

ĐẠI HỌC QUỐC GIA TP.HCM  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN



**Môn học:** KHAI PHÁ DỮ LIỆU TRONG DOANH NGHIỆP

**LỚP:** DS317.P11

## **Báo cáo phân tích bộ dữ liệu**

**GVHD:** ThS. Nguyễn Thị Anh Thư

*Nhóm sinh viên thực hiện:*

Nguyễn Hữu Nam	MSSV: 22520917
Nguyễn Khánh	MSSV: 22520641
Võ Đình Khánh	MSSV: 22520659
Nguyễn Minh Sơn	MSSV: 22521254
Bùi Hồng Sơn	MSSV: 22521246



# Mục lục

<b>1</b>	<b>Báo cáo phân tích bộ dữ liệu</b>	<b>2</b>
1.1	Tìm hiểu dữ liệu . . . . .	2
1.1.1	Giới thiệu bộ dữ liệu sử dụng . . . . .	2
1.1.2	Mô tả về tập dữ liệu . . . . .	4
1.1.3	Nhận xét . . . . .	16
1.1.4	Mục tiêu sử dụng bộ dữ liệu: . . . . .	17
1.2	Chuẩn bị dữ liệu . . . . .	19
1.2.1	Dịch bảng . . . . .	19
1.2.2	Khám phá dữ liệu . . . . .	22
1.2.3	Làm sạch dữ liệu . . . . .	37
1.2.4	Chuyển đổi dữ liệu . . . . .	44
1.3	Phân tích vấn đề . . . . .	46
1.3.1	Câu hỏi nghiên cứu . . . . .	46
1.3.2	Kết quả đề tài . . . . .	47
1.4	Khả năng ứng dụng . . . . .	47



# 1. Báo cáo phân tích bộ dữ liệu

## 1.1. Tìm hiểu dữ liệu

### 1.1.1. Giới thiệu bộ dữ liệu sử dụng

MOOCCubeX là một trong những bộ dữ liệu lớn nhất và chi tiết nhất về MOOCs (Massive Open Online Courses), hỗ trợ các nghiên cứu về hành vi học tập trực tuyến và cá nhân hóa học tập. Bộ dữ liệu được xây dựng bởi Nhóm Kỹ thuật Tri thức (Knowledge Engineering Group) tại Đại học Thanh Hoa (Tsinghua University), Trung Quốc, với sự hợp tác của XuetangX, một nền tảng MOOC lớn tại Trung Quốc. Đây là bộ dữ liệu đa dạng, phục vụ cho nghiên cứu trong các lĩnh vực như học máy, hệ thống học tập thích ứng, phân tích giáo dục, và trí tuệ nhân tạo.

MOOCCubeX bao gồm nhiều loại dữ liệu khác nhau, tập trung vào các khóa học và hành vi học tập của học viên. Các thành phần chính của bộ dữ liệu bao gồm

#### **Courses**

-Số lượng khóa học 4,216

-Nội dung: Mỗi khóa học bao gồm các video giảng dạy, bài tập, và bài kiểm tra. Thông tin về mỗi khóa học bao gồm tiêu đề, mô tả, người hướng dẫn, ngày bắt đầu và ngày kết thúc, ngôn ngữ giảng dạy và lĩnh vực học tập

#### **Video**

-Số lượng: 230,263

-Thông tin: Các video giảng dạy được thu thập từ các khóa học trên nền tảng MOOC. Mỗi video có các thuộc tính như tiêu đề, thời lượng, nội dung được giảng dạy, và số lần xem của học viên

#### **Exercise**

-Số lượng: 258,265

-Thông tin: bao gồm các bài tập tự luyện và kiểm tra đánh giá. Các



bài tập này được thiết kế để giúp học viên ôn luyện kiến thức và kiểm tra khả năng tiếp thu sau mỗi phần học

### **Problem**

- Số lượng: 2,454,397 vấn đề
- Thông tin: Thường là các vấn đề hoặc câu hỏi phức tạp yêu cầu học viên giải quyết bằng cách áp dụng kiến thức học được từ khóa học

### **Student Profile**

- Số lượng: 3,330,294 hồ sơ
- Thông tin: Hồ sơ học viên lưu trữ các thông tin về hành vi học tập, tiến trình học tập và các hoạt động của họ trên nền tảng

### **Video watching behavior**

- Số lượng: 154,332,174 dữ liệu
- Thông tin: Dữ liệu hành vi xem video cung cấp thông tin chi tiết về cách học viên tương tác với video giảng dạy. Dữ liệu này giúp nghiên cứu thói quen học tập của học viên

### **Comment and Reply**

- Số lượng: 8,422,134 bản ghi phản hồi bình luận
- Thông tin: Bình luận và phản hồi là phần quan trọng trong việc đánh giá mức độ tương tác của học viên với khóa học. Là cơ sở để phân tích cảm xúc của học viên, đánh giá mức độ hài lòng và tìm kiếm những khó khăn mà học viên gặp phải trong quá trình học

Bộ dữ liệu MOOCCubeX được cung cấp dưới dạng các tệp tin JSON và CSV, cho phép người dùng dễ dàng tải xuống và sử dụng. Đây là một bộ dữ liệu quý giá cho nghiên cứu về giáo dục trực tuyến và học tập thích ứng. Với khối lượng dữ liệu lớn và đa dạng, bộ dữ liệu này mở ra nhiều cơ hội cho các nhà nghiên cứu trong việc hiểu sâu hơn về hành vi học tập và xây dựng các hệ thống học tập tiên tiến, giúp cải thiện hiệu quả giáo dục trên các nền tảng trực tuyến.



### 1.1.2. Mô tả về tập dữ liệu

#### I. Courses

***Giới thiệu:*** Phần này mô tả về khóa học (course) và các tài nguyên liên quan, bao gồm các file: course.json, video.json, problem.json, school.json, teacher.json, course-field.json, course-school.txt, course-teacher.txt, exercise-problem.txt, video\_id-ccid.txt.

