

KTNLVB 070923

Tìm kiếm thông tin trên văn bản

B1: Phân loại câu hỏi

Nhận diện các thực thể trong câu hỏi

Mở rộng câu truy vấn

B2: Truy tìm kiếm thông tin

TF-IDF càng cao thì từ đó càng quan trọng
corpus: kho ngữ liệu

Tính điểm liên quan của D và Q

D tập từ khoá của document

Q là tập từ khoá của câu truy vấn

Các cách tìm:

- Theo biểu thức logic (ví dụ $Q = \text{Computer} \wedge \text{Network}$, thì D phải có bao gồm cả 2 từ đó)
- BM25 - Những tài liệu càng ngắn thì càng được ưu tiên

BM25: an intuitive view

$$\log \frac{P(D|R=1)}{P(D|R=0)} = \sum_w \left(\frac{d_w(1+k)}{d_w + k(1-b) + \frac{b \cdot \text{avg}(d)}{\text{avg}(w)}} \right) \cdot \log \frac{N - N_w + \frac{1}{2}}{N_w + \frac{1}{2}}$$

Repetitions of query words → good

Common words less important

More words in common with the query → good

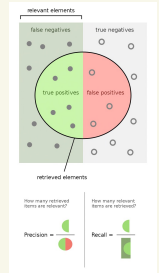
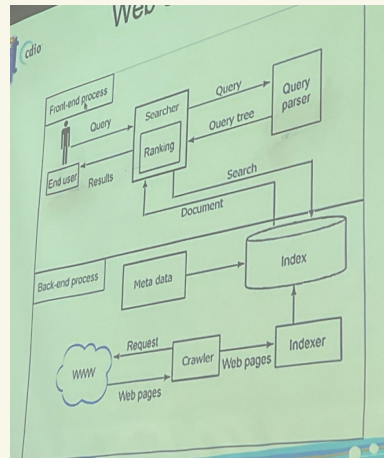
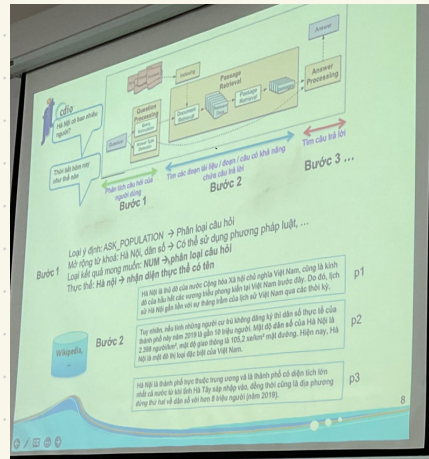
Repetitions less important than different query words

But more important if document is relatively long (wrt. average)

Phương pháp trọng số: tf*idf

- tf (term frequency):
 - tần số xuất hiện của mục từ (term) trong tài liệu
 - Độ quan trọng của từ trong tài liệu
- df (Document frequency):
 - số tài liệu chứa mục từ
 - Phân bố của từ trong kho tài liệu
 - Độ quan trọng của từ trong kho tài liệu
- idf (Inverse document frequency):
 - Mục từ có độ quan trọng hay phổ biến: idf càng lớn thì từ càng đặc trưng

$\text{weight}(t, D) = \text{tf}(t, D) * \text{idf}(t)$



Kết quả của lập chỉ mục

- Mỗi tài liệu được biểu diễn bằng một tập hợp các từ khóa có trọng số (weighted terms):
 - $D1 \rightarrow \{(t1, w1), (t2, w2), \dots\}$
 - v.d. $D1 \rightarrow \{(\text{comput}, 0.2), (\text{architect}, 0.3), \dots\}$
 - $D2 \rightarrow \{(\text{comput}, 0.1), (\text{network}, 0.5), \dots\}$
- Tập tin nghịch đảo (Inverted file):
 - Tập tin nghịch đảo được sử dụng trong truy tìm nhằm tăng hiệu năng → để tìm kiếm nhanh

Index: $\text{comput} \rightarrow \{(D1, 0.2), (D2, 0.1), \dots\}$
 $\text{architect} \rightarrow \{(D1, 0.3)\}$
 $\text{network} \rightarrow \{(D2, 0.5)\}$

Tìm tất cả các tài liệu chứa từ khóa "comput"



MAP (Mean Average Precision)

$$MAP = \frac{1}{n} \sum_{Q_i} \frac{1}{|R_i|} \sum_{D_j \in R_i} \frac{j}{r_{ij}}$$

- r_{ij} = thứ hạng rank of the j -th relevant document for Q_i
- $|R_i|$ = #rel. doc. for Q_i
- n = # test queries

Q1	Q2	j
1	4	1
5	8	2
10		3

$$MAP = \frac{1}{2} \left[\frac{1}{3} \left(\frac{1}{1} + \frac{2}{5} + \frac{3}{10} \right) + \frac{1}{2} \left(\frac{1}{4} + \frac{2}{8} \right) \right]$$

sb_Webui_controller

Photoshop Beta's Generative Fill (uncropt)