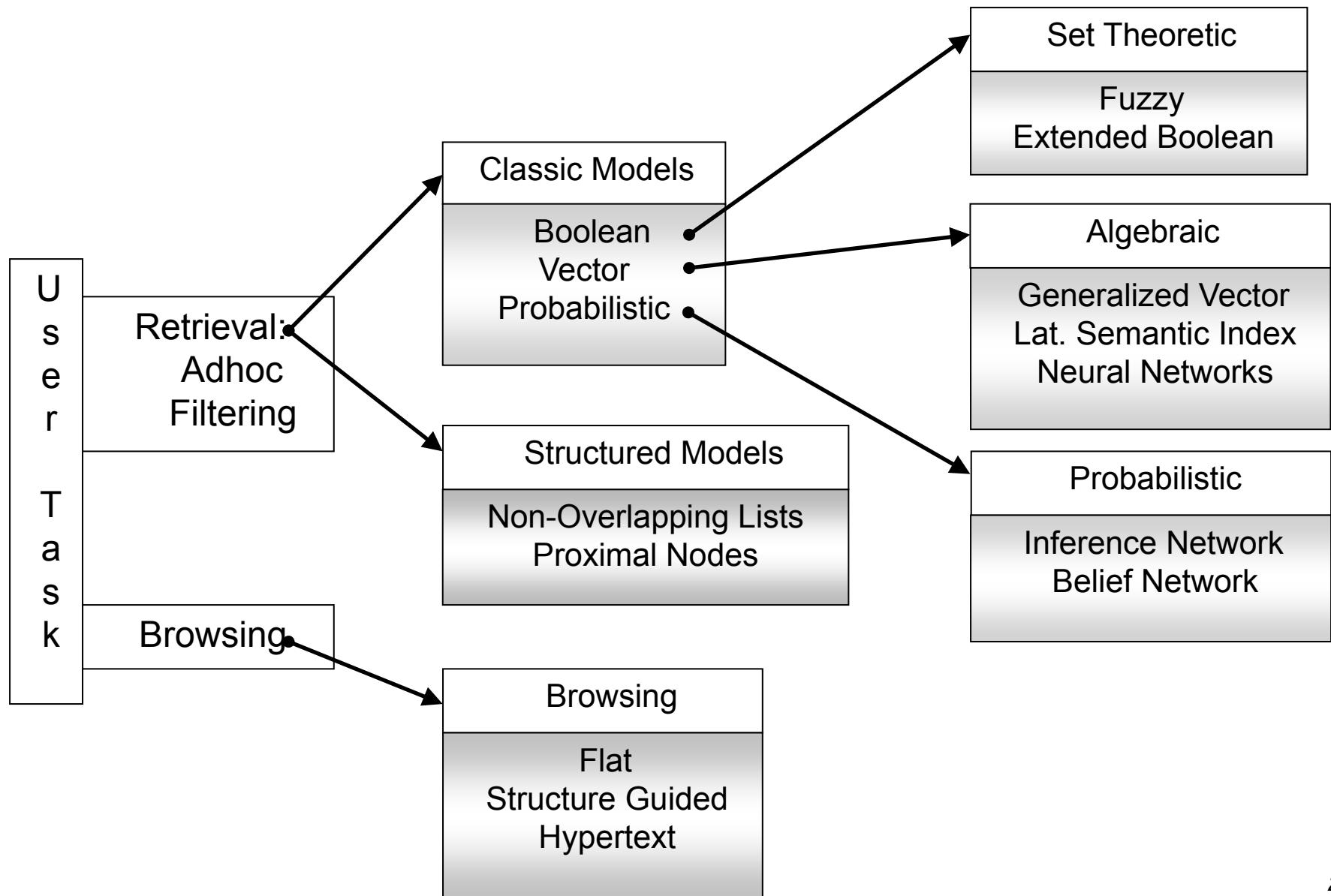


O. front weeks

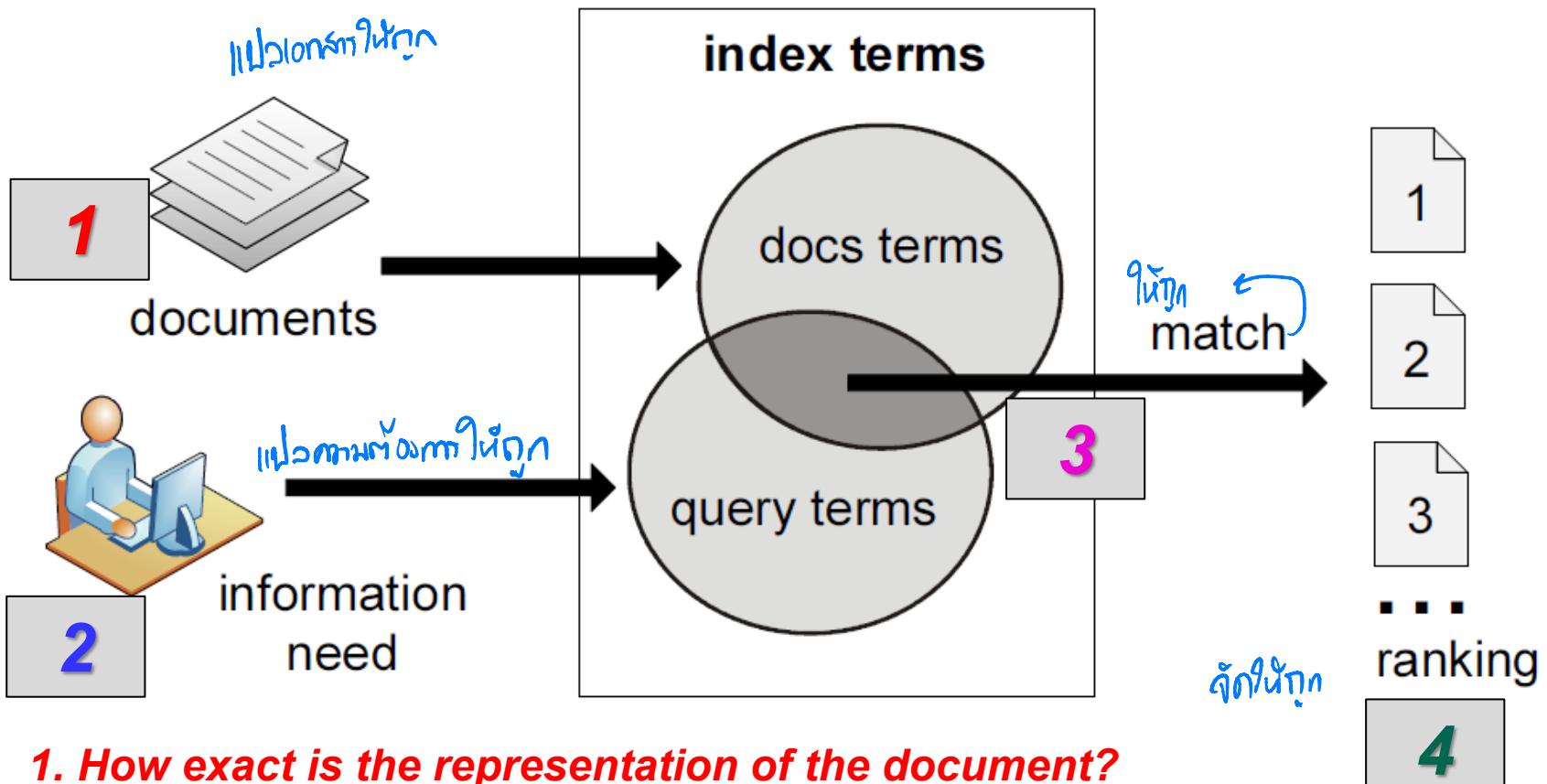
Chapter 02

Modeling

IR Models



IR Problem



1. **How exact is the representation of the document?**
2. **How exact is the representation of the query?** ผู้ใช้งานต้องกำหนดความต้องการของผู้ใช้ตัวเอง
3. **How well is query matched to data?** ข้อมูลที่ได้ ผู้ใช้งานต้องการ
4. **How relevant is the result to the query?** ค่าความถูกต้องของผลลัพธ์ เช่น คำว่า apple ค่าความถูกต้องสูง

Probabilistic Model

វិធានបែកលើអង្គភាពទូទៅ return តម្លៃរាយរាយ
ដើម្បី Doc មានរាយនៅលើ

កំណត់ស្ថាប័ន្ទូយ : ចំណាំបានអាណាពិជ្ជ (សំណងការពេញចាយបានត្រួតពិនិត្យ) → ត្រួត Doc ដែលមានរាយណ៍ (សំណងការពុំង)

- Objective: to capture the IR problem using a probabilistic framework
- Given a user query, there is an **ideal answer set**
- Querying as specification of the properties of this ideal answer set (clustering)
- But, what are these properties?
- **Guess at the beginning** what they could be (i.e., guess initial description of ideal answer set)
- **Improve by iteration**

ໂຄສາກົດ return ຢອດຂາ (ນິຍົມໂຄສາໃຫຍ່)

