

# Tìm kiếm thông tin

## Chương 7. Tìm kiếm thông tin dựa trên xác suất

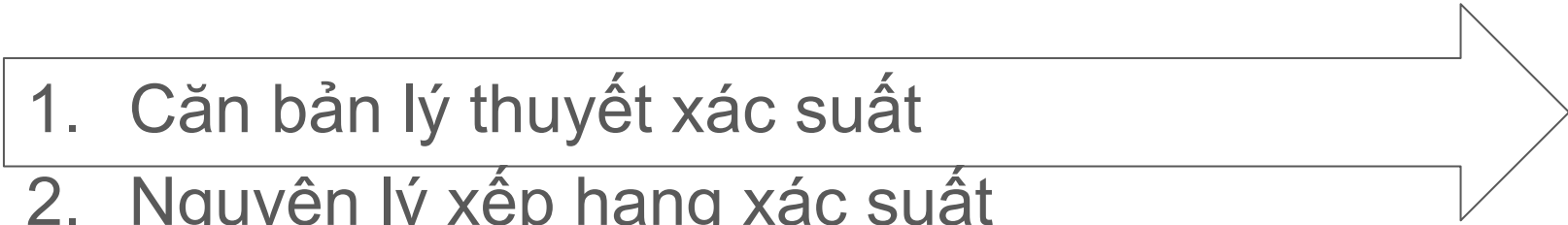
*Soạn bởi: TS. Nguyễn Bá Ngọc*

2021

# Nội dung

1. Căn bản lý thuyết xác suất
2. Nguyên lý xếp hạng xác suất
3. Mô hình nhị phân độc lập
4. Mô hình Okapi BM25
5. Khái niệm mô hình ngôn ngữ
6. Ước lượng xác suất trong mô hình ngôn ngữ
7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
8. Các mở rộng mô hình ngôn ngữ

# Nội dung

- 
1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ

# Ôn lại lý thuyết xác suất cơ bản

- Cho các sự kiện A và B
  - Xác suất xuất hiện đồng thời cả 2 sự kiện được ký hiệu là  $P(A, B)$
  - Xác suất sự kiện A xuất hiện trong điều kiện đã xuất hiện B là  $P(A|B)$  (xác suất có điều kiện).
- Luật chuỗi thiết lập mối quan hệ giữa xác suất có điều kiện và xác suất đồng xuất hiện:  
$$P(A, B) = P(A|B)P(B) \text{ và } P(A, B) = P(B|A)P(A)$$
- Luật phân rã: Nếu B có thể được chia hoàn toàn thành một tập các sự kiện không giao nhau, thì  $P(B)$  là tổng xác suất của các sự kiện đó. Chúng ta xét một trường hợp:

$$B = (B \cap A) \cup (B \cap \bar{A}) \Rightarrow P(B) = P(B, A) + P(B, \bar{A})$$

# Ôn lại lý thuyết xác suất cơ bản<sub>(2)</sub>

- Kết hợp luật Bayes và luật phân rã:

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

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{X \in \{A, \bar{A}\}} P(B|X)P(X)}$$

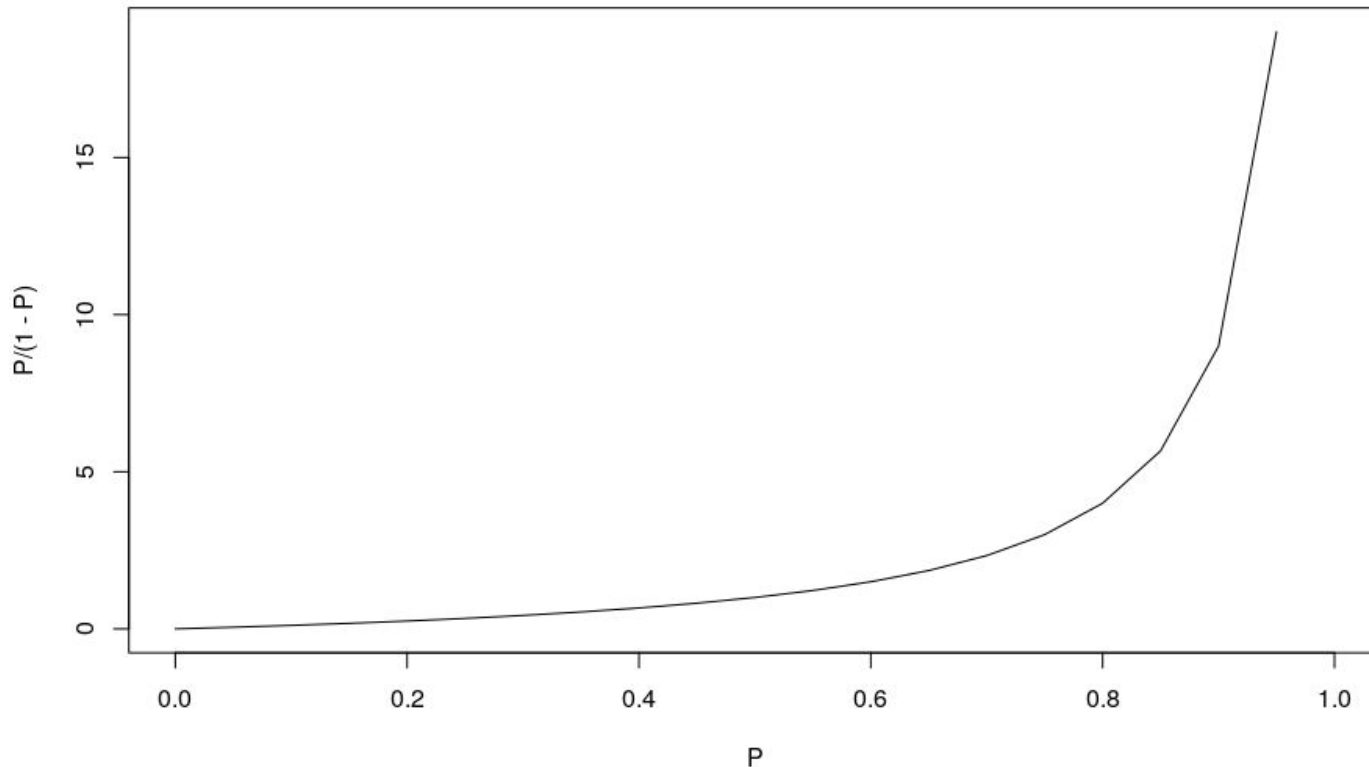
- Giống như một cách cập nhật các xác suất:
  - Bắt đầu với xác suất tiên nghiệm  $P(A)$  (ước lượng xác suất của sự kiện  $A$  trong điều kiện không có bất kỳ thông tin nào khác)
  - Suy diễn một xác suất hậu nghiệm  $P(A|B)$  sau khi đã quan sát sự kiện  $B$  trong 2 trường hợp xuất hiện  $A$  và không xuất hiện  $A$ .

# Ôn lại lý thuyết xác suất cơ bản<sup>(3)</sup>

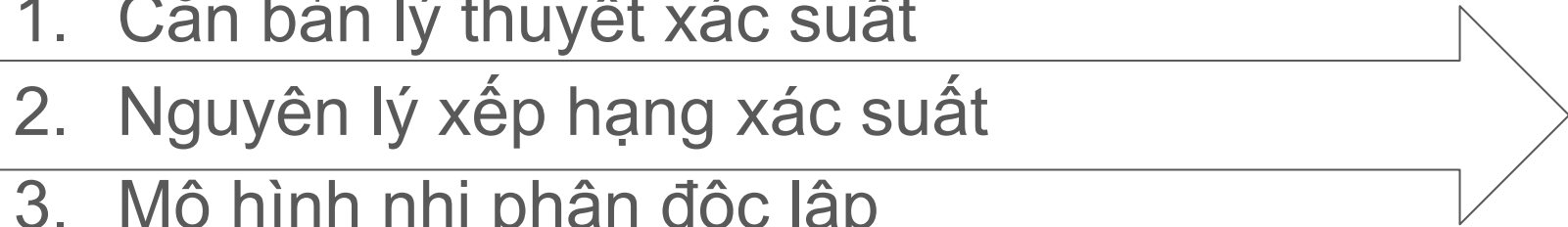
- Cơ hội của một sự kiện là một đại lượng vô hướng bằng tỉ lệ xác suất xảy ra sự kiện/xác suất không xảy ra sự kiện.

$$O(A) = \frac{P(A)}{P(\bar{A})} \quad O(A) = \frac{P(A)}{1 - P(A)}$$

- Là một đại lượng đồng biến với giá trị xác suất



# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# Nguyên lý xếp hạng xác suất (PRP)

- Probability Ranking Principle (PRP)
- PRP giản lược
  - Nếu các văn bản được xếp hạng theo thứ tự giảm dần xác suất phù hợp với câu truy vấn, thì hiệu quả của hệ thống sẽ là cao nhất trong giới hạn có thể đạt được.
- PRP đầy đủ
  - *Nếu kết quả tìm kiếm cho mỗi truy vấn là một danh sách các văn bản được xếp hạng theo thứ tự giảm dần xác suất phù hợp với truy vấn, trong đó các giá trị xác suất đã được ước lượng chính xác nhất có thể dựa trên tất cả những dữ liệu mà hệ thống có cho mục đích này, thì lợi ích mà người dùng nhận được từ hệ thống là cực đại trong phạm vi có thể đạt được dựa trên những dữ liệu đó.*
  - *(Tham khảo thêm về PRP đầy đủ)*



# Xếp hạng văn bản theo xác suất


- Các văn bản được xếp hạng theo thứ tự giảm dần xác phù hợp với truy vấn:

$$\text{RSV}(d, q) = P(R=1|d, q)$$

Trong đó  $R$  là 1 biến nhị phân ngẫu nhiên:  $R = 1$  Nếu văn bản phù hợp với truy vấn;  $R = 0$  nếu ngược lại.

