text{tiger,zoo} = 1/(3+1-1) = 0.33$$

# Example

---

	bird	cat	dog	fish	forest	house	kitchen	kitty	rubber	tiger	zoo
bird	1.00	0.67	0.00	0.33	0.00	0.33	0.50	1.00	0.00	0.25	0.00
cat	0.67	1.00	0.25	0.25	0.00	0.25	0.33	0.67	0.00	0.50	0.33
dog	0.00	0.25	1.00	0.00	0.00	0.33	0.00	0.00	0.50	0.25	0.50
fish	0.33	0.25	0.00	1.00	0.50	0.00	0.00	0.33	0.00	0.67	0.00
forest	0.00	0.00	0.00	0.50	1.00	0.00	0.00	0.00	0.00	0.33	0.00
house	0.33	0.25	0.33	0.00	0.00	1.00	0.50	0.33	0.50	0.00	0.00
kitchen	0.50	0.33	0.00	0.00	0.00	0.50	1.00	0.50	0.00	0.00	0.00
kitty	1.00	0.67	0.00	0.33	0.00	0.33	0.50	1.00	0.00	0.25	0.00
rubber	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	1.00	0.00	0.00
tiger	0.25	0.50	0.25	0.67	0.33	0.00	0.00	0.50	0.00	1.00	0.33
zoo	0.00	0.33	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.33	1.00

# Example

- D1: {bird,cat,bird,cat,tiger,kitty,fish,fish}  
D2: {cat,dog,tiger,zoo}  
D3: {house,kitchen,bird,bird,cat,cat,kitty}  
D4: {dog,rubber,house,dog}  
D5: {tiger,forest,fish}

↑ เก็บเน้นกันทุก keyword ใน Doc และ ทางมากสุด

	Cat	Kitty	Dog
D1	1	1	0.25
D2	1	0.67	1
D3	1	1	0.33
D4	0.25	0.33	1
D5	0.50	0.33	0.25

→ หมายความว่า keyword นี้มีอยู่กับตากากรคำทำนายกับหนังสือ  
กม. กัน keyword นี้ใน Doc 1 มีอยู่ร้อยละ 50  
แล้วก็จะดู

# Example

Query : ”ต้องการเอกสารที่มี cat หรือ kitty และไม่ต้องการเอกสารที่มี dog ”

Query = (cat OR kitty) AND not dog

$$\mu_{\bar{A}}(x) = (1 - \mu_A(x))$$

$$\mu_{A \cup B}(x) = 1 - (1 - \mu_A(x)).(1 - \mu_B(x))$$

$$\mu_{A \cap B}(x) = \mu_A(x). \mu_B(x)$$

	Cat	Kitty	Dog
D1	1	1	0.25
D2	1	0.67	1
D3	1	1	0.33
D4	0.25	0.33	1
D5	0.50	0.33	0.25

∴ Ranking =  $(1 - (1 - \mu_{cat}).(1 - \mu_{kitty})) . (1 - \mu_{dog})$

$$Doc_1 = (1 - (1 - 1) \times (1 - 1)) \times (1 - 0.25) = 1 \times 0.75 = 0.75$$

$$Doc_2 = (1 - (1 - 1) \times (1 - 0.67)) \times (1 - 1) = 1 \times 0 = 0.00$$

$$Doc_3 = (1 - (1 - 1) \times (1 - 1)) \times (1 - 0.33) = 1 \times 0.67 = 0.67$$

$$Doc_4 = (1 - (1 - 0.25) \times (1 - 0.33)) \times (1 - 1) = 0.49 \times 0 = 0.00$$

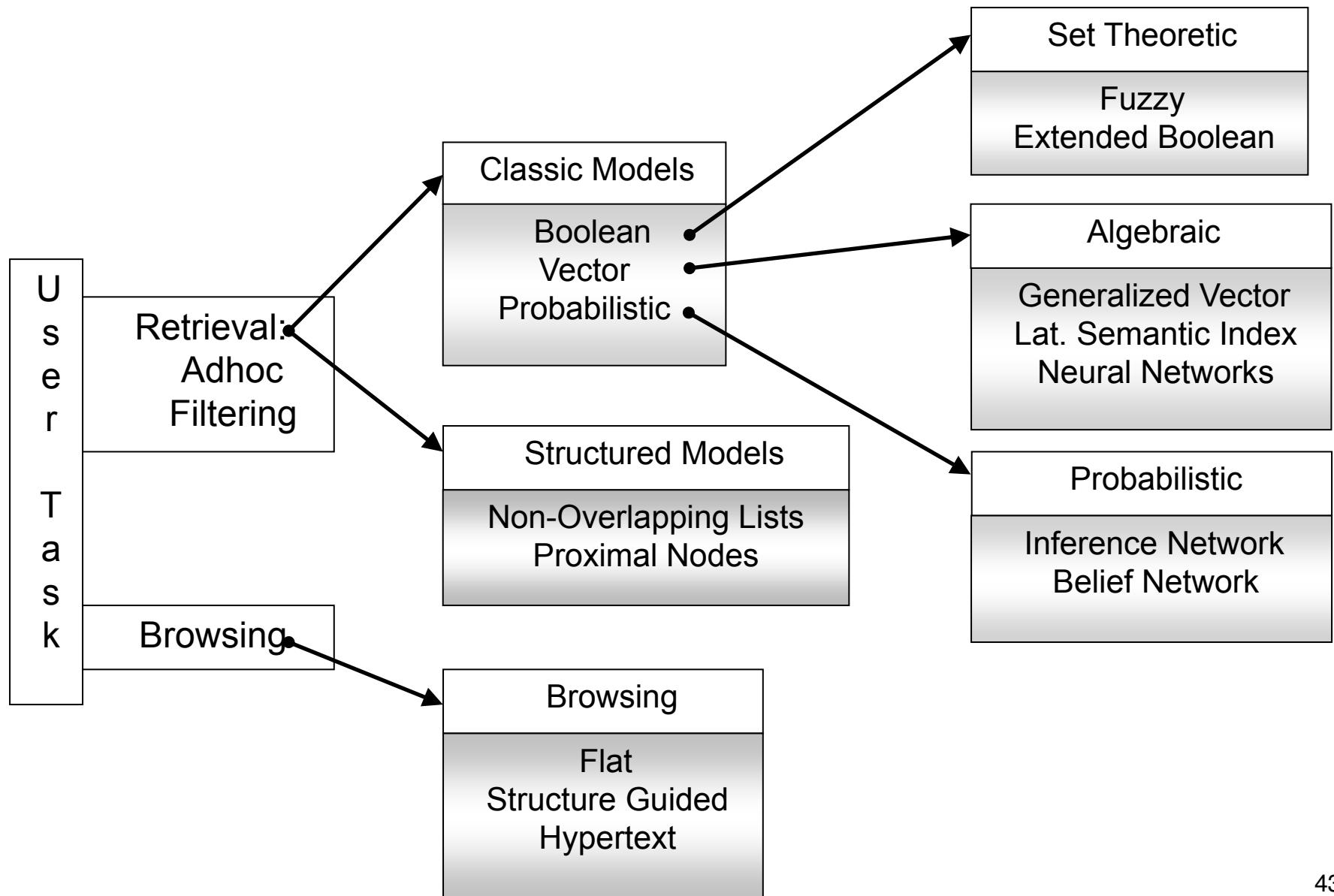
$$Doc_5 = (1 - (1 - 0.50) \times (1 - 0.33)) \times (1 - 0.25) = 0.67 \times 0.75 = 0.50$$

