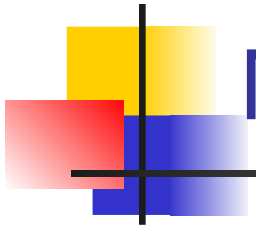
IT4853
Tìm kiếm và trình diễn thông tin

Bài 4. Mô hình không gian vec-tơ

IIR.C6. Scoring, term weighting and the vector space model

Bộ môn Hệ thống thông tin
Viện CNTT & TT



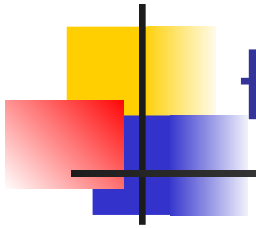
Nội dung chính

- 1. Phương pháp tìm kiếm có xếp hạng
- 2. Trọng số tf.idf
- 3. Mô hình không gian vec-tơ
- 4. Hệ thống SMART

Phương pháp tìm kiếm có xếp hạng

- Trả về những văn bản có khả năng phù hợp cao theo trật tự giảm dần khả năng phù hợp;
- Đại lượng trạng thái tìm kiếm văn bản:
  - Thể hiện khả năng văn bản phù hợp với truy vấn, càng lớn thì văn bản càng có nhiều khả năng là văn bản phù hợp;
  - Ví dụ, độ tương đồng, xác suất phù hợp v.v.
- *"Trong xếp hạng, chỉ quan trọng quan hệ thứ tự giữa các kết quả tìm kiếm, các giá trị cụ thể của đại lượng trạng thái tìm kiếm văn bản không quan trọng."*

Đại lượng trạng thái tìm kiếm: Retrieval Status Value (RSV)



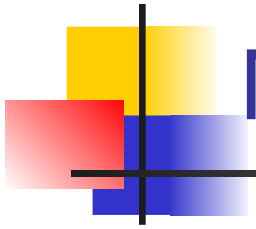
Độ tương đồng

- Đặc điểm:
  - Là giá trị số, thường được chuẩn hóa về [0, 1];
  - Thường được đánh giá trên cơ sở từ vựng:
    - Rất khó đánh giá độ tương đồng ngữ nghĩa;
    - ... Chi phí tính toán lớn, phức tạp v.v.
  - Đánh giá thường được thực hiện trên mô hình:
    - Không gian vec-tơ;
    - Mô hình sinh;
    - ... Hiếm khi sử dụng tài liệu ở nguyên dạng.

Ví dụ, đánh giá độ tương đồng bằng hệ số Jaccard

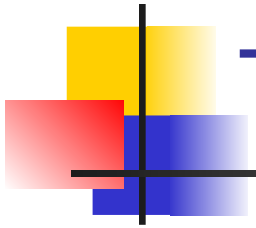
- Biểu diễn các đối tượng cần so sánh bằng các tập đặc trưng;
  - Độ tương đồng tỉ lệ với số lượng đặc trưng chung;
  - ... Từ là đặc trưng tiêu biểu của văn bản.
- Cho hai hai tập đặc trưng A và B:
  - Jaccard(A, B) = $|A \cap B| / |A \cup B|$
  - 0 <= Jaccard(A, B) <= 1
  - Jaccard(A, A) = 1
  - Jaccard(A, B) = 0 nếu A và B không có đặc trưng chung.

Cần xét đến trọng số của các đặc trưng



Nội dung chính

- 1. Phương pháp tìm kiếm có xếp hạng
- 2. Trọng số tf.idf
- 3. Mô hình không gian vec-tơ
- 4. Hệ thống SMART



Trọng số tf.idf

- Trong trường hợp tổng quát, trọng số thể hiện tầm quan trọng của từ đối với văn bản.
  - Nếu coi từ là dấu hiệu tìm kiếm văn bản, thì trọng số thể hiển khả năng phân biệt các văn bản của từ;
- Trọng số tf.idf:
  - Đồng biến với số lần từ được sử dụng trong văn bản;
  - Nghịch biến với số lượng văn bản sử dụng từ.