*Trong các công thức tiếp theo chúng ta thu gọn cách viết, mặc định  $R$  là  $R = 1$ , và  $\bar{R}$  là  $R = 0$ .*

# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# Mô hình nhị phân độc lập (BIM)

## Binary Independence Model (BIM)

### Các giả thuyết:

- “Nhị phân”: Các văn bản và truy vấn được biểu diễn như những vec-tơ đánh dấu sự xuất hiện của từ
  - Văn bản được biểu diễn như vec-tơ đánh dấu  $\vec{x} = (x_1, \dots, x_M)$ ;
  - Trong đó  $x_i = 1$  nếu  $t_i$  xuất hiện trong  $d$ ;  $x_i = 0$  nếu ngược lại;  $M$  là kích thước bộ từ vựng.
  - *!Lưu ý: Các tài liệu khác nhau có thể có cùng biểu diễn.*
- “Độc lập”: Sự xuất hiện của mỗi từ là độc lập với sự xuất hiện của các từ khác
  - Tính đúng đắn của giả thuyết là tương đối.
  - Chúng ta cũng đã sử dụng giả thuyết độc lập trong VSM.

# Vấn đề xếp hạng trong BIM

- Theo PRP chúng ta cần xếp hạng các văn bản theo  $P(R|d, q)$
- Trong BIM  $P(R|d, q)$  được đánh giá dựa trên các biểu diễn:

$$P(R = 1|\vec{x}, \vec{q}) = \frac{P(\vec{x}|R = 1, \vec{q})P(R = 1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(R = 0|\vec{x}, \vec{q}) = \frac{P(\vec{x}|R = 0, \vec{q})P(R = 0|\vec{q})}{P(\vec{x}|\vec{q})}$$

- $P(R=1|\vec{q})$  và  $P(R=0|\vec{q})$  là các xác suất tiên nghiệm theo thứ tự cho các trường hợp văn bản phù hợp và không phù hợp với truy vấn  $q$ .
  - Có thể ước lượng dựa trên tỉ lệ văn bản phù hợp có trong bộ dữ liệu, dữ liệu lịch sử, hoặc đơn giản là sử dụng giá trị đồng nhất, v.v..
  - ... Nhưng là các hằng số đối với 1 truy vấn.

# Suy diễn hàm xếp hạng

- Trong tìm kiếm thông tin người dùng chỉ quan tâm thứ tự tương đối giữa các văn bản (*nguyên lý xếp hạng*, Bài 1).
  - => Có thể sử dụng cơ hội  $O(R|d, q)$  thay thế cho  $P(R|d, q)$ .
- Áp dụng luật Bayes cho các thành phần trong cơ hội:

$$O(R|\vec{x}, \vec{q}) = \frac{P(R=1|\vec{x}, \vec{q})}{P(R=0|\vec{x}, \vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1, \vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0, \vec{q})}{P(\vec{x}|\vec{q})}}$$

Hằng số đối với 1 truy vấn,  
=> Có thể lược bỏ.

$$= \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})} \cdot \frac{P(\vec{x}|R=1, \vec{q})}{P(\vec{x}|R=0, \vec{q})}$$

- Áp dụng giả thuyết độc lập

$$\Rightarrow O(R|\vec{x}, \vec{q}) = O(R|\vec{q}) \cdot \prod_{t=1}^M \frac{P(x_t|R=1, \vec{q})}{P(x_t|R=0, \vec{q})}$$

# Suy diễn hàm xếp hạng<sub>(2)</sub>

Bởi vì  $x_t$  chỉ có thể nhận giá trị 0 hoặc 1, chúng ta có thể tách chuỗi tích thành 2 thành phần:

$$O(R|\vec{x}, \vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1} \frac{P(x_t = 1|R=1, \vec{q})}{P(x_t = 1|R=0, \vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t = 0|R=1, \vec{q})}{P(x_t = 0|R=0, \vec{q})}$$

- Đặt  $p_t = P(x_t = 1|R=1, \vec{q})$  là xác suất từ xuất hiện trong văn bản phù hợp
- Đặt  $u_t = P(x_t = 1|R=0, \vec{q})$  là xác suất từ xuất hiện trong văn bản không phù hợp

Chúng ta có các giá trị xác suất:

Từ/Văn bản		Phù hợp ( $R = 1$ )	Không phù hợp ( $R = 0$ )
Có xuất hiện	$x_t = 1$	$p_t$	$u_t$
Không xuất hiện	$x_t = 0$	$1 - p_t$	$1 - u_t$

# Suy diễn hàm xếp hạng <sup>(3)</sup>

Tiếp tục đặt giả thuyết:

*Giả sử nếu từ không có trong truy vấn thì xác suất xuất hiện trong văn bản phù hợp = xác suất xuất hiện trong văn bản không phù hợp - Nếu  $q_t = 0$  thì  $p_t = u_t$ .*

Chúng ta có:

$$O(R|\vec{x}, \vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

Đại lượng duy nhất cần tính để xếp hạng

# Hàm giá trị trạng thái tìm kiếm

$$RSV(d, q) = \log \prod_{t: x_t = q_t = 1} \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

$$RSV(d, q) = \sum_{t: x_t = q_t = 1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

Đặt  $c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$

Chúng ta có:  $RSV(d, q) = \sum_{t: x_t = q_t = 1} c_t$

$c_t$  có vai trò tương tự trọng số từ và được tích lũy vào giá trị trạng thái tìm kiếm của văn bản.



# Trọng số BIM

$$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

- Trọng số  $c_t$  về bản chất là tỉ lệ của 2 cơ hội:
  - (i) Cơ hội từ xuất hiện trong văn bản phù hợp  $p_t/(1-p_t)$ ,
  - và (ii) Cơ hội từ xuất hiện trong văn bản không phù hợp  $u_t/(1-u_t)$ .
- $c_t = 0$  Nếu các cơ hội bằng nhau;
- $c_t > 0$  Nếu cơ hội xuất hiện trong văn bản phù hợp cao hơn.
- $c_t < 0$  Trong trường hợp còn lại (cơ hội xuất hiện trong văn bản phù hợp nhỏ hơn).

# Ước lượng các giá trị xác suất

- Cho truy vấn  $q$  giả sử **đã biết các văn bản phù hợp** với  $q$ . Với mỗi từ truy vấn  $t$  chúng ta thống kê các đại lượng sau trên toàn tập văn bản:

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	$S$	$N-S$	$N$

$(n = df)$

- Theo khả năng cực đại chúng ta có:

$$p_t = s/S;$$

$$u_t = (n - s)/(N - S);$$

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

*Có thể áp dụng phương pháp làm mịn để tránh giá trị 0.*

# Làm mịn trọng số BIM

- Cộng thêm 0.5 vào mỗi thành phần:

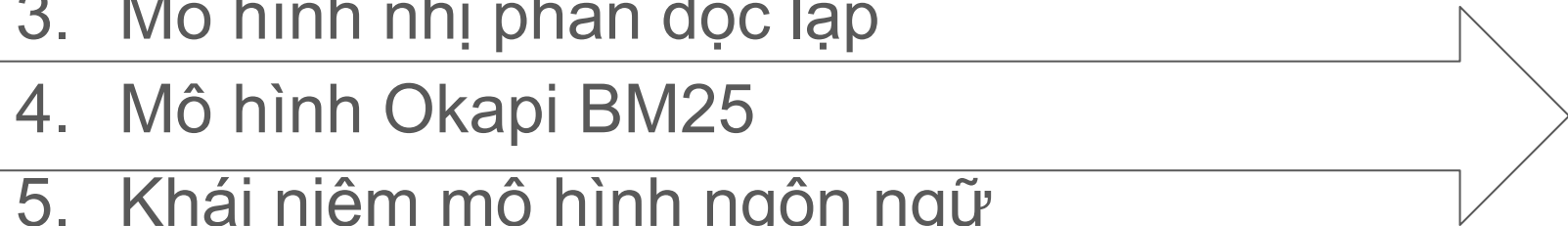
$$c_t = \log \frac{(s + 0.5)(N - S - n + s + 0.5)}{(n - s + 0.5)(S - s + 0.5)}$$

- Trong trường hợp hoàn toàn không biết về các văn bản phù hợp có thể sử dụng công thức giản lược:

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

*Tương tự idf (p - log prob, Chương 4).*

# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# Hạn chế của BIM

- Độ chính xác không cao
  - BIM - Được thiết kế để tìm kiếm văn bản ngắn (tiêu đề hoặc tóm tắt) và số lượng văn bản không quá nhiều.
  - Với biểu diễn dạng vec-tơ nhị phân nhiều văn bản khác nhau có thể có cùng biểu diễn.
  - Không phù hợp cho tìm kiếm với các văn bản dài và số lượng văn bản lớn.
- Để xếp hạng tốt hơn chúng ta cần sử dụng tần suất từ và độ dài văn bản
  - *(Chúng ta đã sử dụng các đại lượng này trong mô hình VSM, Bài 4)*

# Okapi BM25

- BM25 - Best Match 25 (kết quả tốt nhất trong 25 phiên bản)
  - Được phát triển trong hệ thống Okapi
  - Ngày càng được chấp nhận rộng rãi hơn trong ứng dụng
  - Được đánh giá hiệu quả hơn độ tương đồng cosine
- Ngoài thành phần tương tự BIM, BM25 còn sử dụng tần suất từ và độ dài văn bản với các tham số điều chỉnh.