Ranking = Doc<sub>1</sub>, Doc<sub>3</sub>, Doc<sub>5</sub>, Doc<sub>2</sub>, Doc<sub>4</sub>

# Extended Boolean Model

---

# IR Models



# Boolean Model

เอกสาร	เนื้อหาของเอกสาร	ตัวพัฒนาระบบที่ใช้ในเอกสาร
D <sub>1</sub>	สุนขกินแมวอกกับแมวาน	“สุนข” “กิน” “แมว”
D <sub>2</sub>	สุนขไม่ใช่หนู	“สุนข” “หนู”
D <sub>3</sub>	หนูกินไม่มากนัก	“หนู” “กิน”
D <sub>4</sub>	แมวชอบเล่นกับแมวและหนู	“แมว” “เล่น” “งู” “หนู”
D <sub>5</sub>	แมวชอบเล่นแต่ไม่กับแมวด้วยกัน	“แมว” “เล่น”

ถ้าต้องการค้น (แมว AND สุนข) จะได้ผลการค้นเป็น D<sub>1</sub> เท่านั้น

# Boolean Model

---

- เนื่องจาก Boolean Model มีข้อเสียคือการไม่สนใจนำหน้าของ Keyword
- Vector Space Model มีข้อเสียคือการเชื่อมต่อทางตรรกะทำได้ยาก

จึงได้มีความพยายามที่นำข้อดีของทั้งสองมาร่วมกัน ทำเป็น Model ใหม่ขึ้นมา เรียกว่า **Extended Boolean Model**

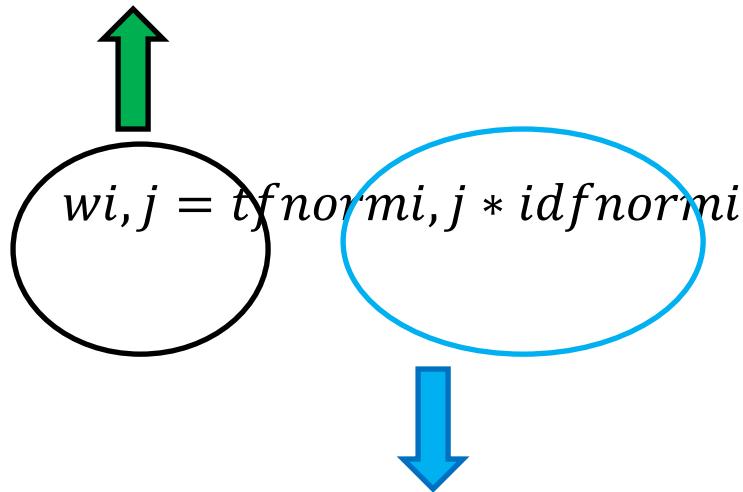


## วิธีคำนวณค่าต่างๆ

# น้ำหนักของ Keyword ในเอกสาร

ต้องการหา  $0 - 1$  เท่านั้น

น้ำหนักของ Keyword “i” ในเอกสาร “j”



