1. **Giới thiệu**

Hệ thống gợi ý bác sĩ giúp bệnh nhân tìm được bác sĩ phù hợp dựa trên lịch sử khám bệnh trước đó. Phương pháp được sử dụng trong hệ thống này là lọc nội dung (content-based filtering), dựa trên các đặc điểm của bác sĩ và lịch sử khám bệnh của bệnh nhân.

- Hạn chế của phương pháp này là không gợi ý được đặc trưng mới của bác sĩ

VD: 1 bác sĩ có các đặc trưng mới không giống đặc trưng của bác sĩ nào trong lịch sử khám bệnh thì khi tính cosin kết quả sẽ bằng 0 và sẽ không được gợi ý

1. **Tổng quan**

**2.1 Tổng quan về phương pháp lọc nội dung**

Phương pháp lọc nội dung so sánh các đặc điểm của bác sĩ (như chuyên khoa, giá khám, cơ sở y tế, tỉnh thành) với các thông tin từ lịch sử khám bệnh của bệnh nhân. Dựa trên sự tương đồng này, hệ thống sẽ gợi ý các bác sĩ phù hợp.

**2.2 Tổng quan về công thức**

**2.2.1 Tính toán TF-IDF**

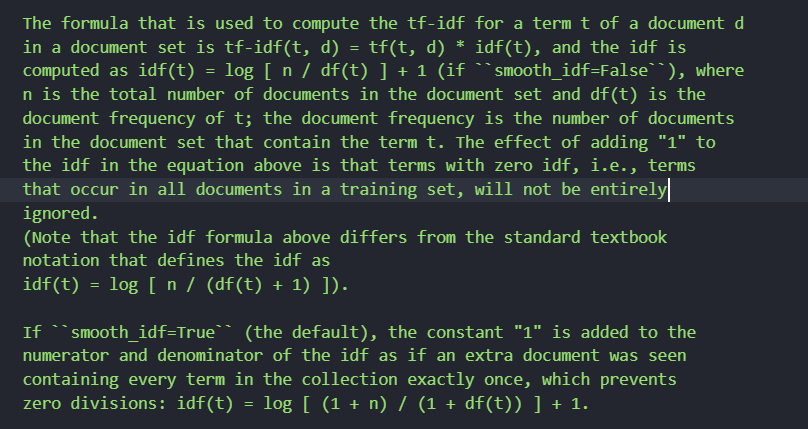
TF-IDF (Term Frequency-Inverse Document Frequency) là một giá trị đo lường tầm quan trọng của một từ trong một văn bản đối với một tập hợp các văn bản.

**Công thức TF-IDF:**

TF-IDF(t,d,D)=TF(t,d)×IDF(t,D)

TF(t, d) = ( số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong văn bản d)

IDF(t, D) = log( Tổng số văn bản trong tập mẫu D + 1/ Số văn bản có chứa từ t + 1) + 1



Tần số từ (TF): Số lần xuất hiện của từ trong một văn bản. Tần số xuất hiện càng cao, TF càng lớn.

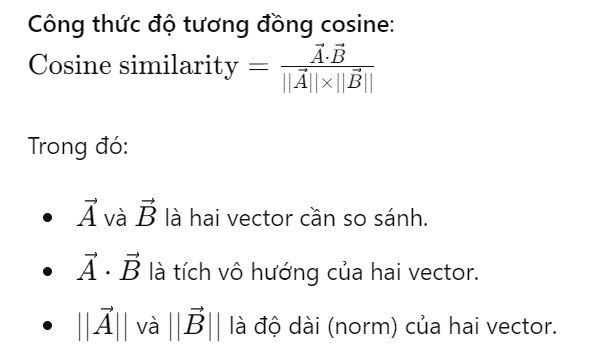
Tần số nghịch đảo văn bản (IDF): Giá trị nghịch đảo của tần số xuất hiện của từ trong toàn bộ tập văn bản. Nếu một từ xuất hiện trong nhiều văn bản, giá trị IDF của nó sẽ thấp, và ngược lại.

**Ý nghĩa của TF-IDF**:

Giá trị TF-IDF càng lớn cho thấy từ đó xuất hiện nhiều trong văn bản hiện tại và ít xuất hiện trong các văn bản khác, làm cho từ đó đặc biệt quan trọng trong văn bản hiện tại.