Đây là bảng sơ lược về các file:

Tên	Loại	Mô tả	Kích thước
course.json	entities	Tổ chức video và bài tập của khóa học.	43MB
video.json	entities	Tên video và phụ đề.	580MB
exercise-problem.txt	relations	Một nhóm các bài tập của khóa học.	129MB
problem.json	entities	Các bài tập thực hành trong một nhóm bài tập.	1.2GB
school.json	entities	Thông tin về trường học.	613KB
teacher.json	entities	Thông tin về giáo viên.	8.7MB
course-field.json	relations	Lĩnh vực mà khóa học thuộc về, được chú thích bởi con người.	62KB

**Bảng course.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
about	Giới thiệu khóa học	string	
id	ID của khóa học	string	Bắt đầu bằng "C_"
field	Danh sách các lĩnh vực của khóa học	list<string>	
name	Tên trường	string	
prerequisites	Nội dung về kiến thức tiên quyết	string	
resource	Danh sách các tài nguyên	list<Resource>*	

**\*Bảng Resource**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
resource_id	ID của tài nguyên.	string	Bắt đầu bằng "V_" nếu là video, "Ex_" nếu là bài tập.
chapter	Số chương.	list<string>	
titles	Danh sách các tiêu đề, bao gồm tiêu đề chương, tiêu đề video, v.v. Có tối đa 3 cấp tiêu đề.	list<string>	

**Bảng video.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
ccid	ID duy nhất của video.	string	
name	Tên của video.	string	
start	Thời gian bắt đầu của từng câu trong phụ đề video.	list<float>	
end	Thời gian kết thúc của từng câu trong phụ đề video.	list<float>	
text	Phụ đề của từng câu trong video.	list<string>	

**Bảng exercise-problem.json**

Mô tả	Định dạng	Kích thước
Câu hỏi của bài tập.	exercise ID\tquestion ID	129MB

**Bảng problem.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID của bài toán.	string	Bắt đầu với "Pm_"
exercise_id	ID của bài tập.	string	Bắt đầu với "Ex_"
language	Ngôn ngữ mô tả của bài toán, tiếng Trung/tiếng Anh.	string	Chinese hoặc English
title	Tiêu đề của bài tập.	string	



content	Mô tả bài toán.	string
option	Lựa chọn của bài toán.	json
answer	Đáp án của câu hỏi.	list<string>
score	Điểm số của câu hỏi.	string
type	Lựa chọn câu hỏi.	int
typetext	Lựa chọn câu hỏi.	string
location	Vị trí chương của bài toán.	string
context_id	leaf_id liên quan đến bài toán.	list<int>

**Bảng school.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID của trường.	string	Bắt đầu với "S_"
name	Tên tiếng Trung của trường.	string	
name_en	Tên tiếng Anh của trường.	string	
sign	Chữ cái đầu của tên tiếng Anh của trường.	string	
about	Giới thiệu về trường.	string	
motto	Khẩu hiệu của trường.	string	



**Bảng teacher.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID của giáo viên	string	Bắt đầu với "T_"
name	Tên tiếng Trung của giáo viên.	string	
name_en	Tên tiếng Anh của giáo viên.	string	
about	Hồ sơ giáo viên.	string	
job_title	Chức danh công việc.	string	
org_name	Cơ quan/đơn vị công tác.	string	

**Bảng course-field.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
course_id	ID của khóa học.	int	
course_name	Tên của khóa học.	string	
field	Danh sách lĩnh vực được gán nhãn thủ công.	list<string>	

**Các mối quan hệ khác**

Tên	Mô tả	Định dạng	Kích thước
course-school.txt	Trường dạy khóa học.	course ID\tschool ID	60KB
course-teacher.txt	Giáo viên dạy khóa học.	course ID\tteacher ID	1.6MB
video_id-ccid.txt	Phụ đề của video	Video ID\tccid	115MB



## II. User

***Giới thiệu:*** Phần này mô tả hành vi người học (user), bao gồm các file: user.json, comment.json, reply.json, course-comment.txt, user-comment.txt, user-reply.txt, comment-reply.txt, user-problem.json, user-video.json, user-xiaomu.json.

Đây là bảng sơ lược về các file:

Tên	Loại	Mô tả	Kích thước
user.json	entities	Thông tin của học sinh (user)	770MB
comment.json	entities	Thông tin bình luận của user lên từng tài nguyên của course	2.1GB
reply.json	entities	Thông tin của phần trả lời bình luận (reply) của user trên từng tài nguyên của courses	50MB
user-problem.json	relations	Thông tin về bài tập mà user làm	50MB
user-video.json	relations	Quá trình của user xem video: số lần tua, giây bắt đầu, giây kết thúc,.	3.0GB
user-xiaomu.json	relations	Tương tác của người dùng với Xiaomu (bot QA của XuetaangX).	50MB

**Bảng user.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Id người dùng	string	bắt đầu bằng "U_"
name	Tên người dùng	string	
gender	Giới tính	int	0, 1, hoặc 2
school	Tên trường	string	
year_of_birth	Năm sinh	list<int>	
course_order	Các mã khóa học đã chọn	Thông tin về bài tập mà user làm	
enroll_time	Thời gian đăng kí tương ứng với từng khoá học	list<DateTime>.	Định dạng DateTime là "YYYY-MM-DD HH:MM:SS"

**Bảng comment.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Comment ID	string	bắt đầu bằng "Cm_"
user_id	ID của người dùng đã bình luận	Int	bắt đầu bằng "U_"
text	Nội dung bình luận	String	
create_time	Thời gian bình luận	DateTime	định dạng "YYYY-MM-DD HH:MM:SS"
resource_id	ID của tài nguyên mà user bình luận	String	Có thể nhận giá trị null

**Bảng reply.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Reply ID	string	bắt đầu bằng “Rp_”
user_id	ID của người dùng đã bình luận	string	bắt đầu bằng “U_”
text	Nội dung phản hồi	string	
create_time	Thời gian phản hồi	DateTime	định dạng “YYYY-MM-DD HH:MM:SS”

**Bảng user-problem.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
log_id	ID của bản ghi câu hỏi của người dùng	string	kết hợp với khóa duy nhất của user_id và problem_id
user_id	ID người dùng	string	bắt đầu bằng “U_”
problem_id	ID vấn đề	string	bắt đầu bằng “Pm_”
is_correct	Câu hỏi có đúng không	bool	0 hoặc 1



attempts	Số lượng câu hỏi đã thử	int	
score	Điểm của người dùng	float	
submit_time	Thời gian làm câu hỏi	DateTime	định dạng “YYYY-MM-DD HH:MM:SS”

**Bảng user-video.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
user_id	ID của user	string	bắt đầu bằng “U_”
seq	Mảng chứa quá trình người dùng xem video, bao gồm thời gian xem video, thời gian bắt đầu và kết thúc của video, và tốc độ xem video, v.v.	list<object>.	Mỗi object sẽ gồm 2 trường video_id (string) và segment (list<object>). Mỗi phần tử trong segment bao gồm các trường start_point (float), end_point (float), speed (float), local_start_time (int)