$$w_{tf.idf}(t, d) = w_{tf}(t,d) \times idf(t)$$

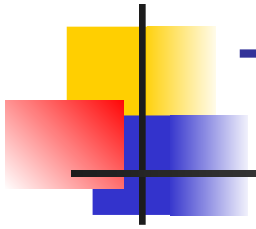
Thành phần tf

- Trọng số:

$$w_{tf}(t,d) = \begin{cases} 1 + log_{10}(tf_{t,d}), nếu\ tf_{t,d} > 0; \\ 0, \qquad\qquad\qquad nếu\ ngược\ lại \end{cases}$$

- Trong đó: $tf_{t,d}$ là tần suất từ *t* trong văn bản *d;*

Tần suất từ: Term Frequency (tf): là số lần từ *t* xuất hiện trong văn bản *d*



Thành phần idf

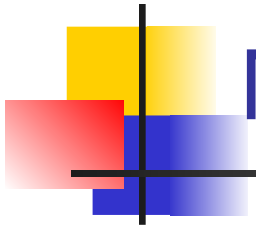
- Thành phần idf(t) được xác định như sau:

$$idf(t) = \log(N/df_t)$$

  - Trong đó N là số văn bản trong bộ dữ liệu; df_t là tần suất văn bản của từ t.

Tần suất văn bản: *document frequency (df)*: là số văn bản chứa từ;

Nghịch đảo tần suất văn bản: *inverse document frequency (idf)*: Đại lượng nghịch đảo của df



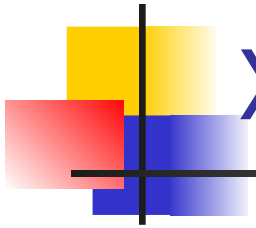
Nội dung chính

- 1. Phương pháp tìm kiếm có xếp hạng
- 2. Trọng số từ
- 3. Mô hình không gian vec-tơ
- 4. Hệ thống SMART

Biểu diễn văn bản và truy vấn

- Coi mỗi thuật ngữ trong bộ từ vựng là một trục của không gian vec-tơ;
  - Không gian M chiều, với M = |V|;
  - M có thể rất lớn
- Mỗi văn bản, truy vấn là một điểm trong không gian
  - Biểu diễn vec-tơ của văn bản và truy vấn là những vec-tơ thưa.

Ký hiệu $\vec{d}, \vec{q}$ tương ứng là biểu diễn vec-tơ của văn bản d và truy vấn q.