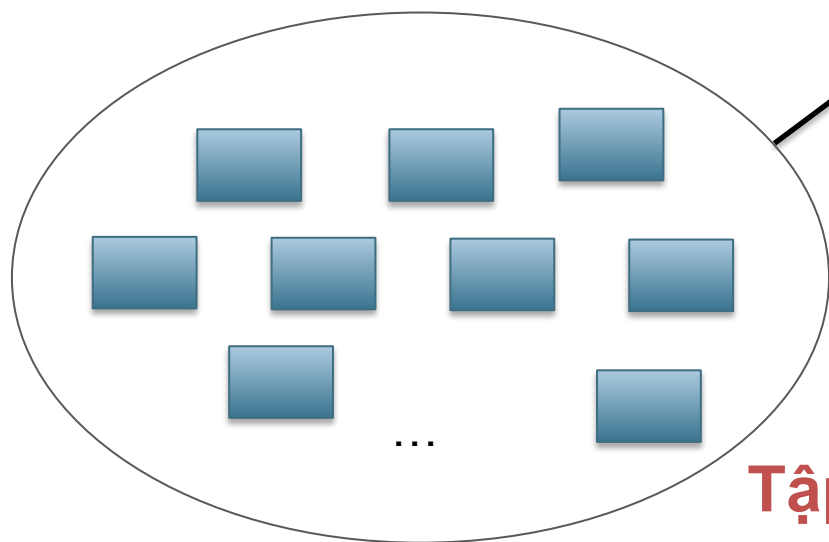
[Robertson and Zaragoza 2009; Spärck Jones et al. 2000]

# Mô hình sinh cho các văn bản

- Các từ được lấy ra độc lập với nhau từ một bộ từ vựng theo 1 phân bố đa thức
- Phân bố tần suất từ (tf) tuân theo một phân bố nhị thức - có thể được mô phỏng bằng phân bố Poisson



**Văn bản**



**Từ ...**



**Tập từ**

# Phân bố Poisson

- Phân bố Poisson mô phỏng xác suất 1 sự kiện xuất hiện  $k$  lần trong một khoảng thời gian hoặc không gian cố định, với tần suất trung bình đã biết  $\lambda$  ( $= cf/T$  đối với từ) và độc lập với sự kiện xuất hiện cuối cùng:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Ví dụ:
  - Số lượng ô tô đến quầy vé trên phút
  - Số lần từ xuất hiện trong văn bản

*“1 khoảng thời gian hoặc không gian cố định” => Độ dài các văn bản là cố định. Cách xử lý vấn đề này sẽ được cung cấp sau.*



# Phân bố Poisson<sub>(2)</sub>

- Nếu T lớn và p nhỏ, thì chúng ta có thể mô phỏng một phân bố nhị thức với 1-Poisson trong đó  $\lambda = Tp$

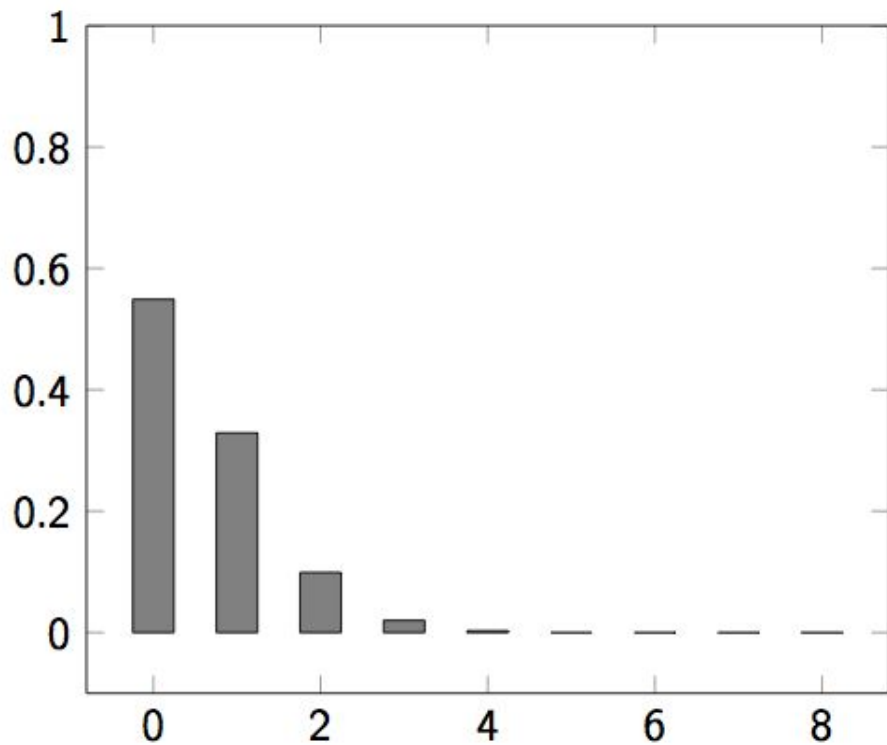
$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Ví dụ  $p = 0.08$ ,  $T = 20$ , xác suất xuất hiện 1 lần:
  - Nhị thức  $P(1) = \binom{20}{1} (.08)^1 (.92)^{19} = .3282$
  - Poisson  $P(1) = \frac{[(20)(.08)]^1}{1!} e^{-(20)(.08)} = \frac{1.6}{1} e^{-1.6} = 0.3230$

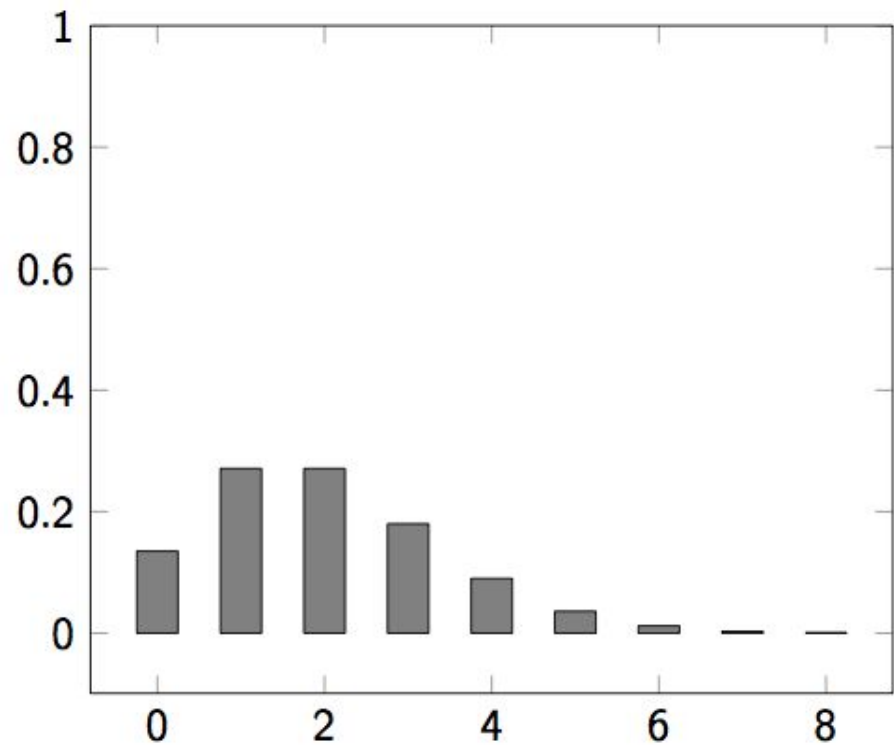
} Tương đối gần

# Ví dụ 7.1. Phân bố Poisson

$\lambda = 0.6$



$\lambda = 2$



# Mô hình 1-Poisson

- Tương đối gần kỳ vọng cho các từ thông dụng
- Tương đối xa với các từ thuộc các chủ đề đặc biệt
  - Thường xuyên có  $p(k)$  cao hơn giá trị được dự đoán.

		Văn bản có từ xuất hiện $k$ lần ( $\lambda = 53/650$ )												
Tần suất	Từ	0	1	2	3	4	5	6	7	8	9	10	11	12
53	Mong đợi	599	49	2										
52	<i>based</i>	600	48	2										
53	<i>conditions</i>	604	39	7										
55	<i>cathexis</i>	619	22	3	2	1	2	0	1					
51	<i>comic</i>	642	3	0	1	0	0	0	0	0	0	1	1	2

[Harter, “A Probabilistic Approach to Automatic Keyword Indexing”, JASIST, 1975 ]

