

BÁO CÁO GIỮA KỲ
NHẬP MÔN NGÔN NGỮ HỌC THỐNG KÊ & ỨNG DỤNG

CHATBOT RAG VỀ CHƯƠNG TRÌNH ĐÀO TẠO FITHCMUS CHƯƠNG TRÌNH CHUẨN

GVHD: Lê Thanh Tùng
Nhóm 5

THÀNH VIÊN NHÓM

| MSSV | Họ và Tên |
|----------|----------------------|
| 20120145 | Đường Yến Ngọc |
| 21120198 | Nguyễn Thị Lan Anh |
| 21120417 | Nguyễn Thị Ngọc Châm |
| 21120426 | Huỳnh Phát Đạt |



NỘI DUNG

I. Giới thiệu đồ án

II. Giới thiệu tập dữ liệu

III. Kiến trúc hệ thống



I. GIỚI THIỆU ĐỒ ÁN

NỘI DUNG ĐỒ ÁN

- **Mục tiêu:** Xây dựng một hệ thống chatbot RAG có thể truy xuất thông tin chương trình đào tạo và trả lời các câu hỏi bằng ngôn ngữ tự nhiên.

NỘI DUNG ĐỒ ÁN

- **Output:** “Bạn cần tích lũy tối thiểu 16 tín chỉ cho các học phần bắt buộc chuyên ngành”

- **Input:** “Tôi cần tích lũy bao nhiêu tín chỉ bắt buộc chuyên ngành?”

PHƯƠNG PHÁP TIẾP CẬN

- Retrieval-based (Dựa trên truy xuất thông tin)
- Generation-based (Dựa trên mô hình sinh câu trả lời)
- Retrieval-Augmented Generation (RAG)



II. GIỚI THIỆU TẬP DỮ LIỆU

NGUỒN DỮ LIỆU

- **Dữ liệu được thu thập và tổng hợp từ các nguồn:**
 - **Đề cương môn học:** Trang web khoa FITHCMUS > Đào tạo > Chương trình đào tạo > Chương trình Chuẩn (<https://www.fit.hcmus.edu.vn/vn/Default.aspx?tabid=36>)
 - **Chương trình đào tạo:** Trang web khoa FITHCMUS > Hệ thống sinh viên > CQuy > Chương trình đào tạo > CQuy Khóa tuyển 2023 (<https://www.fit.hcmus.edu.vn/vn/Default.aspx?tabid=289>)
 - **Câu hỏi và trả lời mẫu:** Trang Q&A FIT (<https://courses.fit.hcmus.edu.vn/q2a/>)

THU THẬP DỮ LIỆU

7. NỘI DUNG CHƯƠNG TRÌNH ĐÀO TẠO

7.1. KIẾN THỨC GIÁO DỤC ĐẠI CƯƠNG

Tích lũy tổng cộng 56 tín chỉ (không kể Ngoại ngữ, Giáo dục thể chất và Giáo dục quốc phòng – an ninh):

7.1.1. Lý luận chính trị – Pháp luật

| STT | MÃ HỌC PHẦN | TÊN HỌC PHẦN | SỐ TC | SỐ TIẾT | | | Loại học phần | Ghi chú |
|-----------|-------------|--------------------------------|-------|-----------|-----------|---------|---------------|---------|
| | | | | Lý thuyết | Thực hành | Bài tập | | |
| 1 | BAA00101 | Triết học Mác – Lênin | 3 | 45 | 0 | 0 | BB | |
| 2 | BAA00102 | Kinh tế chính trị Mác – Lênin | 2 | 30 | 0 | 0 | BB | |
| 3 | BAA00103 | Chủ nghĩa xã hội khoa học | 2 | 30 | 0 | 0 | BB | |
| 4 | BAA00104 | Lịch sử Đảng Cộng sản Việt Nam | 2 | 30 | 0 | 0 | BB | |
| 5 | BAA00003 | Tư tưởng Hồ Chí Minh | 2 | 30 | 0 | 0 | BB | |
| 6 | BAA00004 | Pháp luật đại cương | 3 | 45 | 0 | 0 | BB | |
| TỔNG CỘNG | | | 14 | | | | | |

