# ĐÁNH GIÁ MÔ HÌNH KHÔNG GIAN VECTOR & MÔ HÌNH LSI TRÊN TẬP DỮ LIỆU CRANFIELD

**Phan Thanh Hải**[1]
**18520705**

**Nguyễn Hoàng Long**[2]
**20520239**

**Khoa Khoa học máy tính, Trường ĐH Công nghệ Thông tin**

**CS419.N11: Truy xuất thông tin**

ThS. Nguyễn Trọng Chỉnh

30 tháng 12, năm 2022

TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN
UIT

KHOA HỌC MÁY TÍNH

# Nội dung chính

1. PHÂN TÍCH TẬP DỮ LIỆU
2. CƠ SỞ LÍ THUYẾT
3. KẾT QUẢ THÍ NGHIỆM
4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

# Chương 1. Phân tích tập dữ liệu

# 1. Phân tích tập dữ liệu

Tập tài liệu Cranfield được sử dụng trong đồ án này chứa toàn bộ danh sách tài liệu và câu truy vấn lấy từ trường Đại học Cranfield, Anh Quốc. Tập tài liệu Cranfield bao gồm các thông tin sau:

Tập dữ liệu Cranfield - Chủ đề

- Kho ngữ liệu *1 ... 1400*
- Câu truy vấn *1 ... 225*
- Kết quả truy vấn đúng

# 1. Phân tích tập dữ liệu

## 1.1. Kho ngữ liệu

Tài liệu **798** có kích thước **lớn nhất**: 655

Có 2 tài liệu **rỗng**: **471** và **995**.

**Cấu trúc** và **định dạng**:

| | |
|---|---|
| *experimental investigation of the aerodynamics of a wing in a slipstream .* | Tên |
| *an experimental study of a wing in a propeller slipstream was made in order to [...]* | Tóm tắt |

