

Thứ ngày tháng năm

1) Bước 1: tiến xù lý các tài liệu. Đối với tập tài liệu trên thi cần loại bỏ stop words và sử dụng Porter Stemmer để đưa các term về dạng gốc.

d₁: football cricket

d₂: cricket termite grasshopper

d₃: football football team hockey

d₄: termite team goal

d₅: obama football hockey teams

2) Bước 2:

d₁:

d₂:

Term	chi số doc
football	1
cricket	1

Term	chi số doc
cricket	2
termite	2
grasshopper	2

Thứ ngày tháng năm

d₃

d₄

Term	Chi số doc
football	3
team	3
hockey	3

d₅

d₅

Term	Chi số da
termite	4
team	4
goal	4

Term	Chi số doc
obama	*5
football	5
hockey	5
team	5

Bước 3. Nhập các bảng trên lại và ~~gộp~~ gồm
các dòng theo term, chi số doc

Thứ ngày tháng năm

Term	Chữ số/đọc
football	1.
football	3
football	5
cricket	1.
cricket	2
termite	2
termite	4
grasshopper	2
hockey	3
hockey	5.
team	3
team	4
team	5.
goal	1
obama	5

– Bài 4: Xây dựng vocabulary và posting list

Term	Số doc	$P(td R=0)$	IDF	Chữ số doc	w
football	3	0.6	$\frac{5}{3}$	1	-0.23
cricket	2	0.4	2.5	3	-0.23
termite	2	0.4	2.5	5	-0.23
grasshopper	1	0.2	5	1	0.32
hockey	2	0.4	2.5	2	0.32
team	3	0.6	$\frac{5}{3}$	2	0.32
goal	1	0.2	5	4	0.32
obama	1	0.2	5	2	1.32
				3	0.32
				5	0.32
				3	-0.23
				4	-0.23
				5	-0.23
				4	1.32
				5	1.32

Trọng số w_{td} của mỗi term được tính theo công thức:

$$w_{td} = \log \left(0.5 * \frac{N}{N_{td}} \right)$$

$$= \log (0.5 * IDF)$$

2) q: football cricket hockey termite

- Bước 1: tiến^x xử lý q như các tài liệu và lọc ra các term:

football, cricket, hockey, termite

- Bước 2: tính trọng số của từng term:

$$w_{\text{football}} = \log(0.5 \times \frac{5}{3}) = -0.263$$

$$w_{\text{cricket}} = \log(0.5 \times 2.5) = 0.322$$

$$w_{\text{termite}} = \log(0.5 \times 2.5) = 0.322$$

$$w_{\text{grasshopper}} = \log(0.5 \times 5) = 1.322$$

$$w_{\text{hockey}} = \log(0.5 \times 2.5) = 0.322$$

$$w_{\text{team}} = \log(0.5 \times \frac{5}{3}) = -0.263$$

$$w_{\text{goal}} = \log(0.5 \times 5) = 1.322$$

$$w_{\text{obama}} = \log(0.5 \times 5) = 1.322$$

- Bước 3: độ liên quan giữa q và từng doc:

$$\text{rel}(d_1, q) = -0.263 + 0.322 = 0.059$$

$$\begin{aligned} \text{rel}(d_2, q) &= 0.322 + 0.322 + 1.322 \\ &= 1.966. \end{aligned}$$

$$\text{rel}(d_3, q) = -0.263 + (-0.263) + 0.322$$

$$\text{rel}(d_1, q) = 0.322 + (-0.263) + 1.322 \\ = 1.381$$

$$\text{rel}(d_2, q) = 1.322 + (-0.263) + 0.322 \\ = 1.118$$

- Bước 4: xếp hạng các tài liệu theo thứ
giảm dần của số liên quan:

Tài liệu	rel.
d_2	1.966
d_4	1.381
d_5	1.118
d_1	0.059
d_3	-0.204