7.1.2. Khoa học xã hội – Kinh tế – Kỹ năng

| STT | MÃ HỌC PHẦN | TÊN HỌC PHẦN | SỐ TC | SỐ TIẾT | | | Loại học phần | Ghi chú |
|-----|---|---------------------------|-------|-----------|-----------|---------|---------------|---------|
| | | | | Lý thuyết | Thực hành | Bài tập | | |
| 1 | Chọn 01 học phần (02 tín chỉ) trong các học phần sau: | | | | | | | |
| | BAA00005 | Kinh tế đại cương | 2 | 30 | 0 | 0 | TC | |
| | BAA00006 | Tâm lý đại cương | 2 | 30 | 0 | 0 | TC | |
| | BAA00007 | Phương pháp luận sáng tạo | 2 | 30 | 0 | 0 | TC | |




fit@hcmus
 TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM
KHOA CÔNG NGHỆ THÔNG TIN

ĐĂNG NHẬP
 Webmail | Trang môn học | Liên hệ
 Trang chủ | Góp ý | English


TRANG CHỦ


TIN TỨC


GIỚI THIỆU


NGHIÊN CỨU


ĐÀO TẠO

Lĩnh vực đào tạo

Khoa học máy tính
 Kỹ thuật phần mềm
 Hệ thống thông tin
 Công nghệ thông tin
 Trí tuệ nhân tạo

Kiến thức giáo dục đại cương (bắt buộc)

1. BAA00003 - Tư tưởng Hồ Chí Minh
2. BAA00004 - Pháp luật đại cương
3. BAA00011 - Anh văn 1
4. BAA00012 - Anh văn 2
5. BAA00013 - Anh văn 3
6. BAA00014 - Anh văn 4
7. BAA00021 - Thể dục 1
8. BAA00022 - Thể dục 2
9. BAA00101 - Triết học Mác - Lênin
10. BAA00102 - Kinh tế chính trị Mác - Lênin
11. BAA00103 - Chủ nghĩa xã hội khoa học
12. BAA00104 - Lịch sử Đảng Cộng sản Việt Nam
13. CSC00004 - Nhập môn công nghệ thông tin
14. MTH00021 - Vi tích phân 1
15. MTH00022 - Vi tích phân 2
16. MTH00035 - Đại số tuyến tính
17. MTH00044 - Xác suất thống kê
18. MTH00045 - Toán rời rạc
19. MTH00050 - Toán học tổ hợp

Hình 1. Dữ liệu về chương trình đào tạo được tổng hợp từ trang web chính thức FITHCMUS

THU THẬP DỮ LIỆU

| Các qui định | | |
|--|----------|----------------------|
| Tiêu đề | Category | Modified Date |
| QĐ 2380 sửa đổi bổ sung QĐ 1625 K2022 | | 17/09/2024 Tải Xuống |
| Quy định về việc gia hạn thu học phí QĐ 2165/QĐ-KHTN, 26/10/2023 | | 10/11/2023 Tải Xuống |
| Tiêu chuẩn đăng ký KLTN áp dụng cho Khóa 2011-2014 | | 11/10/2022 Tải Xuống |
| Hướng dẫn cách thức nộp KLTN- Cử nhân Tài năng K2015 | CNTN | 29/09/2021 Tải Xuống |
| Quy định chuẩn đầu ra ngoại ngữ CT Cử nhân tài năng Khóa 2023 về sau | CNTN | 11/03/2024 Tải Xuống |
| DHCQ-Quy chế đào tạo 2021 | QCHV | 11/10/2022 Tải Xuống |
| QĐ 1175-DHCQ-Quy chế đào tạo 2021 | QCHV | 11/10/2022 Tải Xuống |
| Quy chế đào tạo HCTC 2016 | QCHV | 11/10/2022 Tải Xuống |
| QĐ 534 - sửa đổi chuẩn TA và trình độ TA đối với ĐHCQ_ngày 15062020 | QĐ TA | 11/10/2022 Tải Xuống |
| QĐ về xét chuẩn trình độ ngoại ngữ đối với chứng chỉ VNU-EPT - 04/2022 | QĐ TA | 11/10/2022 Tải Xuống |

Hình 2. Các qui định, hướng dẫn chung

THU THẬP DỮ LIỆU

The screenshot shows the Q&A FIT website interface. The top navigation bar includes a menu icon, a list of categories (Danh sách câu hỏi, Mới!, Chưa trả lời, Chuyên mục, Thành viên, Tạo câu hỏi, Đăng nhập bằng tài khoản Google), and a search icon. Below the navigation bar, there's a section titled "Những câu hỏi gần đây" (Recent questions). This section displays a list of five questions, each with a title, a date, a score, and a view count. The questions are:











- Kỳ 2 có thường mở lớp Phát triển ứng dụng web/ mobile/ game nâng cao không?** (Asked ngày 28 tháng 8 năm 2023 in KTPM bởi Trung Nguyên Hồ (120 điểm), 246 đã xem)
- hỏi về các môn tốt nghiệp** (Asked ngày 21 tháng 8 năm 2023 in Giáo vụ đại học bởi Minh Chánh Cái (120 điểm), 268 đã xem)
- Dạ cho em hỏi muốn du học đối với ngành cntt thì cần những điều kiện gì ạ** (Asked ngày 19 tháng 8 năm 2023 in Học phí, học bổng bởi Thắng Nguyễn 1 (120 điểm), 283 đã xem)
- Xét chuyên ngành** (Asked ngày 29 tháng 7 năm 2023 in Giáo vụ đại học bởi Toan NGuyen (120 điểm), 286 đã xem)
- Về cách tính điểm xét chuyên ngành** (Asked ngày 24 tháng 7 năm 2023 in Chương trình chuẩn bởi Kitka Tula (140 điểm), 359 đã xem)
- Có cần trả hết các môn còn nợ hay không?** (Asked ngày 22 tháng 7 năm 2023 in Chương trình chuẩn bởi Thanh Lộc Ngò (120 điểm), 359 đã xem)

The screenshot shows the Q&A FIT website interface, displaying a detailed view of a question and its answer. The top navigation bar is the same as the previous screenshot. The main content area shows a question titled "Dạ cho em hỏi là điều kiện để xét chuyên ngành của nhóm ngành máy tính và công nghệ thông tin cuối năm 2 là gì vậy ạ!". The question is dated ngày 23 tháng 10 năm 2023, in the category "Cố vấn học tập" by Vinh Nguyễn (120 điểm), and has 326 views. Below the question, there are buttons for "trả lời" (answer) and "bình luận" (comment). The answer section is highlighted in green and shows one answer titled "1 Câu trả lời". The answer is dated ngày 18 tháng 1 bởi Giáo Vụ (3.6k điểm) and contains the text: "Chào em, điều kiện đối với sinh viên để được xét chuyên ngành cuối học kỳ 4 của khóa học là sinh viên **đang học** có tổng số tín chỉ tích lũy đạt **ít nhất 50 tín chỉ** (sau 4 học kỳ chính)." Below the answer, there is a button for "bình luận" (comment).

Hình 3. Thu thập câu hỏi - câu trả lời từ web Q&A FIT

THU THẬP DỮ LIỆU

< Data 10 items

| Name | Last modified | File size |
|---|---------------|-----------|
|  BSMS.txt | Dec 4, 2024 | 5 KB |
|  CNTT.txt | Dec 14, 2024 | 47 KB |
|  DKTN.txt | Dec 14, 2024 | 9 KB |
|  HTTT.txt | Dec 14, 2024 | 32 KB |
|  KHMT.txt | Dec 14, 2024 | 67 KB |
|  KTPM.txt | Dec 14, 2024 | 32 KB |
|  MonHoc.txt | Dec 4, 2024 | 474 KB |
|  NN.txt | Dec 14, 2024 | 13 KB |
|  QAdata.xlsx | Dec 14, 2024 | 9 KB |
|  TTNT.txt | Dec 14, 2024 | 33 KB |

Chương trình: Đại học Chính quy

Ngành: Hệ thống thông tin

Khóa tuyển: 2023

(Ban hành kèm theo Quyết định số 1712/QĐ - KHTN ngày 07/9/2023 của Hiệu trưởng Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM)