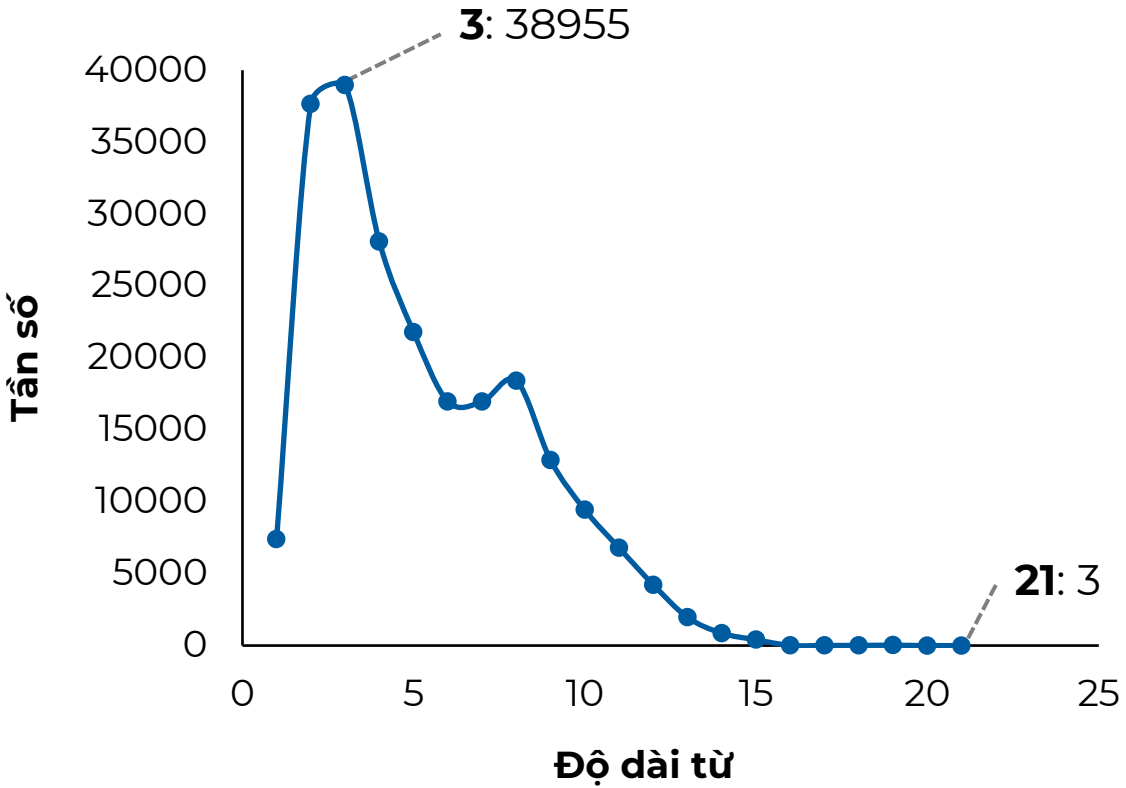
# 1. Phân tích tập dữ liệu

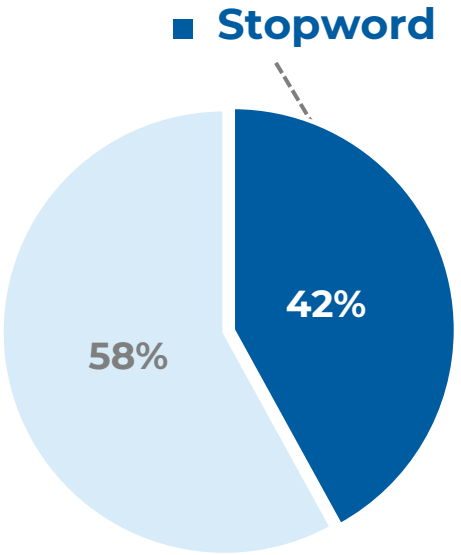
## 1.1. Kho ngữ liệu

| | |
|---|---|
| *Đơn vị đo.* | at speeds up to 25,000 **ft sec** |
| *Biến trong công thức.* | the main flow exceeding **1 1/m** (where **m** is the free stream mach number) |
| *Đánh số đối tượng.* | eqs. **(2)** and **(3)** is rather confusing [...] to table **2** |
| *Dấu nhân.* | varied from 2.35 **x** 10 to 2.99 **x** 10 |

