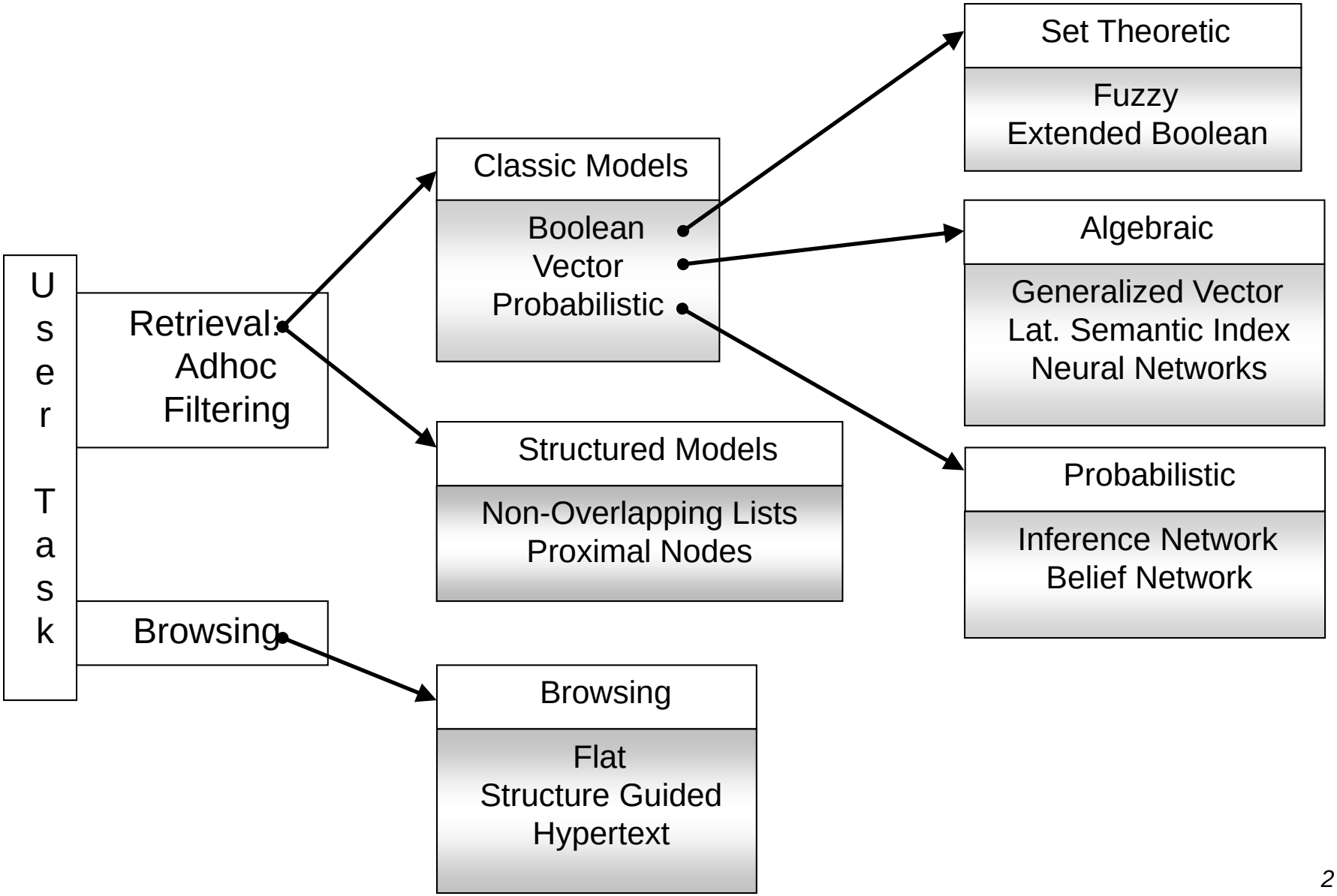
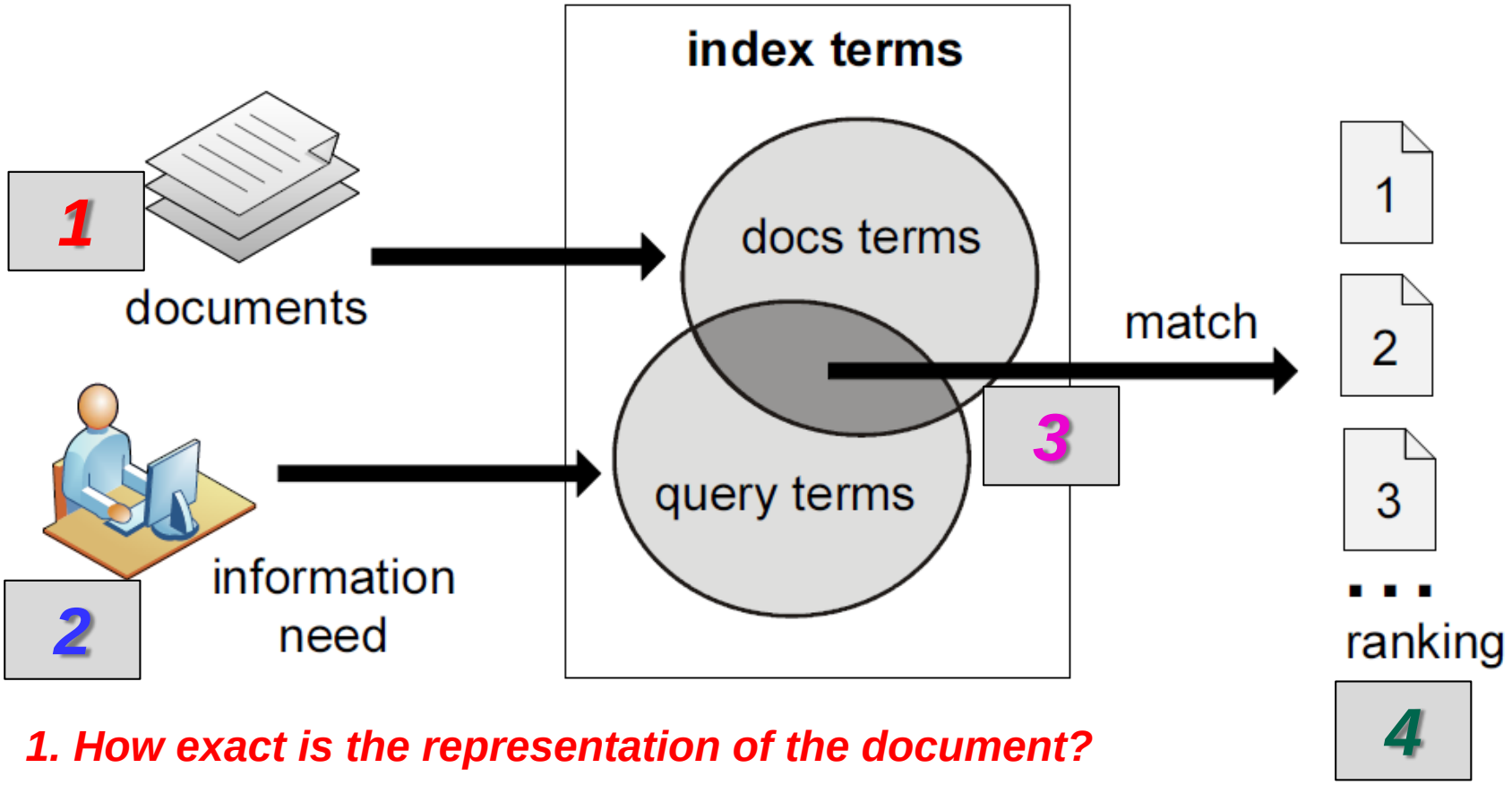
Chapter 02
Modeling