# น้ำหนักของ Keyword ในเอกสาร

$$w_{i,j} = tfnorm_{i,j} * idfnorm_i$$



normalized IDF ของ Keyword “i” ในเอกสารทั้งหมด

# น้ำหนักของ Keyword ในเอกสาร

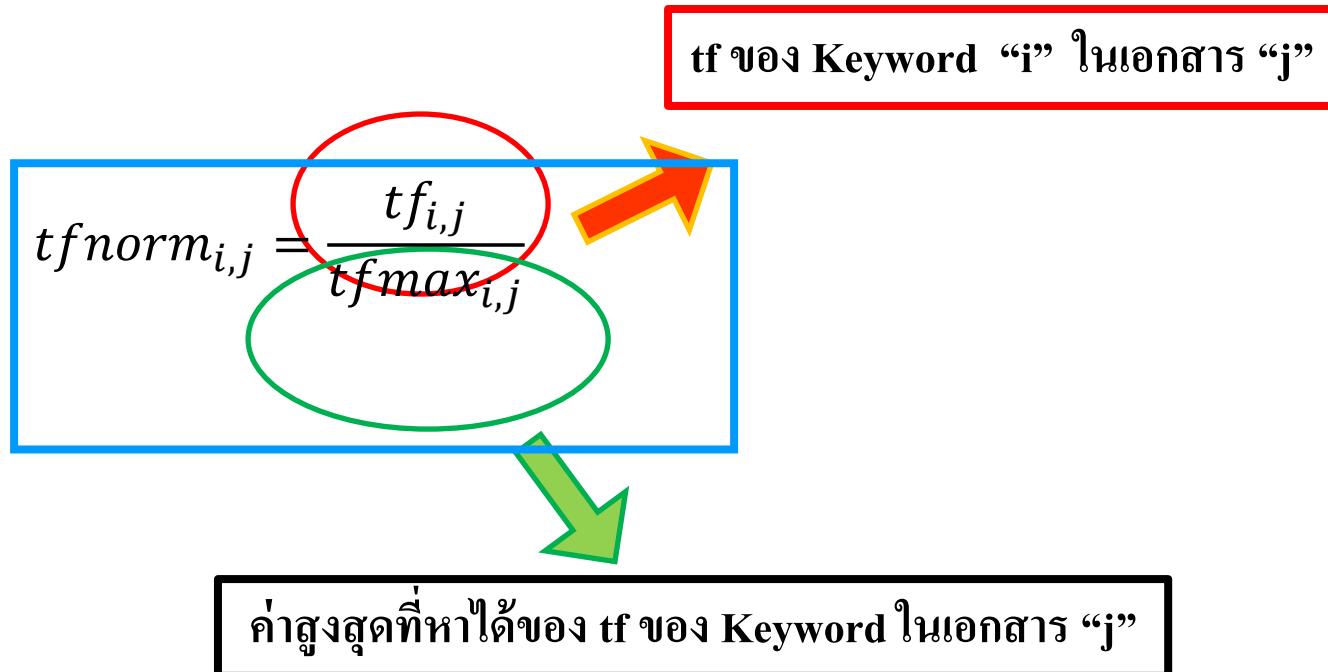
$$w_{i,j} = tfnorm_{i,j} * idfnorm_i$$

$$tfnorm_{i,j} = \frac{tf_{i,j}}{tfmax_{i,j}}$$

$$idfnorm_i = \frac{idf_i}{idfmax_g}$$

# น้ำหนักของ Keyword ในเอกสาร

tf คือจำนวนครั้งที่ Keyword นั้นปรากฏในเอกสารที่สนใจ



# น้ำหนักของ Keyword ในเอกสาร

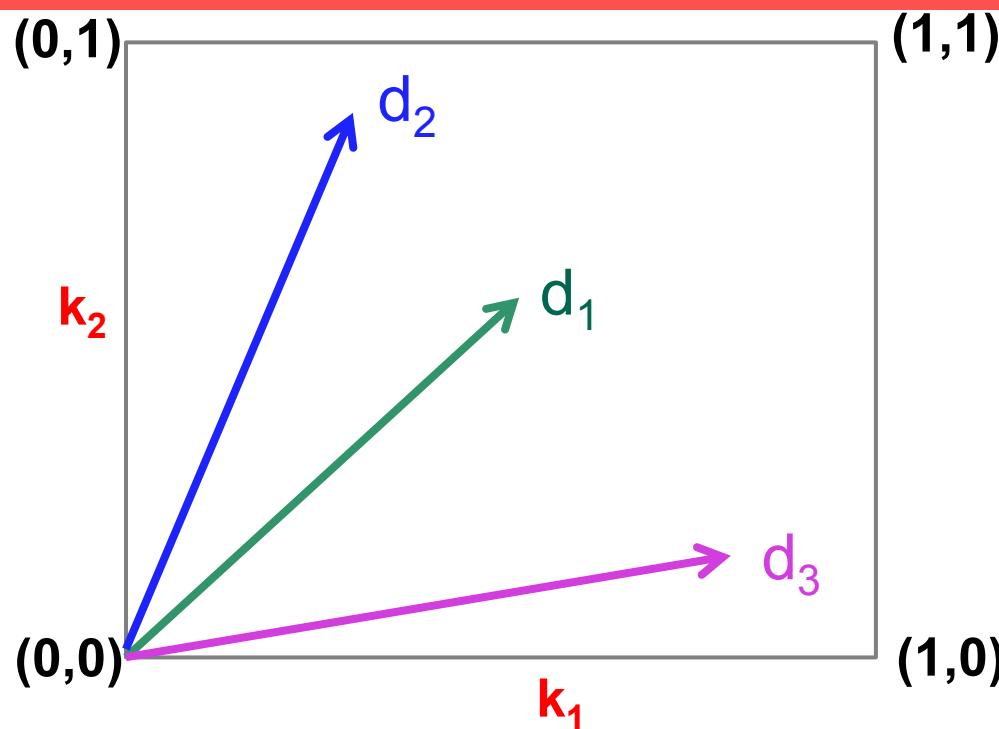
idf คือการรวมของการพน Keyword ที่สนใจ โดยพิจารณาจากเอกสารทั้งหมดในระบบ

idf ของ Keyword “i” ในเอกสารทั้งหมด

$$idfnorm_i = \frac{idf_i}{idfmax_g}$$

ค่าสูงสุดที่หาได้ของ idf ของ Keyword ในเอกสารทั้งหมด

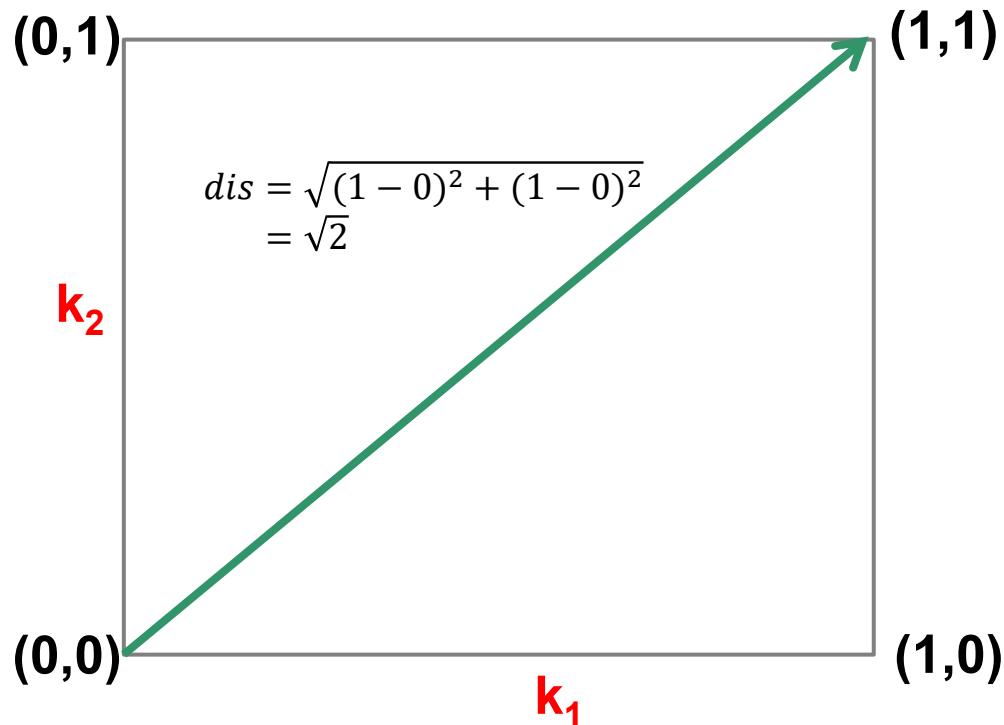
# Relevance



จุด  $(0,0)$  คือจุดที่มีความตรงประเด็นน้อยที่สุด  
จุด  $(1,1)$  คือจุดที่มีความตรงประเด็นมากที่สุด