1. Thông tin chung về chương trình đào tạo:

1.1. Tên ngành:

- Tiếng việt: Hệ thống thông tin

- Tiếng Anh: Information Systems

1.2. Mã ngành: 7480104

1.3. Trình độ đào tạo: Đại học

1.4. Tên chương trình: Cử nhân Hệ thống thông tin

1.5. Loại hình đào tạo: Chính quy

1.6. Thời gian đào tạo: 4 năm

1.7. Tên văn bằng sau khi tốt nghiệp:

- Tiếng Việt: Cử nhân Hệ thống thông tin

- Tiếng Anh: Bachelor of Science in Information Systems

1.8. Ngôn ngữ giảng dạy: tiếng Việt

1.9. Nơi đào tạo:

- Cơ sở 1: 227 Nguyễn Văn Cừ, P4, Q5, Tp. HCM

- Cơ sở 2: Phường Linh Trung, Thành phố Thủ Đức, Tp. HCM

Hình 4. Dữ liệu được chuyển sang dạng text

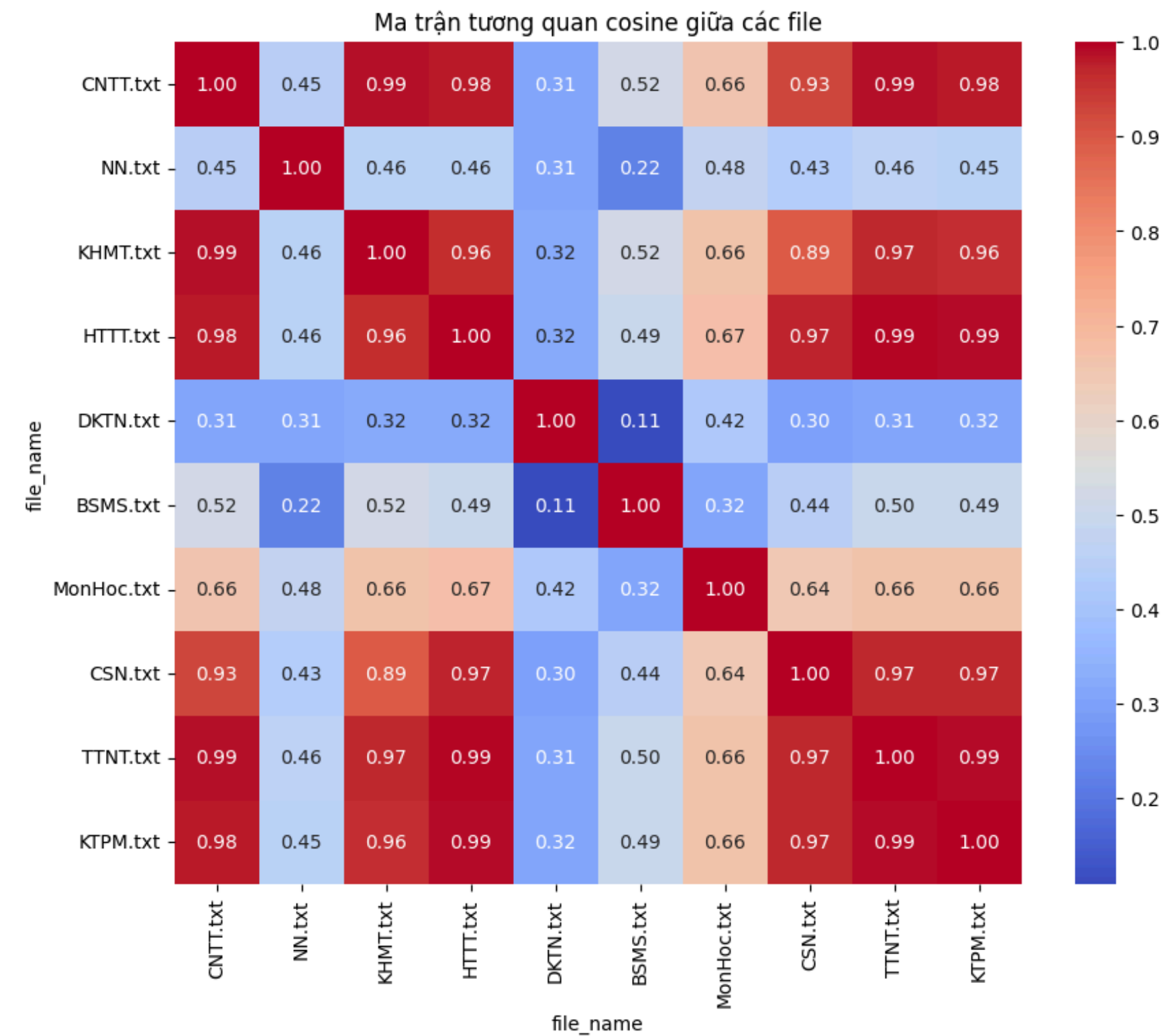
THÔNG TIN DỮ LIỆU

| Tên file | Nội dung |
|----------|--|
| HTTT.txt | - Chương trình đào tạo Ngành Hệ thống thông tin |
| KHMT.txt | - Chương trình đào tạo Ngành Khoa học máy tính |
| CNTT.txt | - Chương trình đào tạo Ngành Công nghệ thông tin |
| CNPM.txt | - Chương trình đào tạo Ngành Công nghệ phần mềm |
| TTNT.txt | - Chương trình đào tạo Ngành Trí tuệ nhân tạo |