IR Models


User Task
Retrieval: Adhoc Filtering
Browsing
Classic Models
Boolean
Vector
Probabilistic
Structured Models
Non-Overlapping Lists
Proximal Nodes
Browsing
Flat
Structure Guided
Hypertext
Set Theoretic
Fuzzy
Extended Boolean
Algebraic
Generalized Vector
Lat. Semantic Index
Neural Networks
Probabilistic
Inference Network
Belief Network

IR Problem



1. How exact is the representation of the document?
2. How exact is the representation of the query?
3. How well is query matched to data?
4. How relevant is the result to the query?

Probabilistic Model

- 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**

K_1 = Cat
K_2 = Dog

N = 20
R = 12
not R = 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***
$d19 = \{0,0\}$
$d20 = \{0,0\}$

K_1 = Cat
K_2 = Dog

N = 20
R = 12
not R = 8

K_1,N = 11
K_1,R = 8
K_1,not R = 3
not K_1,R = 4
not K_1,not R = 5

K_2,N = 11
K_2,R = 7
K_2,not R = 4
not K_2,R = 5
not K_2,not R = 4

Relevance Docs

d1 = {1,1}
d2 = {1,1}
d3 = {1,1}
d4 = {1,1}
d6 = {1,0}
d7 = {1,0}
d8 = {1,0}
d9 = {1,0}
d12={0,1}
d13={0,1}
d14={0,1}
d18={0,0}

K_1 = 8
K_2 = 7

Non Relevance Docs

d5 = {1,1}
d10={1,0}
d11={1,0}
d15={0,1}
d16={0,1}
d17={0,1}
d19={0,0}
d20={0,0}

K_1 = 3
K_2 = 4

Retrieved Document

K_1 = Cat
K_2 = Dog

N = 20
R = 12
not R = 8

K_1,N = 11
K_1,R = 8
K_1,not R = 3
not K_1,R = 4
not K_1,not R = 5

K_2,N = 11
K_2,R = 7
K_2,not R = 4
not K_2,R = 5
not K_2,not R = 4

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}

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

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

$$P(\overline{K_1}|R) = \frac{4}{12} \Rightarrow \frac{R - r_1}{R}$$

$$P(K_1|\bar{R}) = \frac{3}{8} \Rightarrow \frac{n_1 - r_1}{N - R} = \frac{11 - 8}{20 - 12}$$

$$P(\overline{K_1}|\bar{R}) = \frac{5}{8} \Rightarrow \frac{(N - R) - (n_1 - r_1)}{N - R}$$

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

$$P(\overline{K_2}|R) = \frac{5}{12}$$

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

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

$$P(K_i|R) + P(\overline{K_i}|R) = 1$$

$$P(K_i|\bar{R}) + P(\overline{K_i}|\bar{R}) = 1$$

What is $P(\overline{K_1}|R), P(\overline{K_2}|R)$???