**Bảng user-xiaomu.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
user_id	ID của user	string	bắt đầu bằng “U_”
question_type	ID của user	string	
question	Câu hỏi hỏi bởi user	string	

**Các mối quan hệ khác**



Tên	Mô tả	Định dạng	Kích thước
course-comment.txt	Phản hồi bình luận của người dùng lên course	course ID\treview ID	60KB
user-comment.txt	bình luận của người dùng.	user ID\tcomment ID	1.6MB
user-reply.txt	Phản hồi bình luận của người dùng.	user ID\treply ID	1.6MB
comment-reply.txt	Phản hồi bình luận liên quan đến khái niệm (phần concept).	concept ID\treply ID	115MB

### III. Concept

**Giới thiệu:** Phần này mô tả về khái niệm khóa học (course concept) và các file liên quan, bao gồm: concept.json, other.json, paper.json, concept-other.txt, concept-paper.txt, concept-problem.txt, concept-video.txt, concept-comment.txt.

Đây là bảng sơ lược về các file:

Tên	Loại	Mô tả	Kích thước
concept.json	entities	Thông tin về khái niệm khóa học	43MB
other.json	entities	Các tài liệu liên quan được thu thập bên ngoài những khoá học	580MB
paper.json	entities	Những bài báo khoa học liên quan	129MB
concept-other.txt	relations	Khái niệm liên quan tới các nguồn ngoài khóa học	1.2MB
concept-paper.txt	relations	Khái niệm liên quan đến luận án	613KB
concept-problem.txt	relations	Khái niệm liên quan đến vấn đề	8.7MB
concept-video.txt	relations	Khái niệm liên quan đến video	8.7MB
concept-comment.txt	relations	Khái niệm liên quan đến phần bình luận	62KB

**Bảng concept.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID của khái niệm	string	Định dạng là K_concept name_field
name	Tên của khái niệm, và tên này sẽ giống với tên xuất hiện trong id	string	
context	Ngữ cảnh mà khái niệm đó xuất hiện	string	

**Bảng other.json**

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Mã dữ liệu, không có ý nghĩa cụ thể (đơn thuần là một định danh duy nhất cho từng mục dữ liệu).	string	
concept	Khái niệm mà thông tin này liên quan đến hoặc được thu thập dựa trên	string	
type	Nguồn dữ liệu	string	Miền giá trị là ["zhihu", "baike", "wiki"]
content	Nội dung của dữ liệu, có thể là văn bản hoặc thông tin được thu thập từ các nguồn đã nêu	string	



### Các mối quan hệ khác

Tên	Mô tả	Định dạng	Kích thước
concept-other.txt	Lưu trữ mối quan hệ giữa các khái niệm và các tài liệu, tài nguyên ngoại khóa được thu thập từ các nguồn bên ngoài khóa học	concept ID\tresource ID	60KB
concept-paper.txt	Lưu trữ mối quan hệ giữa các khái niệm và các bài báo khoa học có liên quan	concept ID\tpaper ID	1.6MB
concept-problem.txt	Lưu trữ mối quan hệ giữa các khái niệm và các câu hỏi hoặc bài tập liên quan	concept ID\tquestion ID	1.6MB
concept-video.txt	Lưu trữ mối quan hệ giữa các khái niệm và các video liên quan	concept ID\tccid	1.6MB
concept-comment.txt	Lưu trữ mối quan hệ giữa các khái niệm và các bình luận của người dùng có liên quan	concept ID\treview ID	115MB

### IV. Prerequisites

#### Bảng prerequisites/cs.json

- Nội dung: Chú thích và dự đoán về các điều kiện tiên quyết của môn Khoa học máy tính
- Số lượng mẫu: 492,102 mẫu





Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
c1	Khái niệm điều kiện tiên quyết	string	
c2	Khái niệm điều kiện sau sửa chữa	string	
ground_truth	Chỉ ra có mối quan hệ sửa chữa tuần tự hay không	int	Miền giá trị là 0 hoặc 1
text_predict	Cung cấp kết quả dự đoán sử dụng đặc điểm văn bản	list<float>	
graph_predict	Mức độ tin cậy của dự đoán được đạt được bằng các đặc điểm đồ thị	list<float>	

### Bảng prerequisites/math.json

- Nội dung: Chú thích và dự đoán các khái niệm trong lĩnh vực toán học, theo định dạng giống cs.json
- Số lượng mẫu: 331202

### Bảng prerequisites/psy.json

- Nội dung: Chú thích và dự đoán các khái niệm trong lĩnh vực tâm lý học, theo định dạng giống cs.json
- Số lượng mẫu: 757771

### 1.1.3. Nhận xét

Sau khi khảo sát bộ dữ liệu MOOCCubeX, chúng em đã rút ra một số nhận xét như sau:



- Tính đa dạng và phong phú: Bộ dữ liệu MOOCCubeX chứa đựng nhiều loại thông tin khác nhau liên quan đến giáo dục trực tuyến, bao gồm các khóa học, bài giảng video, bài tập, hồ sơ học sinh, cũng như hành vi tương tác của học sinh với các tài nguyên học tập. Đây là một bộ dữ liệu có mức độ đa dạng cao, giúp cung cấp cái nhìn toàn diện về nhiều khía cạnh trong quá trình học tập trực tuyến.
- Quy mô lớn: Bộ dữ liệu có kích thước lớn và chứa đựng hàng triệu điểm dữ liệu, từ đó tạo cơ sở vững chắc cho các bài toán khai thác dữ liệu, học máy, học sâu. Nhờ quy mô này, người nghiên cứu có thể khám phá và áp dụng các phương pháp tiên tiến trong lĩnh vực phân tích dữ liệu giáo dục.
- Tính chi tiết và tổ chức linh hoạt: Mặc dù không đồng nhất về loại dữ liệu, bộ dữ liệu MOOCCubeX được tổ chức bài bản với cấu trúc rõ ràng và chi tiết. Điều này giúp người dùng dễ dàng tìm kiếm và trích xuất các thông tin quan trọng, đồng thời cung cấp sự linh hoạt trong việc áp dụng bộ dữ liệu vào nhiều mục tiêu khác nhau. Các yếu tố như hành vi học tập, bình luận của học sinh, và các tài liệu khóa học đều được ghi nhận chi tiết, tạo nền tảng tốt cho việc xây dựng các hệ thống hỗ trợ học tập thông minh.