# Tính tiêu biểu

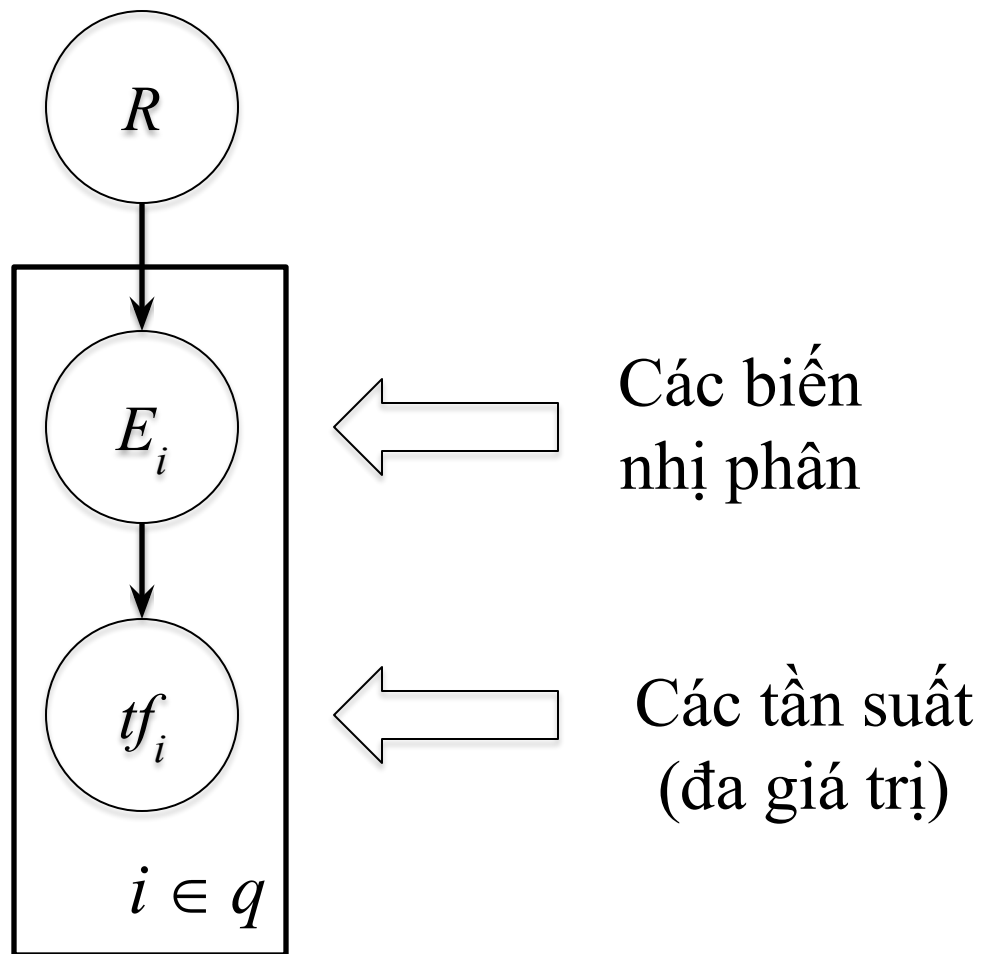
- Mô hình hóa tần suất từ sử dụng *tính tiêu biểu*
- *Tính tiêu biểu/eliteness?*
  - Biến ẩn cho mỗi cặp từ-văn bản, ký hiệu  $E_i$  cho từ  $i$
  - Một từ là *tiêu biểu* đối với một văn bản nếu (ở một mức độ nhất định) nội dung văn bản nói về chủ đề được hàm ý trong từ.
  - *Tính tiêu biểu* là 1 tính chất nhị phân
  - Sự xuất hiện của từ chỉ phụ thuộc vào tính tiêu biểu ...
  - ... và tính tiêu biểu phụ thuộc vào tính phù hợp.

## Ví dụ 7.2. Tính tiêu biểu của từ

Trang bách khoa toàn thư mở (Wikipedia) về công nghệ nano đánh dấu các từ tiêu biểu/elite:

**Công nghệ nano** là việc sử dụng vật chất ở quy mô **nguyên tử**, **phân tử** và **siêu phân tử** cho các mục đích công nghiệp. Mô tả phổ biến sớm nhất về công nghệ nano đề cập đến mục tiêu công nghệ cụ thể là thao tác chính xác các nguyên tử và phân tử để chế tạo các sản phẩm có quy mô vĩ mô, ngày nay còn được gọi là **công nghệ nano phân tử**.<sup>[1][2]</sup> Sau đó, một mô tả khái quát hơn về công nghệ nano đã được thiết lập bởi **Sáng kiến Công nghệ Nano Quốc gia**, định nghĩa công nghệ nano là sự điều khiển vật chất với ít nhất một kích thước có kích thước từ 1 đến 100 **nanomet**. Định nghĩa này phản ánh thực tế rằng **các** hiệu ứng **cơ lượng tử** rất quan trọng ở quy mô **lĩnh vực lượng tử** này, và do đó

# Mô hình sinh với tính tiêu biểu



# Hàm giá trị trạng thái tìm kiếm

- Suy diễn tương tự BIM, chúng ta có:

$$RSV^{elite} = \sum_{i \in q, tf_i > 0} c_i^{elite}(tf_i);$$

trong đó

$$c_i^{elite}(tf_i) = \log \frac{p(TF_i = tf_i | R = 1)p(TF_i = 0 | R = 0)}{p(TF_i = 0 | R = 1)p(TF_i = tf_i | R = 0)}$$

và sử dụng tính tiêu biểu, chúng ta có:

$$\begin{aligned} p(TF_i = tf_i | R) &= p(TF_i = tf_i | E_i = elite)p(E_i = elite | R) \\ &\quad + p(TF_i = tf_i | E_i = \overline{elite})(1 - p(E_i = elite | R)) \end{aligned}$$

# Mô hình 2-Poisson

- Các vấn đề với mô hình 1-Poisson được khắc phục trong phân bố 2-Poisson
- Trong mô hình 2-Poisson, phân bố thay đổi phụ thuộc vào từ có phải là tiêu biểu hoặc không.

$$P(TF_i=k_i|R) = \pi \frac{\lambda^{k_i}}{k_i!} e^{-\lambda} + (1-\pi) \frac{\mu^{k_i}}{k_i!} e^{-\mu}$$

- Trong đó  $\pi$  là xác suất văn bản là tiêu biểu với từ.
- nhưng, vấn đề là chúng ta không biết  $\pi$ ,  $\lambda$ ,  $\mu$



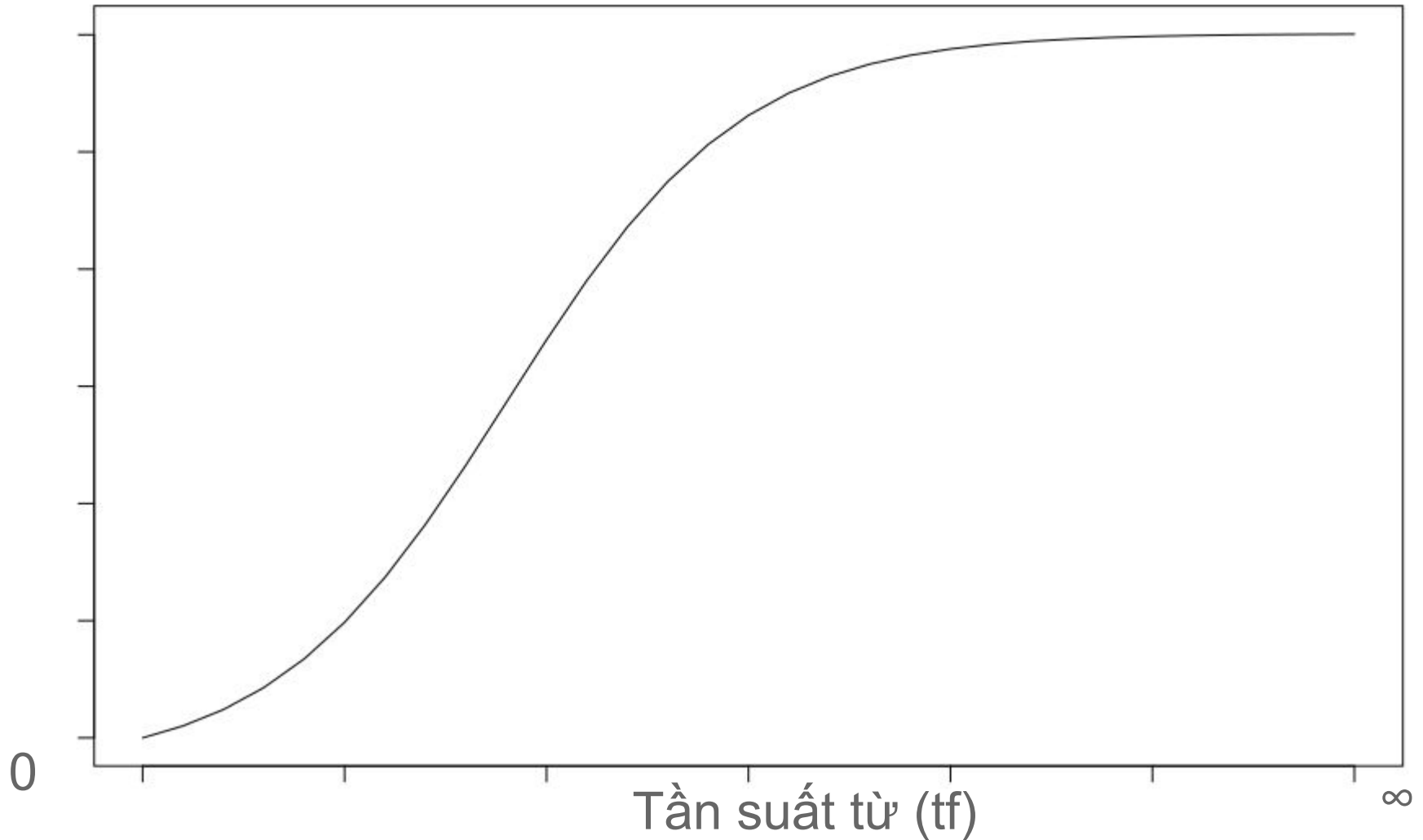
# Trọng số 2-Poisson

- Chúng ta ký hiệu  $E(k) = \frac{\lambda^k}{k!} e^{-\lambda}$  và  $U(k) = \frac{\mu^k}{k!} e^{-\mu}$
- Ký hiệu  $p$  ( $=\pi$ ) = xác suất từ tiêu biểu với văn bản phù hợp và  $q$  = xác suất từ tiêu biểu với văn bản không phù hợp.
- Chúng ta có:

$$c^{\text{elite}}(\text{tf}) = \log \frac{(E(\text{tf}) * p + U(\text{tf}) * (1-p)) * (E(0) * q + U(0) * (1-q))}{(E(0) * p + U(0) * (1-p)) * (E(\text{tf}) * q + U(\text{tf}) * (1-q))}$$

