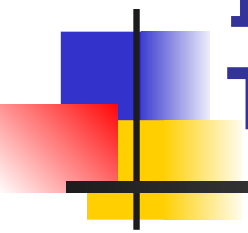
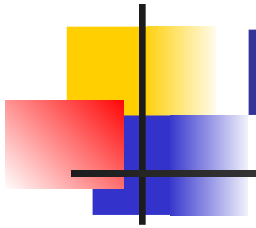


IT4853
Tìm kiếm và trình diễn thông tin

Bài 5. Mô hình nhị phân độc lập

IIR.C11. Probabilistic information retrieval

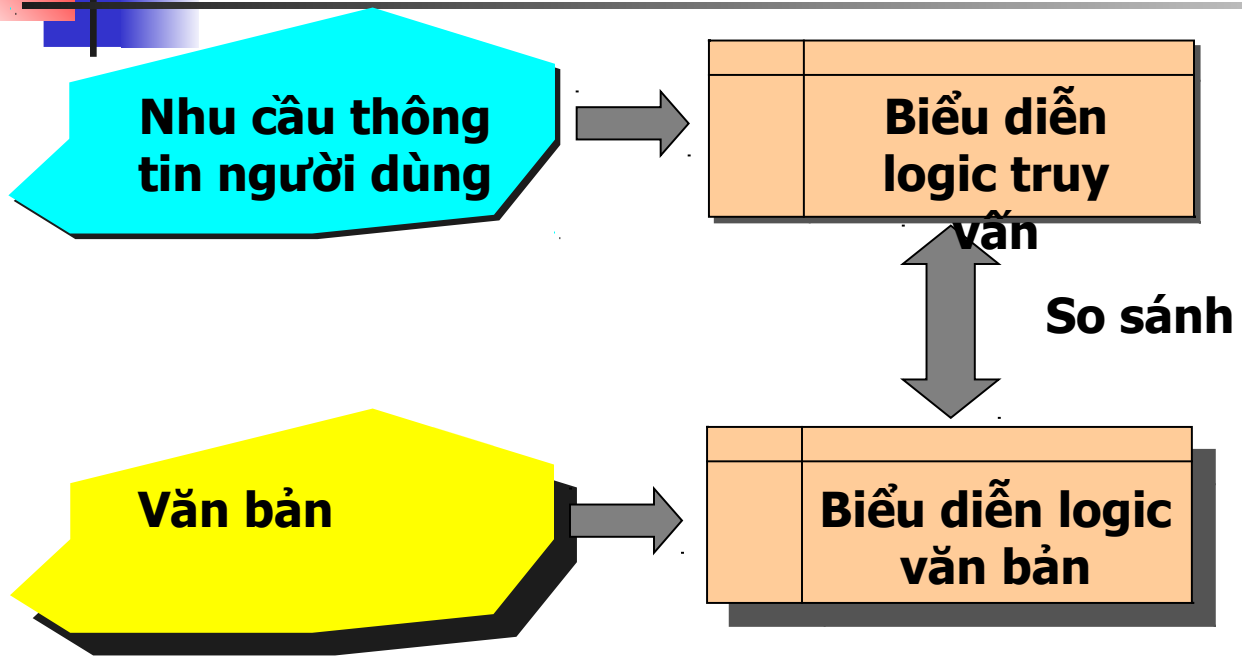
Bộ môn Hệ thống thông tin
Viện CNTT & TT



Nội dung chính

- **Ứng dụng lý thuyết xác suất trong tìm kiếm**
- **Mô hình nhị phân độc lập**
- **Mô hình (Okapi) BM25**

Lý thuyết xác suất trong tìm kiếm thông tin



Không bảo toàn ngữ nghĩa

Văn bản trả về không chắc chắn là văn bản phù hợp

Có thể ứng dụng lý thuyết xác suất trong tìm kiếm thông tin.