#### 1.1.4. Mục tiêu sử dụng bộ dữ liệu:

Với các đặc điểm nêu trên, chúng em định khai thác bộ dữ liệu MOOCCubeX để giải quyết các bài toán thuộc lĩnh vực Cố vấn học tập thông minh. Cụ thể, nhóm đã đưa ra bài toán sau:

#### Bài toán: Hệ khuyến nghị khóa học.

- Mục tiêu: Xây dựng hệ thống khuyến nghị giúp sinh viên chọn lựa môn học hoặc khóa học phù hợp với định hướng chuyên ngành dựa trên hành vi học tập của họ. Điều này có thể bao gồm các yếu tố như các khóa học mà sinh viên đã hoàn thành, kết quả học tập, thời gian dành cho mỗi môn học, và sự tương tác của họ với tài nguyên học tập (Như video, bài tập, và bài kiểm tra).



- Ứng dụng: Hệ thống sẽ hỗ trợ sinh viên đưa ra các quyết định học tập thông minh hơn, giúp họ lựa chọn các môn học phù hợp với năng lực và định hướng cá nhân. Điều này không chỉ giúp tối ưu hóa quá trình học tập mà còn tăng khả năng hoàn thành các chương trình học, đặc biệt trong các môi trường giáo dục trực tuyến hoặc bán trực tuyến.
- Khả năng áp dụng: Bài toán này hoàn toàn có thể được áp dụng trong bối cảnh giáo dục đại học, cụ thể là ở Việt Nam. Đặc biệt là tại các trường có chương trình học trực tuyến hoặc có nhu cầu xây dựng hệ thống cố vấn học tập dựa trên dữ liệu. Hệ thống có thể giúp sinh viên định hướng chuyên ngành, chọn lựa các môn học phù hợp, và điều chỉnh lộ trình học tập dựa trên kết quả học tập và hành vi của họ.

Nhìn chung, việc ứng dụng bộ dữ liệu MOOCCubeX vào các bài toán như vậy có tiềm năng lớn trong việc hỗ trợ sinh viên và nâng cao trải nghiệm học tập trong môi trường giáo dục điểm số.



## 1.2. Chuẩn bị dữ liệu

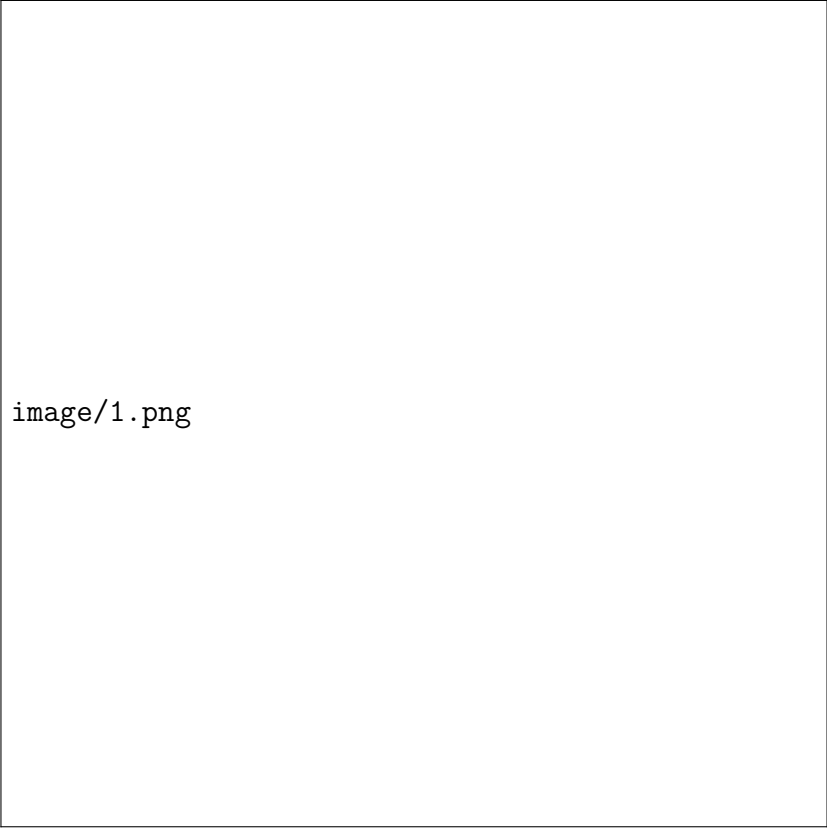
### 1.2.1. Dịch bảng

Trong quá trình chuyển ngữ từ Trung sang Việt, chúng em đã tận dụng thư viện "googletrans" một công cụ Python không mất phí và không giới hạn số lần dịch. Thư viện này vận hành thông qua API Google Translate Ajax để thực hiện các tác vụ như nhận diện ngôn ngữ và dịch thuật.

Do khối lượng dữ liệu lớn, quá trình dịch gặp phải một số thách thức về thời gian và kết nối. Để khắc phục, chúng em đã triển khai các giải pháp sau:

- Lưu lại tiến trình dịch để tránh mất dữ liệu
- Thiết lập cơ chế tự động gửi lại yêu cầu khi mất kết nối
- Ứng dụng thư viện "asyncio" cho phép gửi đồng thời nhiều API, giúp tối ưu tốc độ xử lý

Đây là một phần code mẫu đã sử dụng phương pháp đã nêu trên:



image/1.png

Ngoài ra, chúng em nhận thấy không cần thiết phải dịch toàn bộ các trường dữ liệu lớn để huấn luyện mô hình vì một số trường dữ liệu không hỗ trợ cho việc huấn luyện mô hình. Thay vào đó, chúng em chỉ tập trung dịch 1 số trường sau đây:

**-course.json:** dịch cột “name”, “field”, “prerequisites” và “about”

**-user.json:** dịch cột “school”

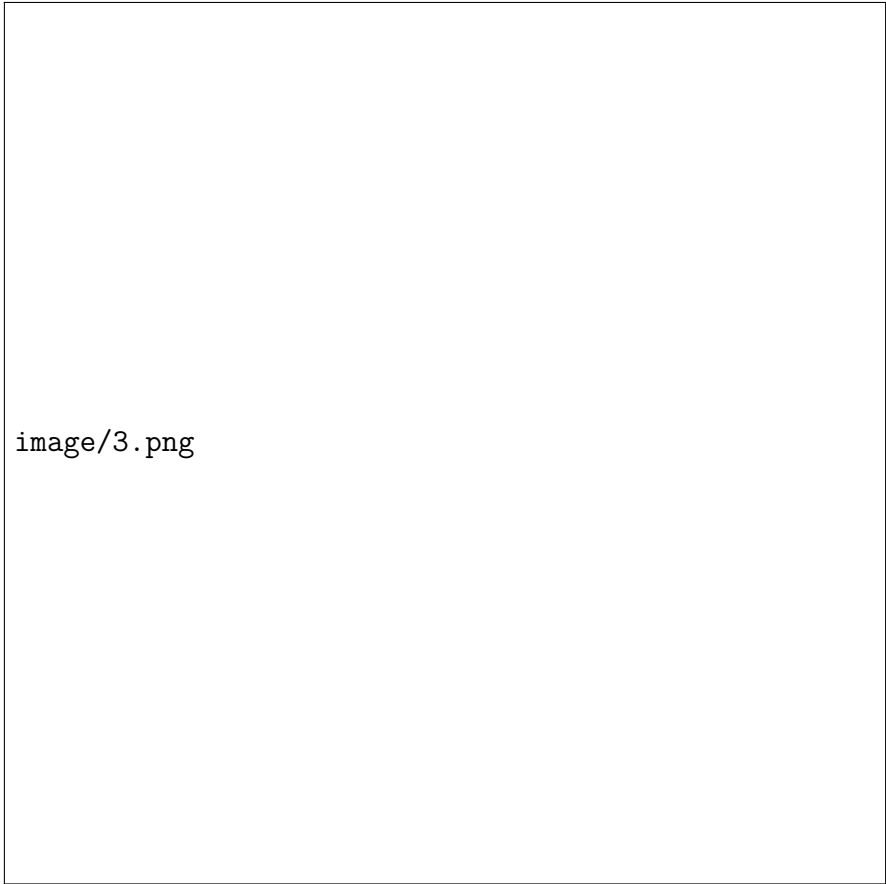


image/2.png

**-teacher.json:** Tiến hành dịch tất cả (trừ “id” và “name”)

**-concept.json:** Dịch tất cả các cột của bảng này vì toàn bộ đều ở dạng chuỗi

**-course-field.json:** Tiến hành dịch cột course\_name và field mang các thông tin dưới dạng chuỗi của bảng.



image/3.png

### 1.2.2. Khám phá dữ liệu

#### a) Bảng course.json

Ta xem qua bảng course.json:

Ta xét độ dài của 3 cột “about”, “name\_trans” và “resource”:



image/4.png

Ta có thể thấy được 1 số thông tin từ dữ liệu trên:

- Có những dòng dữ liệu không tồn tại cột “about”, tồn tại giá trị ngoại lệ ở cột “about” vì mean là 393 mà max lên đến 2293. Ta thể hiện trên boxplot độ dài của cột “about”:
- Có thể thấy thật sự nhiều giá trị ngoại lệ cần được xử lí.
- Có những dòng dữ liệu không có resource\_length, mean cũng rất ngắn (71) chứng tỏ ít thông tin về khoá học.





image/5.png

Ta phân tích sâu cột “resource”:

Mỗi resource trong bảng 2 là 1 tập hợp các video hay một tập các exercise. Mỗi resource sẽ có thêm 1 resource\_id là id của resource, chapter là chương chứa resource trong khóa học, titles gồm các tiêu đề như tiêu đề chương, video chương.

Thông tin của resource có thể tìm thấy trong file course.json. Một resource có 2 loại: Video và Exercise. Nếu loại tài nguyên là video, nó được xác định bằng ID video bắt đầu bằng ký tự V\_. Nhiều video\_id khác nhau tương ứng với một ccid, và ccid xác định duy nhất một video. Các video\_id này tương ứng với việc



image/6.png

hiển thị cùng một video ccid tại các thời gian bắt đầu khác nhau. Mỗi liên hệ giữa video\_id và ccid được lưu trong relations/video\_id-ccid.txt. Phụ đề video có thể được tìm thấy trong tệp entities/video.json thông qua ccid.

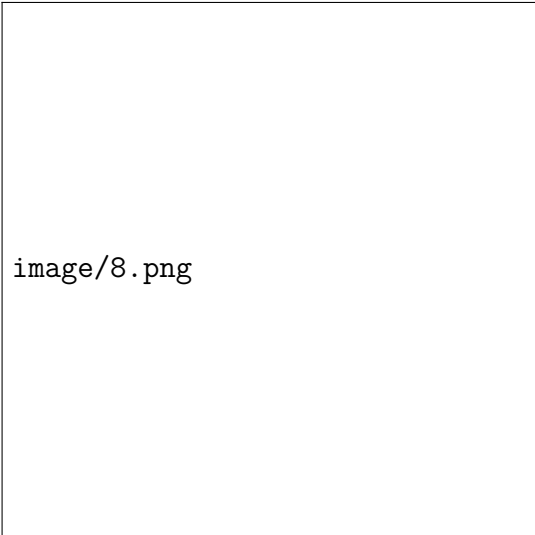
Ta sẽ kiểm tra xem có bao nhiêu ID video không hợp lệ để phục vụ cho quá trình xử lý dữ liệu sau này:

Có 2397 video ID không tồn tại, ta sẽ lọc đi hỗ trợ cho hiển thị thông tin trong tương lai.