Retrieved Document

K_1 = Cat
K_2 = Dog

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

$$P(\overline{K_1}|R) = \frac{4}{12}$$

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

$$P(\overline{K_1}|\bar{R}) = \frac{5}{8}$$

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

$$P(\overline{K_2}|R) = \frac{5}{12}$$

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

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

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}

We need ??? ⇨ $sim(d_j, q) = ???$

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

Bayes' rule

$$\frac{P(R|\vec{d_j})}{P(\bar{R}|\vec{d_j})} = \frac{P(\vec{d_j}|R).P(R)}{P(\vec{d_j}|\bar{R}).P(\bar{R})}$$

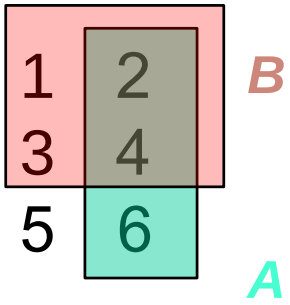
Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5



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

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

$$P(A|B) = \frac{2}{4}$$ ⇒ ความน่าจะเป็นที่ลูกเต๋าจะออกเลขคู่จากเหตุการณ์ที่ลูกเต๋ามีแต้มน้อยกว่า 5

$$P(B|A) = \frac{2}{3}$$ ⇒ ความน่าจะเป็นที่ลูกเต๋าจะมีค่าน้อยกว่า 5 จากเหตุการณ์ที่ลูกเต๋ามีแต้มเป็นเลขคู่

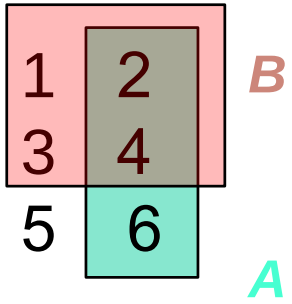
Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5



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

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

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

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

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

$$= \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{2}{6} * \frac{6}{4} = \frac{1}{2}$$

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

$$= \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{6} * \frac{6}{3} = \frac{2}{3}$$

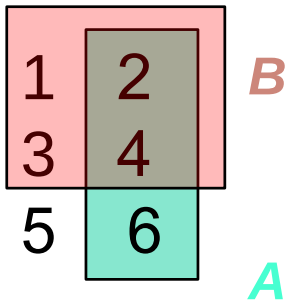
Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5



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

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

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

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

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

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

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

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

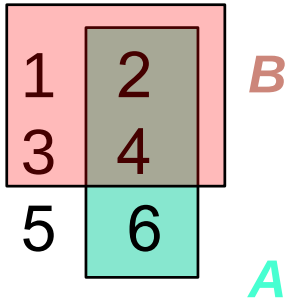
Bayes' Rule

เหตุการณ์โยนลูกเต๋า 1 ลูก

N คือเหตุการณ์ทั้งหมด (6 หมายเลข)

A คือเหตุการณ์ที่ได้เลขคู่

B คือเหตุการณ์ที่ได้เลขน้อยกว่า 5



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

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

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

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

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

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

Probabilistic Model

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

$$= \frac{\frac{P(\vec{d_j}|R) * P(R)}{P(\vec{d_j})}}{\frac{P(\vec{d_j}|\bar{R}).P(\bar{R})}{P(\vec{d_j})}}$$

$$= \frac{P(\vec{d_j}|R) * P(R)}{P(\vec{d_j})} * \frac{P(\vec{d_j})}{P(\vec{d_j}|\bar{R}).P(\bar{R})}$$

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

Bayes' Rule

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

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

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

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

$$P(\overline{K_1}|R) = \frac{4}{12}$$

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

$$P(\overline{K_1}|\bar{R}) = \frac{5}{8}$$

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

$$P(\overline{K_2}|R) = \frac{5}{12}$$

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

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

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}

Simple Probabilistic Method

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

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

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

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

d21={1,0} ➡ **sim(d21,q) = ???**

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

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

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

$$P(\overline{K_1}|R) = \frac{4}{12}$$

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

$$P(\overline{K_1}|\bar{R}) = \frac{5}{8}$$

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

$$P(\overline{K_2}|R) = \frac{5}{12}$$

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

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

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}

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

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

d1 = {1,1}

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

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

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

d6 = {1,0}

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

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

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

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

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

$P(\overline{K_1}|R) = \frac{4}{12}$

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

$P(\overline{K_1}|\bar{R}) = \frac{5}{8}$

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

$P(\overline{K_2}|R) = \frac{5}{12}$

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

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

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}

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

d12 = {0,1}

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

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

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

d18 = {0,0}

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

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

Probabilistic Model

$d_j \rightarrow \{1,1\}$ $\quad sim(d_j, q) = \frac{28}{9} \Rightarrow sim(d_j, q) = \frac{28}{37} = 0.757$