Lý thuyết xác suất trong tìm kiếm thông tin (2)

- **Bài toán tìm kiếm thông tin:**
  - **Cho một câu truy vấn và một biểu diễn của bộ dữ liệu văn bản, hệ thống phải xác định liệu văn bản có đáp ứng nhu cầu thông tin hay không;**
  - **Mô hình Boolean lựa chọn những văn bản thỏa mãn biểu thức truy vấn; mô hình không gian vec-tơ xếp hạng theo độ tương đồng cosine.**
- **Hệ thống tìm kiếm nắm bắt nhu cầu thông tin người dùng ở mức độ không chắc chắn, và không chắc chắn về khả năng văn bản đáp ứng nhu cầu thông tin;**
- **Lý thuyết xác suất là nền tảng suy diễn trong điều kiện không chắc chắn nói chung, và đưa ra quyết định văn bản là văn bản phù hợp trong các mô hình dựa trên xác suất nói riêng.**

Tổng quan các mô hình xác suất

- **Các mô hình xác suất cổ điển:**
  - **Nguyên tắc xếp hạng xác suất**
    - **Mô hình nhị phân độc lập, BestMatch25(Okapi)**
- **Tìm kiếm văn bản sử dụng mạng Bayes;**
- **Các mô hình ngôn ngữ**
  - **Hướng nghiên cứu mới, hiệu năng cao;**

Phương pháp xác suất là một trong những phương pháp đã tồn tại từ lâu nhưng vẫn là đề tài nóng trong tìm kiếm thông tin hiện đại.

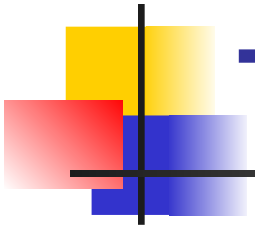
Xếp hạng xác suất

- **Ký hiệu $R_{d,q}$: là một biến nhị phân ngẫu nhiên:**
  - **$R_{d,q}$ = 1 nếu d phù hợp với q;**
  - **$R_{d,q}$ = 0, nếu ngược lại.**
- **Theo phương pháp xếp hạng xác suất, các văn bản được trả về theo thứ tự giảm dần giá trị xác suất văn bản phù hợp với truy vấn: P(R=1|d, q).**

Nguyên tắc xếp hạng xác suất

- **PRP giản lược :**
  - **Thứ tự giảm dần xác suất văn bản phù hợp với truy vấn là thứ tự tối ưu cho danh sách kết quả tìm kiếm.**
- **PRP đầy đủ:**
  - **IIR 11.2**

Nguyên tắc xếp hạng xác suất: PRP: The Probability Ranking Principle



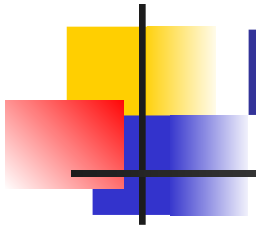
Trọng số từ

"Từ xuất hiện trong những văn bản đã biết là phù hợp phải có trọng số cao hơn so với trọng số của từ đó trong trường hợp không biết những văn bản phù hợp này."

"Có thể xây dựng cách tính trọng số từ dựa trên giả thuyết về phân bố từ vựng và luật Bayes."

[Van Rijsbergen]

Xếp hạng xác suất: Probabilistic Ranking



Nội dung chính

- Ứng dụng lý thuyết xác suất trong tìm kiếm
- **Mô hình nhị phân độc lập**
- Mô hình (Okapi) BM25

Lý thuyết xác suất căn bản

- **Quy tắc nhân xác suất (luật chuỗi)*:***

$$p(A,B) = p(A \wedge B)$$

$$p(A,B) = p(A|B)\,p(B)$$

$$p(A,B) = p(B|A)\,p(A)$$

- **Luật Bayes**

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Lý thuyết xác suất căn bản (2)

- **Quy tắc phân tích xác suất (luật phân tích):**

$$p(B) = p(A,B) + p(\overline{A},B)$$

- **Kết hợp luật Bayes và luật phân tích**

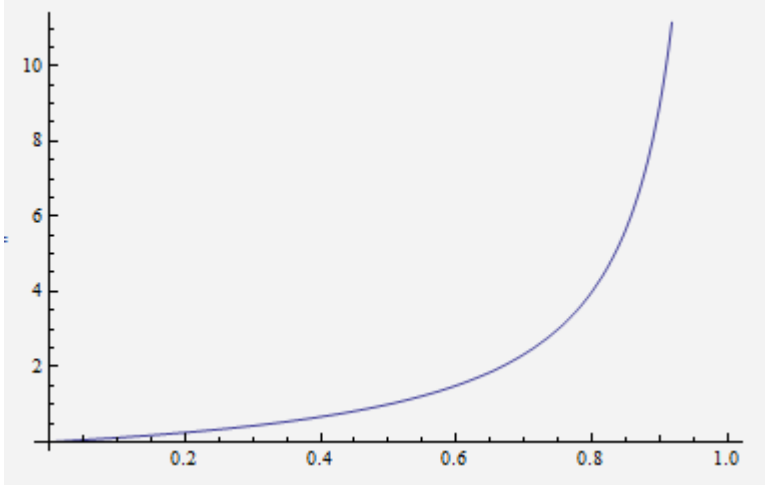
$$p(A|B) = \left[\frac{p(B|A)}{\sum p(B|X)\,p(X)} \right] p(A)$$