THÔNG TIN DỮ LIỆU

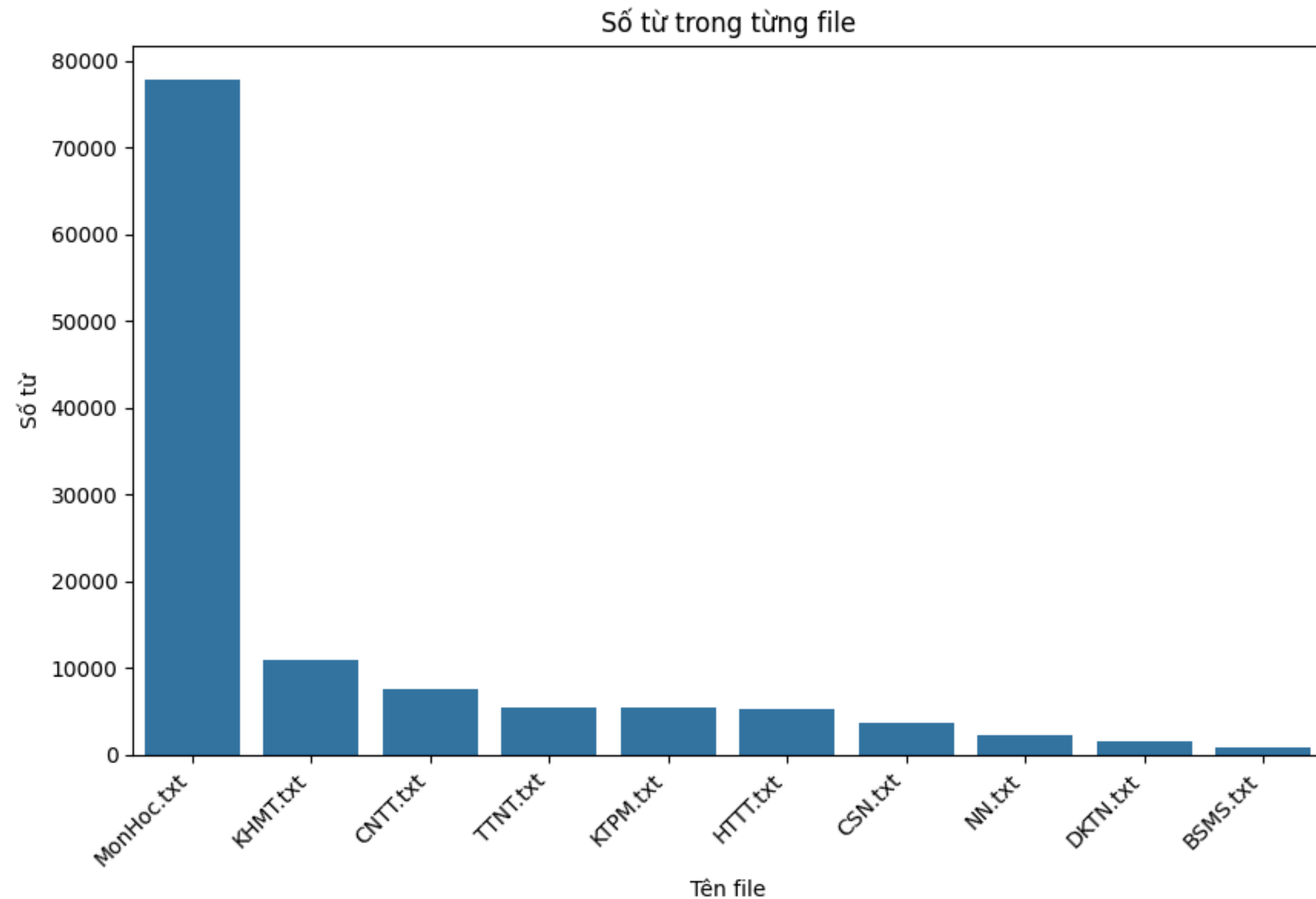
| Tên file | Nội dung |
|-------------|---|
| BSMS.txt | - Danh sách môn học liên thông Đại học - Thạc sĩ |
| DKTN.txt | - Điều kiện và quy trình thực hiện đề tài tốt nghiệp |
| MonHoc.txt | - Đề cương môn học các ngành/chuyên ngành |
| NN.txt | - Quy định về Chuẩn đầu ra ngoại ngữ |
| QAdata.xlsx | - Các cặp câu hỏi - câu trả lời về chương trình đào tạo |