# 1. Phân tích tập dữ liệu

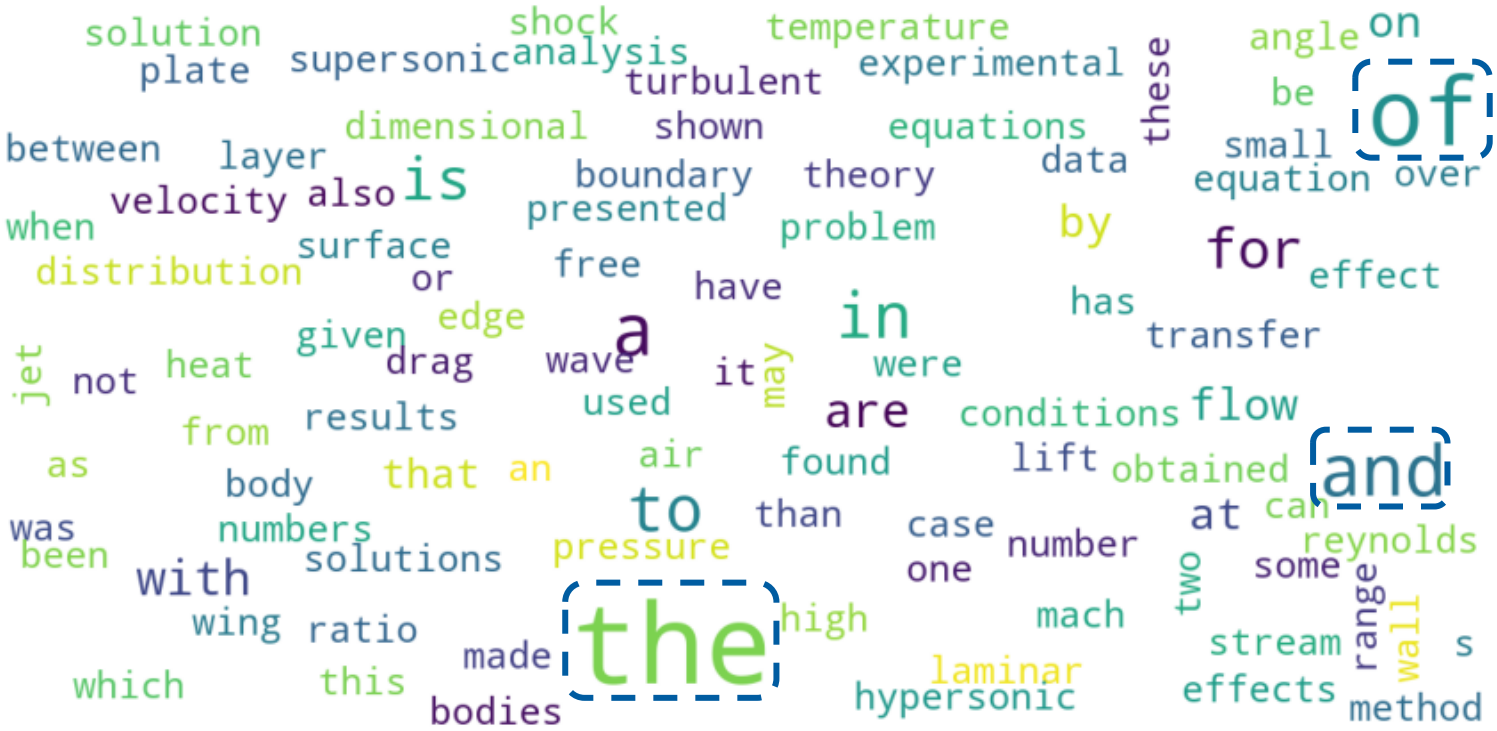
## 1.1. Kho ngữ liệu



# 1. Phân tích tập dữ liệu

## 1.1. Kho ngữ liệu




Word cloud

# 1. Phân tích tập dữ liệu

## 1.2. Câu truy vấn

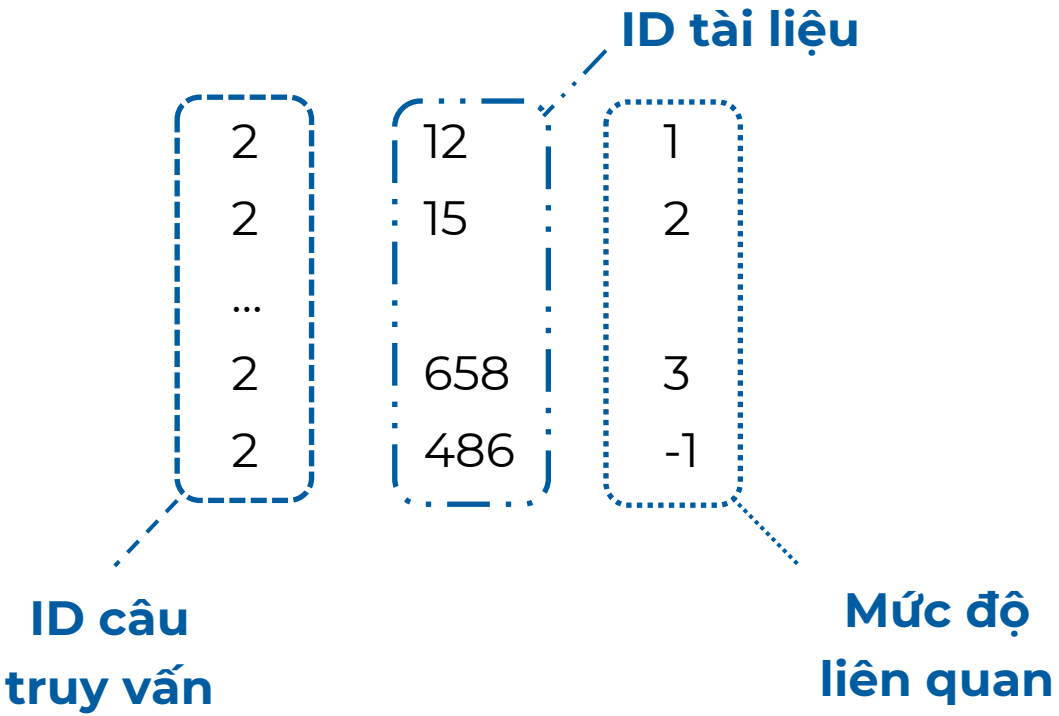
Mỗi câu truy vấn trên một dòng.