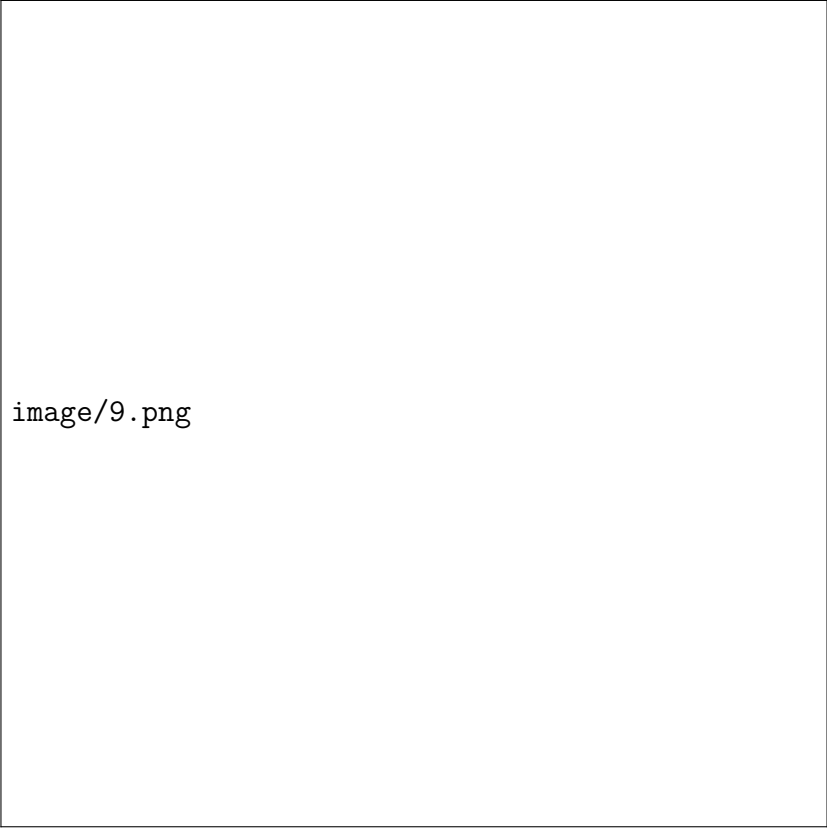
image/7.png

Ta bắt đầu tiến hành đếm số khoá học trong cột “name\_trans”, chia bởi lĩnh vực (cột “field”):



image/8.png

Ta thấy có tổng 3781 khoá học và 81 lĩnh vực, với “computer science and technology” đứng đầu với 63 khoá học, chiếm 6.8% trên tổng khoá học. Ta cũng kiểm tra với mỗi khoá học được xếp bao nhiêu lĩnh vực (cột “field”):



image/9.png

Ta có thể thấy có rất nhiều khoá học không thuộc lĩnh vực nào, có rất nhiều khóa học không có field nào, có thể cột “field” sẽ không đóng góp nhiều trong xây dựng thuật toán hoặc cần xử lí.

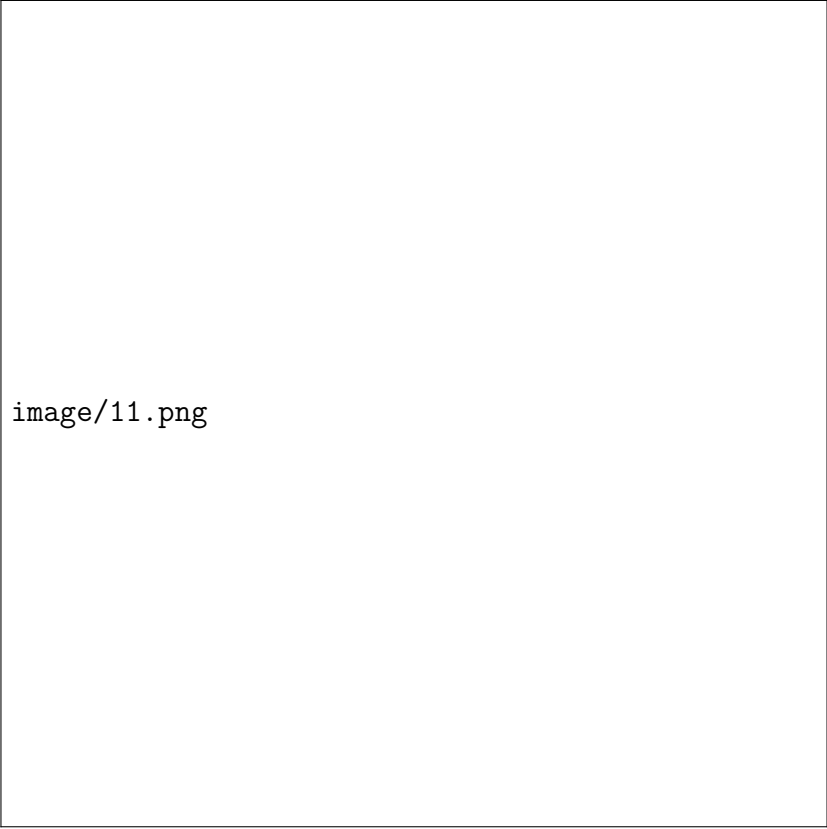
**b) Bảng user.json**

Đầu tiên, ta đọc dữ liệu và quan sát dữ liệu thông qua dạng bảng (DataFrame):



image/10.png

Ta tiến hành thống kê đặc điểm từng cột có trong bảng:



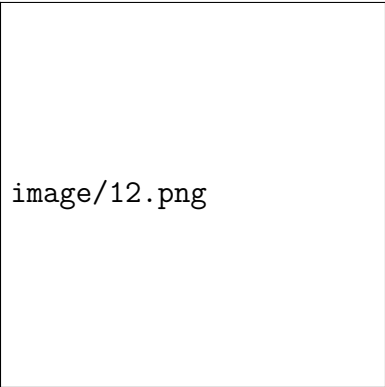
image/11.png

**c) Bảng teacher.json**

Sau đây là các thống số cơ bản của bảng

Tham khảo phân phối của top 10 tên việc xuất hiện nhiều nhất trong bảng

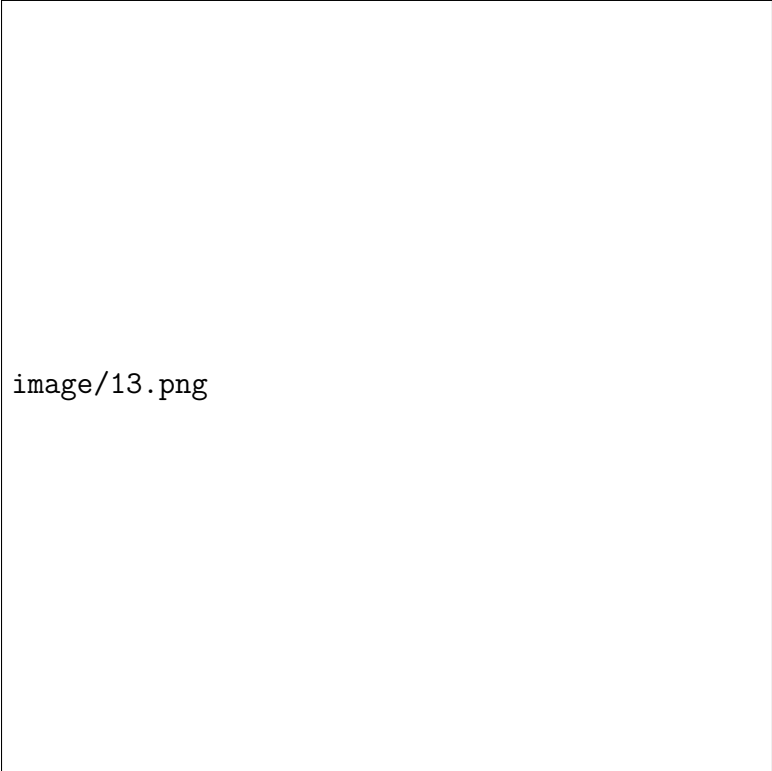
Tham khảo phân phối của top 10 tổ chức xuất hiện nhiều nhất trong bảng