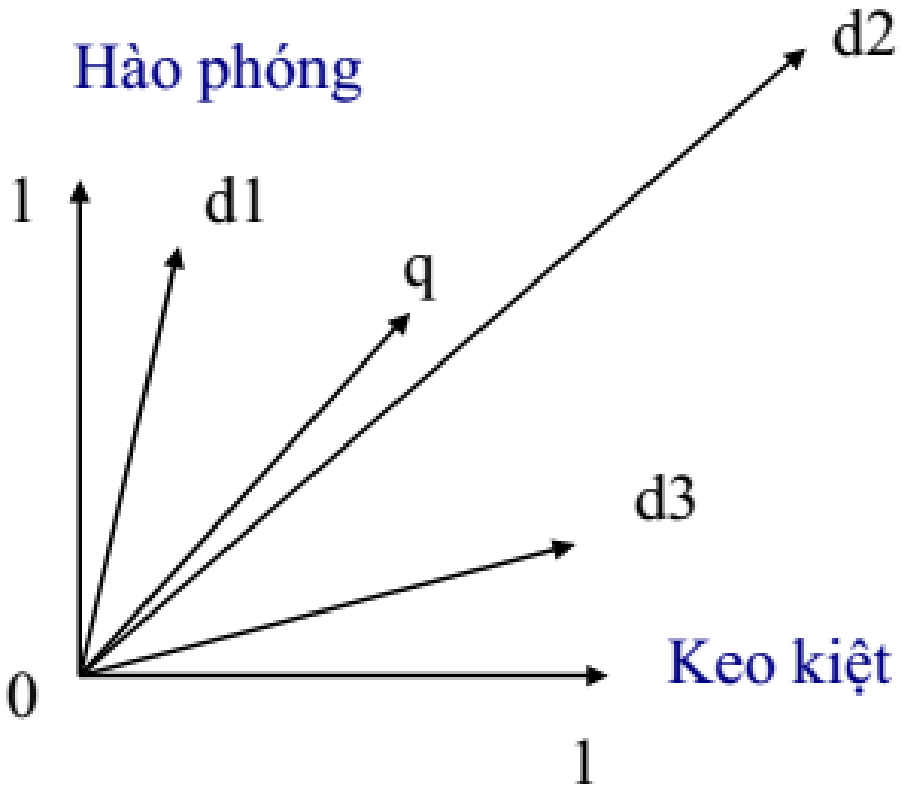
Xác định độ tương đồng

- Tương đồng là đặc tính nghịch của sự khác biệt.
- Trong không gian vec-tơ có thể đo khoảng cách Euclide giữa hai điểm biểu diễn hai văn bản bất kỳ và sử dụng giá trị khoảng cách này đại diện cho sự khác biệt giữa hai văn bản tương ứng.

Xếp hạng văn bản theo thứ tự tăng dần khoảng cách Euclide?

Thử nghiệm 1: Sử dụng khoảng cách Euclide

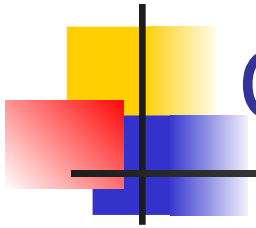
- Khoảng cách Euclide giữa biểu diễn vec-tơ của q và d_2 tương đối lớn mặc dù phân bố từ rất giống nhau



Thử nghiệm 2: Sử dụng khoảng cách góc

- Từ văn bản d thiết lập d′ bằng cách lặp lại nội dung của d
  - Về mặt nội dung thì d và d′ là tương đương. Văn bản d′ tuy dài hơn nhưng không cung cấp thông tin mới.
  - Khoảng cách Euclide giữa biểu diễn vec-tơ của d và d′ có thể rất lớn
  - Góc giữa biểu diễn vec-tơ của d và d′ bằng 0 thể hiện mức tương đồng cực đại

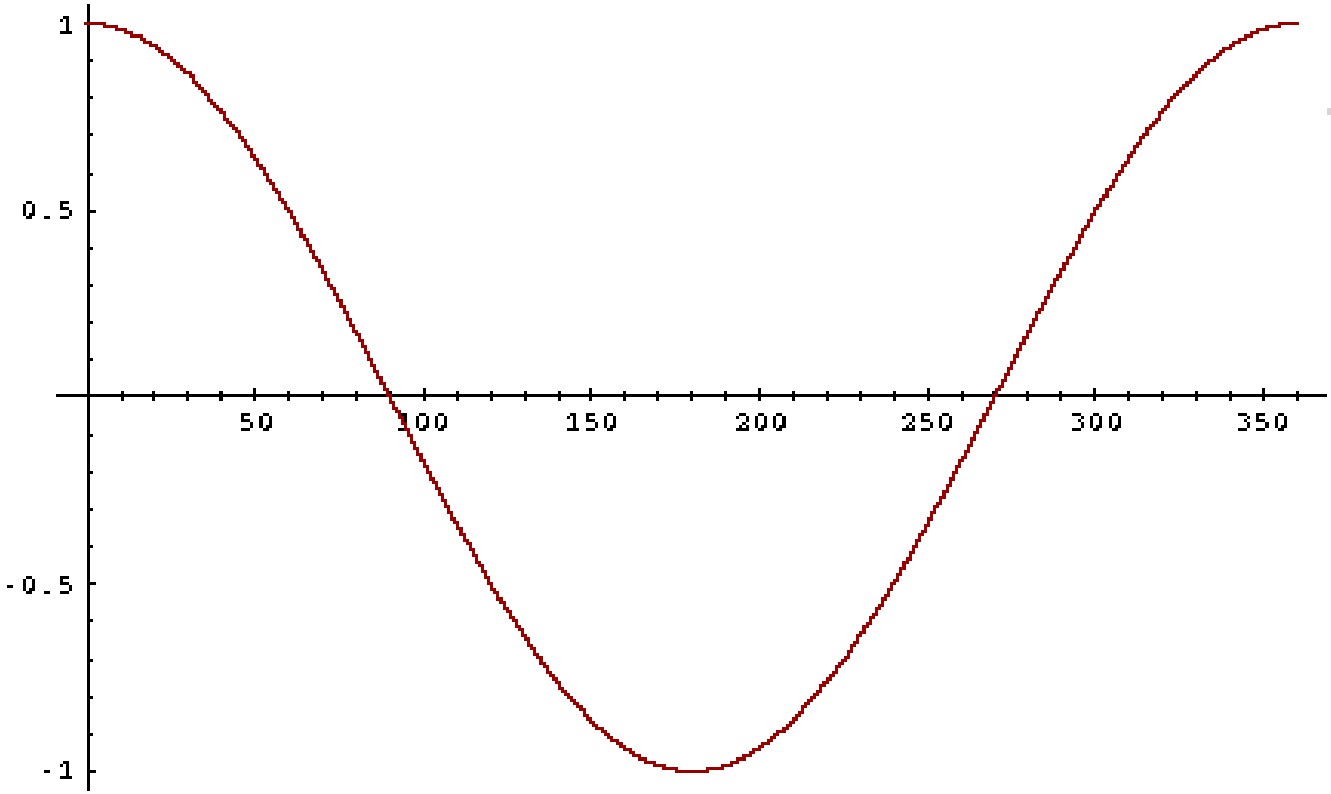
Xếp hạng văn bản theo thứ tự tăng dần của khoảng cách góc?



