Ex. ไม่มี Doc ที่ตรงประเด็น

Example 3

ความถี่แต่ละ keyword ใน query = 1

- Query Q = “omega mike golf” (qf = 1)
- มีเอกสารทั้งหมด 6,200,000 ฉบับ N
- คำว่า “omega” ปรากฏในเอกสารทั้งหมด 500,000 เอกสาร (n_1=500,000) n_{mega}
- คำว่า “mike” ปรากฏในเอกสารทั้งหมด 314 เอกสาร (n_2 = 314) n_{mike}
- คำว่า “golf” ปรากฏในเอกสารทั้งหมด 80,000 เอกสาร (n_3 = 80,000) n_{golf}
- คำว่า “omega” ปรากฏ 21 ครั้งในเอกสารที่สนใจ (f_1 = 21) (เช่น Doc ที่ 800)
- คำว่า “mike” ปรากฏ 14 ครั้งในเอกสารที่สนใจ (f_2 = 14)
- คำว่า “golf” ปรากฏ 90 ครั้งในเอกสารที่สนใจ (f_3 = 90)

} อาจเป็น Doc เดียวกันที่ไหนก็ได้

- ขนาดของเอกสารที่สนใจต่อขนาดเฉลี่ยของเอกสารทั้งหมดเท่ากับ 0.4 $(\frac{dl}{avdl})$ (ถ้าทราบค่าของ Doc ต่อ)
- *กำหนดให้ k_1 = 1.25, b = 0.75, k_2 = 200*

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

(ตัวหลัง = 1)

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

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

$$\therefore K = 0.688$$

Query Q = "omega mike golf"

Example 3

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

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

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

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

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

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

Example 4

- Query Q = "lincoln lincoln" (qf = 2)
- มีเอกสารทั้งหมด 200,000 ฉบับ $N = 200{,}000$
- 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 $k_1 = 1.25$, $b = 0.75$, $k_2 = 200$

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

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

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

$$\because K = 0.781$$

Example 4

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

ตัดสูตรทิ้ง เพราะ: ไม่มี Doc ก่อนหน้านี้

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

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

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

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

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

BM25

ข้อดี

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

Doc ที่ถูก return / Doc ที่อยู่ในระบบ (ไม่ถูก return)

ข้อเสีย cat tiger ตัวอย่าง อย่างพอก ตำแหน่ง, หรือ, ที่แล้วไม่ได้ผล

- รองรับ Query อย่างง่ายเท่านั้น
- การ Ranking เปลี่ยนตาม Document ในระบบ

จำนวน N, ค่าเฉลี่ยคำ m, mark R

(การเพิ่มลดเอกสาร, การเพิ่มลด R)

- ไม่สนใจ Relationship ของ Keyword

ความสัมพันธ์