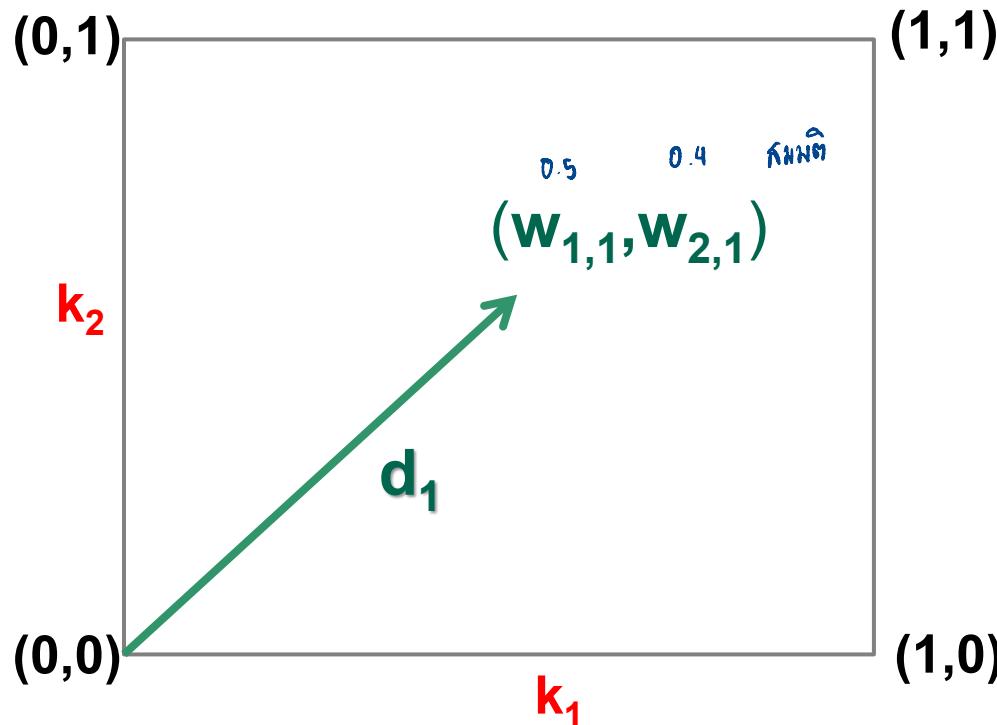
หมาย สาระเด็ก ก้มงาบกูด ด้วย

## ระยะห่างสูงสุดของความตรงประเด็น (Relevance)



# OR

คำนวณระยะห่างจากจุด $(0,0)$ ไปที่ $(w_1, w_2)$  ของเอกสารที่สนใจ

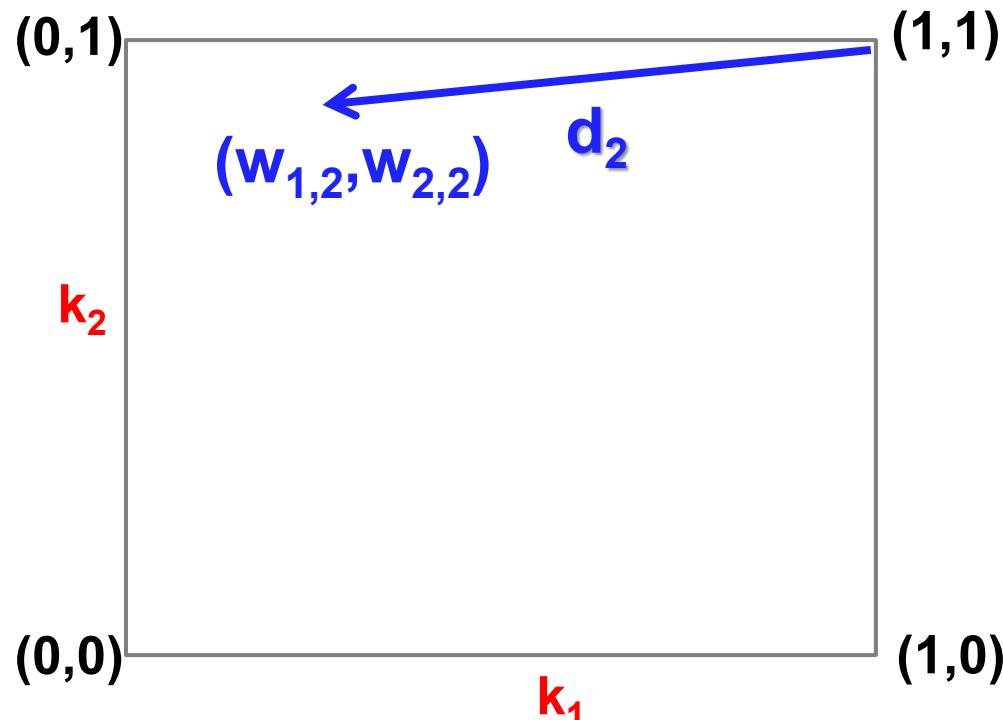


$$dis_{or} = \sqrt{(W_{1,j} - 0)^2 + (W_{2,j} - 0)^2}$$

คือ กำหนดว่า  $\rightarrow$  ผ่านตอกความตามที่ต้อง

# AND

คำนวณระยะห่างจากจุด  $(1,1)$  ไปที่  $(w_1, w_2)$  ของเอกสารที่สนใจ



$$dis_{and} = \sqrt{2} - \sqrt{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}$$

# ความต่างประเด็นของ Query

OR

$$dis_{or} = \sqrt{{W_{1,j}}^2 + {W_{2,j}}^2}$$

หมายเหตุ

ระยะห่าง  $\sqrt{2}$

ระยะห่าง  $dis$

ทรงประเด็น  $1.00$

ทรงประเด็น  $\frac{dis}{\sqrt{2}}$

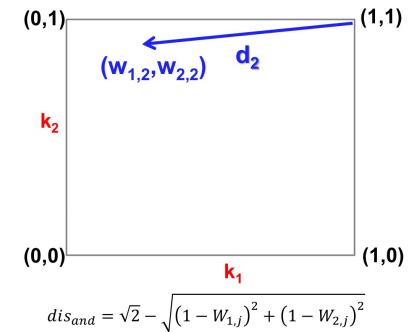
กรณีที่มีค่าความคลาสสิกมากกว่า 0.5

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

$j$  เกิน 100%

→ เท่านาก็ยังหัน  $\sqrt{2}$   
 пример:  $\sqrt{2}$  ถือความคลาสสิก  $(1,1)$   
 อยู่ 100 %.

# ความตรงประเด็นของ Query



**AND**

ให้ vector ห่างจาก 1,1 เท่าไร

$$dis_{and} = \sqrt{2} - \sqrt{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}$$

ได้ผลลัพธ์ 100 %.

ระยะห่าง  $\sqrt{2}$   
ระยะห่าง  $dis$

ตรงประเด็น 1.00  
ตรงประเด็น  $\frac{dis}{\sqrt{2}}$

ให้ห้องน้ำแบบ  
สองปี รอดู

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

# Question

Q = Cat OR Not dog ???

$$sim(q_{or}, dj) = \sqrt{\frac{W_{cat,j}^2 + (1 - W_{dog,j})^2}{2}}$$

Q = (Dog AND Cat) OR Tiger ???

$$sim(q_{or}, dj) = \sqrt{\frac{W_{1,j}^2 + W_{tiger,j}^2}{2}} \quad \text{OR}$$
$$sim(q_{or}, dj) = \sqrt{\frac{\left(1 - \sqrt{\frac{(1 - W_{dog,j})^2 + (1 - W_{cat,j})^2}{2}}\right)^2 + W_{tiger,j}^2}{2}}$$

# EXAMPLE

ข้อ 2. สมมติในระบบมีเอกสาร 10 เอกสารดังนี้ (bird, cat, dog, tiger คือ Keyword ซึ่งไม่สัมพันธ์กัน)

- 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}

เด็กหญิงดาวิกาส่งคำเรียกด้วย “รักแมวและสุนัข แต่ไม่รักเสือ” เข้าไปในระบบ จงตอบคำถาม

2.1 เพื่อให้ได้คำตอบในคำถาม 2.2 เด็กหญิงดาวิกาควรเลือกใช้โมเดลใดเพื่ออะไร (เลือกได้เฉพาะตัวเลือกที่ให้มา)

A) Probabilistic Model B) ~~Fuzzy Model~~ C) Extend Boolean Model D) ~~Vector Model~~ *แนว query ไม่แน่นอน*

2.2 ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ตามที่เด็กหญิงดาวิกาต้องการ

(33 คะแนน) ข้อสอบ 1/2559

### Answer

**2.1** เลือกใช้ Extend Boolean Model เนื่องจากลักษณะของ Query เป็นแบบ Boolean และโจทย์กำหนดให้ Keyword “ไม่สัมพันธ์กัน”

# ขั้นตอนที่ 1

## หาหัวคำ Keyword

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** = รักแมวและสุนัข แต่ไม่รักเสือ