Dạng câu truy vấn:

| | |
|---|---|
| *Câu hỏi.* | **is** it possible to predict when and how it [...] |
| *Cụm danh từ.* | material **properties** of photoelastic [...] |
| *Cụm động từ.* | **find** a calculation procedure applicable to [...] |

# 1. Phân tích tập dữ liệu

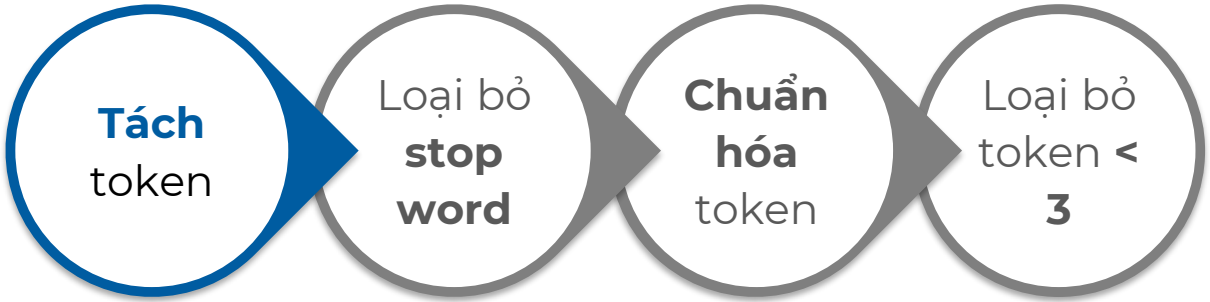
## 1.3. Kết quả truy vấn đúng



# Chương 2. Xác định term của tài liệu


**Term** *là đơn vị cơ bản của tìm kiếm.*
**Tài liệu** *là tập hợp các* **term**.

# 2. Xác định term của tài liệu




## 2.1. Tách token

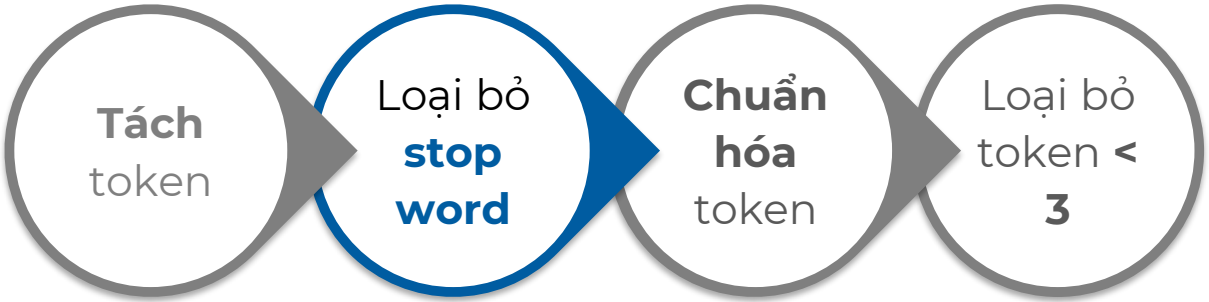
**Tách** tài liệu thành các đơn vị nhỏ hơn **token** + **loại bỏ** kí tự nhất định