Retrieved Document

$K_1 = \text{Cat}$

$K_2 = \text{Dog}$

$N = 20$

$R = 12$ mark ດ້ວຍມີເຄີຍ

not $R = 8$

$d1 = \{1,1\} R$

$d2 = \{1,1\} R$

$d3 = \{1,1\} R$

$d4 = \{1,1\} R$

$d5 = \{1,1\}$

$d6 = \{1,0\} R$

$d7 = \{1,0\} R$

$d8 = \{1,0\} R$

$d9 = \{1,0\} R$

$d10 = \{1,0\}$

$d11 = \{1,0\}$

$d12 = \{0,1\} R$

$d13 = \{0,1\} R$

$d14 = \{0,1\} R$

$d15 = \{0,1\}$

$d16 = \{0,1\}$

$d17 = \{0,1\}$

$d18 = \{0,0\} R$

$d19 = \{0,0\}$

$d20 = \{0,0\}$

ການທັງໝົດ ການຕາມປະດຳ ດີ

$$\frac{12}{20}$$

20 Doc

Doc និង mark រវាងព័ត៌មាន

លក្ខណៈកសាង

Doc និង mark រវាងព័ត៌មាន

Relevance Docs

$K_1 = \text{Cat}$

$K_2 = \text{Dog}$

$N = 20$

$R = 12$

$\text{not } R = 8$

នៃ all Doc និង keyword 1 នៅក្នុង Doc

$K_1, N = 11$ (8+3)
នៃ all mark Doc ដែលមាននឹង keyword 1 នៅក្នុង Doc

$K_1, R = 8$
នៃ all Doc ដែលមាននឹង keyword 1 នៅក្នុង Doc

$K_1, \text{not } R = 3$

$\text{not } K_1, R = 4$

$\text{not } K_1, \text{not } R = 5$

$K_2, N = 11$

$K_2, R = 7$

$K_2, \text{not } R = 4$

$\text{not } K_2, R = 5$

$\text{not } K_2, \text{not } R = 4$

$d_1 = \{1, 1\}$

$d_2 = \{1, 1\}$

$d_3 = \{1, 1\}$

$d_4 = \{1, 1\}$

$d_6 = \{1, 0\}$

$d_7 = \{1, 0\}$

$d_8 = \{1, 0\}$

$d_9 = \{1, 0\}$

$d_{12} = \{0, 1\}$

$d_{13} = \{0, 1\}$

$d_{14} = \{0, 1\}$

$d_{18} = \{0, 0\}$

$d_5 = \{1, 1\}$

$d_{10} = \{1, 0\}$

$d_{11} = \{1, 0\}$

$d_{15} = \{0, 1\}$

$d_{16} = \{0, 1\}$

$d_{17} = \{0, 1\}$

$d_{19} = \{0, 0\}$

$d_{20} = \{0, 0\}$

$K_1 = 3$

$K_2 = 4$

$K_1 = 8$

→ ចំនួនការពែង keyword 1 នៃកសាង

$K_2 = 7$

នៃ keyword 9 នូវការបញ្ជាក់ទំនាក់ទំនង

N → ចំនួន តម្លៃការកំណត់ (តើចាប់ពីរាយការកំណត់ឡើង)

R → ចំនួន Doc និង mark រវាងព័ត៌មាន

n_n → ចំនួន all Doc និង k_n នៃ Doc

r_n → ចំនួនកសាង និង mark រវាងព័ត៌មាន នៃ Doc និង k_n

Retrieved Document

$$K_1 = \text{Cat}$$

$$K_2 = \text{Dog}$$

$$N = 20$$

$$R = 12$$

$$\text{not } R = 8$$

$$K_1, N = 11$$

$$K_1, R = 8$$

$$K_1, \text{not } R = 3$$

$$\text{not } K_1, R = 4$$

$$\text{not } K_1, \text{not } R = 5$$

$$K_2, N = 11$$

$$K_2, R = 7$$

$$K_2, \text{not } R = 4$$

$$\text{not } K_2, R = 5$$

$$\text{not } K_2, \text{not } R = 4$$

$$d1 = \{1,1\} \textcolor{red}{R}$$

$$d2 = \{1,1\} \textcolor{red}{R}$$

$$d3 = \{1,1\} \textcolor{red}{R}$$

$$d4 = \{1,1\} \textcolor{red}{R}$$

$$d5 = \{1,1\}$$

$$d6 = \{1,0\} \textcolor{red}{R}$$

$$d7 = \{1,0\} \textcolor{red}{R}$$

$$d8 = \{1,0\} \textcolor{red}{R}$$

$$d9 = \{1,0\} \textcolor{red}{R}$$

$$d10 = \{1,0\}$$

$$d11 = \{1,0\}$$

$$d12 = \{0,1\} \textcolor{red}{R}$$

$$d13 = \{0,1\} \textcolor{red}{R}$$

$$d14 = \{0,1\} \textcolor{red}{R}$$

$$d15 = \{0,1\}$$

$$d16 = \{0,1\}$$

$$d17 = \{0,1\}$$

$$d18 = \{0,0\} \textcolor{red}{R}$$

$$d19 = \{0,0\}$$

$$d20 = \{0,0\}$$

$$P(R|N) = \frac{12}{20} \Rightarrow P(R)$$

$$\text{ความน่าจะเป็น Doc ที่มี } k_1, \text{ ความน่าจะเป็น } \frac{8}{12}$$

$$P(K_1|R) = \frac{8}{12} \Rightarrow \frac{r_1}{R}$$

$$\text{ความน่าจะเป็น all Doc ที่มี } k_1,$$

$$\Rightarrow \frac{R - r_1}{R} = \frac{12 - 8}{12}$$

$$\text{ความน่าจะเป็นทั้งหมด, ต่อไปนี้จะเรียกว่า: } \bar{R}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$\text{ความน่าจะเป็นทั้งหมด, ต่อไปนี้จะเรียกว่า: } \bar{N}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$\text{ความน่าจะเป็นทั้งหมด, ต่อไปนี้จะเรียกว่า: } \bar{N} - R$$

$$P(K_1|R) + P(\bar{K}_1|R) = 1$$

$$P(K_1|\bar{R}) + P(\bar{K}_1|\bar{R}) = 1$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

$$P(K_2|R) + P(\bar{K}_2|R) = 1$$

$$P(K_2|\bar{R}) + P(\bar{K}_2|\bar{R}) = 1$$

What is

$P(\bar{K}_1|R), P(\bar{K}_2|R) ???$

Retrieved Document

$$K_1 = \text{Cat}$$

$$K_2 = \text{Dog}$$

$$P(K_1|R) = \frac{8}{12}$$

$$P(\bar{K}_1|R) = \frac{4}{12}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

$$d1 = \{1,1\} \textcolor{red}{R}$$

$$d2 = \{1,1\} \textcolor{red}{R}$$

$$d3 = \{1,1\} \textcolor{red}{R}$$

$$d4 = \{1,1\} \textcolor{red}{R}$$

$$d5 = \{1,1\}$$

$$d6 = \{1,0\} \textcolor{red}{R}$$

$$d7 = \{1,0\} \textcolor{red}{R}$$

$$d8 = \{1,0\} \textcolor{red}{R}$$

$$d9 = \{1,0\} \textcolor{red}{R}$$

$$d10 = \{1,0\}$$

$$d11 = \{1,0\}$$

$$d12 = \{0,1\} \textcolor{red}{R}$$

$$d13 = \{0,1\} \textcolor{red}{R}$$

$$d14 = \{0,1\} \textcolor{red}{R}$$

$$d15 = \{0,1\}$$

$$d16 = \{0,1\}$$

$$d17 = \{0,1\}$$

$$d18 = \{0,0\} \textcolor{red}{R}$$

$$d19 = \{0,0\}$$

$$d20 = \{0,0\}$$

We need ??? 

⇒ អត្ថបទរឿង query នៃវត្ថុ Doc

$$\text{sim}(d_j, q) = ???$$

$$\text{sim}(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\bar{R}|\vec{d}_j)}$$

រាយការ. ភ័ព្យប្រាក់លិខិត

រាយការ. ចូលរួមចំណាំរាយការ

Bayes' rule

$$\frac{P(R|\vec{d}_j)}{P(\bar{R}|\vec{d}_j)} = \frac{P(\vec{d}_j|R) \cdot P(R)}{P(\vec{d}_j|\bar{R}) \cdot P(\bar{R})}$$

Bayes' Rule

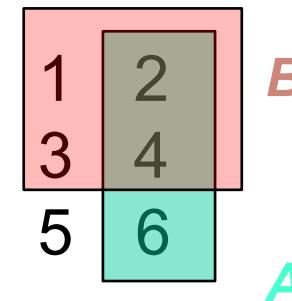
ทำม

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5



$$P(A|N) = \frac{3}{6} \rightarrow P(A)$$

$$P(B|N) = \frac{4}{6} \rightarrow P(B)$$

$$P(A|B) = \frac{2}{4} \rightarrow \text{ความน่าจะเป็นที่ลูกเต๋าจะออกเลขคู่จากเหตุการณ์ที่ลูกเต่ามีแต้มน้อยกว่า 5}$$

$$P(B|A) = \frac{2}{3} \rightarrow \text{ความน่าจะเป็นที่ลูกเต๋าจะมีค่าน้อยกว่า 5 จากเหตุการณ์ที่ลูกเต่ามีแต้มเป็นเลขคู่}$$

Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5

$$P(A) = \frac{3}{6}$$

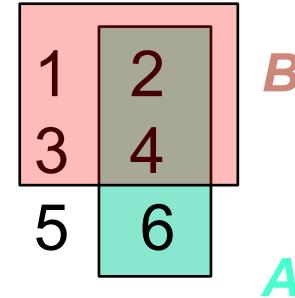
$$P(B) = \frac{4}{6}$$

$$P(A|B) = \frac{2}{4}$$

$$P(B|A) = \frac{2}{3}$$

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{2}{6} * \frac{6}{4} = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} \\ &= \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{6} * \frac{6}{3} = \frac{2}{3} \end{aligned}$$



Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5

$$P(A) = \frac{3}{6}$$

$$P(B) = \frac{4}{6}$$

$$P(A|B) = \frac{2}{4}$$

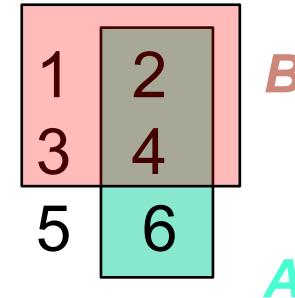
$$\mathbf{P(B|A) = ???}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A|B) = \frac{P(B \cap A)}{P(B)}$$

$$P(B \cap A) = P(A|B) * P(B)$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \rightarrow P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

$$P(B|A) = \frac{\frac{2}{4} * \frac{4}{6}}{\frac{3}{6}} = \frac{2}{6} * \frac{6}{3} = \frac{2}{3}$$



Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5

$$P(A) = \frac{3}{6}$$

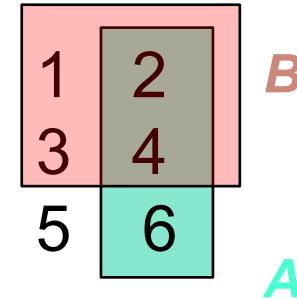
$$P(B) = \frac{4}{6}$$

$$P(B|A) = \frac{2}{3}$$

$$\mathbf{P(A|B) = ???}$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$P(A|B) = \frac{\frac{2}{3} * \frac{3}{6}}{\frac{4}{6}} = \frac{2}{6} * \frac{6}{4} = \frac{1}{2}$$



Probabilistic Model

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\bar{R}|\vec{d}_j)}$$

ຕາມປະເທດ
ຖຸນຕາມປະເລກ

$$\begin{aligned} &= \frac{P(\vec{d}_j|R) * P(R)}{P(\vec{d}_j)} \\ &= \frac{P(\vec{d}_j|R) * P(R)}{P(\vec{d}_j|\bar{R}) * P(\bar{R})} \\ &= \frac{P(\vec{d}_j|R) * P(R)}{P(\vec{d}_j)} * \frac{P(\vec{d}_j)}{P(\vec{d}_j|\bar{R}) * P(\bar{R})} \end{aligned}$$

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) * P(R)}{P(\vec{d}_j|\bar{R}) * P(\bar{R})}$$

Bayes' Rule

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

$$P(R) = \frac{12}{20}$$

$$P(\bar{R}) = \frac{8}{20}$$

$$P(K_1|R) = \frac{8}{12}$$

$$P(\bar{K}_1|R) = \frac{4}{12}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

Retrieved Document

- $d_1 = \{1,1\} \textcolor{red}{R}$
- $d_2 = \{1,1\} \textcolor{red}{R}$
- $d_3 = \{1,1\} \textcolor{red}{R}$
- $d_4 = \{1,1\} \textcolor{red}{R}$
- $d_5 = \{1,1\}$
- $d_6 = \{1,0\} \textcolor{red}{R}$
- $d_7 = \{1,0\} \textcolor{red}{R}$
- $d_8 = \{1,0\} \textcolor{red}{R}$
- $d_9 = \{1,0\} \textcolor{red}{R}$
- $d_{10} = \{1,0\}$
- $d_{11} = \{1,0\}$
- $d_{12} = \{0,1\} \textcolor{red}{R}$
- $d_{13} = \{0,1\} \textcolor{red}{R}$
- $d_{14} = \{0,1\} \textcolor{red}{R}$
- $d_{15} = \{0,1\}$
- $d_{16} = \{0,1\}$
- $d_{17} = \{0,1\}$
- $d_{18} = \{0,0\} \textcolor{red}{R}$
- $d_{19} = \{0,0\}$
- $d_{20} = \{0,0\}$



Simple Probabilistic Method

គាយករណ៍នេះ តើអាមេរិក ទុង Doc នៃកីឡា, កីឡា 4 all k₁, k₂ តម្លៃរង់ទូលាយ

$$P(R|(1, 1)) = \frac{4}{5} \text{ all កីឡា, កីឡា}$$

$$P(R|(1, 0)) = \frac{4}{6}$$

$$P(R|(0, 1)) = \frac{3}{6}$$

$$P(R|(0, 0)) = \frac{1}{3}$$

$d_{21} = \{1,0\} \rightarrow \text{sim}(d_{21}, q) = ???$

$$P(R) = \frac{12}{20}$$

$$P(\bar{R}) = \frac{8}{20}$$

$$P(K_1|R) = \frac{8}{12}$$

$$P(\bar{K}_1|R) = \frac{4}{12}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

Retrieved Document

- $d_1 = \{1,1\} \textcolor{red}{R}$
- $d_2 = \{1,1\} \textcolor{red}{R}$
- $d_3 = \{1,1\} \textcolor{red}{R}$
- $d_4 = \{1,1\} \textcolor{red}{R}$
- $d_5 = \{1,1\}$
- $d_6 = \{1,0\} \textcolor{red}{R}$
- $d_7 = \{1,0\} \textcolor{red}{R}$
- $d_8 = \{1,0\} \textcolor{red}{R}$
- $d_9 = \{1,0\} \textcolor{red}{R}$
- $d_{10} = \{1,0\}$
- $d_{11} = \{1,0\}$
- $d_{12} = \{0,1\} \textcolor{red}{R}$
- $d_{13} = \{0,1\} \textcolor{red}{R}$
- $d_{14} = \{0,1\} \textcolor{red}{R}$
- $d_{15} = \{0,1\}$
- $d_{16} = \{0,1\}$
- $d_{17} = \{0,1\}$
- $d_{18} = \{0,0\} \textcolor{red}{R}$
- $d_{19} = \{0,0\}$
- $d_{20} = \{0,0\}$

$d_1 = \{1,1\}$

$d_6 = \{1,0\}$

$\rightarrow \vec{q} \text{ Vector = normalizing keyword}$

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) * P(R)}{P(\vec{d}_j|\bar{R}) * P(\bar{R})}$$

$$sim(d_1, q) = \frac{P(\vec{d}_1|R) * P(R)}{P(\vec{d}_1|\bar{R}) * P(\bar{R})}$$

$$= \frac{P(K_1|R) * P(K_2|R)}{P(K_1|\bar{R}) * P(K_2|\bar{R})} * \frac{12}{8}$$

$$= \frac{\frac{8}{12} * \frac{7}{12}}{\frac{3}{8} * \frac{4}{8}} * \frac{12}{8} = \frac{28}{9}$$

$$sim(d_6, q) = \frac{P(\vec{d}_6|R) * P(R)}{P(\vec{d}_6|\bar{R}) * P(\bar{R})}$$

$$= \frac{P(K_1|R) * P(\bar{K}_2|R)}{P(K_1|\bar{R}) * P(\bar{K}_2|\bar{R})} * \frac{12}{8}$$

$$= \frac{\frac{8}{12} * \frac{5}{12}}{\frac{3}{8} * \frac{4}{8}} * \frac{12}{8} = \frac{20}{9}$$

$$P(R) = \frac{12}{20}$$

$$P(\bar{R}) = \frac{8}{20}$$

$$P(K_1|R) = \frac{8}{12}$$

$$P(\bar{K}_1|R) = \frac{4}{12}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

Retrieved Document

d1 = {1,1} **R**
 d2 = {1,1} **R**
 d3 = {1,1} **R**
 d4 = {1,1} **R**
 d5 = {1,1}
 d6 = {1,0} **R**
 d7 = {1,0} **R**
 d8 = {1,0} **R**
 d9 = {1,0} **R**
 d10 = {1,0}
 d11 = {1,0}
 d12 = {0,1} **R**
 d13 = {0,1} **R**
 d14 = {0,1} **R**
 d15 = {0,1}
 d16 = {0,1}
 d17 = {0,1}
 d18 = {0,0} **R**
 d19 = {0,0}
 d20 = {0,0}

$$sim(d_{12}, q) = \frac{P(\vec{d_{12}}|R) * P(R)}{P(\vec{d_{12}}|\bar{R}) * P(\bar{R})}$$

$$= \frac{P(\bar{K}_1|R) * P(K_2|R)}{P(\bar{K}_1|\bar{R}) * P(K_2|\bar{R})} * \frac{12}{8}$$

$$= \frac{\frac{4}{12} * \frac{7}{12}}{\frac{5}{8} * \frac{4}{8}} * \frac{12}{8} = \frac{28}{30}$$

$$sim(d_{18}, q) = \frac{P(\vec{d_{18}}|R) * P(R)}{P(\vec{d_{18}}|\bar{R}) * P(\bar{R})}$$

$$= \frac{P(\bar{K}_1|R) * P(\bar{K}_2|R)}{P(\bar{K}_1|\bar{R}) * P(\bar{K}_2|\bar{R})} * \frac{12}{8}$$

$$= \frac{\frac{4}{12} * \frac{5}{12}}{\frac{5}{8} * \frac{4}{8}} * \frac{12}{8} = \frac{10}{15}$$