Giống quá trình cập nhật xác suất:
+ Bắt đầu với xác suất tiền nghiệm p(A);
+ Suy diễn xác suất p(A|B) sau khi quan sát B trong hai trường hợp xuất hiện A hoặc không.

Lý thuyết xác suất căn bản (3)

- **Cơ hội (Odds):**

$$O(A) = \frac{p(A)}{p(\overline{A})} = \frac{p(A)}{1 - p(A)}$$



Liên hệ giữa O và p

Mô hình nhị phân độc lập

- **Nhị phân: Văn bản được biểu diễn như vec-tơ nhị phân đánh dấu sự xuất hiện của từ**

$$\vec{d} = (x_1, \ldots, x_n)$$

  - **x_i = 1 nếu thuật ngữ thứ i xuất hiện trong d, 0 nếu ngược lại**
- **Độc lập: Sự xuất hiện của mỗi từ trong văn bản là độc lập với những từ còn lại;**
- **Những văn bản khác nhau có thể có cùng một biểu diễn vec-tơ.**

Mô hình nhị phân độc lập (1)

- **Cho truy vấn q**
  - **Với mỗi văn bản d cần tính p(R=1|q, d)**
  - **Chỉ quan tâm tới thứ hạng**
- **Sử dụng cơ hội (Odds) và luật Bayes**

$$O(R|q,d) = \frac{p(R=1|q,d)}{p(R=0|q,d)} = \frac{\frac{p(R=1|q)\,p(d|R=1,q)}{p(d|q)}}{\frac{p(R=0|q)\,p(d|R=0,q)}{p(d|q)}}$$

Mô hình nhị phân độc lập (2)

$$O(R|q,d) = \frac{p(R=1|q,d)}{p(R=0|q,d)} = \frac{p(R=1|q)}{p(R=0|q)} \cdot \frac{p(d|R=1,q)}{p(d|R=0,q)}$$

Hằng số với một truy vấn

Cần xác định

- **Sử dụng giả thuyết độc lập**

$$\frac{p(d|R=1,q)}{p(d|R=0,q)} = \prod_{i=1}^{n} \frac{p(x_i|R=1,q)}{p(x_i|R=0,q)}$$

$$O(R|q,d) = O(R|q) \cdot \prod_{i=1}^{n} \frac{p(x_i|R=1,q)}{p(x_i|R=0,q)}$$

Mô hình nhị phân độc lập (3)

$$O(R|q,d) = O(R|q) \cdot \prod_{i=1}^{n} \frac{p(x_i|R=1,q)}{p(x_i|R=0,q)}$$

Vì x_i chỉ nhận giá trị 1 hoặc 0

$$O(R|q,d) = O(R|q) \cdot \prod_{x_i=1} \frac{p(x_i=1|R=1,q)}{p(x_i=1|R=0,q)} \prod_{x_i=0} \frac{p(x_i=0|R=1,q)}{p(x_i=0|R=0,q)}$$