**Query** = (Cat AND Dog) AND NOT Tiger

### Ranking

Doc8

ภาคคองกรีต

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	

# ขั้นตอนที่ 1

$$tf_{bird} = \frac{3}{3} = 1.000$$

$$tf_{cat} = \frac{2}{3} = 0.667$$

$$tf_{dog} = \frac{2}{3} = 0.667$$

$$tf_{tiger} = \frac{0}{3} = 0.000$$

Only Doc1

normalize  
เร็วๆ นี้

	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	

→ จำนวนเอกสารทั้งหมด

$$idf_{bird} = \log\left(\frac{10}{5}\right) = 0.301$$

$$idf_{cat} = \log\left(\frac{10}{8}\right) = 0.097$$

$$idf_{dog} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{tiger} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{norm, bird} = \frac{0.301}{0.301} = 1.000$$

$$idf_{norm, cat} = \frac{0.097}{0.301} = 0.322$$

$$idf_{norm, dog} = \frac{0.155}{0.301} = 0.515$$

$$idf_{norm, tiger} = \frac{0.155}{0.301} = 0.515$$

$$w_{bird} = 1.000 * 1.000 = 1.000$$

$$w_{cat} = 0.667 * 0.322 = 0.215$$

$$w_{dog} = 0.667 * 0.515 = 0.343$$

$$w_{tiger} = 0.000 * 0.515 = 0.000$$

# ขั้นตอนที่ 1

นำหน้าของแต่ละ **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 = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger

AND

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

	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

$$sim(q_{and}, dj) = 1 - \sqrt{\left(1 - \left(1 - \sqrt{\frac{(1 - WCat, j)^2 + (1 - Wdog, j)^2}{2}}\right)\right)^2 + (1 - (1 - WTiger, j))^2}$$

ลองคำนวณ 1

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

$$sim(q_{and}, d_1) = 0.488$$

# ขั้นตอนที่ 3

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger  
ถ้า degree ต่ำ Ranking

	Sim
Doc1	0.488
Doc2	0.465
Doc3	0.377
Doc4	0.295
Doc5	0.291
Doc6	0.320
Doc7	0.447
Doc8	0.583
Doc9	0.447
Doc10	0.205

Ranking	Sim
Doc8	0.583
Doc1	0.488
Doc2	0.465
Doc7	0.447
Doc9	0.447
Doc3	0.377
Doc6	0.320
Doc4	0.295
Doc5	0.291
Doc10	0.205

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

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

หมายเหตุ keyword จำนวน 1 ชุด:  
 rank ของ Doc ที่มี keyword มากที่สุด

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

# 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	

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat OR Dog) AND NOT Tiger

Ranking

Doc8

Doc1

...

# ขั้นตอนที่ 1

Only Doc1

$$tf_{bird} = \frac{3}{3} = 1.000$$

$$tf_{cat} = \frac{2}{3} = 0.667$$

$$tf_{dog} = \frac{2}{3} = 0.667$$

$$tf_{tiger} = \frac{0}{3} = 0.000$$

	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	

$$idf_{bird} = \log\left(\frac{10}{5}\right) = 0.301$$

$$idf_{cat} = \log\left(\frac{10}{8}\right) = 0.097$$

$$idf_{dog} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{tiger} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{norm, bird} = \frac{0.301}{0.301} = 1.000$$

$$idf_{norm, cat} = \frac{0.097}{0.301} = 0.322$$

$$idf_{norm, dog} = \frac{0.155}{0.301} = 0.515$$

$$idf_{norm, tiger} = \frac{0.155}{0.301} = 0.515$$

$$w_{bird} = 1.000 * 1.000 = 1.000$$

$$w_{cat} = 0.667 * 0.322 = 0.215$$

$$w_{dog} = 0.667 * 0.515 = 0.343$$

$$w_{tiger} = 0.000 * 0.515 = 0.000$$

# ขั้นตอนที่ 1

## คำนักของแต่ละ 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 = รักแมวและสุนัข แต่ไม่รักเสือ**

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

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

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

$$sim(q_{and}, dj) = 1 - \sqrt{\frac{\left(1 - \left(\sqrt{\frac{(WCat,j)^2 + (Wdog,j)^2}{2}}\right)\right)^2 + (1 - (1 - W_{Tiger,j}))^2}{2}}$$

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

$$sim(q_{and}, d_1) = 0.488$$

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

## ขั้นตอนที่ 3

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger

	Sim
Doc1	0.488
Doc2	0.465
Doc3	0.377
Doc4	0.295
Doc5	0.291
Doc6	0.320
Doc7	0.447
Doc8	0.583
Doc9	0.447
Doc10	0.205

Ranking	Sim
Doc8	0.583
Doc1	0.488
Doc2	0.465
Doc7	0.447
Doc9	0.447
Doc3	0.377
Doc6	0.320
Doc4	0.295
Doc5	0.291
Doc10	0.205

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

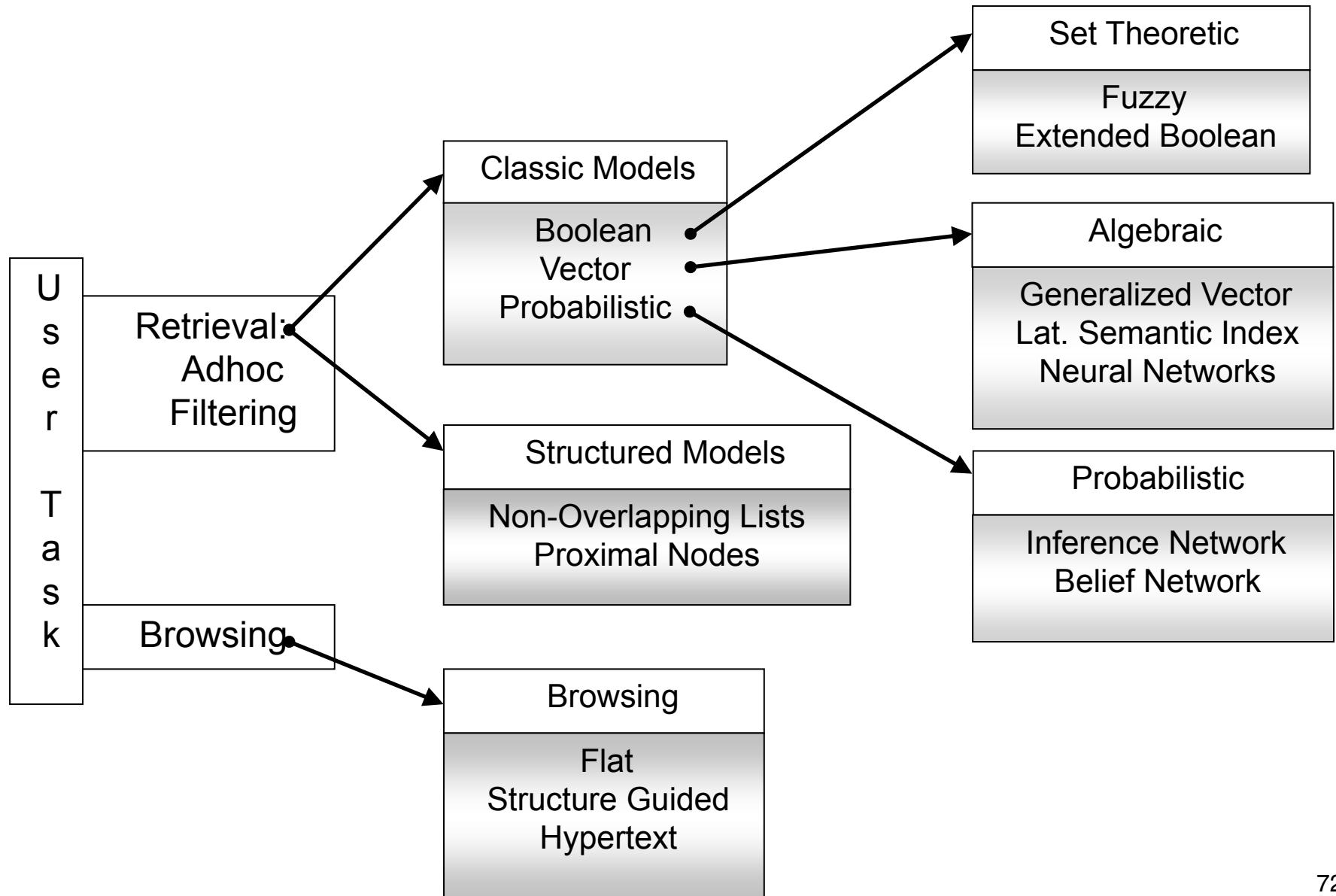
- D1: {bird, cat, bird, cat, dog, dog, dog, bird}  
D2: {cat, tiger, cat, dog}  
D3: {dog, bird, bird}  
D4: {cat, tiger}  
D5: {tiger, tiger, dog, tiger, cat}  
D6: {bird, cat, bird, cat, tiger, tiger, bird}  
D7: {bird, tiger, cat, dog}  
D8: {dog, cat, bird}  
D9: {cat, dog, tiger}  
D10: {tiger,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