$d_j \rightarrow \{1,0\}$ $\quad sim(d_j, q) = \frac{20}{9} \Rightarrow sim(d_j, q) = \frac{20}{29} = 0.690$

$d_j \rightarrow \{0,1\}$ $\quad sim(d_j, q) = \frac{28}{30} \Rightarrow sim(d_j, q) = \frac{28}{58} = 0.483$

$d_j \rightarrow \{0,0\}$ $\quad sim(d_j, q) = \frac{10}{15} \Rightarrow sim(d_j, q) = \frac{10}{25} = 0.400$

Probabilistic value $\in [0,1]$

Then

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

Binary Independence Retrieval Model (BIR)

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

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

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

$$P(\overline{K_1}|R) = \frac{4}{12}$$

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

$$P(\overline{K_1}|\bar{R}) = \frac{5}{8}$$

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

$$P(\overline{K_2}|R) = \frac{5}{12}$$

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

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

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

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

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

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

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

Probabilistic Ranking Principle

- 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?

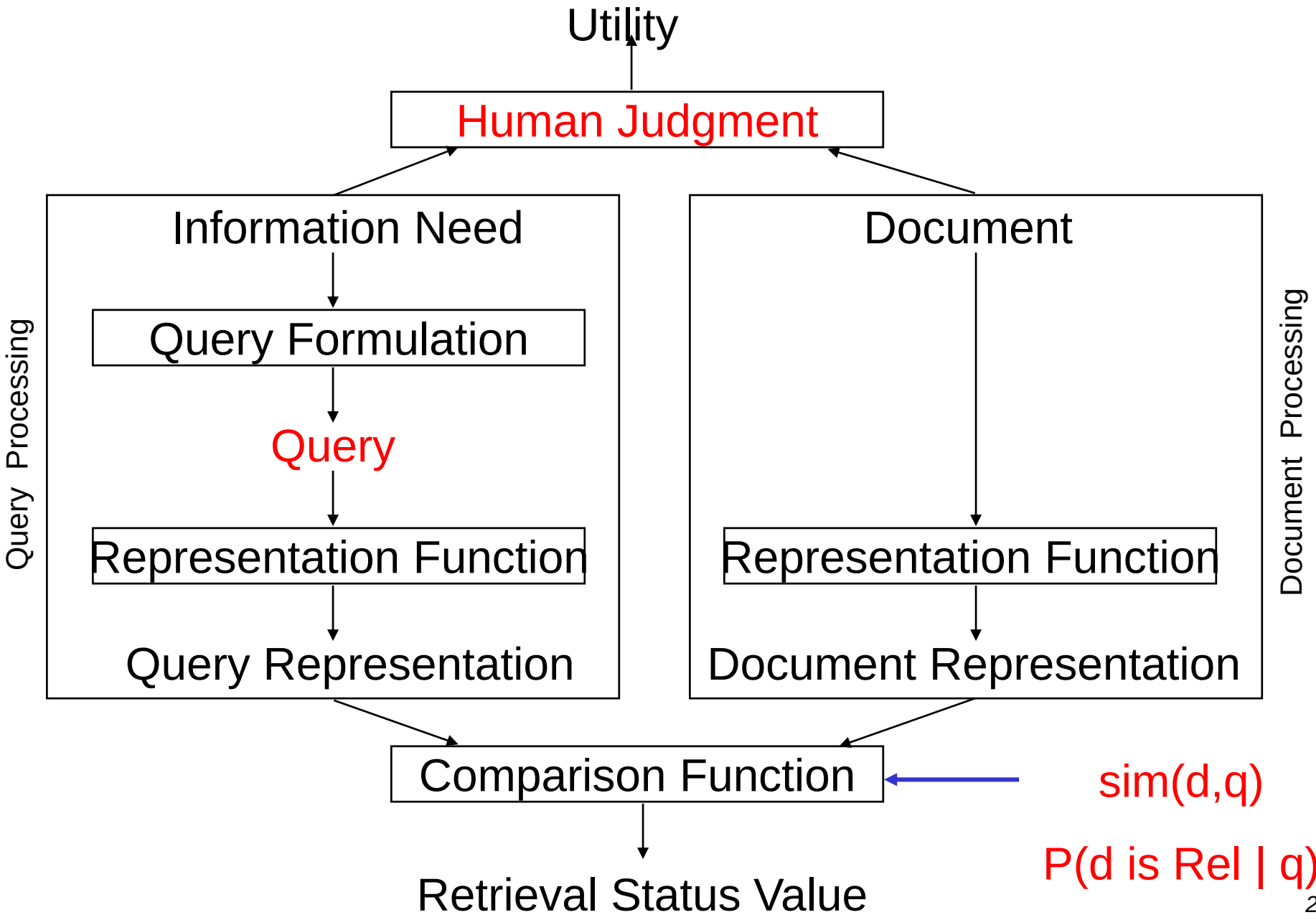
The Ranking

- Probabilistic ranking computed as:
  - *sim(d_j,q) = P(d_j relevant-to q) / P(d_j non-relevant-to q)*

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

  - This is the **odds** of the document d_j being relevant
  - Taking the **odds** minimize the probability of an erroneous judgement
- Definition:
  - *wij* $\in$ *{0,1}*
  - $P(R|\vec{d_j})$: probability that given doc is relevant
  - $P(\bar{R}|\vec{d_j})$: probability doc is not relevant

Where do the probabilities fit?