- **Đặt:**

$$p_i = p(x_i=1|R=1,q); \quad r_i = p(x_i=1|R=0,q);$$

- **Giả sử với thuật ngữ không có trong truy vấn thì $p_i = r_i$**

¿ ¿

Các đại lượng xác suất cơ bản

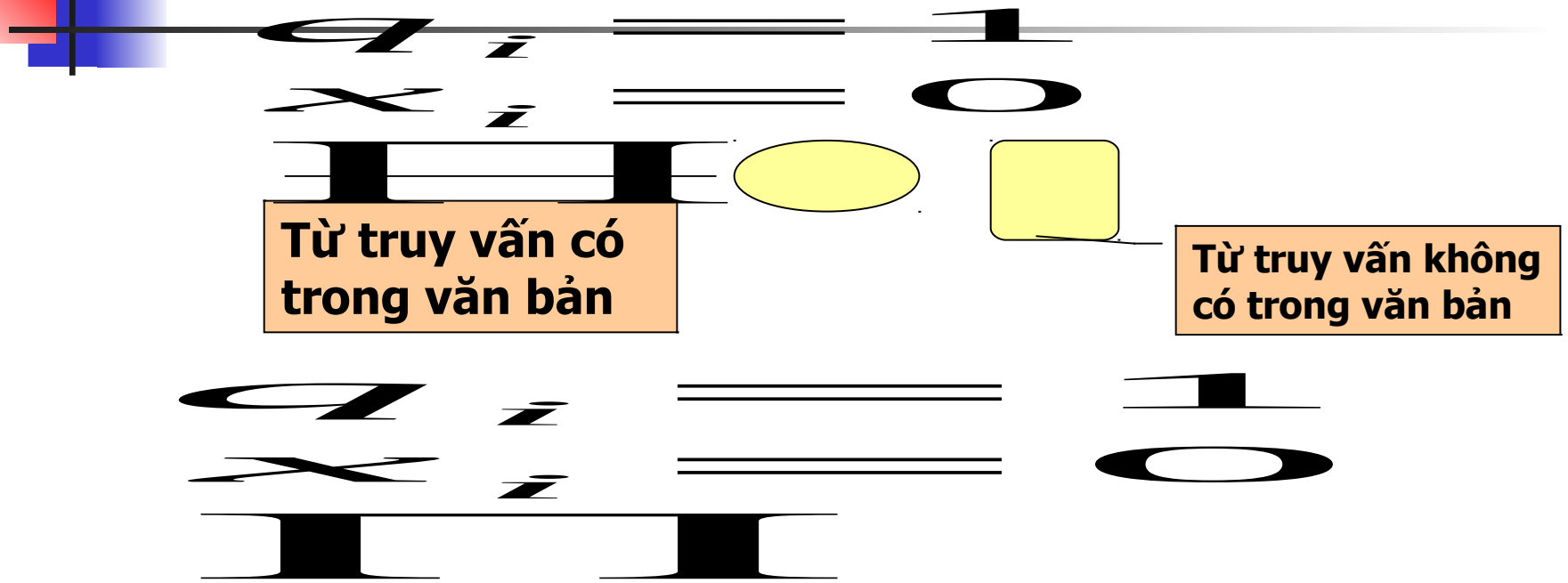
$$p_i = p(x_i = 1 | R = 1, q)$$

$$1 - p_i = p(x_i = 0 | R = 1, q)$$

$$r_i = p(x_i = 1 | R = 0, q)$$

$$1 - r_i = p(x_i = 0 | R = 0, q)$$

Mô hình nhị phân độc lập (4)



Từ truy vấn có trong văn bản

Từ truy vấn không có trong văn bản

$$O(R|q,d) = O(R|q) \cdot \prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \cdot \prod_{q_i=1} \frac{1-p_i}{1-r_i}$$

Từ truy vấn có trong văn bản

Tất cả từ truy vấn

Mô hình nhị phân độc lập (5)

$$O(R|q,d) = O(R|q) \cdot \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} \cdot \prod_{q_i = 1} \frac{1-p_i}{1-r_i}$$

Hằng số với một truy vấn

Đại lượng duy nhất cần xác định cho mục đích xếp hạng

Hàm xếp hạng

$$Rank(d,q) = \log \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Mô hình nhị phân độc lập (6)

- **Kết quả tìm kiếm được xác định dựa trên Rank**

$$RSV(d,q) = \log \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

$$RSV(d,q) = \sum_{x_i = q_i = 1} c_i; \qquad c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

c_i có vai trò như trọng số thuật ngữ trong mô hình này

Tính c_i ntn từ bộ dữ liệu sẵn có.

Những số liệu thống kê cơ bản

Đại lượng thống kê ứng với từ thứ *i*:

Từ/văn bản	Phù hợp	Không phù hợp	Tổng
$x_i=1$	s	$n-s$	n
$x_i=0$	$S-s$	$N-n-S+s$	$N-n$
Tổng	S	$N-S$	N

$$p_i \approx \frac{s}{S} \qquad r_i \approx \frac{n-s}{N-S} \qquad p_i = p(x_i=1|R=1,q); \quad r_i = p(x_i=1|R=0,q);$$

$$c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)} \qquad c_i \approx K(N,n,S,s) = \log \frac{s/(S-s)}{(n-s)/(N-n-S+s)}$$

Làm mịn trọng số

- **Có thể thêm 0.5 vào mỗi tham số để đảm bảo các trọng số không trở thành vô cùng khi S, s nhỏ:**

$$c_t = \log \frac{(s+0.5)(N-S-n+s+0.5)}{(n-s+0.5)(S-s+0.5)}$$

Bắt đầu thực hiện truy vấn

- **Hoàn toàn không biết về R**

$$c_t = \log \frac{N - n + 0.5}{n + 0.5}$$

Tương tự trọng số idf.