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

---

# Generalizes Vector Space Model (GVSM)

# IR Models



# Basic Vector Space Model

---

## Example

$$D1 = (2, 1, 0, 0)$$

$$D2 = (5, 1, 0, 0)$$

$$D3 = (1, 1, 1, 1)$$

$$D4 = (0, 0, 2, 2)$$

$$D5 = (0, 1, 1, 2)$$

$$D6 = (0, 0, 1, 1)$$

$$D7 = (0, 0, 1, 0)$$

$$D8 = (1, 1, 0, 0)$$

$$D9 = (2, 1, 1, 1)$$

$$D10 = (0, 2, 2, 2)$$

$$D11 = (1, 0, 2, 0)$$

$$D12 = (0, 0, 2, 1)$$

$$q = 2k_1 + 3k_2 - k_3$$

$$W_{ij} = tf_{ij} * idfi = tfij * \log\left(\frac{N}{n_i}\right)$$

$$W_{iq} = \left(0.5 + \frac{0.5 * freqi_q}{Max(freqi_q)}\right) * \log\left(\frac{N}{n_i}\right)$$

$$sim(q, dj) = \frac{\sum_{i=1}^t (w_{ij} * wiq)}{\sqrt{\sum_{i=1}^t w_{ij}^2 * \sum_{i=1}^t w_{iq}^2}}$$

→ for vector អាជីវកម្ម

# Generalizes Vector Space Model

ឯកសារ key និង keyword តាមរយៈរាយការណ៍

រាយការណ៍ 0 ពាក្យសាខា

4 keyword ដែលមាន 15 រាយ

$$D_1 = (2, 1, 0, 0)$$

$$D_2 = (5, 1, 0, 0)$$

$$D_3 = (1, 1, 1, 1)$$

$$D_4 = (0, 0, 2, 2)$$

$$D_5 = (0, 1, 1, 2)$$

$$D_6 = (0, 0, 1, 1)$$

$$D_7 = (0, 0, 1, 0)$$

$$D_8 = (1, 1, 0, 0)$$

$$D_9 = (2, 1, 1, 1)$$

$$D_{10} = (0, 2, 2, 2)$$

$$D_{11} = (1, 0, 2, 0)$$

$$D_{12} = (0, 0, 2, 1)$$

$$D_1 = (2, 1, 0, 0)$$

$$D_2 = (5, 1, 0, 0)$$

$$D_8 = (1, 1, 0, 0)$$

$$D_3 = (1, 1, 1, 1)$$

$$D_9 = (2, 1, 1, 1)$$

$$D_4 = (0, 0, 2, 2)$$

$$D_6 = (0, 0, 1, 1)$$

$$D_{12} = (0, 0, 2, 1)$$

$$D_5 = (0, 1, 1, 2)$$

$$D_{10} = (0, 2, 2, 2)$$

$$D_7 = (0, 0, 1, 0)$$

$$D_{11} = (1, 0, 2, 0)$$

ដោយ key 1-2

$$q = 2k_1 + 3k_2 - k_3$$

# ข้อเสียของ Vector Space Model

---

- Keyword บางส่วนอาจจะเป็นอิสระต่อกัน บางส่วนอาจจะเกี่ยวข้องกันและบางส่วนอาจจะเกี่ยวข้องกันมาก
- การอนุมานว่า Keyword เป็นอิสระจากกัน จึงเป็นการกระทำไม่ได้สอดคล้องกับความเป็นจริง
- Keyword ที่มีความหมายเดียวกัน จัดอยู่ในกลุ่มเดียวกัน

# Generalizes Vector Space Model (GVSM)

---

- Keyword จะไม่ได้เป็นอิสระต่อกัน แต่จะเกี่ยวข้องกันในลักษณะใดลักษณะหนึ่ง โดยสังเกตจากการปรากฏร่วมกัน
- การปรากฏของ Keyword จะนำมาซึ่งการเปรียบเทียบความคล้ายหรือความต่างของเอกสารกับคำเรียกค้นที่เข้ามา
- GVSM ใช้หลักการเดียวกับ VSM ด้วยการคำนวณหาค่าความสอดคล้องของคำเรียกค้นกับเอกสารในระบบ แต่บน Vector Space ใหม่

## GVSM Definition

- Definition Given the set  $\{k_1, k_2, \dots, k_t\}$  of index terms in a collection, as before, let  $w_{i,j}$  be the weight associated with the term-document pair  $[k_i, d_j]$ . If the  $w_{i,j}$  weights are **all binary**, then all possible patterns of term co-occurrence (inside documents) can be represented by a set of  **$2^t$  minterms** given by

$$m_1 = (0, 0, \dots, 0),$$

then minterms / 11th

$$m_2 = (1, 0, \dots, 0)$$

,...,

$$m_{2^t} = (1, 1, \dots, 1)$$

Let  $g_i(m_j)$  return the weight  $\{0, 1\}$  of the index term  $k_i$  in the minterm  $m_j$

# GVSM Definition

---

- Definition Let us define the following set of vectors
  - $m_1 = (0, 0, \dots, 1)$
  - $m_2 = (0, 0, \dots, 1, 0)$
  - .....
  - $m_{2^{t-1}} = (1, 1, \dots, 1)$  <sup>15</sup>

where each vector  $m_i$  is associated with the respective minterm  $m_i$ .