KHÁM PHÁ TẬP DỮ LIỆU



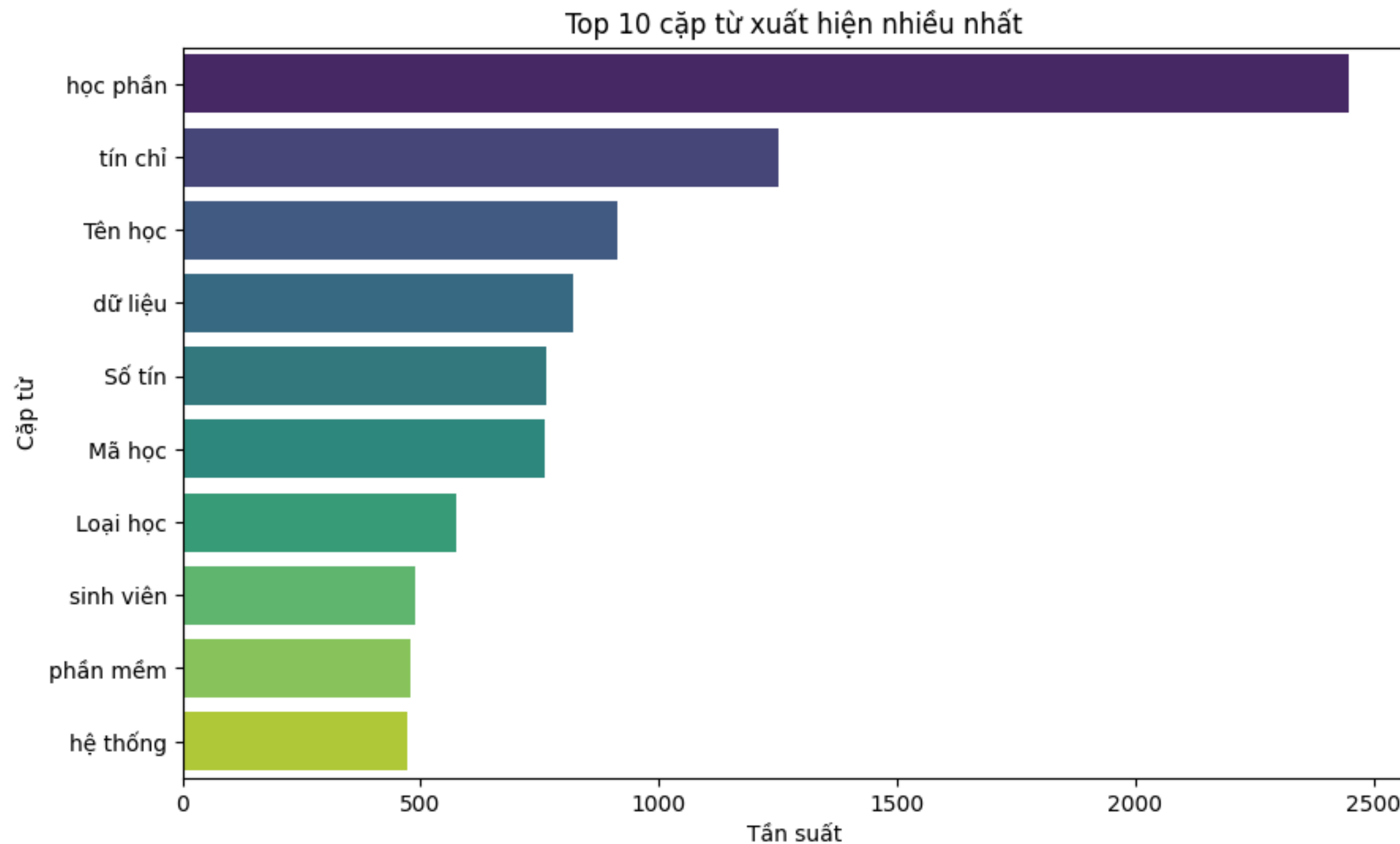
Hình 5. Ma trận tương quan giữa các file dữ liệu