Cosine vs. khoảng cách góc

- Hai xếp hạng sau là tương đương
  - Xếp hạng văn bản theo thứ tự tăng dần góc giữa các biểu diễn vec-tơ của văn bản và truy vấn
  - Xếp hạng văn bản theo thứ tự giảm dần cosine góc giữa các biểu diễn vec-tơ của văn bản và truy vấn.
- Cosine là hàm đơn điệu giảm trong khoảng $[0^o, 180^o]$

Cosine vs. khoảng cách góc (2)

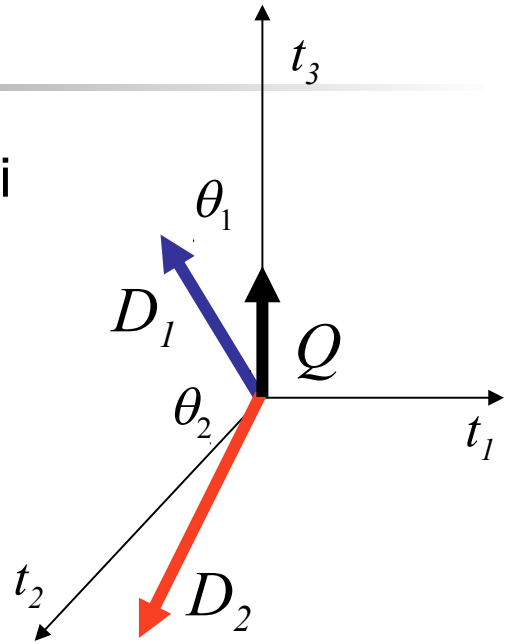


Tính cosine như thế nào? Ưu điểm sử dụng cosine so với góc là gì?

Độ tương đồng Cosine

- Độ tương đồng cosine là cosine góc giữa hai vec-tơ
  - Bằng tích vô hướng chia tích độ dài các vec-tơ

$$Sim_{\cos}(d,q) = \frac{\vec{d} \cdot \vec{q}}{\|\vec{d}\| \cdot \|\vec{q}\|} = \frac{\sum_{i=1}^{|V|} (w_{i,d} \cdot w_{i,q})}{\sqrt{\sum_{i=1}^{|V|} w_{i,d}^2} \cdot \sqrt{\sum_{i=1}^{|V|} w_{i,q}^2}}$$



$D_1 = 2T_1 + 3T_2 + 5T_3$ $\quad \text{Sim}_{\cos}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$

$D_2 = 3T_1 + 7T_2 + 1T_3$ $\quad \text{Sim}_{\cos}(D2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 phù hợp với truy vấn hơn D_2 6 lần theo độ tương đồng cosine nhưng chỉ hơn 5 lần theo tích vô hướng.