Probabilistic Model

$$\begin{array}{lll} d_j \rightarrow \{1,1\} & sim(d_j, q) = \frac{28}{9} & \xrightarrow{\hspace{1cm}} sim(d_j, q) = \frac{28}{37} = 0.757 \\ d_j \rightarrow \{1,0\} & sim(d_j, q) = \frac{20}{9} & \xrightarrow{\hspace{1cm}} sim(d_j, q) = \frac{20}{29} = 0.690 \\ d_j \rightarrow \{0,1\} & sim(d_j, q) = \frac{28}{30} & \xrightarrow{\hspace{1cm}} sim(d_j, q) = \frac{28}{58} = 0.483 \\ d_j \rightarrow \{0,0\} & sim(d_j, q) = \frac{10}{15} & \xrightarrow{\hspace{1cm}} sim(d_j, q) = \frac{10}{25} = 0.400 \end{array}$$

Probabilistic value $\in [0,1]$

Then

normalize 100% probability

$$sim(d_j, q) = \frac{sim(d_j, q)}{1 + sim(d_j, q)}$$

Binary Independence Retrieval Model (BIR)

$$P(R) = \frac{12}{20}$$

$$P(\bar{R}) = \frac{8}{20}$$

$$P(K_1|R) = \frac{8}{12}$$

$$P(\bar{K}_1|R) = \frac{4}{12}$$

$$P(K_1|\bar{R}) = \frac{3}{8}$$

$$P(\bar{K}_1|\bar{R}) = \frac{5}{8}$$

$$P(K_2|R) = \frac{7}{12}$$

$$P(\bar{K}_2|R) = \frac{5}{12}$$

$$P(K_2|\bar{R}) = \frac{4}{8}$$

$$P(\bar{K}_2|\bar{R}) = \frac{4}{8}$$

Retrieved Document

d1 = {1,1} **R**
 d2 = {1,1} **R**
 d3 = {1,1} **R**
 d4 = {1,1} **R**
 d5 = {1,1}
 d6 = {1,0} **R**
 d7 = {1,0} **R**
 d8 = {1,0} **R**
 d9 = {1,0} **R**
 d10 = {1,0}
 d11 = {1,0}
 d12 = {0,1} **R**
 d13 = {0,1} **R**
 d14 = {0,1} **R**
 d15 = {0,1}
 d16 = {0,1}
 d17 = {0,1}
 d18 = {0,0} **R**
 d19 = {0,0}
 d20 = {0,0}

	{1,1}	{1,0}	{0,1}	{0,0}
Simple	0.800	0.667	0.500	0.333
BIR	0.757	0.690	0.483	0.400

Probabilistic Model

Smooth Tuning

ถ้าเกิด เมื่อ 0 วิธีแก้ 1 เที่ยงตัวต่อไป

$$P(K_i|R) = \frac{r_i}{R} \quad \rightarrow \quad P(K_i|R) = \frac{r_i + 0.5}{R + 1} \quad \rightarrow \quad P(K_i|R) = \frac{r_i + \frac{n_i}{N}}{R + 1}$$

วิธีแก้ 2 กรณีที่มีผลลัพธ์ที่ดีกว่าเดิม

$$P(\bar{K}_i|R) = 1 - P(K_i|R)$$

$$P(K_i|\bar{R}) = \frac{n_i - ri}{N - R} \quad \rightarrow \quad P(K_i|\bar{R}) = \frac{n_i - ri + 0.5}{N - R + 1} \quad \rightarrow \quad P(K_i|\bar{R}) = \frac{n_i - ri + \frac{n_i}{N}}{N - R + 1}$$

$$P(\bar{K}_i|\bar{R}) = 1 - P(K_i|\bar{R})$$

Probabilistic Ranking Principle

ឬ ផ្លូវការ ទូនាស់ output ដែលខ្លួន → កំណត់តារ (Smooth ស្នើសុំប័ណ្ណការបង្ហី) → លទ្ធផល

- Given a user query q and a document d_j , the probabilistic model tries to estimate the probability that the user will find the document d_j interesting (i.e., relevant). The model assumes that this probability of relevance **depends on the query and the document representations only**. Ideal answer set is referred to as R and should maximize the probability of relevance. Documents in the set R are predicted to be relevant.
- But,
 - how to compute probabilities?
 - what is the sample space?

The Ranking

- Probabilistic ranking computed as:

- $\square \text{sim}(d_j, q) = P(d_j \text{ relevant-to } q) / P(d_j \text{ non-relevant-to } q)$

$$\text{sim}(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\bar{R}|\vec{d}_j)}$$

- \square This is the **odds** of the document d_j being relevant
- \square Taking the **odds** minimize the probability of an erroneous judgement

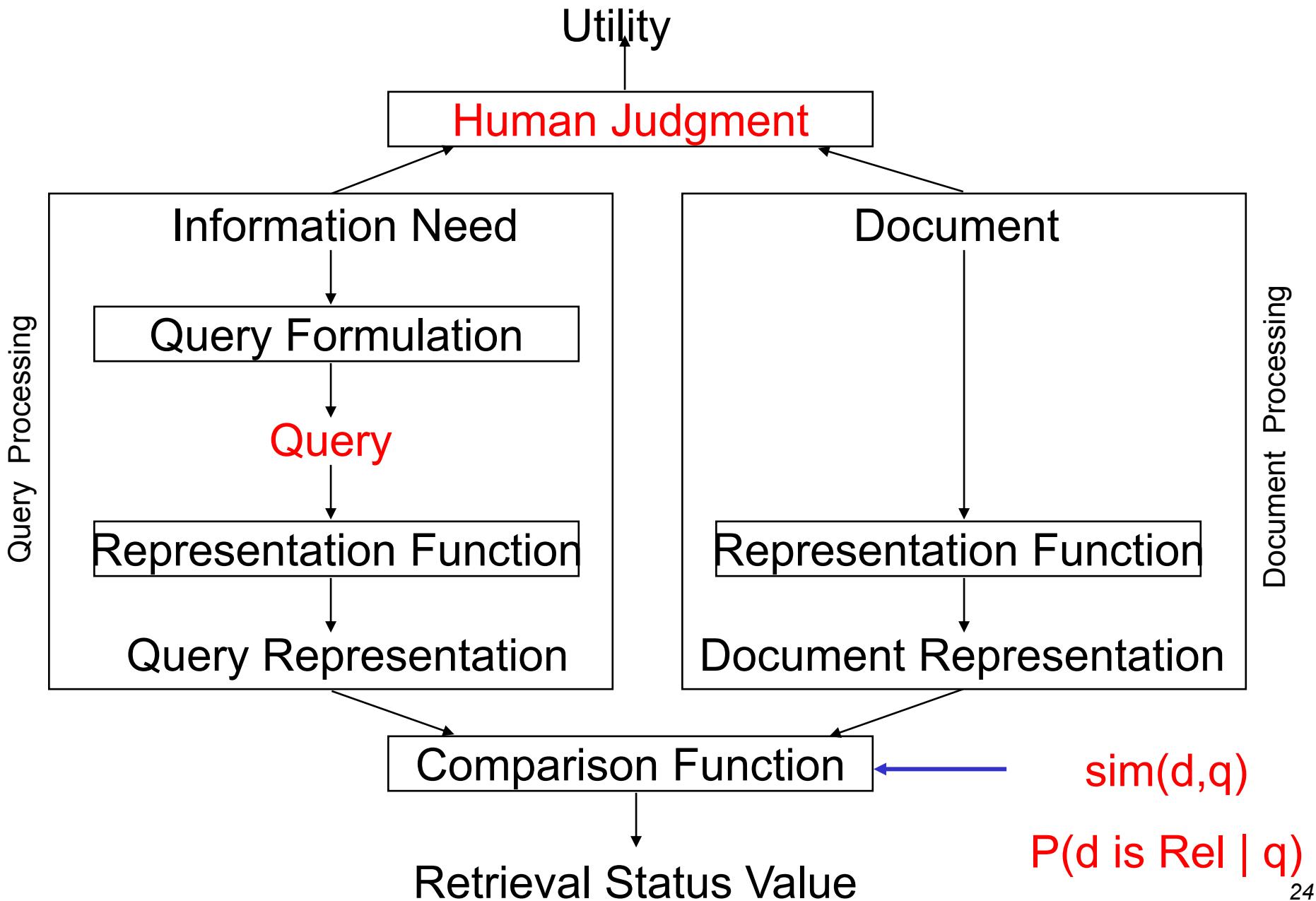
- Definition:

- $\square w_{ij} \in \{0, 1\}$

- $\square P(R|\vec{d}_j)$: probability that given doc is relevant

- $\square P(\bar{R}|\vec{d}_j)$: probability doc is not relevant

Where do the probabilities fit?