KHÁM PHÁ TẬP DỮ LIỆU



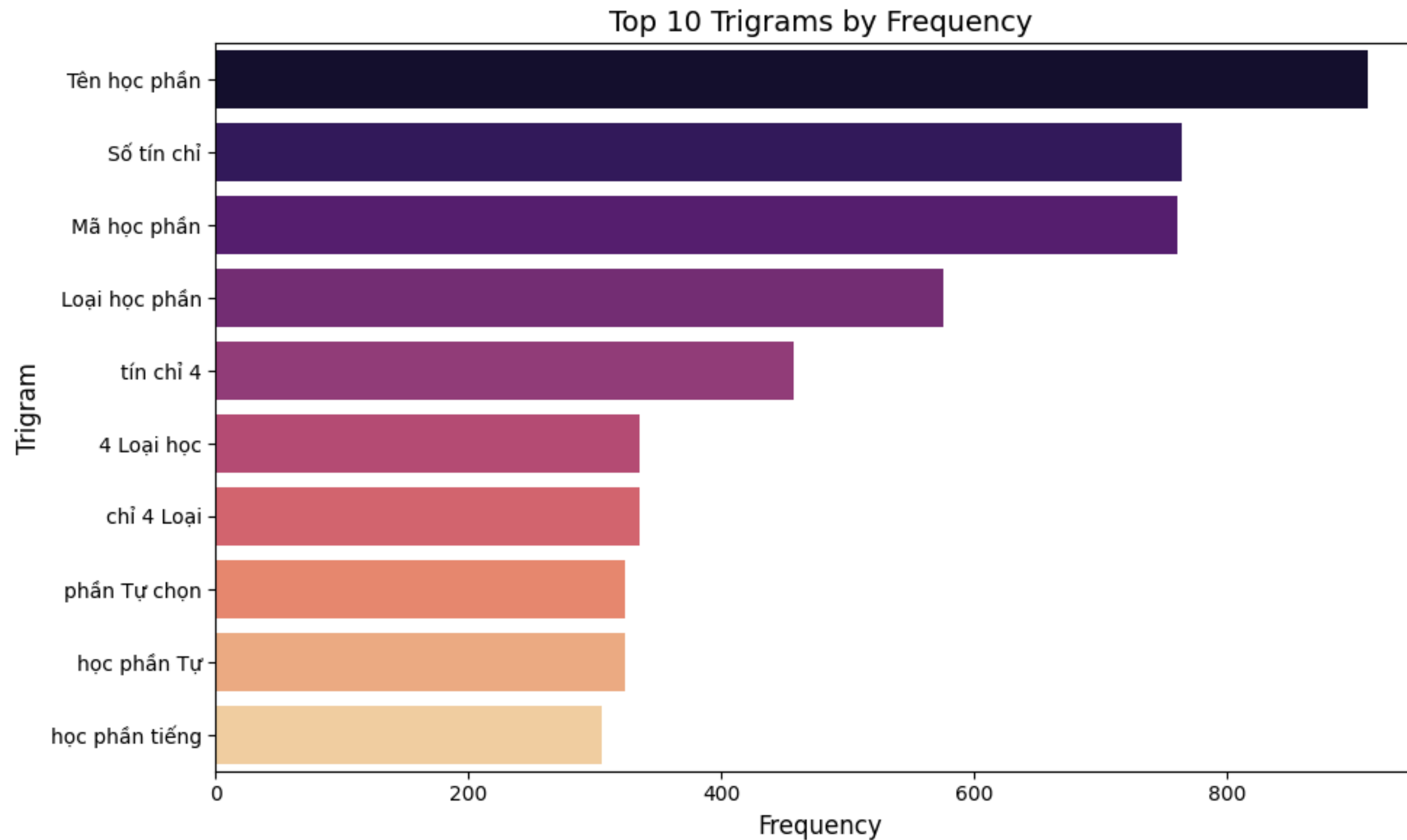
Hình 6. Thống kê số lượng từ trong từng file