Pluses and Minuses

- Advantages:
  - Docs ranked in decreasing order of probability of relevance
- Disadvantages:
  - need to guess initial estimates for $P(k_i \mid R)$
  - method does not take into account tf and idf factors

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

$D_3,D_{10},D_2,D_5,D_9,D_6,D_1$

เมื่อนำผลลัพธ์ที่ได้มาวิเคราะห์ และนำไปจัดลำดับความตรงประเด็นของเอกสารทั้งหมดอีกครั้ง ลำดับของความตรงประเด็นใหม่เป็นเท่าใด (จงแสดงวิธีคำนวณ)

BM25 (Best Matching 25) Extended Probabilistic Model

BM25

Goals

- *All Documents (not only retrieved documents)*
- *Term frequency in each document*
- *Term frequency in query*

BM25

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

Inverse document frequency

Document term frequency

Query term frequency

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

Variables

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

Inverse document frequency

$$\frac{\text{ความน่าจะเป็นที่เอกสารจะตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสารจะไม่ตรงประเด็น}}$$

$$\text{ความน่าจะเป็นที่เอกสารจะตรงประเด็น} = \frac{\text{ความน่าจะเป็นที่เอกสารมี } Keyword \text{ แล้วตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสารไม่มี } Keyword \text{ แล้วตรงประเด็น}}$$

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

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

$$\frac{\text{จำนวนเอกสารที่มี } Keyword \text{ ในเอกสารที่กำหนดว่าตรงประเด็น}}{\text{จำนวนเอกสารที่ตรงประเด็นทั้งหมด}}$$

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

$$\frac{\text{จำนวนเอกสารที่ไม่มี } Keyword \text{ ในเอกสารที่กำหนดว่าตรงประเด็น}}{\text{จำนวนเอกสารที่ตรงประเด็นทั้งหมด}}$$

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

Inverse document frequency

$$\frac{\text{ความน่าจะเป็นที่เอกสารจะตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสารจะไม่ตรงประเด็น}}$$

$$\text{ความน่าจะเป็นที่เอกสารจะไม่ตรงประเด็น} = \frac{\text{ความน่าจะเป็นที่เอกสารมี } Keyword \text{ แล้วไม่ตรงประเด็น}}{\text{ความน่าจะเป็นที่เอกสารไม่มี } Keyword \text{ แล้วไม่ตรงประเด็น}}$$

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

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

$$\frac{\text{จำนวนเอกสารที่มี } Keyword \text{ ในเอกสารที่กำหนดว่าไม่ตรงประเด็น}}{\text{จำนวนเอกสารที่ไม่ตรงประเด็นทั้งหมด}}$$

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

$$\frac{\text{จำนวนเอกสารที่ไม่มี } Keyword \text{ ในเอกสารที่กำหนดว่าไม่ตรงประเด็น}}{\text{จำนวนเอกสารที่ไม่ตรงประเด็นทั้งหมด}}$$

$$\frac{\text{จำนวนเอกสารที่ไม่มี } Keyword \text{นั้นทั้งหมด} - \text{จำนวนเอกสารที่ไม่มี } Keyword \text{นั้นแล้วตรงประเด็น}}{\text{จำนวนเอกสารที่ไม่ตรงประเด็นทั้งหมด}}$$

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

Inverse document frequency

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

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

R – *จำนวนเอกสารที่ตรงประเด็น*

N – *จำนวนเอกสารทั้งหมด*

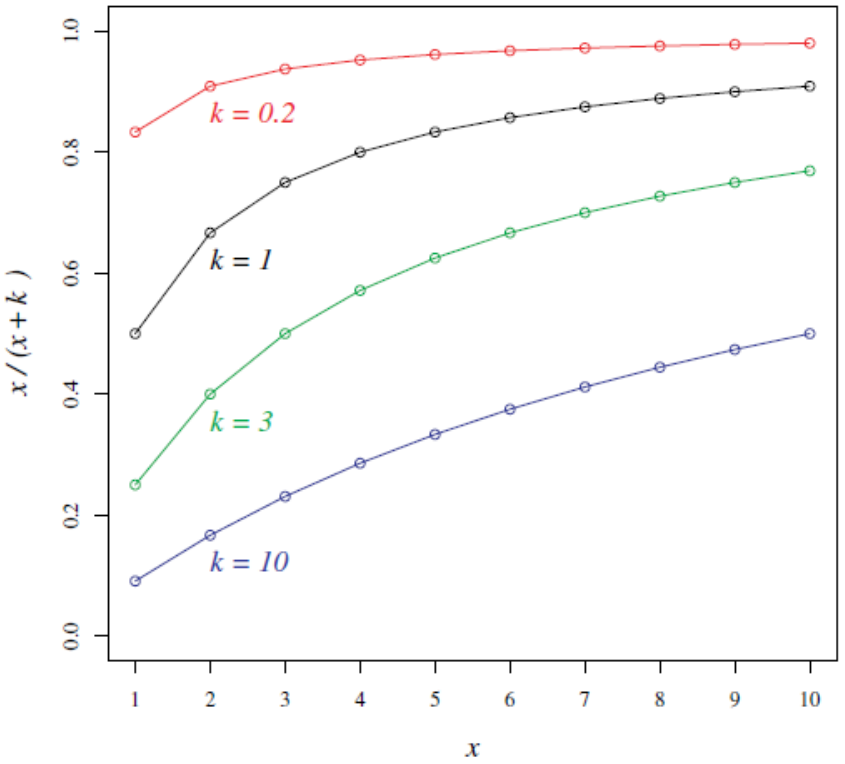
r_i – *จำนวนเอกสารที่ตรงประเด็นที่มี keyword i*