Pluses and Minuses

- Advantages: វិនិត្យ នៃការស្វែងរកបាន និង តាមលទ្ធផល ត្រូវបាន និរន័យ។
 - Docs ranked in decreasing order of probability of relevance
- Disadvantages: មិនត្រឹមត្រូវ
 - need to guess initial estimates for $P(k_i | R)$
 - method does not take into account tf and idf factors
ដែលអាចការពារ តាម រាយការ បាន ប៉ុណ្ណោះ ប៉ុណ្ណោះ បាន ប៉ុណ្ណោះ បាន ប៉ុណ្ណោះ

ตัวอย่างโจทย์

$K_n = \{Cat, Dog, Tiger\}$

เอกสารทั้งหมดในระบบมีดังนี้

$D_1 = \{1, 0, 0\}$

$D_2 = \{0, 0, 1\}$

$D_3 = \{1, 0, 1\}$

$D_4 = \{1, 1, 0\}$

$D_5 = \{0, 1, 0\}$

$D_6 = \{0, 1, 1\}$

$D_7 = \{0, 1, 1\}$

$D_8 = \{1, 1, 1\}$

$D_9 = \{1, 0, 0\}$

$D_{10} = \{1, 0, 1\}$

ในการส่ง Query = {1, 0, 1} เข้าไปในระบบ มีผลลัพธ์คือ

$D_3, D_{10}, D_2, D_5, D_9, D_6, D_1$ (ลำดับตามตัวอักษร: เด็กแมว,)

เมื่อนำผลลัพธ์ที่ได้มาวิเคราะห์ และนำไปจัดลำดับความตรงประเด็นของเอกสาร
ทั้งหมดอีกครั้ง ลำดับของความตรงประเด็นใหม่เป็นเท่าใด (จงแสดงวิธีคำนวณ)

ข้อ 2. สมมติในระบบมีเอกสาร 12 เอกสารดังนี้ (bird, cat, dog, tiger คือ Keyword)

- 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}
- D11: {???, ???, ???} } 9 ลักษณะ keyword ของข้อมูล (1 ต่อ 3 keyword)
- D12: {???, ???, ???}

*** เอกสารหมายเลข 11 และ 12 ให้น.ศ.กำหนด keyword เอง อย่างน้อยเอกสารละ 3 keyword

เมื่อส่ง Query เข้าไปในระบบ เอกสารที่ถูกส่งออกมาก็ D1, D8, D2, D3, D4, D7, D6, D9, D10, D11

เด็กหญิงค่าวิเคราะห์เอกสารที่ตรงประเด็นคือ D1, D8, D2, D3, D4, D7, D6, D9 จงตอบคำถาม

2.1 เพื่อให้ได้คำตอบในคำตาม 2.2 เด็กหญิงค่าวิเคราะห์ “การเลือกใช้โมเดลใดเพราะอะไร”

- (A) Probabilistic Model B) Generalize Vector Model C) Extend Boolean Model D) Vector Model

2.2 ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารทุกเอกสาร ในระบบ ตามที่เด็กหญิงค่าวิเคราะห์มา

2.3 หากระบบกำหนดให้ D1 มีความตรงประเด็นมากกว่า D9 ข้อสรุปของเด็กหญิงค่าวิเคราะห์ถูกต้องหรือไม่ อย่างไร หากผิดควรทำอย่างไร

(36 คะแนน) ส่งคำตอบที่ : **IR.CE.KMITL@gmail.com**

ชื่อเมล์: Quiz4_รหัสนักศึกษา เช่น Quiz4_64010109_64010198 ไฟล์แนบชื่อไดกีได
เขียนด้วยลายมือ ส่งภายใน 15/08/2566 ไม่เกินเที่ยงคืน (คัดลอกกันถือว่าผิดจริยธรรม)

หากว่าใน query ไม่ได้ตั้ง query
แต่ว่าการนับก็จะนับจำนวนของรากที่

ပုဂ္ဂန်၏သတေသနများ pro

BM25 (Best Matching 25) Extended Probabilistic Model

BM25

Goals

ទីផ្សារក្នុងសរុប

- All Documents (*not only retrieved documents*) តម្លៃគឺ ≠ នៅ
- Term frequency in each document សរុបចំណាំគឺជាបញ្ជី ... $q_{in\ Doc}$
- Term frequency in query $q_{in\ query}$ 

BM25

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

↑
Inverse document frequency
idf

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 $0 - 1000$

↑
จำนวน keyword ที่มี
จำนวนคำที่ keyword ซ้ำกันในเอกสาร

Document term frequency
↑
Query term frequency

จงหา Doc bird cat tiger
query bird cat
จำนวนคราวเดียวกันในหนึ่ง / 2 หน้า

BM25

Variables

- Inverse document frequency* (จำนวนของเอกสารที่มี Keyword)
- Term frequency* (ความถี่ของ Keyword ในแต่ละเอกสาร)
- Document length normalization* (ความยาวของเอกสาร “จำนวนคำ”)
- Query term frequency* (ความถี่ของ Keyword ในแต่ล่ะ Query)

Inverse document frequency

ความน่าจะเป็นที่เอกสารจะตรงประเด็น

ความน่าจะเป็นที่เอกสารจะ **ไม่**ตรงประเด็น

$$\text{ความน่าจะเป็นที่เอกสารจะตรงประเด็น} = \frac{\text{ความน่าจะเป็นที่เอกสารมี } Keyword \text{ และตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสาร } \textcolor{red}{\text{ไม่มี}} \text{ } Keyword \text{ และตรงประเด็น}}$$

$$= \frac{(r_i + 0.5)/(R + 0.5)}{(R - r_i + 0.5)/(R + 0.5)}$$

$$= \frac{(r_i + 0.5)}{(R - r_i + 0.5)}$$

จำนวนเอกสารที่มี *Keyword* ในเอกสารที่กำหนดค่าตรงประเด็น

จำนวนเอกสารที่ตรงประเด็นทั้งหมด

$$\frac{r_i}{R}$$

กับปีหนัง ทำดังนี้

$$\frac{r_i + 0.5}{R + 0.5}$$



จำนวนเอกสารที่**ไม่มี** *Keyword* ในเอกสารที่กำหนดค่าตรงประเด็น

จำนวนเอกสารที่ตรงประเด็นทั้งหมด

$$\frac{R - r_i}{R} \rightarrow \frac{R - r_i + 0.5}{R + 0.5}$$

Inverse document frequency

ความน่าจะเป็นที่เอกสารจะตรงประเด็น

ความน่าจะเป็นที่เอกสารจะ **ไม่**ตรงประเด็น

$$\text{ความน่าจะเป็นที่เอกสารจะ } \text{ไม่}\text{ตรงประเด็น} = \frac{\text{ความน่าจะเป็นที่เอกสารมี Keyword และ } \text{ไม่}\text{ตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสาร } \text{ไม่มี} \text{ Keyword และ } \text{ไม่}\text{ตรงประเด็น}}$$

$$= \frac{(n_i - ri + 0.5)/(N - R + 0.5)}{(N - ni - R + ri + 0.5)/(N - R + 0.5)}$$

$$= \frac{n_i - ri + 0.5}{N - ni - R + ri + 0.5}$$

จำนวนเอกสารที่มี **Keyword** ในเอกสารที่กำหนดค่า **ไม่**ตรงประเด็น

จำนวนเอกสารที่ **ไม่**ตรงประเด็นทั้งหมด

$$\frac{n_i - ri}{N - R} \rightarrow \frac{n_i - ri + 0.5}{N - R + 0.5}$$

จำนวนเอกสารที่ **ไม่มี** **Keyword** ในเอกสารที่กำหนดค่า **ไม่**ตรงประเด็น

จำนวนเอกสารที่ **ไม่**ตรงประเด็นทั้งหมด

จำนวนเอกสารที่ **ไม่มี** **Keyword** นั้นทั้งหมด — จำนวนเอกสารที่ **ไม่มี** **Keyword** นั้นแล้วตรงประเด็น

$$\frac{N - ni - (R - ri)}{N - R} \rightarrow \frac{N - ni - R + ri + 0.5}{N - R + 0.5}$$

↑ จำนวนเอกสารที่ **ไม่**ตรงประเด็นทั้งหมด
↓ จำนวนเอกสารที่ **ไม่มี** **Keyword** นั้นแล้วตรงประเด็น

Inverse document frequency

ความน่าจะเป็นที่เอกสารจะตรงประเด็น
ความน่าจะเป็นที่เอกสารจะ **ไม่** ตรงประเด็น

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

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

↓
ดูอยู่ 9 แฟ้ม ศักราช / หนังสือ แม้ว่า ฐาน 10 ก็ตาม มาก-น้อย ต่ำสูง

R – จำนวนเอกสารที่ตรงประเด็น

N – จำนวนเอกสารทั้งหมด

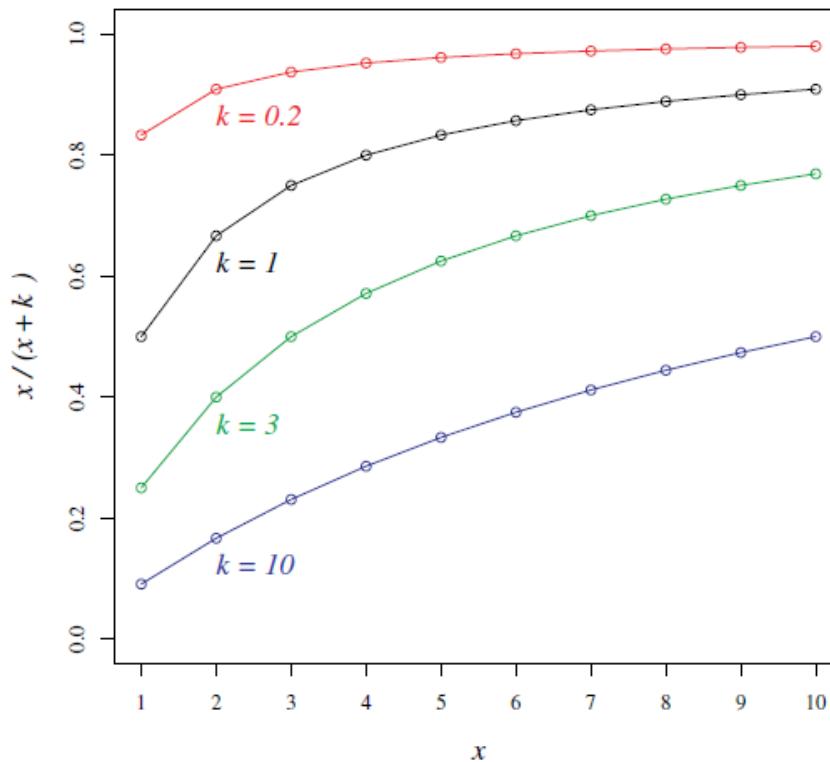
r_i – จำนวนเอกสารที่ตรงประเด็นที่มี keyword i