# GVSM Definition

คำนวน keyword ที่ q<sub>1</sub> วิ่ง.... ที่ n น น คำนวน key 1 ที่ index 1 = 2+5+1 = 8

$$c_{i,r} = \sum_{d_j | g_l(d_j) = g_l(m_r), \text{for all } l} w_{i,j}$$

D1 = (2, 1, 0, 0)
D2 = (5, 1, 0, 0)
D8 = (1, 1, 0, 0)

key word ที่ 1 ถึง ที่ 8

$$k_i = \frac{\sum_{\forall r, g_i(m_r)} c_{i,r} m_r}{\sqrt{\sum_{\forall r, g_i(m_r)} c_{i,r}^2}}$$

$$k_i \bullet k_j = \sum_{\forall r | g_i(m_r) = 1 \wedge g_j(m_r) = 1} c_{i,r} \times c_{j,r}$$

# GVSM Definition

$$d_j = \sum_i w_{i,j} k_i \quad \longrightarrow \quad d_j = \sum_r s_{j,r} m_r$$

$$q_j = \sum_i w_{i,q} k_i \quad \longrightarrow \quad q_j = \sum_r s_{q,r} m_r$$

$$\text{sim}(q, d_j) = \frac{\sum_{i=1}^t w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2 \cdot \sum_{i=1}^t w_{i,q}^2}}$$



$$\text{sim}(q, d_j) = \frac{\sum_r S_{d,r} \cdot S_{q,r}}{\sqrt{\sum_r S_{d,r}^2 \cdot \sum_r S_{q,r}^2}}$$

# Example

①

ຕັບແນວມູນຄົນ ນາງ ກອນ (tf / idf)

$$D_1 = \underline{(2, 1, 0, 0)} m_1$$

② ດັດເໜີມາດນຸ່ງໆໄປໜັກ

$$D_2 = (5, 1, 0, 0) m_1$$

$$D_3 = (1, 1, 1, 1) m_2$$

$$D_4 = (0, 0, 2, 2) m_3$$

$$D_5 = (0, 1, 1, 2) m_4$$

$$D_6 = (0, 0, 1, 1) m_3$$

$$D_7 = (0, 0, 1, 0) m_5$$

$$D_8 = (1, 1, 0, 0) m_1$$

$$D_9 = (2, 1, 1, 1) m_2$$

$$D_{10} = (0, 2, 2, 2) m_4$$

$$D_{11} = (1, 0, 2, 0) m_6$$

$$D_{12} = (0, 0, 2, 1) m_3$$

$$q = 2k_1 + 3k_2 - k_3$$

$$m_1 = (1, 1, 0, 0)$$

$$m_2 = (1, 1, 1, 1)$$

$$m_3 = (0, 0, 1, 1)$$

$$m_4 = (0, 1, 1, 1)$$

$$m_5 = (0, 0, 1, 0)$$

$$m_6 = (1, 0, 1, 0)$$

Minterms: 6 minterms

All Weight → TF\*IDF First (คำนวณจาก vector model)

$m1=(1, 1, 0, 0)$

$m2=(1, 1, 1, 1)$

$m3=(0, 0, 1, 1)$

$m4=(0, 1, 1, 1)$

$m5=(0, 0, 1, 0)$

$m6=(1, 0, 1, 0)$

$D1 = (2, 1, 0, 0) \text{ m1}$

$D2 = (5, 1, 0, 0) \text{ m1}$

$D3 = (1, 1, 1, 1) \text{ m2}$

$D4 = (0, 0, 2, 2) \text{ m3}$

$D5 = (0, 1, 1, 2) \text{ m4}$

$D6 = (0, 0, 1, 1) \text{ m3}$

$D7 = (0, 0, 1, 0) \text{ m5}$

$D8 = (1, 1, 0, 0) \text{ m1}$

$D9 = (2, 1, 1, 1) \text{ m2}$

$D10 = (0, 2, 2, 2) \text{ m4}$

$D11 = (1, 0, 2, 0) \text{ m6}$

$D12 = (0, 0, 2, 1) \text{ m3}$

$k_i = \frac{\sum_{\forall r, g_i(m_r)} c_{i,r} m_r}{\sqrt{\sum_{\forall r, g_i(m_r)} c_{i,r}^2}}$

$k_4 = \frac{c_{1,1}m_1 + c_{1,2}m_2 + c_{1,6}m_6}{\sqrt{c_{1,1}^2 + c_{1,2}^2 + c_{1,6}^2}}$

$$\begin{aligned} k_1 &= \frac{8m_1 + 3m_2 + m_6}{\sqrt{64 + 9 + 1}} \\ &= \frac{8m_1 + 3m_2 + m_6}{\sqrt{74}} \end{aligned} \quad (1)$$

$k_2 = \frac{c_{2,1}m_1 + c_{2,2}m_2 + c_{2,4}m_4}{\sqrt{c_{2,1}^2 + c_{2,2}^2 + c_{2,4}^2}}$

$k_2 = \frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}} \quad (2)$

$c_{i,r} = \sum_{d_j | g_l(d_j) = g_l(m_r), \text{for.all.l}} w_{i,j}$

$c_{1,1} = w_{1,1} + w_{1,2} + w_{1,8} = 2 + 5 + 1 = 8$

$c_{1,2} = w_{1,3} + w_{1,9} = 1 + 2 = 3$

$c_{1,6} = w_{1,11} = 1$

$c_{2,1} = w_{2,1} + w_{2,2} + w_{2,8} = 1 + 1 + 1 = 3$

$c_{2,2} = w_{2,3} + w_{2,9} = 1 + 1 = 2$

$c_{2,4} = w_{2,5} + w_{2,10} = 1 + 2 = 3$

**m1=(1, 1, 0, 0)**  
**m2=(1, 1, 1, 1)**  
**m3=(0, 0, 1, 1)**  
**m4=(0, 1, 1, 1)**  
**m5=(0, 0, 1, 0)**  
**m6=(1, 0, 1, 0)**

$$k_3 = \frac{c_{3,2}m_2 + c_{3,3}m_3 + c_{3,4}m_4 + c_{3,5}m_5 + c_{3,6}m_6}{\sqrt{c_{3,2}^2 + c_{3,3}^2 + c_{3,4}^2 + c_{3,5}^2 + c_{3,6}^2}} \quad (3)$$

$$c_{3,2} = w_{3,3} + w_{3,9} = 1 + 1 = 2$$

$$\begin{aligned} c_{3,3} &= w_{3,4} + w_{3,6} + w_{3,12} \\ &= 2 + 1 + 2 = 5 \end{aligned}$$