**2.2.2 Tính toán độ tương đồng cosine**

Độ tương đồng cosine đo lường mức độ tương đồng giữa hai vector



**Ý nghĩa của độ tương đồng cosine:**

* Giá trị cosine gần với 1: Hai vector rất giống nhau (góc giữa chúng nhỏ).
* Giá trị cosine gần với 0: Hai vector không giống nhau (góc giữa chúng gần như vuông góc).
* Giá trị cosine gần với -1: Hai vector đối nhau (góc giữa chúng lớn).

**Kết luận hai công thức:**

* TF-IDF càng lớn: Cho thấy từ đó đặc biệt quan trọng trong văn bản hiện tại, và ít xuất hiện trong các văn bản khác. Điều này có nghĩa từ đó mang nhiều thông tin đặc trưng cho văn bản đó.
* Góc cosine càng lớn (giá trị cosine similarity càng nhỏ): Cho thấy hai vector (hay hai văn bản) ít tương đồng với nhau. Nếu góc cosine bằng 0, điều đó có nghĩa là hai văn bản hoàn toàn không giống nhau. Nếu giá trị cosine gần với -1, điều đó có nghĩa là hai văn bản đối lập nhau.Ngược lại, góc cosine càng nhỏ (giá trị cosine similarity càng lớn), điều đó có nghĩa là hai văn bản càng giống nhau. Giá trị cosine similarity bằng 1 cho thấy hai văn bản hoàn toàn giống nhau về mặt nội dung.

3. Quy trình thực hiện

3.1. Tiền xử lý dữ liệu

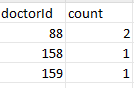
Thông tin bác sĩ: Chứa các đặc điểm như chuyên khoa, giá khám, cơ sở y tế, tỉnh thành.

Lịch sử khám bệnh của bệnh nhân: Chứa thông tin về các bác sĩ mà bệnh nhân đã khám và số lần khám bác sĩ.

3.2. Tính toán TF-IDF

- Số lần khám mỗi bác sĩ

**Table1:**



- TF-IDF của bác sĩ trong lịch sử khám bệnh của bệnh nhân

**Table2:**

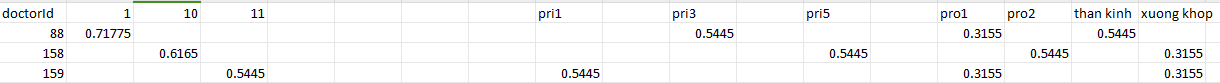
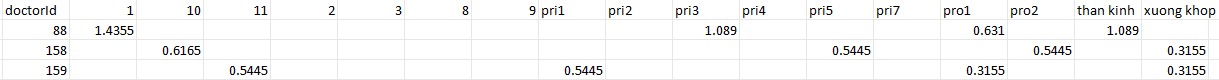