KHÁM PHÁ TẬP DỮ LIỆU



Hình 7. Top 10 cặp từ xuất hiện nhiều nhất

KHÁM PHÁ TẬP DỮ LIỆU



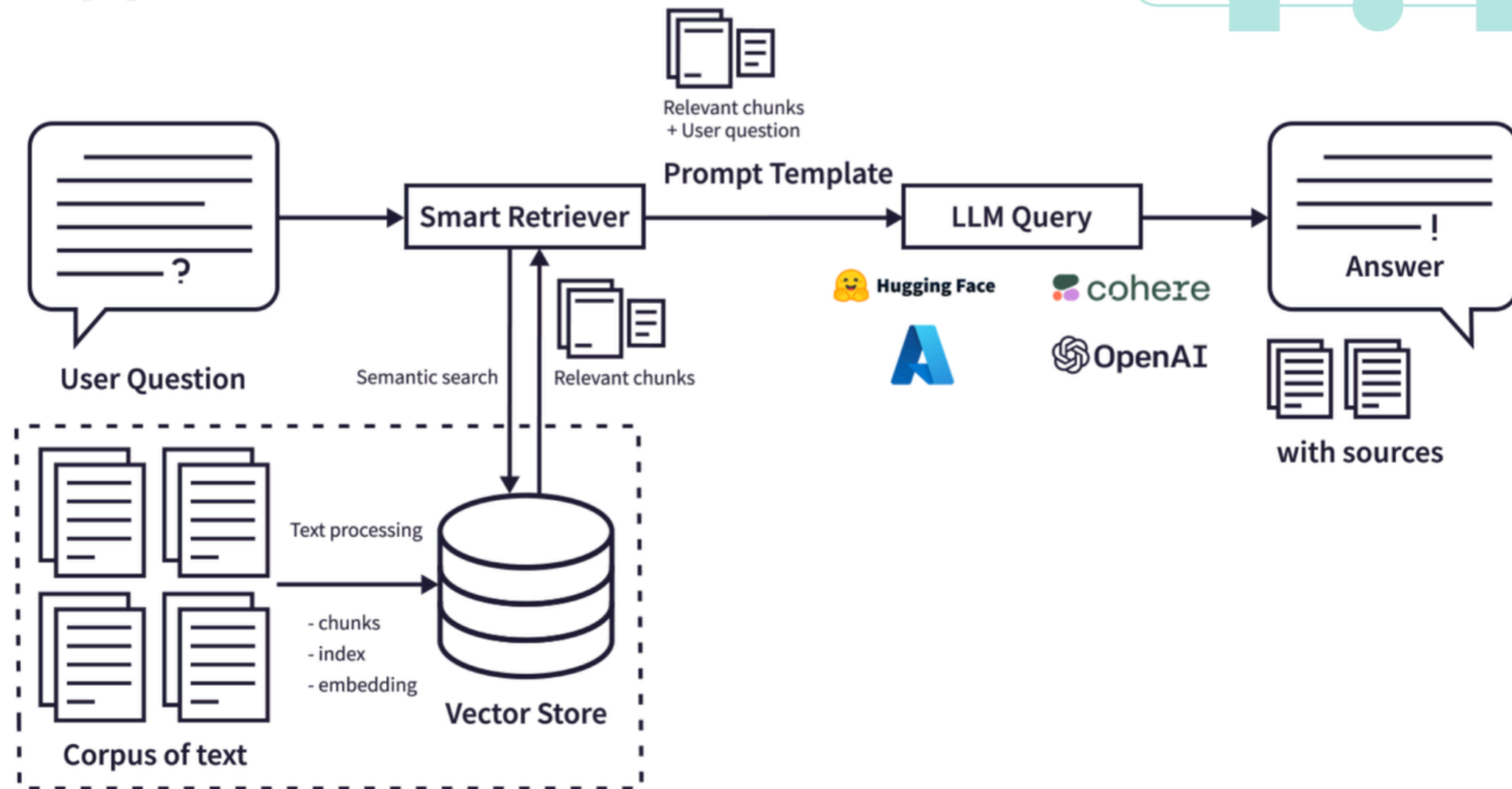
Hình 8. Top 10 bộ ba xuất hiện nhiều nhất



III. KIẾN TRÚC HỆ THỐNG

RAG PIPELINE

RAG pipeline





**CẢM ƠN THẦY VÀ CÁC BẠN
ĐÃ THEO DÕI**