n_i – *จำนวนเอกสารทั้งหมดที่มี keyword i*

Document term frequency

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

ใช้ $f_{i,j}$ มีปัญหา → $\frac{f_{i,j}}{f_{i,j}+1}$ → $\frac{f_{i,j}}{f_{i,j}+k}$



Document term frequency

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

dl – จำนวนคำของเอกสาร j

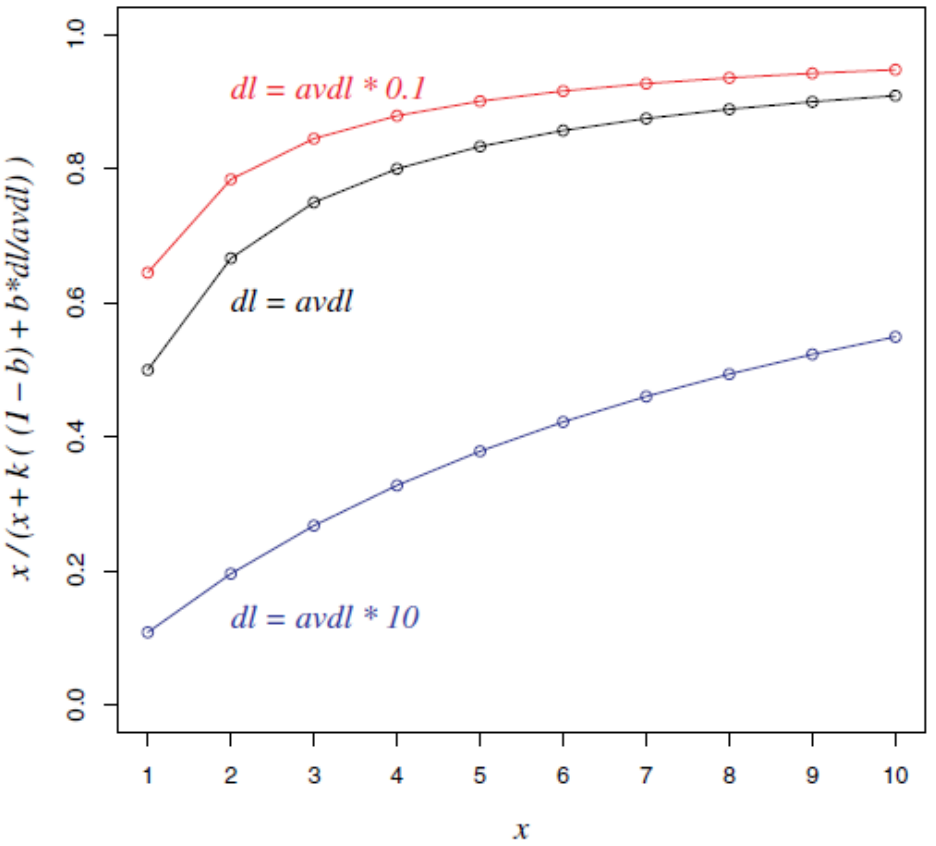
$avdl$ – จำนวนคำเฉลี่ยของทุกเอกสาร

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

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

Document term frequency

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

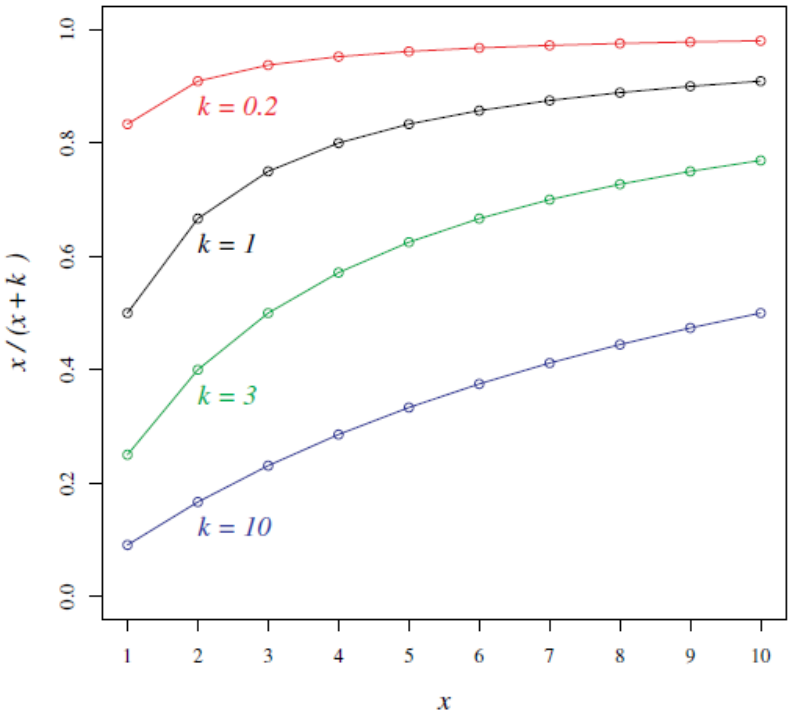


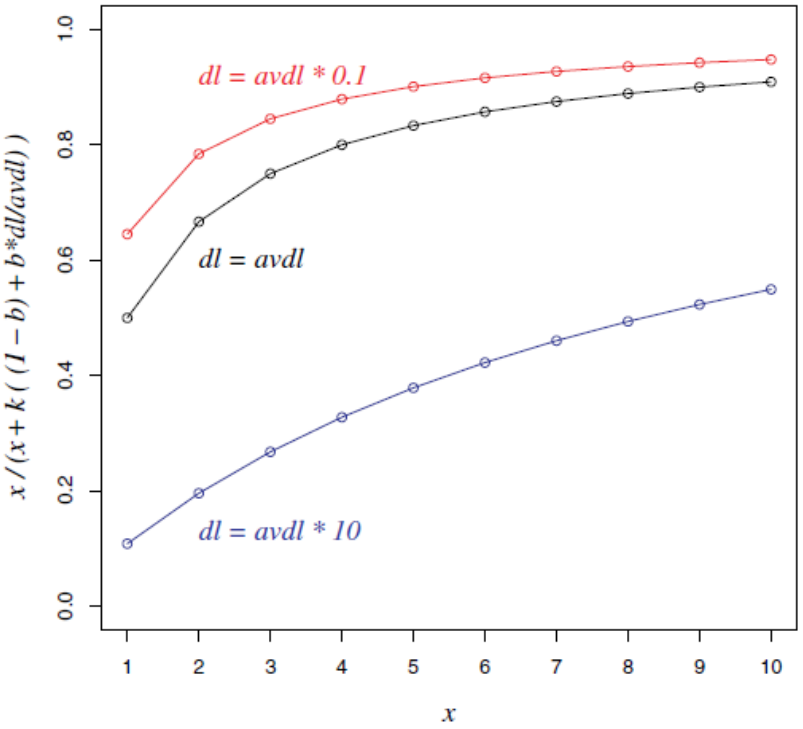
k=1, b=0.5

Document term frequency

$$\frac{f_{i,j}}{f_{i,j} + k}$$

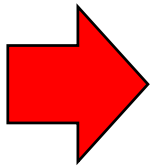
$$\frac{f_i}{k\left((1-b) + b \cdot \frac{dl}{avdl}\right) + f_i}$$





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

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

k_2 – ค่าคงที่โดยปกติจะอยู่ในช่วง 0 - 1000
qf_i – ความถี่ของ $keyword$ i ใน $query$

- พิจารณาจากความถี่ของแต่ละ Keyword ใน Query
- ให้ความสำคัญน้อยหรือไม่ให้ความสำคัญเลย
- มีผลน้อยกว่าความถี่ของ Keyword ในเอกสาร

BM25

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

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*

ตัวอย่างการใช้ BM25

Example (retrieved docs)

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** althought **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 ,althought Honda have more scientist than **Toyota** but **Toyata** have car in production line more Honda.