n_i – จำนวนเอกสารทั้งหมดที่มี keyword i

Document term frequency

จำนวนครั้งที่ Keyword ปรากฏในเอกสาร (ความถี่) $\rightarrow f_{i,j}$

เมื่อ $f_{i,j}$ มีปัญหา $\rightarrow \frac{f_{i,j}}{f_{i,j} + 1} \rightarrow \frac{f_{i,j}}{f_{i,j} + k}$



Document term frequency

ตัวอย่าง

$$\frac{f_{i,j}}{f_{i,j} + k} \rightarrow \boxed{\frac{f_i}{k \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i}}$$

dl – จำนวนคำของเอกสาร j

$avdl$ – จำนวนคำเฉลี่ยของทุกเอกสาร

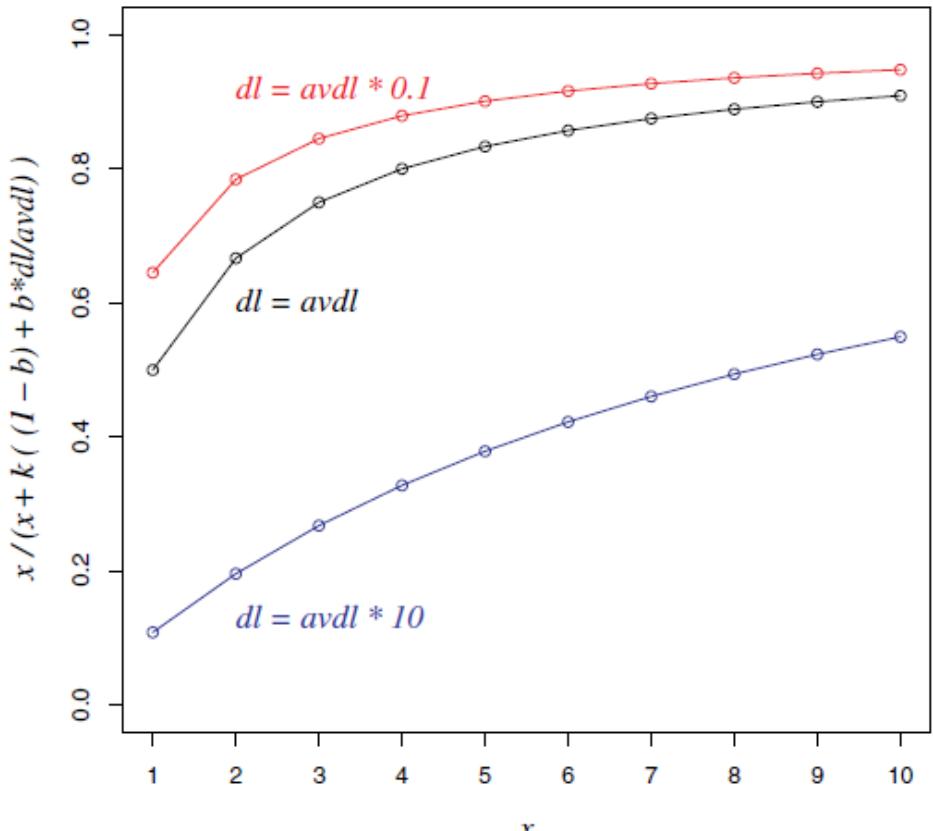
b – ค่าคงที่โดยตาม $TREC$ จะใช้ค่า 0.75 ($0.5 < b < 0.8$) } หนึ่งในสองค่า

k – ค่าคงที่โดยตาม $TREC$ จะใช้ค่า 1.25 ($1.2 < k < 2$) } หนึ่งในสองค่า

ถ้าหาก $model$ นี้

Document term frequency

$$\frac{f_i}{k \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i}$$

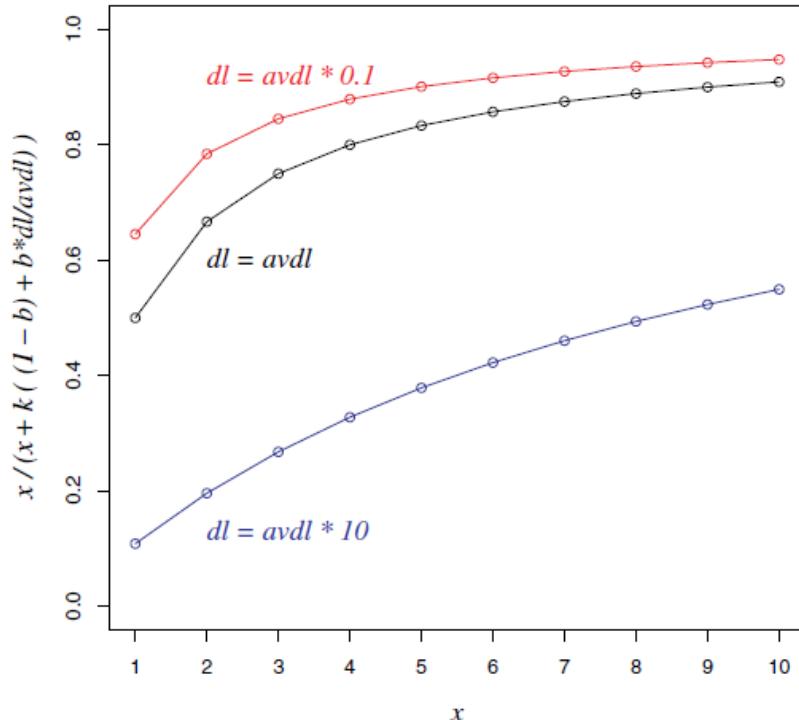
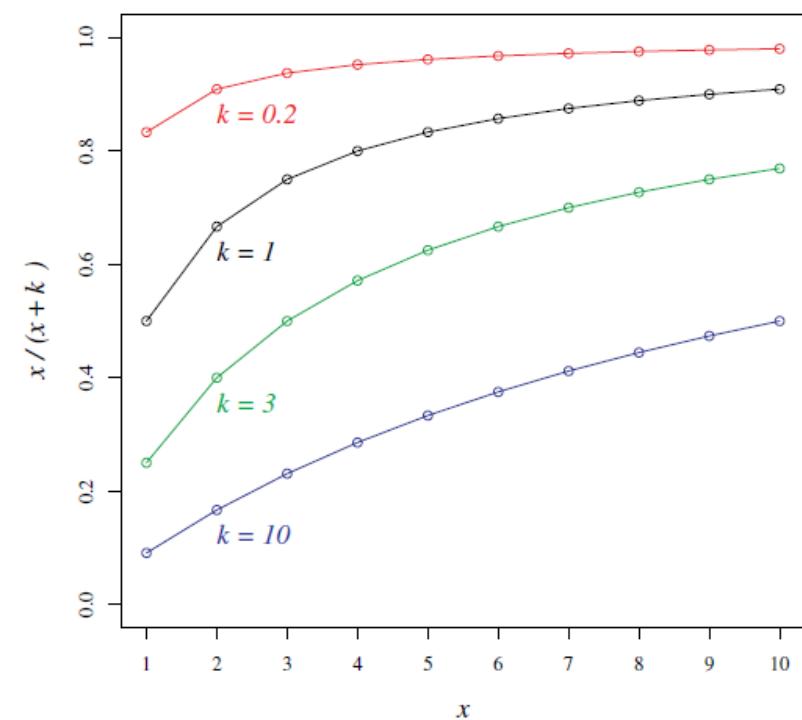


$k=1, b=0.5$

Document term frequency

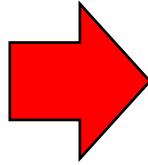
$$\frac{f_{ij}}{f_{ij} + k}$$

$$\frac{f_i}{k \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i}$$



Document term frequency

$$\frac{f_i}{k \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i}$$



$$\frac{(k + 1)f_i}{k \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i}$$

dl – จำนวนคำของเอกสาร j

$avdl$ – จำนวนคำเฉลี่ยของทุกเอกสาร

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

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

Query term frequency

$$\frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

ฝึกหัด 1 หน้า = 1 , ฝึกหัด 2 หน้า ↑
จุดนี้สำคัญมาก

k_2 - ค่าคงที่โดยปกติจะอยู่ในช่วง $0 - 1000$

qf_i - ความถี่ของ keyword i ใน query

- พิจารณาจากความถี่ของแต่ละ Keyword ใน Query
- ให้ความสำคัญน้อยหรือไม่ให้ความสำคัญเลย
- มีผลน้อยกว่าความถี่ของ Keyword ในเอกสาร