Chuẩn hóa cosine

- Chia mỗi thành phần vec-tơ cho độ dài vec-tơ, độ dài vec-tơ được xác định như sau:

$$\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$$

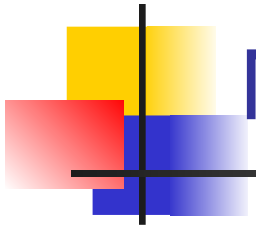
- Độ dài vec-tơ đã chuẩn hóa bằng 1, vì vậy mỗi văn bản là một điểm trên bề mặt siêu cầu có bán kính 1 đơn vị.
- Chuẩn hóa làm mờ sự khác biệt trọng số giữa các văn bản dài và ngắn

Cosine cho vec-tơ đã chuẩn hóa

- Cosine góc giữa các vec-tơ đã chuẩn hóa bằng tích vô hướng của các vec-tơ này:

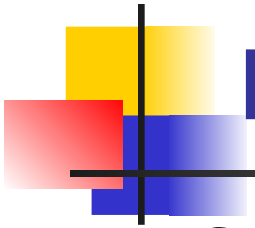
$$\cos(\vec{q},\vec{d}) = \vec{q}\cdot\vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

Với $\vec{d}$ và $\vec{q}$ là những vec-tơ đã chuẩn hóa



Nội dung chính

- 1. Phương pháp tìm kiếm có xếp hạng
- 2. Trọng số từ
- 3. Mô hình không gian vec-tơ
- 4. Hệ thống SMART



Hệ thống SMART

- SMART là một hệ thống tìm kiếm thông tin được xây dựng dựa trên lý thuyết đại số;
- Cung cấp nhiều cách đánh giá trọng số tf.idf khác nhau;
- Sử dụng phương pháp xếp hạng tương tự như mô hình không gian vec-tơ.

SMART – System for the Mechanical Analysis and Retrieval of Text

Hệ ký hiệu SMART

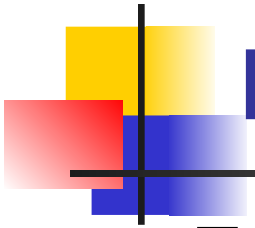
Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Trong đó:

u là số lượng từ duy nhất trong văn bản

CharLength : số ký tự trong văn bản

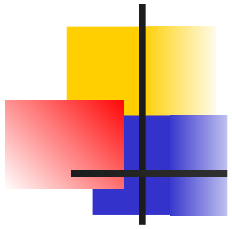
Vì sao cơ số hàm log để trống?



Phương pháp xếp hạng

- Trong hệ SMART văn bản và truy vấn có thể được biểu diễn theo những cách khác nhau;
- Một phương pháp xếp hạng được ký hiệu ngắn gọn bằng một bộ 6 ký tự theo định dạng **ddd.qqq**
- Phương pháp xếp hạng mặc định là **lnc.ltc**:
  - Đối với văn bản: Lấy log tf, không sử dụng idf, và chuẩn hóa cosine
  - Đối với truy vấn: Lấy log tf, idf, chuẩn hóa cosine
  - Xếp hạng theo tích vô hướng hai vec-tơ.

Ví dụ phương pháp lnc.ltc

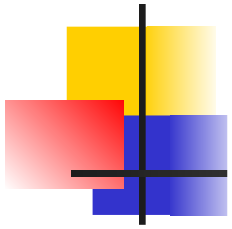


- Văn bản: *Bảo hiểm ô tô bảo hiểm xe máy*
- Truy vấn: *bảo hiểm ô tô tốt nhất*

Thuật ngữ	Truy vấn						Văn bản				Tích
	tf-raw	tf-wt	df	idf	wt	n'lize	tf-raw	tf-wt	wt	n'lize	
xe máy	0		5000	2.3			1				
tốt nhất	1		50000	1.3			0				
ô tô	1		10000	2.0			1				
bảo hiểm	1		1000	3.0			2				

Số văn bản N = ?