Input: extended up to 400,000 btu slug, corresponding to

Output: | extended | up | to | btu | slug | corresponding | to |

Kết quả: **7052** token

# 2. Xác định term của tài liệu




## 2.2. Loại bỏ stopword

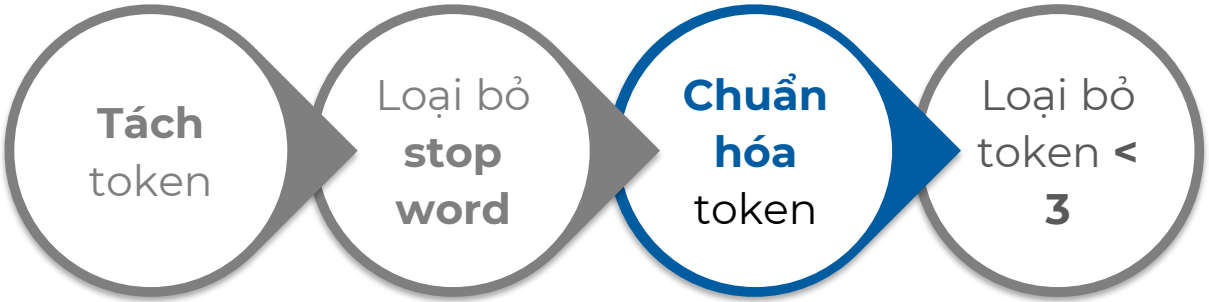
**Ít giá trị** khi chọn tài liệu phù hợp với nhu cầu của người dùng.

**Giảm** số lượng postings lưu trữ.

Đa số trường hợp **không** gây ảnh hưởng tới kết quả tìm kiếm.

Dùng danh sách stopword trong **NLTK**.

# 2. Xác định term của tài liệu




## 2.3. Chuẩn hóa token

Chuẩn hóa cùng một token về định dạng nguyên bản của nó.

*Ví dụ.* experiment – experiments **cùng nghĩa** | **token khác nhau**.

# 2. Xác định term của tài liệu

## 2.3. Chuẩn hóa token

**Stemming.**

Sử dụng **heuristic** & **lược bỏ** kí tự cuối.

*study* vs *studies*

Thời gian xử lí nhanh.

Giảm số lượng terms nhiều hơn.

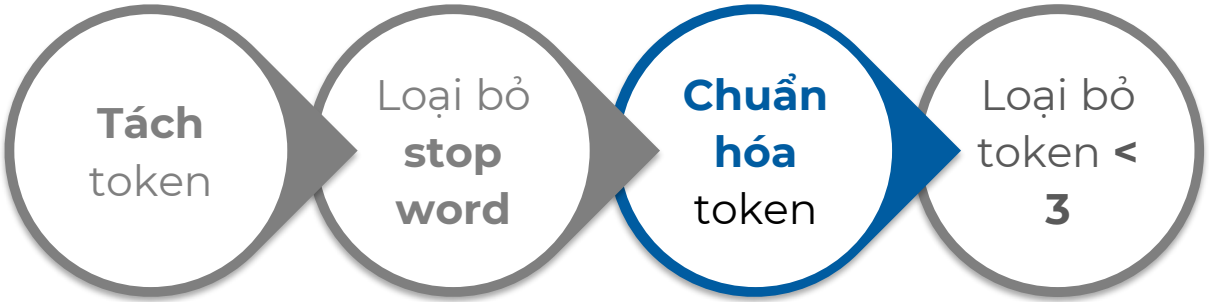
Tạo ra những từ không có thực.

**Lemmatization.**

Xem xét đến **ngữ nghĩa** và **cách dùng** của từ rồi mới lược bỏ/đưa về dạng gốc.