BM25

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

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 $0 - 1000$

ตัวอย่างการใช้ BM25

Example (retrieved docs)

ໂທນຳກິດຈົນ return

Query = Honda , Toyota , Isuzu

d1		My father is thinking about the car that she want to buy between Isuzu D-MAX X-series, Isuzu D-MAX V-Cross 4 door, Isuzu D-MAX V-Cross 2 door, Isuzu Mu-7, Isuzu Mu-X, Isuzu D-MAX Hi-Lander, Toyota Hilux vigo, Toyota Hilux revo or Toyota Innova.
d2		My uncle suggest My father that he should buy Toyota camry, Toyota vios, Toyota yaris or Toyota corolla altis.
d3	R	My mother will buy a car for me and my brother, we think we should Honda city, Honda brio, Honda CR-V, Honda BR-V, Honda civic, Honda accord, Toyota Yaris, Toyota vios.
d4		When i was young, My father driven Toyota Mighty-X but now He want to sell it and he will buy Toyota vigo, Isuzu D-MAX, Isuzu Mu-7, Isuzu Mu-X. But my mother do not want to sell it.
d5	R	A silver Honda Accord pulled up and the window rolled down after black Toyota yaris passed the Toyota hilux vigo in front of Toyota yaris.
d6	R	Isuzu D-MAX more popular than Honda and Toyota although Toyota hilux vigo are cheaper than Isuzu D-Max and Honda accord. So , Isuzu have much more profit than Toyota and Honda.
d7	R	Finally , I am decide to buy Isuzu Mu-7 because it can carry people than Toyota camry and Honda accord inspite of Honda accord has beautiful than Isuzu Mu-7 and Toyota camry ,but Isuzu mu-7 has the most power consumed
d8	R	A new generation of car are leading by Honda Toyota and Isuzu and Toyota have most car in production line ,although Honda have more scientist than Toyota but Toyata have car in production line more Honda.

1. ກຳ term frequency ໃນແຕ່ລະ document

f	Honda	Toyota	Isuzu
	n_1 / r_1	n_2 / r_2	n_3 / r_3
d1	0	3	6
d2	0	4	0
d3	6	2	0
d4	0	2	3
d5	1	3	0
d6	3	3	2
d7	2	2	3
d8	3	4	1

$$N = 8, R = 5$$

$$r_1 = 5$$

$$r_2 = 5$$

$$r_3 = 3$$

$$n_1 = 5$$

$$n_2 = 8$$

$$n_3 = 5$$

R

R

R

R

R

2. หา document length และ average length

Length (จำนวนคำในเอกสาร)

$$d1 = 42 \quad d5 = 25$$

$$d2 = 19 \quad d6 = 31$$

$$d3 = 31 \quad d7 = 39$$

$$d4 = 37 \quad d8 = 36$$

$$\text{AVR} = 32.5$$

3. ໜ້າ Inverse document frequency

$$R = 5$$

$$r_{Honda} = 5$$

$$r_{Toyota} = 5$$

$$r_{Isuzu} = 3$$

$$N = 8$$

$$n_{Honda} = 5$$

$$n_{Toyota} = 8$$

$$n_{Isuzu} = 5$$

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

$$idf_{honda} = \log \frac{(5 + 0.5)/(5 - 5 + 0.5)}{(5 - 5 + 0.5)/(8 - 5 - 5 + 5 + 0.5)} = 1.89$$

$$idf_{toyota} = \log \frac{(5 + 0.5)/(5 - 5 + 0.5)}{(8 - 5 + 0.5)/(8 - 8 - 5 + 5 + 0.5)} = 0.20$$

$$idf_{isuzu} = \log \frac{(3 + 0.5)/(5 - 3 + 0.5)}{(5 - 3 + 0.5)/(8 - 5 - 5 + 3 + 0.5)} = -0.08$$

4. หา sim ของ BM25 ของ document ที่ต้องการ

ต้องการหา sim ของ document d10,d20,d30 และ d40 มีผลลัพธ์ที่รับ回来แล้ว return
แต่ต้องคำนึงถึงความต้องการเดิม 9 ครั้งที่มี
idf

	ความถี่ Honda	ความถี่ Toyota	ความถี่ Isuzu	จำนวนคำในเอกสาร
d10	0	4	2	21
d20	9	15	2	55
d30	11	7	5	35
d40	6	6	6	25

4. หา sim ของ BM25 ของ document ที่ต้องการ

Query = Honda , Toyota , Isuzu

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

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 ($0.5 < b < 0.8$)

k₁ - ค่าคงที่โดยตาม TREC จะใช้ค่า 1.25 ($1.2 < k_1 < 2$)

k₂ - ค่าคงที่โดยปกติจะอยู่ในช่วง 0 - 1000

4. หา sim ของ BM25 ของ document ที่ต้องการ

query ที่ต้องการ keyword กี่ครั้ง กี่ครั้ง \rightarrow keyword frequency = 1

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}$$

$$\cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_{30}, q) = 1.89 \cdot \frac{(2.25)11}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{35}{32.5} \right) + 11}$$

$$+ 0.20 \cdot \frac{(2.25)7}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{35}{32.5} \right) + 7}$$

$$- 0.08 \cdot \frac{(2.25)5}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{35}{32.5} \right) + 5}$$

$$= 3.789 + 0.371 - 0.135$$

$$= 4.026$$

$= 1$ จำนวน keyword
 $idf_{honda} = 1.89$
 $idf_{toyota} = 0.20$
 $idf_{isuzu} = -0.08$

จำนวน keyword

5. จัดลำดับความตรงประเด็น และแสดงผลลัพธ์การ query

	ความถี่ Honda	ความถี่ Toyota	ความถี่ Isuzu	จำนวนคำใน เอกสาร	<i>sim</i>
d30	11	7	5	35	4.026
d40	6	6	6	25	3.852
d20	9	15	2	55	3.810
d10	0	4	2	21	0.242

រាជ្យប៊ូល នៃកសាង
ភ័ព្យាគ្រេងៗ
អនុញ្ញាតក្នុង

Example (all docs)

Query = Honda , Toyota , Isuzu

អនុញ្ញាតក្នុង

d1	My father is thinking about the car that she want to buy between Isuzu D-MAX X-series, Isuzu D-MAX V-Cross 4 door, Isuzu D-MAX V-Cross 2 door, Isuzu Mu-7, Isuzu Mu-X, Isuzu D-MAX Hi-Lander, Toyota Hilux vigo, Toyota Hilux revo or Toyota Innova.
d2	My uncle suggest My father that he should buy Toyota camry, Toyota vios, Toyota yaris or Toyota corolla altis.
d3	My mother will buy a car for me and my brother, we think we should Honda city, Honda brio, Honda CR-V, Honda BR-V, Honda civic, Honda accord, Toyota Yaris, Toyota vios.
d4	When i was young, My father driven Toyota Mighty-X but now He want to sell it and he will buy Toyota vigo, Isuzu D-MAX, Isuzu Mu-7, Isuzu Mu-X. But my mother do not want to sell it.
d5	A silver Honda Accord pulled up and the window rolled down after black Toyota yaris passed the Toyota hilux vigo in front of Toyota yaris.
d6	Isuzu D-MAX more popular than Honda and Toyota although Toyota hilux vigo are cheaper than Isuzu D-Max and Honda accord. So , Isuzu have much more profit than Toyota and Honda.
d7	Finally , I am decide to buy Isuzu Mu-7 because it can carry people than Toyota camry and Honda accord inspite of Honda accord has beautiful than Isuzu Mu-7 and Toyota camry ,but Isuzu mu-7 has the most power consumed
d8	A new generation of car are leading by Honda Toyota and Isuzu and Toyota have most car in production line ,althought Honda have more scientist than Toyota but Toyata have car in production line more Honda.