# Phác thảo đồ thị

Vẽ  $c_i^{\text{elite}}(\text{tf})$  cho 1 số từ (các tham số 2-Poisson khác nhau)



# Các thuộc tính nhận biết

- $c_i^{\text{elite}}(0) = 0$
- $c_i^{\text{elite}}(tf)$  tăng đơn điệu cùng với  $tf$
- ... có xu hướng tiến tới một giá trị cực đại khi  $tf \rightarrow \infty$

*[không đúng với các trọng số theo  $tf$  như trong VSM]*

- Với giới hạn tiệm cận

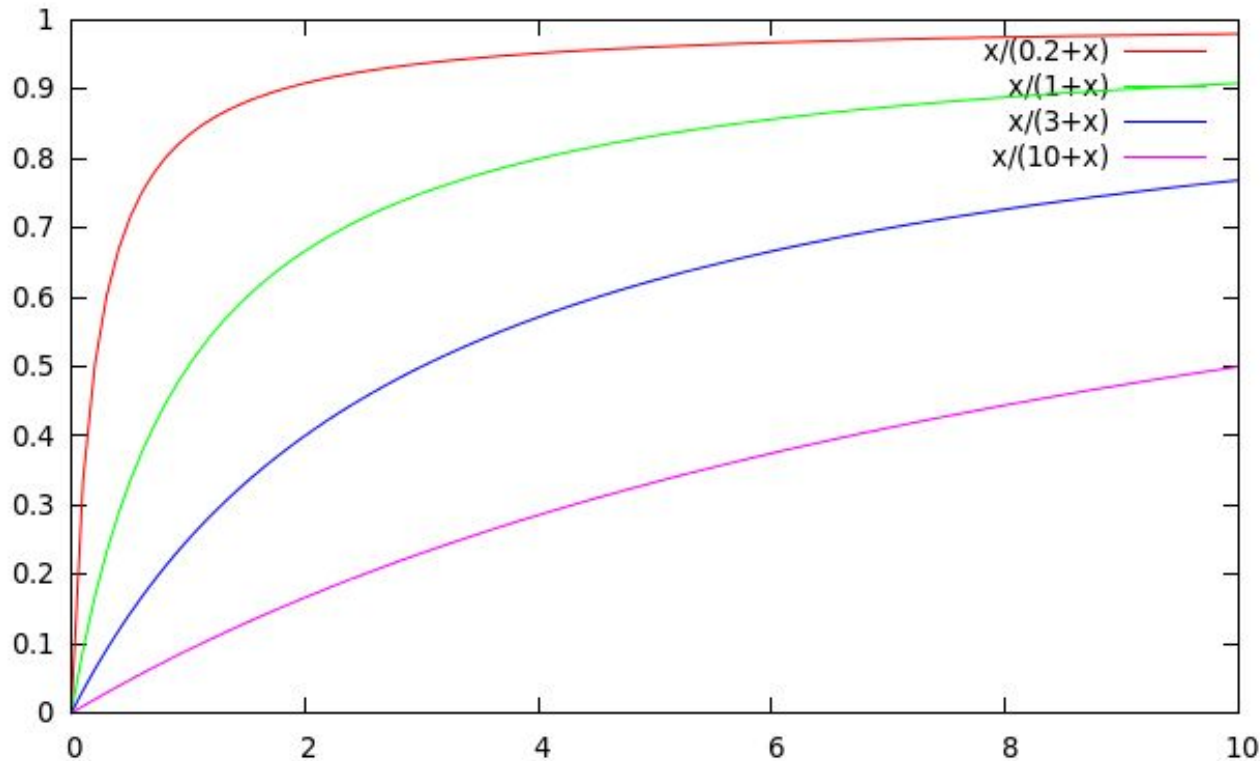
$$\log \frac{p^*(1-q)}{(1-p)^*q} \leftarrow \text{Trọng số của các đặc trưng tiêu biểu}$$

# Ước lượng hàm bão hòa

- Không có cách đơn giản để ước lượng các tham số cho mô hình 2-Poisson
- ... Vì vậy chúng ta sẽ sử dụng một đường cong tương tự có cùng các thuộc tính nhận biết

$$\frac{tf}{k_1 + tf}$$

# Hàm bão hòa



- Với giá trị  $k_1$  lớn, tăng  $tf_i$  tiếp tục đóng góp đáng kể vào điểm xếp hạng
- Tỷ lệ đóng góp giảm nhanh đáng kể với giá trị  $k_1$  nhỏ

# Các phiên bản đầu tiên của BM25

- Phiên bản 1: Sử dụng hàm bão hòa

$$c_i^{BM25v1}(tf_i) = c_i^{BIM} \frac{tf_i}{k_1 + tf_i}$$

- Phiên bản 2: Giảm lược BIM thành IDF và tăng giá trị cho thành phần tf.

$$c_i^{BM25v2}(tf_i) = \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf_i}{k_1 + tf_i}$$

- Thành phần  $k_1 + 1$  không làm thay đổi xếp hạng, nhưng khiến thành phần chứa  $tf_i$  bằng 1 khi  $tf_i = 1$

*Tương tự tf-idf, nhưng trọng số bị giới hạn.*

# Chuẩn hóa độ dài

- Các văn bản dài có xu hướng có giá trị tf<sub>i</sub> lớn
- Vì sao văn bản có thể dài hơn?
  - Quá chi tiết: Cho rằng tf<sub>i</sub> quan sát được quá lớn
  - Phạm vi rộng hơn: Cho rằng tf<sub>i</sub> quan sát được là hợp lý.
- Một tập văn bản thực có thể có cả 2 trường hợp
- ... vì vậy cần chuẩn hóa một phần

# Chuẩn hóa độ dài văn bản

- Chúng ta ký hiệu độ dài văn bản là:

$$dl = \sum_{i \in V} tf_i$$

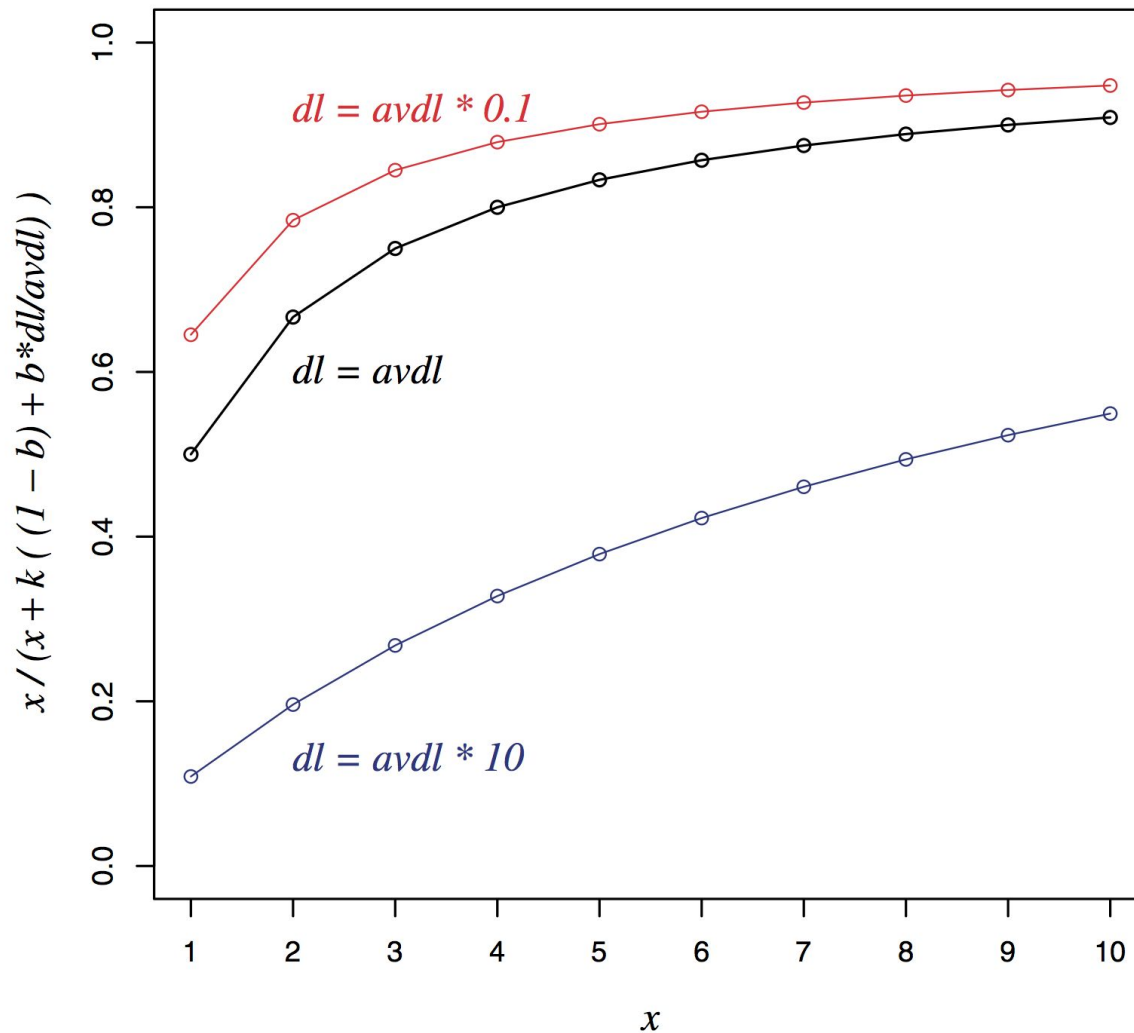
- $avdl$ : Trung bình độ dài văn bản trên toàn tập văn bản
- Thành phần chuẩn hóa độ dài

$$B = \left( (1 - b) + b \frac{dl}{avdl} \right), \quad 0 \leq b \leq 1$$

- $b = 1$  chuẩn hóa với toàn bộ độ dài văn bản
- $b = 0$  không chuẩn hóa độ dài văn bản