1. หา term frequency ในแต่ละ document

f	Honda	Toyota	Isuzu	
d1	0	3	6	
d2	0	4	0	
d3	6	2	0	R
d4	0	2	3	
d5	1	3	0	R
d6	3	3	2	R
d7	2	2	3	R
d8	3	4	1	R

2. หา document length และ average length

Length (จำนวนคำในเอกสาร)

d1 = 42	*d5 = 25*
d2 = 19	*d6 = 31*
d3 = 31	*d7 = 39*
d4 = 37	*d8 = 36*

AVR = 32.5

3. หา Inverse document frequency

$$R = 5 \qquad N = 8$$
$$r_{Honda} = 5 \qquad n_{Honda} = 5$$
$$r_{Toyota} = 5 \qquad n_{Toyota} = 8$$
$$r_{Isuzu} = 3 \qquad n_{Isuzu} = 5$$

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

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

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

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

4. หา sim ของ BM25 ของ document ที่ต้องการ

ต้องการหา sim ของ document d10,d20,d30 และ d40

	ความถี่ **Honda**	ความถี่ **Toyota**	ความถี่ **Isuzu**	จำนวนคำใน เอกสาร
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

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

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*

4. หา sim ของ BM25 ของ document ที่ต้องการ

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

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

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

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

$$= 3.789 + 0.371 - 0.135$$

$$= 4.026$$

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

5. จัดลำดับความตรงประเด็น แล้วแสดงผลลัพธ์การ query

	ความถี่ Honda	ความถี่ Toyota	ความถี่ Isuzu	จำนวนคำในเอกสาร	sim
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

Example (all docs)