1. หา document length และ average length

Length (จำนวนคำในเอกสาร)

$$d_1 = 42 \quad d_5 = 25$$

$$d_2 = 19 \quad d_6 = 31$$

$$d_3 = 31 \quad d_7 = 39$$

$$d_4 = 37 \quad d_8 = 36$$

$$\textbf{AVR} = 32.5 \quad \text{ร้อยละคลิ้บ}$$

2. ឧប Inverse document frequency

$$R = 0$$

$$r_{Honda} = 0$$

$$r_{Toyota} = 0$$

$$r_{Isuzu} = 0$$

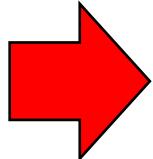
$$N = 8$$

$$n_{Honda} = 5$$

$$n_{Toyota} = 8$$

$$n_{Isuzu} = 5$$

វិមាន Doc រាយការណ៍ នឹងត្រួតពិនិត្យ

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

$$idf_{honda} = \log \frac{(0 + 0.5)/(0 - 0 + 0.5)}{(5 - 0 + 0.5)/(8 - 5 - 0 + 0 + 0.5)} = -0.20$$

$$idf_{toyota} = \log \frac{(0 + 0.5)/(0 - 0 + 0.5)}{(8 - 0 + 0.5)/(8 - 8 - 0 + 0 + 0.5)} = -1.23$$

$$idf_{isuzu} = \log \frac{(0 + 0.5)/(0 - 0 + 0.5)}{(5 - 0 + 0.5)/(8 - 5 - 0 + 0 + 0.5)} = -0.20$$

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

3. หา sim ของ BM25 ของ document ที่ต้องการ

ต้องการหา sim ของ document d4 และ d8

	ความถี่ Honda	ความถี่ Toyota	ความถี่ Isuzu	จำนวนคำในเอกสาร
d4	0	2	3	37
d8	3	4	1	36

4. หา sim ของ BM25 ของ document ที่ต้องการ

Query = Honda , Toyota , Isuzu

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

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 ($0.5 < b < 0.8$)

k₁ - ค่าคงที่โดยตาม TREC จะใช้ค่า 1.25 ($1.2 < k_1 < 2$)

k₂ - ค่าคงที่โดยปกติจะอยู่ในช่วง 0 - 1000

5. หา sim ของ BM25 ของ document ที่ต้องการ

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_4, q) = -0.20 \cdot \frac{(2.25)0}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{37}{32.5} \right) + 0}$$

$$-1.23 \cdot \frac{(2.25)2}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{37}{32.5} \right) + 2}$$

$$-0.20 \cdot \frac{(2.25)3}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{37}{32.5} \right) + 3}$$

$$= 0.000 - 1.638 - 0.308$$

$$= -1.941$$

$idf_{honda} = -0.20$
 $idf_{toyota} = -1.23$
 $idf_{isuzu} = -0.20$

5. หา sim ของ BM25 ของ document ที่ต้องการ

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_8, q) = -0.20 \cdot \frac{(2.25)3}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{36}{32.5} \right) + 3}$$

$$-1.23 \cdot \frac{(2.25)4}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{36}{32.5} \right) + 4}$$

$$-0.20 \cdot \frac{(2.25)1}{1.25 \left((1 - 0.75) + 0.75 \cdot \frac{36}{32.5} \right) + 1}$$

$$= -0.310 - 2.069 - 0.191$$

$$= -2.562$$

$idf_{honda} = -0.20$
 $idf_{toyota} = -1.23$
 $idf_{isuzu} = -0.20$

5. จัดลำดับความตรงประเด็น และแสดงผลลัพธ์การ query

	ความถี่ Honda	ความถี่ Toyota	ความถี่ Isuzu	จำนวนคำในเอกสาร	<i>sim</i>
d4	0	2	3	37	-1.941
d8	3	4	1	36	-2.562

ถ้า keyword มาก = ความปิงเพิงมาก

↑
น้อย มากเท่าไร

$$\log\left(\frac{N}{n}\right)$$

↑ มากเท่าไร = +
 ↓ น้อยกว่า N เท่ากัน = -

Example 3

ตามที่ตั้ง keyword 9 int query = 1

- Query Q = “omega mike golf” (qf = 1)
- มีเอกสารทั้งหมด 6,200,000 ฉบับ
- คำว่า “omega” ปรากฏในเอกสารทั้งหมด 500,000 เอกสาร ($n_1 = 500,000$)
- คำว่า “mike” ปรากฏในเอกสารทั้งหมด 314 เอกสาร ($n_2 = 314$)
- คำว่า “golf” ปรากฏในเอกสารทั้งหมด 80,000 เอกสาร ($n_3 = 80,000$)
- คำว่า “omega” ปรากฏ 21 ครั้ง ในเอกสารที่สนใจ ($f_1 = 21$)
- คำว่า “mike” ปรากฏ 14 ครั้ง ในเอกสารที่สนใจ ($f_2 = 14$)
- คำว่า “golf” ปรากฏ 90 ครั้ง ในเอกสารที่สนใจ ($f_3 = 90$)
- ขนาดของเอกสารที่สนใจต่อขนาดเฉลี่ยของเอกสารทั้งหมดเท่ากับ 0.4 ($\frac{dl}{avdl}$)
- กำหนดให้ $k_1 = 1.25$, $b = 0.75$, $k_2 = 200$

$$\therefore K = k_1((1-b) + b \cdot \frac{dl}{avdl})$$

$$\therefore K = 1.25((1 - 0.75) + 0.75 \cdot 0.4)$$

$$\therefore K = 0.688$$

ร้อยละ 68.8%

Example 3

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{N - n_i + 0.5}{(n_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{K + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

} ក្នុងពាក្យល់នៃ Doc តើអាមេរិក

$$\begin{aligned} sim_{bm25}(d_{\text{Doc}}, q) &= \log \frac{(6,200,000 - 500,000 + 0.5)}{(500,000 + 0.5)} \times \frac{(1.25 + 1)21}{0.688 + 21} \times \frac{(200 + 1)1}{200 + 1} \\ &+ \log \frac{(6,200,000 - 314 + 0.5)}{(314 + 0.5)} \times \frac{(1.25 + 1)14}{0.688 + 14} \times \frac{(200 + 1)1}{200 + 1} \\ &+ \log \frac{(6,200,000 - 80,000 + 0.5)}{(80,000 + 0.5)} \times \frac{(1.25 + 1)90}{0.688 + 90} \times \frac{(200 + 1)1}{200 + 1} \end{aligned}$$

$$sim_{bm25}(d_{\text{Doc}}, q) = 2.303 + 9.211 + 4.206$$

$$sim_{bm25}(d_{\text{Doc}}, q) = 15.720$$

= 1

$K = 0.688$

$k_1 = 1.25$

$k_2 = 200$

$b = 0.75$

$N = 6,200,000$

$n_1 = 500,000$

$n_2 = 314$

$n_3 = 80,000$

$f_1 = 21$

$f_2 = 14$

$f_3 = 90$

Example 4

- Query $Q = "lincoln lincoln"$ ($qf = 2$)
- มีเอกสารทั้งหมด 200,000 ฉบับ
- คำว่า “*lincoln*” ปรากฏในเอกสารทั้งหมด 80,000 เอกสาร ($n_1 = 80,000$)
- คำว่า “*lincoln*” ปรากฏ 90 ครั้งใน เอกสารที่สนใจ ($f_1 = 90$)
- ขนาดของเอกสารที่สนใจต่อขนาดเฉลี่ยของเอกสารทั้งหมดเท่ากับ 0.5 ($\frac{dl}{avdl}$)
- กำหนดให้ $k_1 = 1.25$, $b = 0.75$, $k_2 = 200$

$$\therefore K = k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right)$$

$$\therefore K = 1.25((1 - 0.75) + 0.75 \cdot 0.5)$$

$$\therefore K = 0.781$$

Example 4

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_j, q) = \sum_{i \in q} \log \frac{N - n_i + 0.5}{(n_i + 0.5)} \cdot \frac{(k_1 + 1)f_i}{K + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$$sim_{bm25}(d_{100}, q) = \log \frac{(200,000 - 80,000 + 0.5)}{(80,000 + 0.5)} \times \frac{(1.25 + 1)90}{0.781 + 90} \times \frac{(200 + 1)2}{200 + 2}$$

$$sim_{bm25}(d_{100}, q) = 0.176 \times 2.231 \times 1.990$$

$K = 0.781$

$k_1 = 1.25$

$k_2 = 200$

$b = 0.75$

$N = 200,000$

$n_1 = 80,000$

$f_1 = 90$

$$sim_{bm25}(d_{100}, q) = 0.782$$

BM25

ข้อดี

- จัดลำดับละเอียดกว่า BIR (ความถี่ของ Keyword ในเอกสาร, Query)
- ใช้กับเอกสารทั้งหมดหรือเฉพาะเอกสารที่ได้รับจากการเรียกคืน (all docs, retrieved docs)
Doc ที่ถูก return / Doc ที่อยู่ในระบบ (ไม่ถูก return)

- ข้อเสีย cat tiger หนาๆ, อย่างเดียว คำว่า; นาร์, น้ำเต้าหู้ อันดับ
- รองรับ Query อย่างง่ายเท่านั้น
 - การ Ranking เป็นไปตาม Document ในระบบ
ตาม $\text{tf}(t, d) \cdot \text{df}(t)$ ตาม mark R
(การเพิ่มลดเอกสาร, การเพิ่มลด R)
 - ไม่สนใจ Relationship ของ Keyword
ตามส่วนตัว

BM25

ເອກສາຣອ້າງອີງ

- <http://www.cs.cornell.edu/courses/cs4300/2013fa/lectures/retrieval-models-2-4pp.pdf>
- https://en.wikipedia.org/wiki/Okapi_BM25
- <http://xapian.org/docs/bm25.html>
- <https://dato.com/learn/userguide/feature-engineering/bm25.html>
- http://www.staff.city.ac.uk/~sb317/papers/foundations_bm25_review.pdf
- <http://homepages.inf.ed.ac.uk/vlavrenk/doc/pmir-1x2.pdf>
- <https://pdfs.semanticscholar.org/524b/35f49e854f0cec5b829ee6cea143e9f27a47.pdf>
- [http://berlin.csie.ntnu.edu.tw/Courses/Information%20Retrieval%20and%20Extraction/2015S_Lectures/IR2015S-Lecture05-Modeling-II\(Set,%20Algebra%20&%20Probabilistic\).pdf](http://berlin.csie.ntnu.edu.tw/Courses/Information%20Retrieval%20and%20Extraction/2015S_Lectures/IR2015S-Lecture05-Modeling-II(Set,%20Algebra%20&%20Probabilistic).pdf)