*good* vs *well*

# 2. Xác định term của tài liệu

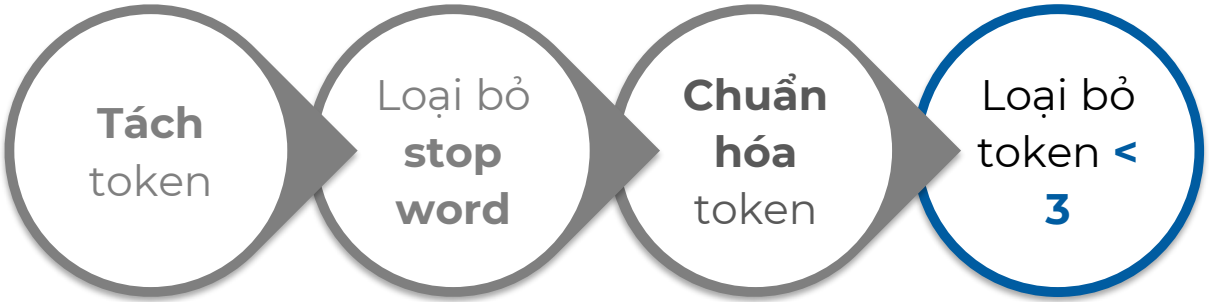



## 2.3. Chuẩn hóa token

Dùng **Snowball Stemmer** trong **NLTK** – **Porter Stemmer** cải tiến.

| | |
|---|---|
| **Cranfield (tài liệu 3)** | the boundary layer in simple shear flow past a flat plate . the boundary layer equations are presented for steady incompressible flow with no pressure gradient . |
| **Kết quả xử lí** | boundari layer simpl shear flow past flat plate boundari layer equat present steadi incompress flow pressur gradient |

# 2. Xác định term của tài liệu




2.4. Loại bỏ token < 3

**Loại bỏ** những token có độ dài < 3 kí tự.

Kết quả cuối cùng: **4147**/7052 (~ 58%).

# 2. Xác định term của tài liệu

03

Chương 3.

# LẬP CHỈ MỤC.

*Chỉ mục là **cấu trúc dữ liệu** chuyên biệt để **tối ưu hóa tốc độ** thực hiện truy vấn.*

# 3. Lập chỉ mục

## 3.1. Mô hình không gian vector

Dùng **chỉ mục đảo ngược**:

- Với mỗi term *t* lưu trữ **tài liệu** có chứa *t*.
- Tài liệu được biểu diễn bằng 1 con số (bắt đầu từ 1).

**'experiment'** → [1, 11, 12, 16, 17, 191, 25, 29]

Danh sách tài liệu có chứa *t*: **posting list**.

Posting list cần được **sắp xếp**.

- **Q:** Cách **lưu trữ** chỉ mục đảo ngược **hiệu quả**?
- **Q:** Cách **tra cứu nhanh** phần tử?

# 3. Lập chỉ mục

## 3.1. Mô hình không gian vector

Lưu trữ hiệu quả: **bảng băm** và **cây**.

**Bảng băm.**

- Python: **dictionary** ~ bảng băm.
- Ưu điểm: tốc độ truy xuất **nhanh**, độ phức tạp **O(1)**.
- Biểu diễn: `inverted_index = {`

```
<term1>: {     <chỉ số tài liệu 1>: <thông tin term>,
               <chỉ số tài liệu 2>: <thông tin term>,
        … }
… }
```

# 3. Lập chỉ mục

## 3.2. Mô hình chỉ mục ngữ nghĩa ngầm

Dùng **ma trận tài liệu**:

- Hàng tương ứng với term.
- Cột tương ứng với tài liệu.
- Phần tử ***A[i, j]*** tương ứng với mức độ quan trọng của term: trọng số TF-IDF. (dựa trên thông tin ở mục **3.1.**)