# Chuẩn hóa độ dài văn bản



# Okapi BM25

- Chuẩn hóa tf sử dụng độ dài văn bản

$$tf'_i = \frac{tf_i}{B}$$

$$\begin{aligned} c_i^{BM25}(tf_i) &= \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf'_i}{k_1 + tf'_i} \\ &= \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i} \end{aligned}$$

- Hàm giá trị trạng thái tìm kiếm BM25

$$RSV^{BM25} = \sum_{i \in q} c_i^{BM25}(tf_i);$$

# Okapi BM25<sub>(2)</sub>

$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

- $k_1$  Điều khiển thành phần tần suất văn bản
  - $k_1 = 0$  là mô hình nhị phân;  $k_1$  lớn là tần suất từ ở dạng thô.
- $b$  điều khiển chuẩn hóa độ dài văn bản
  - $b = 0$  là không chuẩn hóa độ dài;  $b = 1$  là tần suất tương đối (hiệu chỉnh hoàn toàn bằng độ dài văn bản)
- Thông thường,  $k_1$  được thiết lập trong khoảng 1.2-2 và  $b$  thường được thiết lập = 0.75

Tham khảo [IIR.11.4.3]

# Okapi BM25<sub>(3)</sub>

IIR. 11.4 bổ xung thêm thành phần điều khiển trọng số từ truy vấn:

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

- $k_3$  là tham số điều khiển thành phần trọng số từ truy vấn.
  - $k_3 = 0$  - mô hình nhị phân, thành phần trọng số truy vấn không làm ảnh hưởng tới kết quả xếp hạng.
  - $k_3$  rất lớn - tf ở dạng thô.
  - Một số nghiên cứu cho thấy  $k_3$  ít có ảnh hưởng đến kết quả xếp hạng (*tần suất từ truy vấn thường = 1*).

## Ví dụ 7.3. BM25 và tf-idf

- Giả sử truy vấn của bạn là [học máy]
- Giả sử bạn có 2 văn bản với số lượng từ (tf):
  - d1: học 1024; máy 1
  - d2: học 16; máy 8
- tf-idf:  $(1 + \log_2 \text{tf}) * \log_2 (N/\text{df})$ 
  - doc1:  $11 * 7 + 1 * 10 = 87$
  -
- BM25:  $k_1 = 2$ 
  - doc1:  $7 * 3 + 10 * 1 = 31$
  -

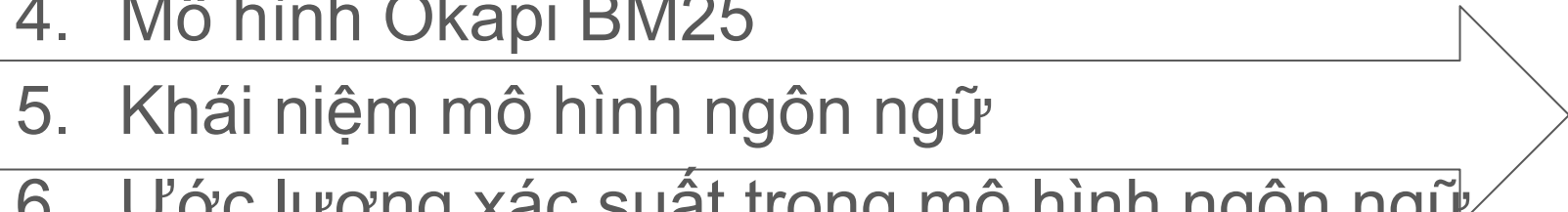
$$k_1 = ? \quad b = ?$$

## Ví dụ 7.3. BM25 và tf-idf<sub>(2)</sub>

- Giả sử truy vấn của bạn là [học máy]
- Giả sử bạn có 2 văn bản với số lượng từ (tf):
  - d1: học 1024; máy 1
  - d2: học 16; máy 8
- tf-idf:  $(1 + \log_2 \text{tf}) * \log_2 (N/\text{df})$ 
  - doc1:  $11 * 7 + 1 * 10 = 87$
  - doc2:  $5 * 7 + 4 * 10 = 75$
- BM25:  $k_1 = 2$ 
  - doc1:  $7 * 3 + 10 * 1 = 31$
  - doc2:  $7 * 2.67 + 10 * 2.4 = 42.7$

*(Sử dụng  $k_1 = 2$ ;  $b = 0$ )*

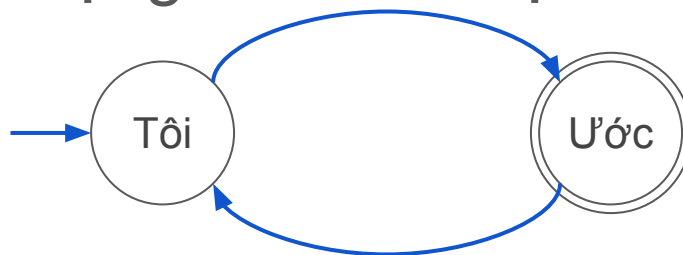
# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# Mô hình ngôn ngữ là gì?

*Mô hình ngôn ngữ/Language Model (LM)*

- Xét 1 máy trạng thái hữu hạn với như sơ đồ:

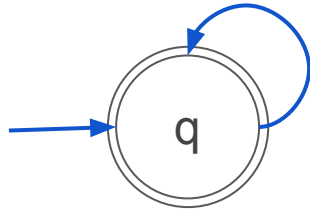


- Có thể sinh: Tôi Ước Tôi Ước ... nhưng không thể sinh: “Ước Tôi Ước” hoặc “Tôi Ước Tôi”.
- *(Các máy trạng thái tương tự chỉ có thể sinh các văn bản theo quy luật.)* Tập hợp tất cả các nội dung có thể được sinh được gọi là ngôn ngữ của máy trạng thái.

*Nếu thay từ cụ thể trong mỗi nút bằng 1 phân bố xác suất lấy ra 1 từ trong 1 bộ từ vựng thì kết quả thu được là 1 mô hình ngôn ngữ.*



## Ví dụ 7.4. Mô hình ngôn ngữ



w	P(w q)	w	P(w q)
STOP	0.2	thì	0.01
nắng	0.2	bay	0.03
cao	0.1	mưa	0.02
chuồn chuồn	0.01	thấp	0.04
		...	...

- Xét 1 máy trạng thái hữu hạn với 1 trạng thái như trong hình vẽ. Máy hoạt động theo cơ chế xác suất:
  - Ở 1 thời điểm máy có thể xuất ra 1 từ hoặc dừng hoạt động.
  - Các giá trị xác suất là xác suất có điều kiện phụ thuộc vào trạng thái  $q_1$ .
  - Phân bố xác suất được cho trong bảng: STOP là dừng hoạt động, còn lại là các từ. Tổng các giá trị xác suất = 1.
- Ví dụ, với  $d = \underline{\text{chuồn chuồn}} \text{ bay } \underline{\text{thấp}} \text{ thì } \underline{\text{mưa}} \textbf{ STOP}$   
$$P(d) = 0.01 * 0.03 * 0.04 * 0.01 * 0.02 * \mathbf{0.2} = 4.8\text{E-}10$$

# Ví dụ 7.5. Xếp hạng mô hình ngôn ngữ

Xét 2 mô hình ngôn ngữ  $M_1$  và  $M_2$  tương tự như trong ví dụ 7.4.

$M_1$

w	P(w q)	w	P(w q)
STOP	0.2	thì	0.01
nắng	0.2	bay	0.03
cao	0.1	mưa	0.02
chuồn chuồn	0.01	thấp	0.04
		...	...

$M_2$

w	P(w q)	w	P(w q)
STOP	0.2	thì	0.02
nắng	0.15	bay	0.03
cao	0.08	mưa	0.02
chuồn chuồn	0.01	thấp	0.05
		...	...

Với  $s = \underline{\text{chuồn chuồn}} \text{ bay } \underline{\text{thấp}} \text{ thì } \underline{\text{mưa}} \text{ STOP}$

$$P(s|M_1) = 0.01 * 0.03 * 0.04 * 0.01 * 0.02 * 0.2 = 4.8E-10$$

$$P(s|M_2) = 0.01 * 0.03 * 0.05 * 0.02 * 0.02 * 0.2 = 1.2E-9$$

$\Rightarrow P(s|M_1) < P(s|M_2)$  - mô hình  $M_2$  có khả năng sinh chuỗi  $s$  cao hơn mô hình  $M_1$ .

# Mô hình ngôn ngữ trong IR

*Cho truy vấn q, các văn bản được xếp hạng theo  $P(d|q)$*

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)}$$

- $P(q)$  là hằng số đối với 1 truy vấn;
- $P(d)$  là xác suất tiên nghiệm:
  - (Có thể thiết lập giá trị tiên nghiệm cho các văn bản theo giá trị chất lượng tĩnh, ví dụ PageRank, v.v..;)
  - Thường được thiết lập bằng nhau cho tất cả văn bản và có thể lược bỏ khỏi các tính toán.
  - => Xếp hạng theo  $P(q|d)$  và  $P(d|q)$  cho kết quả như nhau.