image/12.png

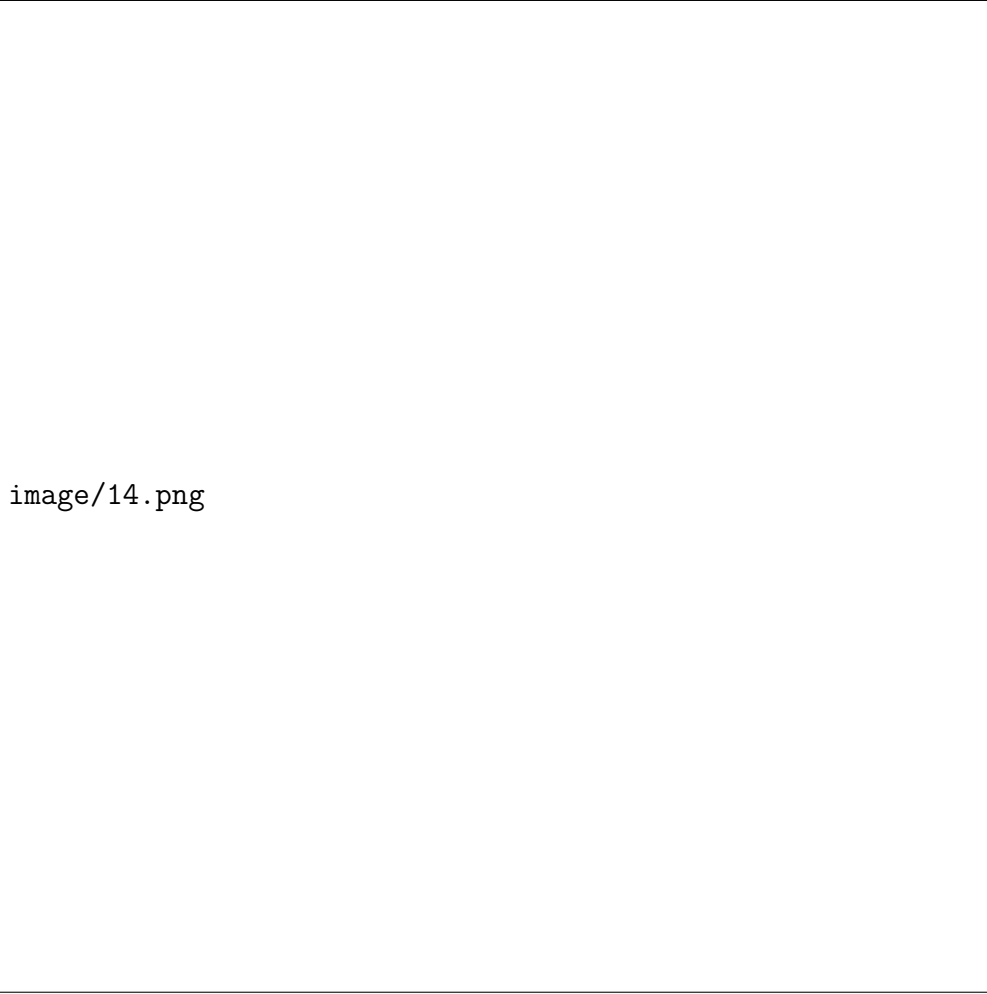
Ta thực hiện phân tích mối quan hệ giữa ba chức vụ (job titles) có số lượng giáo viên nhiều nhất và mười tổ chức (organizations) có số lượng giáo viên cao nhất





image/13.png

Sau khi lọc bỏ các liên kết có khóa học hoặc teacher không tồn tại dựa vào file course-teacher.txt, số hàng còn lại là 35593. Các thông tin được trực quan hóa như sau



image/14.png

**d) Bảng school.json**

Ta đếm dữ liệu ở từng cột, đếm các giá trị đặc biệt, giá trị xuất hiện nhiều nhất với tần số của nó:

Kiểm tra kiểu dữ liệu của từng cột:



image/15.png

Ta tạo 2 cột mới là “about\_length” và “motto\_length” để lần lượt thể hiện độ dài của giá trị dữ liệu ở 2 cột “about” và “motto”:



Có 2 cột ta cần là “about\_length” và “motto\_length” để ta tìm phân bố độ dài của giá trị lên đồ thị:

Dựa vào biểu đồ ta có thể nhận xét rằng mô tả của các trường đều rất chi tiết, số lượng trường với số lượng từ phần mô tả  $> 2000$  chiếm phần lớn. Tuy nhiên thông tin này có vẻ không hữu ích với hệ thống khuyến nghị.

Hầu hết các trường đại học đều có một khẩu hiệu ngắn gọn dưới 20 từ vì chủ yếu khẩu hiệu sẽ đơn giản nhất có thể để truyền đạt tầm nhìn và mục tiêu của trường một cách trực tiếp ngắn gọn, đọng lại trong trí nhớ người xem. Một phần nhỏ hơn các trường có khẩu hiệu tương đối dài với 40 đến 88 chữ.



---

e) Bảng course-field.json



### 1.2.3. Làm sạch dữ liệu

#### a) Bảng `course.json`

Ta kiểm tra dữ liệu thiếu, dữ liệu không nhất quán, dữ liệu trùng lặp và dữ liệu trống:

Đầu tiên ta thấy được có 647 giá trị ở cột “name\_trans” bị trùng lặp cho dù id không bị trùng, chứng tỏ có sự lỗi nhất định trong bộ dữ liệu, cũng như này đã thống kê ta thấy được có rất nhiều giá trị trống ở cột “field\_trans”.

Ta kiểm tra kĩ hơn về các dòng có giá trị trong cột “name” bị trùng lặp:



Ta thấy được đa số dữ liệu trong này cột “field” đa số bị trống và trùng lặp, cũng như các cột khác không có ý nghĩa hoặc trùng với các cột khác, thực hiện chi square test, ta có được kết quả với P-value rất thấp, chứng tỏ các giá trị phụ thuộc với nhau chứ không hề có giá trị mới. Chứng tỏ ta có thể xoá được các dòng dữ liệu này, cũng như các khoá học không tồn tại trong “course-field.json”.