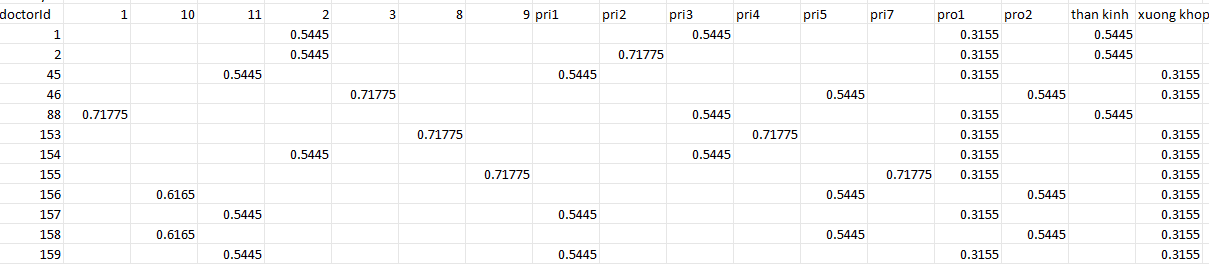
Table1 X Table2



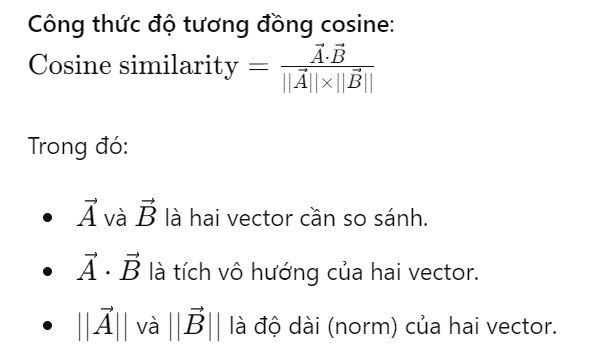
- Chuẩn hóa các đặc trưng của bác sĩ



-TF-IDF của bác sĩ trong database



3.3. Tính toán độ tương đồng cosine

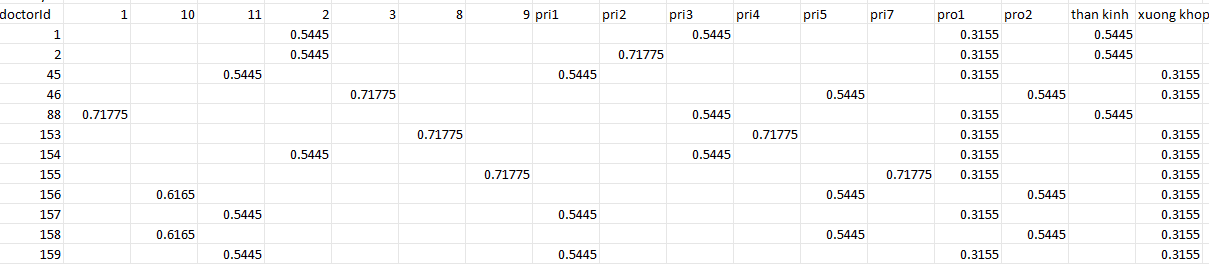


- Tính cosine của chuẩn hóa các đặc trưng của bác sĩ với TF-IDF của bác sĩ trong database

- Chuẩn hóa các đặc trưng của bác sĩ

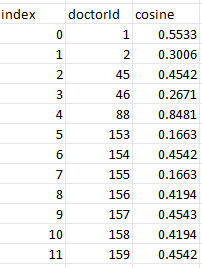


-TF-IDF của bác sĩ trong database



- Kết quả:

+ tính tay chưa sắp xếp



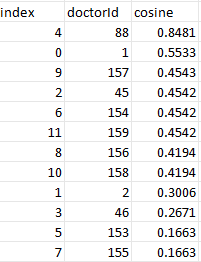
+ dùng thư viện tính chưa sắp xếp

[0.53614334 0.29439221 0.50963022 0.27599275 0.81668691 0.17556312

0.45431422 0.17556312 0.43053581 0.50963022 0.43053581 0.50963022]

-Sắp xếp độ tương đồng giảm dần

- Kết quả tính tay

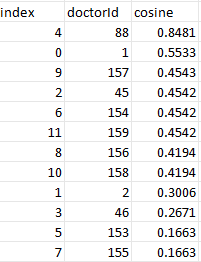


- Kết quả dùng thư viện

[(4, 0.8166869113958024), (0, 0.5361433409977497), (2, 0.5096302208617777), (9, 0.5096302208617777), (11, 0.5096302208617777), (6, 0.45431422148231526), (8, 0.43053580750125886), (10, 0.43053580750125886), (1, 0.29439220787757714), (3, 0.2759927498985814), (5, 0.17556312074751923), (7, 0.17556312074751923)]

[4, 0, 2, 9, 11, 6, 8, 10, 1, 3, 5, 7]

--> các chỉ số trong mảng là chỉ số index tương ứng với doctorId trong kết quả tính tay



- Ví dụ tính cosin với hai bác sĩ doctorId1 và doctorId2

- Tính tích vô hướng 2 vector

A\*doctorId1 = (0.27225\*0.5445)+(0.3155\*0.2366)+(0.5445\*0.27225) = 0.37112755

- Độ dài vector A

||A|| = Math.sqrt(0.3577\*\*2+0.1541\*\*2+0.1361\*\*2+0.1361\*\*2+0.2722\*\*2+0.1361\*\*2+0.2366\*\*2+0.1361\*\*2+0.2722\*\*2+0.1577\*\*2) = 0.6744

- Độ dài vector doctorId1

||doctorId1|| = Math.sqrt(0.5445\*\*2+0.5445\*\*2+0.3155\*\*2+0.5445\*\*2) = 0.9944

- Tích độ dài 2 vector

||A|| \* ||doctorId1|| = 0.6744 \* 0.9944 = 0.6706

- Độ tương đồng cosin của vector A và vector doctorId1

A\*doctorId1 / ||A|| \* ||doctorId1|| = 0.3711 / 0.6706 = 0.5533

A\*doctorId2 = (0.3155\*0.2366)+(0.5445\*0.27225) = 0.2228

||doctorId2|| = Math.sqrt(0.5445\*\*2+0.7177\*\*2+0.3155\*\*2+0.5445\*\*2) = 1.0989

||A|| \* ||doctorId2|| = 0.6744 \* 1.0989 = 0.7410

A\*doctorId2 / ||A|| \* ||doctorId2|| = 0.2228 / 0.7410 = 0.3006