*Tiếp theo chúng ta sẽ tính  $P(q|d)$  - xác suất mô hình văn bản d sinh truy vấn q.*

# Mô hình sinh truy vấn

**Ý tưởng:** *Coi mỗi văn bản như 1 mô hình ngôn ngữ và xếp hạng các văn bản theo xác suất sinh câu truy vấn.*

**Giả thuyết cơ sở:** *Người dùng có một văn bản trong suy nghĩ và câu truy vấn được sinh theo văn bản đó.*

- $P(q|d)$  thể hiện khả năng văn bản  $d$  chính là văn bản có trong suy nghĩ của người dùng.
- Các vấn đề cần xử lý:
  - Ước lượng các tham số và thiết lập các mô hình ngôn ngữ;
  - Làm mịn để tránh giá trị 0;
  - Tính xác suất sinh truy vấn;
  - Trả về các văn bản có xác suất sinh truy vấn cao nhất.

*Có nhiều điểm tương đồng với Naive Bayes (sẽ học trong phân lớp văn bản)*

# Nội dung

1. Căn bản lý thuyết xác suất
2. Nguyên lý xếp hạng xác suất
3. Mô hình nhị phân độc lập
4. Mô hình Okapi BM25
5. Khái niệm mô hình ngôn ngữ
6. Ước lượng xác suất trong mô hình ngôn ngữ
7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
8. Các mở rộng mô hình ngôn ngữ

# Ước lượng các giá trị xác suất

- Với giả thuyết độc lập như trong BIM, chúng ta có:

$$P(q|M_d) = \infty \prod_{t \in q} P(t|M_d)$$

- Sử dụng khả năng cực đại chúng ta có:

$$P(t|M_d) = tf_{t,d}/|d|,$$

Trong đó  $|d|$  là độ dài (số lượng từ) của văn bản  $d$ .

- Vấn đề với các giá trị 0:
  - Với 1 từ truy vấn  $t$  không có trong  $d \Rightarrow P(t|M_d) = 0 \Rightarrow P(q|M_d) = 0$
  - - Vấn đề quá vừa trong học máy (*overfitting*)
  - Chúng ta cần làm mịn để tránh vấn đề với các giá trị 0.

*Hệ số đa thức =  $L_q!/(tf_{t1,q}! * tf_{t2,q}! * \dots * tf_{tM,q}!)$  là hằng số đối với 1 văn bản và có thể được bỏ qua trong xếp hạng.*

# Làm mịn

- Ký hiệu:  $M_c$  - mô hình được thiết lập từ tập văn bản  $C$ ;
  - $cf_t$  - số lần từ  $t$  xuất hiện trong  $C$ ;
  - $T$  - Tổng số từ trong  $C$ ;  $T = \sum cf_t$
- Chúng ta có thể sử dụng  $P(t|M_c)$  để làm mịn  $P(t|d)$ :
$$P(t|d) = \lambda P(t|M_d) + (1 - \lambda)P(t|M_c)$$
  - Trong trường hợp từ truy vấn không xuất hiện trong  $d$  nhưng có trong bộ dữ liệu thì  $P(t|d) = (1 - \lambda)P(t|M_c)$

*Từ truy vấn vẫn có thể không xuất hiện trong bộ dữ liệu văn bản, có cần điều chỉnh giá trị 0 trong trường hợp đó?*

# Mô hình kết hợp

$$P(q|M_d) = \propto \prod_{t \in q} (\lambda P(t|M_d) + (1 - \lambda)P(t|M_c))$$

- $\lambda$  là tham số kết hợp,  $\lambda \in [0, 1]$ :
  - $\lambda$  lớn - Thành phần mô hình văn bản có tỉ lệ lớn, hiệu ứng làm mịn yếu, giá trị 0 có ảnh hưởng lớn.
  - $\lambda$  nhỏ - Thành phần mô hình văn bản có ảnh hưởng nhỏ, hiệu ứng làm mịn mạnh, giá trị 0 có ảnh hưởng nhỏ.

$$RSV(d, q) = \propto \prod_{t \in q} (\lambda * tf_{t,d}/|d| + (1 - \lambda)cf_t/T)$$



## Ví dụ 7.6. Xếp hạng theo LM

Cho tập văn bản bao gồm 2 văn bản  $d_1$  và  $d_2$

$d_1$ : Lập trình C cơ bản, *Ngôn ngữ lập trình C*

$d_2$ : Kỹ thuật lập trình C++ và lập trình hướng đối tượng  
và truy vấn  $q$ : Lập trình C

**Yêu cầu:**

Xếp hạng  $d_1$  và  $d_2$  theo **mô hình kết hợp** với  $\lambda = 1/2$

## Ví dụ 7.6. Xếp hạng theo $LM_{(2)}$

Cho tập văn bản bao gồm 2 văn bản  $d_1$  và  $d_2$

$d_1$ : Lập trình C cơ bản, *Ngôn ngữ lập trình C*

$d_2$ : Kỹ thuật lập trình C++ và lập trình hướng đối tượng  
và truy vấn  $q$ : Lập trình C

**Yêu cầu:**


Xếp hạng  $d_1$  và  $d_2$  theo **mô hình kết hợp** với  $\lambda = 1/2$

$$P(q|d_1) = [(2/6 + 4/18)/2] * [(2/6 + 2/18)/2] \approx 0.062$$

$$P(q|d_2) = [(2/6 + 4/18)/2] * [(0/6 + 2/18)/2] \approx 0.015$$

Xếp hạng:  $d_1 > d_2$

# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# VSM (tf-idf) vs. LM

Rec.	precision		%chg	significant?
	tf-idf	LM		
0.0	0.7439	0.7590	+2.0	
0.1	0.4521	0.4910	+8.6	
0.2	0.3514	0.4045	+15.1	*
0.4	0.2093	0.2572	+22.9	*
0.6	0.1024	0.1405	+37.1	*
0.8	0.0160	0.0432	+169.6	*
1.0	0.0028	0.0050	+76.9	
11-point average	0.1868	0.2233	+19.6	*

[Ponte & Croft]

*Trong nghiên cứu này LM hoạt động tốt hơn VSM.*

# LM vs. VSM

*LM được phát triển dựa trên nền tảng XSTK, còn VSM được phát triển dựa trên nền tảng đại số tuyến tính.*

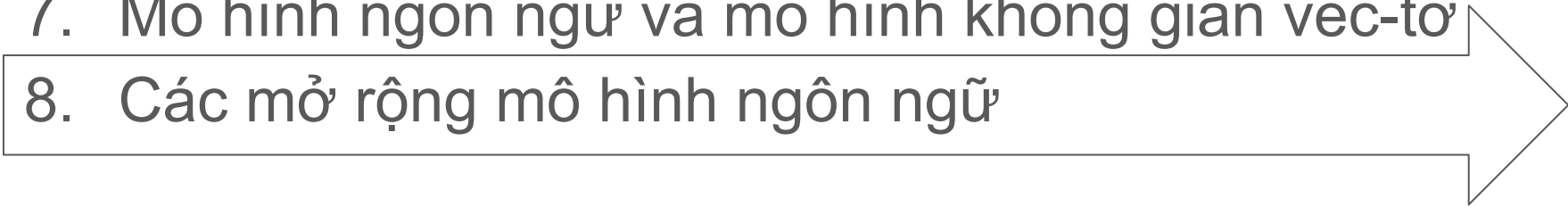
Tuy nhiên vẫn có thể nhận thấy một số điểm chung trong các công thức xếp hạng của các mô hình:

- Cả 2 mô hình đều sử dụng tần suất từ (tf);
  - Nhưng LM không hiệu chỉnh, tương đương với tần suất ở dạng thô trong VSM.
- Giá trị xác suất trong LM về hình thức giống như chuẩn hóa độ dài trong VSM
  - Trong LM các giá trị tần suất được chia cho số lượng từ.
  - Trong VSM với chuẩn hóa cosine các trọng số được chia cho độ dài vec-tơ.