$$c_{3,4} = w_{3,5} + w_{3,10} = 1 + 2 = 3$$

$$c_{3,5} = w_{3,7} = 1$$

$$c_{3,6} = w_{3,11} = 2$$

D1 =(2, 1, 0, 0) m1

D2 =(5, 1, 0, 0) m1

D3 =(1, 1, 1, 1) m2

D4 =(0, 0, 2, 2) m3

D5 =(0, 1, 1, 2) m4

D6 =(0, 0, 1, 1) m3

D7 =(0, 0, 1, 0) m5

D8 =(1, 1, 0, 0) m1

D9 =(2, 1, 1, 1) m2

D10=(0, 2, 2, 2) m4

D11=(1, 0, 2, 0) m6

D12=(0, 0, 2, 1) m3

$$k_3 = \frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}}$$

$$k_4 = \frac{c_{4,2}m_2 + c_{4,3}m_3 + c_{4,4}m_4}{\sqrt{c_{4,2}^2 + c_{4,3}^2 + c_{4,4}^2}} \quad (4)$$

$$c_{4,2} = w_{4,3} + w_{4,9} = 1 + 1 = 2$$

$$c_{4,3} = w_{4,4} + w_{4,6} + w_{4,12} = 2 + 1 + 1 = 4$$

$$c_{4,4} = w_{4,5} + w_{4,10} = 2 + 2 = 4$$

$$k_4 = \frac{2m_2 + 4m_3 + 4m_4}{6}$$

D1 = (2, 1, 0, 0) m1  
 D2 = (5, 1, 0, 0) m1  
 D3 = (1, 1, 1, 1) m2  
 D4 = (0, 0, 2, 2) m3  
 D5 = (0, 1, 1, 2) m4  
 D6 = (0, 0, 1, 1) m3  
 D7 = (0, 0, 1, 0) m5  
 D8 = (1, 1, 0, 0) m1  
 D9 = (2, 1, 1, 1) m2  
 D10 = (0, 2, 2, 2) m4  
 D11 = (1, 0, 2, 0) m6  
 D12 = (0, 0, 2, 1) m3

query

$$q = 2k_1 + 3k_2 - k_3$$

using key 3

$$q = 2 * \left( \frac{8m_1 + 3m_2 + m_6}{\sqrt{74}} \right) + 3 * \left( \frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}} \right) - \left( \frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}} \right)$$

in m 10th

$$q = 3.779m_1 + 1.672m_2 - 0.762m_3 + 1.461m_4 - 0.152m_5 - 0.073m_6$$

$$d_1 = 2k_1 + k_2$$

$$d_1 = 2 * \left( \frac{8m_1 + 3m_2 + m_6}{\sqrt{74}} \right) + \left( \frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}} \right)$$

$$d_1 = 2.50m_1 + 1.124m_2 + 0.640m_4 + 0.232m_6$$

$$\text{sim}(q, d_j) = \frac{\sum_r S_{d,r} \cdot S_{q,r}}{\sqrt{\sum_r S_{d,r}^2 \cdot \sum_r S_{q,r}^2}}$$

$m_1 \times m_1$	$m_2 \times m_2$	$m_4 \times m_4$	$m_6 \times m_6$
------------------	------------------	------------------	------------------

$$\text{sim}(q, d_1) = \frac{2.50 * 3.779 + 1.124 * 1.672 + 0.640 * 1.461 - 0.232 * 0.073}{\sqrt{(2.50^2 + 1.124^2 + 0.640^2 + 0.232^2) * (3.779^2 + 1.672^2 + 0.762^2 + 1.461^2 + 0.152^2 + 0.073^2)}}$$

in degree

min terms

min query

D1 = (2, 1, 0, 0) m1  
 D2 = (5, 1, 0, 0) m1  
 D3 = (1, 1, 1, 1) m2  
 D4 = (0, 0, 2, 2) m3  
 D5 = (0, 1, 1, 2) m4  
 D6 = (0, 0, 1, 1) m3  
 D7 = (0, 0, 1, 0) m5  
 D8 = (1, 1, 0, 0) m1  
 D9 = (2, 1, 1, 1) m2  
 D10 = (0, 2, 2, 2) m4  
 D11 = (1, 0, 2, 0) m6  
 D12 = (0, 0, 2, 1) m3

$$q = 3.779m_1 + 1.672m_2 - 0.762m_3 + 1.461m_4 - 0.152m_5 - 0.073m_6$$

$$d_4 = 2k_3 + 2k_4$$

$$d_4 = 2 * \left( \frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}} \right) + 2 * \left( \frac{2m_2 + 4m_3 + 4m_4}{6} \right)$$

$$d_4 = 1.277m_2 + 2.858m_3 + 2.248m_4 + 0.305m_5 + 0.610m_6$$

$$sim(q, d_j) = \frac{\sum_r S_{d,r} \cdot S_{q,r}}{\sqrt{\sum_r S_{d,r}^2 \cdot \sum_r S_{q,r}^2}}$$

$$\begin{aligned}
 sim(q, d_4) &= \frac{1.277 * 1.672 - 2.858 * 0.762 + 2.248 * 1.461 - 0.305 * 0.152 - 0.610 * 0.073}{\sqrt{(1.277^2 + 2.858^2 + 2.248^2 + 0.305^2 + 0.610^2) * (3.779^2 + 1.672^2 + 0.762^2 + 1.461^2 + 0.152^2 + 0.073^2)} \\
 &= 0.181
 \end{aligned}$$

# Degree of similarity

Doc	Sim
D1	0.974
D2	0.952
D3	0.697
D4	0.181
D5	0.419
D6	0.181
D7	0.124
D8	0.981
D9	0.806
D10	0.485
D11	0.494
D12	0.162

## Rank

D8	m1
D1	
D2	
D9	m2
D3	
D11	m6
D10	m4
D5	
D4	m3
D6	
D12	
D7	m5

និង Pattern / នៃព័ត៌មាននេះ  
គឺមានរយៈការ  
ដែលមានរយៈការ / m1 m2 m3 m4 m5 m6

$$D1 = (2, 1, 0, 0) \text{ m1}$$

$$D2 = (5, 1, 0, 0) \text{ m1}$$

$$D3 = (1, 1, 1, 1) \text{ m2}$$

$$D4 = (0, 0, 2, 2) \text{ m3}$$

$$D5 = (0, 1, 1, 2) \text{ m4}$$

$$D6 = (0, 0, 1, 1) \text{ m3}$$

$$D7 = (0, 0, 1, 0) \text{ m5}$$

$$D8 = (1, 1, 0, 0) \text{ m1}$$

$$D9 = (2, 1, 1, 1) \text{ m2}$$

$$D10 = (0, 2, 2, 2) \text{ m4}$$

$$D11 = (1, 0, 2, 0) \text{ m6}$$

$$D12 = (0, 0, 2, 1) \text{ m3}$$

$$q = 2k_1 + 3k_2 - k_3$$

# Generalizes Vector Space Model (GVSM)

---

## Conclusions

- เอกสารที่มี minterm เดียวกันจะมีความตรงประเด็นใกล้เคียงกัน  
เนื่องจากลักษณะการปรากฏของ Keyword มีความคล้ายกัน
- Keyword อาจมีความเกี่ยวข้องกันได้ เช่น Cat และ Tiger (มีโอกาสเป็นไปได้สูงที่เอกสารที่มี Cat อาจมี Tiger อยู่ด้วย) ซึ่งหากมีการเรียกคืน Keyword ได้ Keyword หนึ่ง เอกสารที่มีอีก Keyword หนึ่งก็จะตรงประเด็นด้วย

ภาษาไทย สำหรับ GVSM