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	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** althought **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)

d1 = 42	*d5 = 25*
d2 = 19	*d6 = 31*
d3 = 31	*d7 = 39*
d4 = 37	*d8 = 36*

AVR = 32.5

2. หา Inverse document frequency

$R = 0$ $\quad N = 8$

$r_{Honda} = 0 \quad n_{Honda} = 5$

$r_{Toyota} = 0 \quad n_{Toyota} = 8$

$r_{Isuzu} = 0 \quad n_{Isuzu} = 5$

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

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

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

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

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

3. หา sim ของ BM25 ของ document ที่ต้องการ

ต้องการหา sim ของ document d4 และ d8

	ความถี่ **Honda**	ความถี่ **Toyota**	ความถี่ **Isuzu**	จำนวนคำใน เอกสาร
d4	0	2	3	37
d8	3	4	1	36

4. หา sim ของ BM25 ของ document ที่ต้องการ

Query = Honda , Toyota , Isuzu

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

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*

5. หา sim ของ BM25 ของ document ที่ต้องการ

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

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

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

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

$$= 0.000 - 1.638 - 0.308$$

$$= -1.941$$

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

5. หา sim ของ BM25 ของ document ที่ต้องการ

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

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

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

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

$$= -0.310 - 2.069 - 0.191$$

$$= -2.562$$

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

5. จัดลำดับความตรงประเด็น แล้วแสดงผลลัพธ์การ query

	ความถี่ **Honda**	ความถี่ **Toyota**	ความถี่ **Isuzu**	จำนวนคำในเอกสาร	sim
d4	0	2	3	37	-1.941
d8	3	4	1	36	-2.562

Example 3

- 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 ครั้งในเอกสารที่สนใจ (f_1 = 21)
- คำว่า “mike” ปรากฏ 14 ครั้งในเอกสารที่สนใจ (f_2 = 14)
- คำว่า “golf” ปรากฏ 90 ครั้งในเอกสารที่สนใจ (f_3 = 90)
- ขนาดของเอกสารที่สนใจต่อขนาดเฉลี่ยของเอกสารทั้งหมดเท่ากับ 0.4 $(\frac{dl}{avdl})$
- *กำหนดให้ k_1 = 1.25, b = 0.75, k_2 = 200*

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

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

$$\because K = 0.688$$

Example 3

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

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

$$\begin{aligned} sim_{bm25}(d_1, q) = & \log \frac{(6{,}200{,}000 - 500{,}000 + 0.5)}{(500{,}000 + 0.5)} \times \frac{(1.25 + 1)21}{0.688 + 21} \times \frac{(200 + 1)1}{200 + 1} \\ & + \log \frac{(6{,}200{,}000 - 314 + 0.5)}{(314 + 0.5)} \times \frac{(1.25 + 1)14}{0.688 + 14} \times \frac{(200 + 1)1}{200 + 1} \\ & + \log \frac{(6{,}200{,}000 - 80{,}000 + 0.5)}{(80{,}000 + 0.5)} \times \frac{(1.25 + 1)90}{0.688 + 90} \times \frac{(200 + 1)1}{200 + 1} \end{aligned}$$

$$sim_{bm25}(d_1, q) = 2.303 + 9.211 + 4.206$$

$$sim_{bm25}(d_1, q) = 15.720$$

$K = 0.688$
$k_1 = 1.25$
$k_2 = 200$
$b = 0.75$
$N = 6{,}200{,}000$
$n_1 = 500{,}000$
$n_2 = 314$
$n_3 = 80{,}000$
$f_1 = 21$
$f_2 = 14$
$f_3 = 90$

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 ครั้งในในเอกสารที่สนใจ (f_1 = 90)
- ขนาดของเอกสารที่สนใจต่อขนาดเฉลี่ยของเอกสารทั้งหมดเท่ากับ 0.5 $(\frac{dl}{avdl})$
- กำหนดให้ k_1 = 1.25, b = 0.75, k_2 = 200

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

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

$$\because K = 0.781$$

Example 4

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

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

$$sim_{bm25}(d_1, q) = \log \frac{(200{,}000 - 80{,}000 + 0.5)}{(80{,}000 + 0.5)} \times \frac{(1.25 + 1)90}{0.781 + 90} \times \frac{(200 + 1)2}{200 + 2}$$

$$sim_{bm25}(d_1, q) = 0.176 \text{ x } 2.231 \text{ x } 1.990$$

$$sim_{bm25}(d_1, q) = 0.782$$

$K = 0.781$
$k_1 = 1.25$
$k_2 = 200$
$b = 0.75$
$N = 200{,}000$
$n_1 = 80{,}000$
$f_1 = 90$

BM25

ข้อดี

- *จัดลำดับละเอียดกว่า BIR (ความถี่ของ Keyword ในเอกสาร,Query)*
- *ใช้กับเอกสารทั้งหมดหรือเฉพาะเอกสารที่ได้รับจากการเรียกค้น(all docs,retrieved docs)*

ข้อเสีย

- *รองรับ Query อย่างง่ายเท่านั้น*
- *การ Ranking เปลี่ยนตาม Document ในระบบ*
- *ไม่สนใจ Relationship ของ Keyword*

BM25

เอกสารอ้างอิง

- *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*