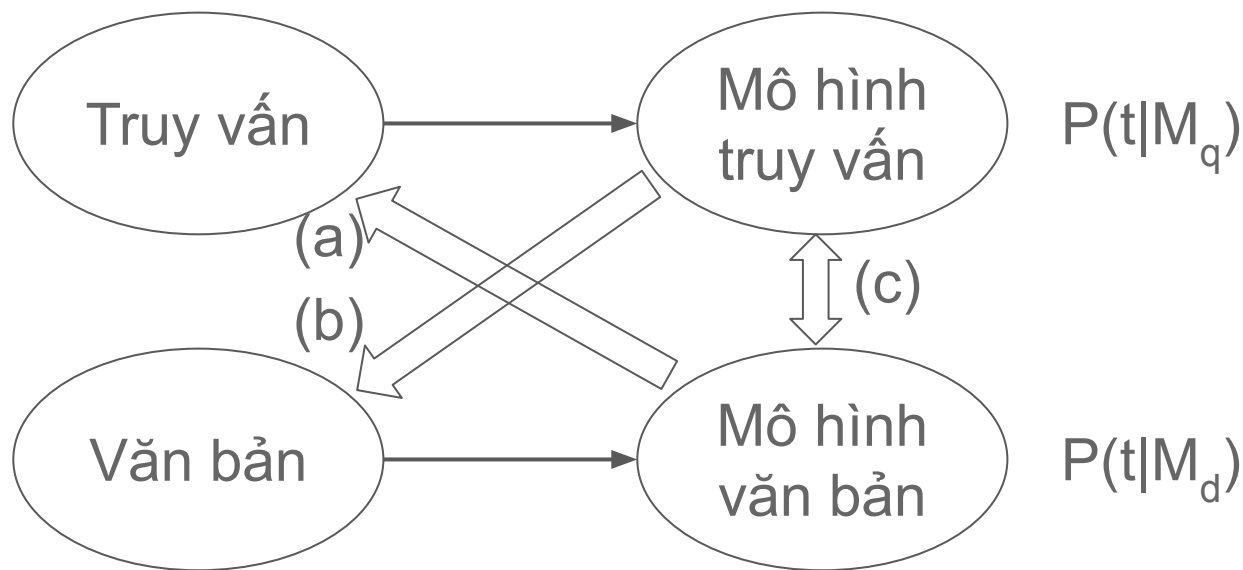
# LM vs. VSM<sub>(2)</sub>

- Kết hợp mô hình văn bản với mô hình bộ dữ liệu có hiệu ứng tương tự sử dụng idf
  - Các từ hiếm trong bộ dữ liệu, nếu xuất hiện nhiều trong văn bản sẽ có ảnh hưởng lớn đến xếp hạng
  - Các từ phổ biến trong bộ dữ liệu có giá trị xác suất lớn theo mô hình bộ dữ liệu, tuy nhiên thành phần xác suất này là bằng nhau cho tất cả các văn bản, vì vậy ít ảnh hưởng đến kết quả xếp hạng.
- Bên cạnh đó vẫn có 1 số điểm khác biệt rõ nét:
  - LM sử dụng tf và cf, còn VSM sử dụng tf và df.
  - VSM áp dụng nhiều hàm hiệu chỉnh các giá trị tần suất khác nhau và các công thức chuẩn hóa độ dài khác nhau.
  - v.v..

# Nội dung

1. Căn bản lý thuyết xác suất
  2. Nguyên lý xếp hạng xác suất
  3. Mô hình nhị phân độc lập
  4. Mô hình Okapi BM25
  5. Khái niệm mô hình ngôn ngữ
  6. Ước lượng xác suất trong mô hình ngôn ngữ
  7. Mô hình ngôn ngữ và mô hình không gian vec-tơ
  8. Các mở rộng mô hình ngôn ngữ
- 

# Các lựa chọn mô hình hóa ngôn ngữ



- a) Xác suất sinh truy vấn;
- b) Xác suất sinh văn bản;
- c) So sánh các mô hình.



# So sánh mô hình

- Ước lượng các mô hình truy vấn và văn bản sau đó so sánh
- Đại lượng phù hợp là hệ số phân rã Kullback-Leibler
$$R(d; q) = KL(M_d || M_q) = \sum_{t \in V} P(t|M_q) \log \frac{P(t|M_q)}{P(t|M_d)}$$
  - Tương đương với xác suất sinh truy vấn nếu phân bố thực nghiệm đơn giản được sử dụng cho mô hình truy vấn.
- Kết quả thu được tốt hơn so với các tiếp cận xác suất sinh truy vấn hoặc xác suất sinh văn bản.

*(Tham khảo [Zhai and Lafferty 2001])*

# Các phương pháp ước lượng xác suất

Mô hình ngôn ngữ đã tìm hiểu trong bài giảng này là mô hình đơn từ,  $P_{\text{uni}}(s) \propto P(t_1) * P(t_2) * \dots * P(t_n)$

- Một số nghiên cứu khác tìm cách biểu diễn sự phụ thuộc từ bằng các xác suất có điều kiện:

- Mô hình 2-từ (từ sau phụ thuộc vào từ trước)

$$P_{\text{bi}}(s) \propto P(t_1) * P(t_2|t_1) * \dots * P(t_n|t_{n-1})$$

- Mô hình chuỗi từ

$$P_{\text{chain}}(s) \propto P(t_1) * P(t_2|t_1) * \dots * P(t_n|t_1 t_2 \dots t_{n-1})$$

- Ví dụ, với nội dung  $s = \text{“Hôm nay là Thứ Sáu”}$

- Mô hình 1-từ:  $P_{\text{uni}}(s) \propto P(\text{hôm nay}) * P(\text{là}) * P(\text{thứ sáu})$

- Mô hình 2-từ:

$$P_{\text{bi}}(s) \propto P(\text{hôm nay}) * P(\text{là}|\text{hôm nay}) * P(\text{thứ sáu}|\text{là})$$

- Mô hình chuỗi từ:

$$P_{\text{chain}} \propto P(\text{hôm nay}) * P(\text{là}|\text{hôm nay}) * P(\text{thứ sáu}|\text{là, hôm nay})_{66}$$

# Các phương pháp làm mịn

- Có vai trò quan trọng để tránh vấn đề quá vừa và đảm bảo mô hình thu được là hữu ích trong thực tế
  - Quá vừa/overfitting: Là vấn đề mô hình được huấn luyện gắn chặt với dữ liệu huấn luyện, nhưng không có khả năng xử lý dữ liệu mới/chưa thấy.
  - *Vấn đề xác suất 0 trong LM là 1 trường hợp quá vừa.*
- Có nhiều phương pháp làm mịn khác nhau:
  - Laplace
  - Jelinek-Mercer
  - Dirichlet tiên nghiệm
  - Katz
  - Good-Turing
  - ..., và sự kết hợp của các phương pháp.

*Các lựa chọn và tùy chỉnh được thực hiện gần như hoàn toàn dựa trên thực nghiệm*

# Làm mịn 2-bước

*Chúng ta xét 1 trường hợp đơn giản:*

Truy vấn = “Các giải thuật cho khai phá dữ liệu”

$d_1$ :            0.04   0.001            0.02    0.002    0.003

$d_2$ :            0.02   0.001            0.01    0.003    0.004

$P(\text{giải thuật}|d_1) = P(\text{giải thuật}|d_2)$

$P(\text{dữ liệu}|d_1) < P(\text{dữ liệu}|d_2)$

$P(\text{khai phá}|d_1) < P(\text{khai phá}|d_2)$

Nhưng  $P(q|d_1) > P(q|d_2)$  ??

*Cần giảm sự biệt xác suất sinh các từ “các” và “cho” bởi các văn bản.*

[Zhai/Lafferty]

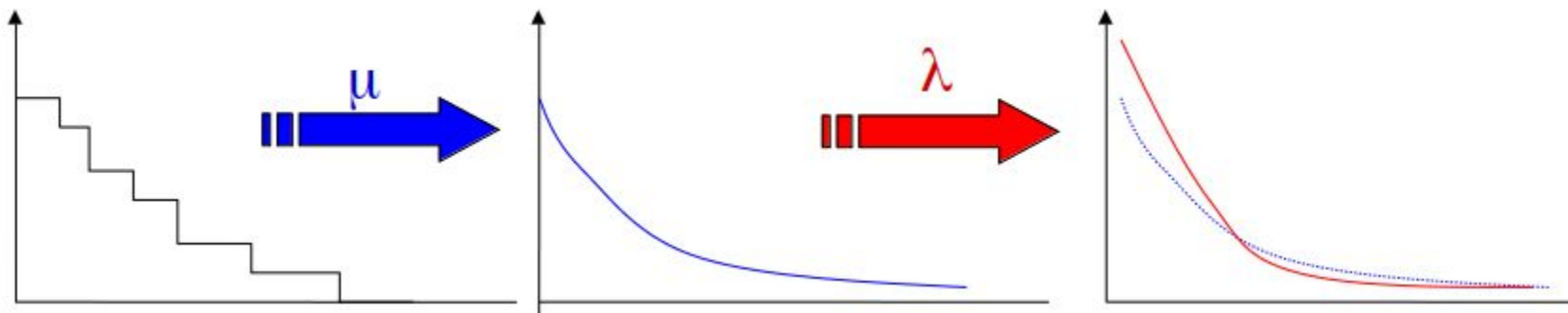
# Làm mịn 2-bước<sub>(2)</sub>

## Bước 1

- Các từ chưa thấy  
=> Tiên nghiệm Dirichlet

## Bước 2

- Nhiều trong truy vấn  
=> Jelinek-Mercer



$$P(t|d) = (1 - \lambda) \frac{tf_{t,d} + \mu P(t|C)}{|d| + \mu} + \lambda p(w|U)$$

*U: Mô hình hồ sơ người dùng, hoặc  $M_C$*

Kết hợp Dirichlet (hiệu quả với truy vấn ngắn)  
và làm mịn Jelinek-Mercer (tốt với các truy vấn chi tiết)

# Bài tập 7.1. Trọng số Okapi BM25

Cho bộ dữ liệu 6 văn bản với các đại lượng thống kê như trong bảng:

<b>d</b>	<b>Biểu diễn vec-tơ văn bản</b>										<b>dl</b>
	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

Truy vấn q: h k;

và các trọng số:  $k_1 = 1.2$ ,  $k_3 = 0$  và  $b = 0.75$ .

**Yêu cầu:** Tính RSV cho  $d_2$  và  $d_5$  theo Okapi BM25.

## Bài tập 7.2. Xếp hạng theo LM

Cho bộ dữ liệu gồm 2 văn bản  $d_1$  và  $d_2$

$d_1$ : Mô hình ngôn ngữ, xác suất sinh, xác suất phù hợp

$d_2$ : Mô hình VSM, đại số tuyến tính, đô tương đồng

Truy vấn  $q$ : Mô hình xác suất

**Yêu cầu:** Xếp hạng các văn bản theo mô hình ngôn ngữ kết hợp với  $\lambda = 0.75$