Có thể sử dụng giá trị này để tính hạng ban đầu

Ví dụ mô hình xác suất

d	Biểu diễn vec-tơ văn bản									
	a	b	c	d	e	f	g	h	k	l
1	1			1				1	1	
2								1	1	1
3		1				1	1			
4	1			1						1
5								1	1	
6			1		1					

w_t

$$w_t = \log\left(\frac{N - n + 0.5}{n + 0.5}\right)$$

Cải thiện xếp hạng

- **Nếu người dùng phản hồi về văn bản phù hợp**
- **Xác định lại p_i và r_i dựa trên thông tin này**
- **Hoặc có thể kết hợp với thông tin mới**

$$p_i^{(2)} = \frac{|VR_i| + \kappa p_i^{(1)}}{|VR| + \kappa} = \frac{s + \kappa p_i^{(1)}}{S + \kappa}$$

κ là trọng số đã biết

- **Lặp lại để xác định chính xác hơn những văn bản phù hợp**

Xác định p_i và r_i nhờ vòng lặp

Phù hợp phản hồi giả lập

- **1. Giả sử p_i là hằng số với mọi x_i trong truy vấn. *Ví dụ,* $p_i = 0.5$ với văn bản bất kỳ**
- **2. Giả sử tập V với những văn bản được xếp hạng cao nhất theo mô hình này là những văn bản phù hợp.**
- **3. Cần xác định lại p_i và r_i, sử dụng phân bố từ trong V. Đặt V_i là tập văn bản có chứa x_i, chúng ta có $p_i = |V_i| / |V|$**
- **4. Giả sử không được trả về đồng nghĩa với không phù hợp, $r_i = (n_i - |V_i|) / (N - |V|)$**
- **5. Lặp các bước 2-4 cho tới khi hội tụ và trả về kết**

Ví dụ trọng số phù hợp

d	Biểu diễn vec-tơ văn bản									
	a	b	c	d	e	f	g	h	k	l
1	1			1				1	1	
2								1	1	1
3		1				1	1			
4	1			1						1
5								1	1	
6			1		1					

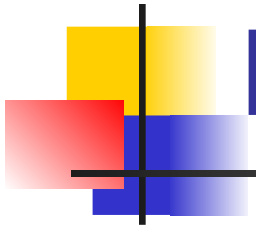
w_t

Văn bản số 2 là văn bản phù hợp

$$w_t = \log \frac{(s+0.5)(N-S-n+s+0.5)}{(n-s+0.5)(S-s+0.5)}$$

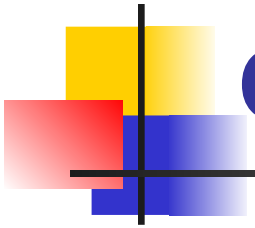
Tổng kết mô hình BIM

- Mô hình xác suất dựa trên lý thuyết xác suất để mô hình hóa sự không chắc chắn trong quá trình tìm kiếm
- Sử dụng các giả thuyết về sự độc lập trong quá trình ước lượng giá trị xác suất
- Từ không xuất hiện trong truy vấn không ảnh hưởng tới tính phù hợp (có $p_i = r_i$)
- Trọng số ban đầu của thuật ngữ khi chưa có thông tin về văn bản phù hợp được xác định tương tự *idf*.
- Phù hợp phản hồi giả lập có thể giúp cải thiện xếp hạng bằng cách xác định lại xác suất thuật ngữ
- Không sử dụng các tần suất thuật ngữ nội văn bản



Nội dung chính

- Ứng dụng lý thuyết xác suất trong tìm kiếm
- Mô hình nhị phân độc lập
- **Mô hình (Okapi) BM25**