# Chương 4.
# Mô hình truy xuất thông tin

- Mô hình **không gian vector** *(Vector Space Model)*
- Mô hình **chỉ mục ngữ nghĩa ngầm** *(Latent Semantic Indexing)*

# 4. Mô hình truy xuất thông tin

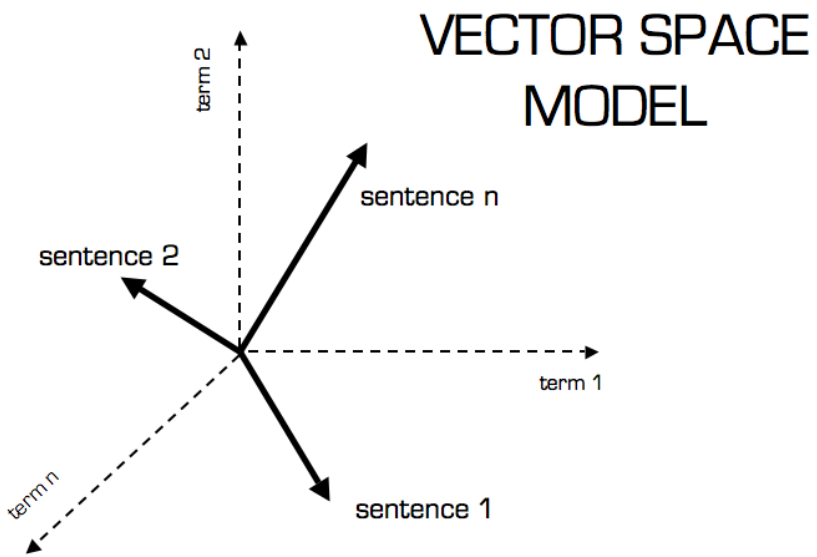
**(D, Q, F, R)**

Trong đó:

- **D:** Cách biểu diễn văn bản
- **Q:** Cách biểu diễn truy vấn
- **F:** Nền tảng lí thuyết (toán học) tương thích với **D** và **Q**, giữ vai trò cơ sở để thực hiện các suy diễn xếp hạng
- **R(D, Q):** Hàm xếp hạng, là hàm định lượng mức độ phù hợp giữa văn bản và truy vấn

# 4. Mô hình truy xuất thông tin

## 4.1. Mô hình không gian vector




**D:** Văn bản ~ 1 vector thưa, phần tử ~ mức độ quan trọng của term

**Q:** Như **D**

**F:** Lí thuyết toán học về **không gian vector**

**R(D, Q):** Hàm xếp hạng dựa trên độ tương đồng/giống nhau giữa 2 vector trong không gian:

- Độ đo **khoảng cách Euclide**?
- Độ đo **góc**/độ đo **cosin**?