Ví dụ phương pháp lnc.ltc



- Văn bản: *Bảo hiểm ô tô bảo hiểm xe máy*
- Truy vấn: *bảo hiểm ô tô tốt nhất*

Thuật ngữ	Truy vấn						Văn bản				Tích
	tf-raw	tf-wt	df	idf	wt	n'lize	tf-raw	tf-wt	wt	n'lize	
xe máy	0	0	5000	2.3	0	0	1	1	1	0.52	0
tốt nhất	1	1	50000	1.3	1.3	0.34	0	0	0	0	0
ô tô	1	1	10000	2.0	2.0	0.52	1	1	1	0.52	0.27
bảo hiểm	1	1	1000	3.0	3.0	0.78	2	1.3	1.3	0.68	0.53

Độ dài văn bản $= \sqrt{1^2 + 0^2 + 1^2 + 1.3^2} \approx 1{,}92$

Độ dài truy vấn $= \sqrt{1{,}3^2 + 0^2 + 2{,}0^2 + 3.0^2} \approx 3{,}83$

$N = 10^2 * 10000 = 1000\,000$ Score = 0+0+0.27+0.53 = 0.8

Ví dụ 2, phương pháp lnc.lnc

Tần suất từ (tf)

Từ	d1	d2	d3
a	115	58	20
b	10	7	11
c	2	0	6
d	0	0	38

Trong ví dụ này, không tính idf (idf = 1).

Ví dụ 2, phương pháp lnc.lnc (2)

Log tần suất từ

Từ	d1	d2	d3
a	3,06	2,76	2,30
b	2,00	1,85	2,04
c	1,30	0	1,78
d	0	0	2,58

Sau khi chuẩn hóa

Từ	d1	d2	d3
a	0,789	0,832	0,524
b	0,515	0,555	0,465
c	0,335	0	0,405
d	0	0	0,588

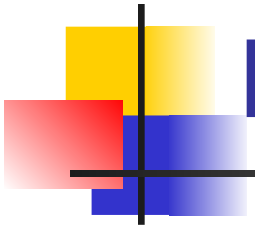
Score(d1,d2) ≈
0.789 × 0.832 + 0.515 × 0.555 + 0.335 × 0.0 + 0.0 × 0.0
≈ 0.94
Score(d1,d3) ≈ 0.79
Score(d2,d3) ≈ 0.69

Bài tập 4.1

- Khoảng cách Euclide (hoặc khoảng cách L_2) giữa hai vec-tơ được xác định như sau:

$$\left\|\vec{x}-\vec{y}\right\| = \sqrt{\sum_{i=1}^{M}\left(x_i - y_i\right)^2}$$

- Cho truy vấn q và các văn bản d_1, d_2, ... Hãy chứng minh rằng nếu biểu diễn vec-tơ $\vec{q}$ và các văn bản $\vec{d_i}$ đều được chuẩn hóa thành vec-tơ đơn vị thì kết quả xếp hạng theo thứ tự tăng dần khoảng cách Euclide giống kết quả xếp hạng theo thứ tự giảm dần mức tương đồng cosine



Bài tập 4.2

- a) Trọng số idf của từ xuất hiện trong mọi văn bản bằng bao nhiêu? So sánh ảnh hưởng của trọng số idf với thao tác lọc từ dừng?
- b) Trọng số tf-idf của một từ có thể vượt quá 1 hay không?

Bài tập 4.3

- Cho dữ liệu tf và df như sau:

tf(t, d)	Doc1	Doc2	Doc3
xe máy	27	4	24
ô tô	3	33	0
bảo hiểm	0	33	29
tốt nhất	14	0	17

df(t)	df	idf
xe máy	18 165	
ô tô	6 723	
bảo hiểm	19 241	
tốt nhất	25 235	

Cho N = 806 791:

a) Hãy tính ma trận tf.idf

b) Xếp hạng cho truy vấn "bảo hiểm ô tô tốt nhất "

(i) nnn.atc; (ii) ntc.atc