Okapi BM25

- **BM25 "Best Match 25"**
  - **Được phát triển trong hệ thống Okapi (City University London)**
  - **Hiệu quả đã được xác nhận trong thực nghiệm**
- **Sử dụng tần suất từ và độ dài văn bản, nhưng không bổ xung quá nhiều tham số so với BIM**

(Robertson and Zaragoza 2009; Spärck Jones et al. 2000)

Trọng số Okapi

$$RSV_d = \sum_{t \in q} [\log \left[\frac{(|VR_t| + 1/2)/(|VNR_t| + 1/2)}{(df_t - |VR_t| + 1/2)/(N - df_t - |VR| + |VR_t| + 1/2)} \right] \times ¿$$

$$¿ \; ¿ \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b \times (L_d/L_{ave}))+tf_{t,d}} \times \frac{(k_3+1)tf_{t,q}}{k_3+tf_{t,q}}] \; ¿ ¿$$

VR_t – tập văn bản phù hợp có chứa t

VNR_t – không chứa t

$$RSV_d = \sum_{t \in q} [\log \left[\frac{(s+1/2)/(S-s+1/2)}{(n-s+1/2)/(N-n-S+s+1/2)} \right] \times ¿$$

$$¿ \; ¿ \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b \times (L_d/L_{ave}))+tf_{t,d}} \times \frac{(k_3+1)tf_{t,q}}{k_3+tf_{t,q}}] \; ¿ ¿$$

Trọng số Okapi BM25

- **Khi từ xuất hiện trong quá nửa số văn bản và S = s = 0, thành phần:**

$$\log\left[\frac{(s+1/2)/(S-s+1/2)}{(n-s+1/2)/(N-n-S+s+1/2)}\right]$$

có thể nhận giá trị âm

- **Trong trường hợp không có thông tin về văn bản phù hợp, có thể sử dụng công thức:**

$$RSV_d=\sum_{t\in q}\left[\log\frac{N}{df_t}\right]\cdot\frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b\times(L_d/L_{ave}))+tf_{t,d}}\times\frac{(k_3+1)tf_{t,q}}{k_3+tf_{t,q}}$$

Trọng số Okapi

- **Trọng số Okapi sử dụng**
  - **thành phần "tf" tương tự như VSM**
  - **chuẩn hóa độ dài văn bản và độ dài truy vấn độc lập**
  - **một vài tham số phụ thuộc bộ dữ liệu**

$$RSV_d = \sum_{t \in q} \left[\log \left[\frac{(s+1/2)/(S-s+1/2)}{(n-s+1/2)/(N-n-S+s+1/2)} \right] \times \right.$$

$$\left. \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b\times(L_d/L_{ave}))+tf_{t,d}} \times \frac{(k_3+1)tf_{t,q}}{k_3+tf_{t,q}} \right]$$

Tính trọng số Okapi BM25

d	Biểu diễn vec-tơ văn bản										dl
	a	b	c	d	e	f	g	h	k	l	
1	2			1				2	1		6
2								1	1	1	3
3		1				1	1				3
4	1			2						1	4
5								3	1		4
6			1		1						2

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{df_t} \right] \cdot \frac{(k_1+1) tf_{t,d}}{k_1 \left((1-b) + b \times (L_d / L_{ave}) \right) + tf_{t,d}} \times \frac{(k_3+1) tf_{t,q}}{k_3 + tf_{t,q}}$$

k1 = 1.2

k3 = 7

b = 0.75

Khi có thông tin về văn bản phù hợp

d	Biểu diễn vec-tơ văn bản										dl
	a	b	c	d	e	f	g	H	k	l	
1	2			1				2	1		6
2								1	1	1	3
3		1				1	1				3
4	1			2						1	4
5								3	1		4
6			1		1						2

$$RSV_d = \sum_{t \in q} [\log \left[\frac{(s+1/2)/(S-s+1/2)}{(n-s+1/2)/(N-n-S+s+1/2)} \right] \times ¿$$

¿

$$¿¿ \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b\times(L_d/L_{ave}))+tf_{t,d}} \times \frac{(k_3+1)tf_{t,q}}{k_3+tf_{t,q}}] ¿¿$$

k1 = 1.2 **k3 = 7**

b = 0.75

(L_{ave}) avdl = 3.66

Bài tập 5.1

- **So sánh sự khác biệt giữa trọng số tf-idf của mô hình không gian vec-tơ và trọng số c_i của mô hình BIM (trong trường hợp không có thông tin về văn bản phù hợp).**