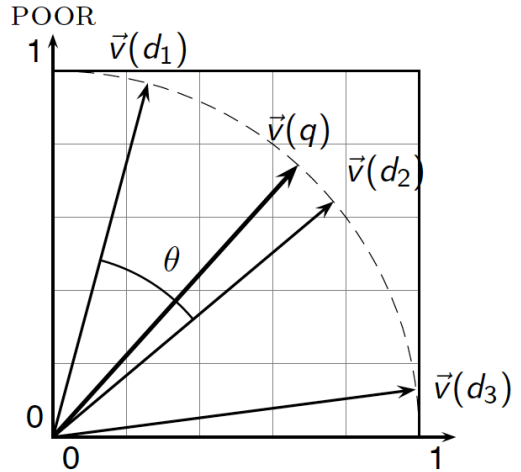
# 4. Mô hình truy xuất thông tin

## 4.1. Mô hình không gian vector

**Độ đo tương đồng**: **cosin** góc giữa 2 vector.

$$cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

$$= \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$




**Độ đo tương đồng** không ảnh hưởng bởi chiều dài vector.

# 4. Mô hình truy xuất thông tin

## 4.1. Mô hình không gian vector

**VECTOR SPACE MODEL** (*inverted_index*)

1. ∀ *term* **in** *inverted_index*:
2. Tính $idf(t,d) = \log \frac{N}{1 + \{d \in D, t \in D\}}$
3. ∀ *doc* **in** *inverted_index*: // $f_{t,d}$ = *inverted_index[term][doc][freq]*
4. Tính $tf(t,d) = \log\left(1 + f_{t,d}\right)$
5. Tính $tf\text{-}idf(t,d) = tf(t,d) \cdot idf(t,d)$
6. *inverted_index*[*term*][*doc*][*weight*] = $tf\text{-}idf(t,d)$

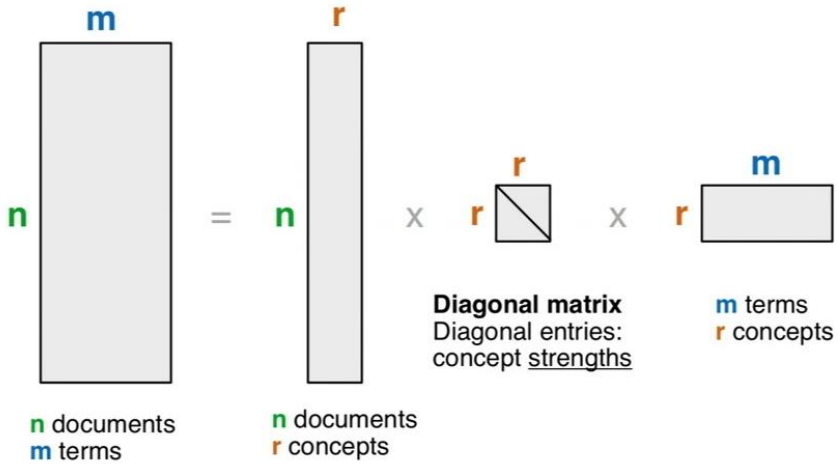
# 4. Mô hình truy xuất thông tin

## 4.1. Mô hình không gian vector

```
CosineScore(q)
 1  float Scores[N] = 0
 2  float Length[N]
 3  for each query term t
 4  do calculate w_{t,q} and fetch postings list for t
 5      for each pair(d, tf_{t,d}) in postings list
 6      do Scores[d] += w_{t,d} × w_{t,q}
 7  Read the array Length
 8  for each d
 9  do Scores[d] = Scores[d]/Length[d]
10  return Top K components of Scores[]
```

# 4. Mô hình truy xuất thông tin

## 4.2. Mô hình chỉ mục ngữ nghĩa ngầm




**D:** Ma trận term-doc

**Q:** Ma trận 1 chiều

**F:** Lí thuyết toán học về **ma trận**

**R(D, Q):** Sử dụng độ đo cosin.

# 4. Mô hình truy xuất thông tin

## 4.2. Mô hình chỉ mục ngữ nghĩa ngầm

**LATENT SEMANTIC INDEXING**(*matrix*)

1. Vector hoá tập tài liệu có bằng tần số của term trong tài liệu đó.
2. Chuyển vị ma trận vector tài liệu
3. Phân tích ma trận tài liệu thành ba ma trận theo phương pháp SVD

$$A = USV^T$$

4. Vector hóa câu truy vấn
5. Tính độ liên quan giữa tài liệu và câu truy vấn
6. Lấy ra những tài liệu liên quan nhất

# Chương 5. Kết quả thử nghiệm

# 5. Kết quả thử nghiệm

**Bảng 1.** Kết quả đánh giá mô hình

| Mô hình | Vector Space Model | Latent Semantic Indexing |
|---|---|---|
| **Precision** | **0.16** | 0.09 |
| **Recall** | **0.51** | 0.22 |
| **F1-score** | **0.24** | 0.12 |

# - HẾT -

GIẢI ĐÁP THẮC MẮC
HỎI XOÁY ĐÁP XOAY