#### **b) Bảng user.json**

Ta thấy cột “year\_of\_birth” bị thiếu dữ liệu hơn 97% trong khi các cột còn lại tỉ lệ % thiếu là rất thấp. Ta tiến hành loại bỏ cột này, sau đó ta sẽ tiến hành xử lý dữ liệu nhiều trên cột gender với 2 giá trị nhiều là 232 và 3

#### **c) Bảng concept.json**

Xử lý dữ liệu thiếu giúp cải thiện độ chính xác của mô hình, đảm bảo tính toàn vẹn của phân tích, tránh lỗi tính toán và giảm độ thiên lệch. Một số cách xử lý phổ biến gồm:

- Loại bỏ hàng/cột: Áp dụng khi dữ liệu thiếu quá nhiều.
- Điền giá trị thay thế: Điền trung bình, trung vị, hoặc giá trị dự đoán vào chỗ thiếu.
- Dùng mô hình dự đoán: Áp dụng các thuật toán để dự đoán giá trị thiếu.

Việc xử lý phù hợp giúp dữ liệu chính xác và đáng tin cậy hơn trong phân tích và dự đoán.



Xử lý dữ liệu trùng lặp là bước quan trọng trong tiền xử lý dữ liệu nhằm loại bỏ các bản ghi trùng lặp để đảm bảo tính chính xác và hiệu quả của mô hình. Dữ liệu trùng lặp có thể gây sai lệch và làm chậm quá trình xử lý.

Các phương pháp xử lý dữ liệu trùng lặp phổ biến bao gồm:

- Xóa các bản ghi trùng lặp: Loại bỏ các hàng hoàn toàn trùng lặp trong DataFrame bằng hàm `drop_duplicates()` trong Pandas.
- Giữ lại bản ghi đầu tiên hoặc cuối cùng: Nếu cần giữ lại một bản ghi đại diện, có thể chỉ xóa các bản trùng lặp sau hoặc trước.
- Xác định tiêu chí trùng lặp: Tìm và xóa bản ghi trùng lặp dựa trên một số cột cụ thể thay vì toàn bộ hàng.

Loại bỏ dữ liệu trùng lặp giúp dữ liệu trở nên nhất quán, giảm dung lượng và cải thiện độ chính xác của phân tích và mô hình.

Loại bỏ dữ liệu với phương thức `drop_duplicates()`:



**d) Bảng course-field.json**

Sử dụng `isnull().sum()` để tính số lượng giá trị thiếu trong từng cột. Sau đó loại bỏ hàng chứa giá trị thiếu bằng cách sử dụng `dropna()`



Dữ liệu văn bản thường chứa nhiều thông tin nhiễu chẳng hạn như các ký tự không mong muốn: Các ký tự đặc biệt, dấu câu, hoặc ký tự không phải chữ cái có thể làm giảm chất lượng phân tích. Ở đây chúng ta sẽ tiến hành loại bỏ các ký tự không cần thiết, các khoảng trắng dư thừa và thường hóa các ký tự viết hoa

Để kiểm tra dữ liệu trùng lặp, chúng ta sử dụng phương thức `uplicated()` trong `pandas`. Đầu tiên xác định các bản ghi trùng lặp, sau đó đếm số lượng và hiển thị các bảng ghi trùng lặp đó. Sau đó tiến hành xóa bản ghi trùng lặp bằng cách sử dụng `drop_duplicates()`

**e) Bảng school.json**

Ta xoá cột “name” đi vì trùng với ý nghĩa với cột “name\_en” (tên nhưng trong Tiếng Anh)

Ta thống nhất cột “sign” (kí hiệu đại diện cho trường) đều là tất cả in hoa:



Vì ở đây tên trường (“name\_en”) cũng như kí hiệu (“sign”) là chìa khoá chính, hay nói cách khác là giá trị duy nhất nên không thể có dòng trùng với nhau, ta tiến hành xoá các dòng trùng giá trị:

**g) Bảng teacher.json**

Ở đây có cột name\_en bị thiếu nên điền vào cột đó bằng cách lấy phiên âm của cột name là được. Để làm việc này có thể sử dụng thư viện pypinyin để lấy phát âm dùng cho tên tiếng anh.



#### 1.2.4. Chuyển đổi dữ liệu

**Feature Engineering:** Nhóm sẽ chọn các bảng và thuộc tính có thể sử dụng để tạo ra feature các mô hình khuyến nghị dựa trên bộ dữ liệu đã xử lý và làm sạch trước đó:

Các bảng được chọn và thuộc tính sử dụng:



- Với ‘user.json’: ‘course\_order’ gồm các khóa học mà user đã đăng ký với khóa học sau cùng là khóa học gần đây nhất, dùng để tạo liên kết giữa ‘user.json’ và ‘course.json’.
- Với ‘course.json’: Đây là table quan trọng chứa thông tin về các khóa học như ‘name’, ‘about’ và ‘field’.
- Với ‘teacher.json’, ‘school.json’: dùng để tạo relation với ‘course.json’ chứa thông tin về trường tổ chức khóa học và giáo viên giảng dạy.
- Với ‘course-field.json’: chứa các field của mỗi khóa học, dùng để kiểm tra với trường ‘field’ trong ‘course.json’.
- Với ‘concept.json’: id theo quy ước ‘K\_concept namefield’, tạo thêm feature concept-name\_field với mỗi khoá học.

Tạo knowledge graph:

Tạo interaction giữa người dùng với khóa học: sử dụng 5-core filtering, lọc người dùng với ít hơn 5 khóa học và những khóa học có số lượng đăng ký dưới 5.

Kết quả: Vì data đã được xử lý trước đó nên ta thấy không có thay đổi đáng kể

Trước khi filter	Sau khi filter
1.183.774 interactions	1.182.745 interactions



Tạo relation giữa các entities: course-relation-attribute. Sau đó ta tiến hành lọc theo tiêu chí, số lần course xuất hiện tối thiểu là 5 và số lần xuất hiện tối thiểu của một relation là 25.

Kết quả:

Trước khi filter	Sau khi filter
376.093 interactions	71.787 interactions

### 1.3. Phân tích vấn đề

Hệ thống học tập trực tuyến MOOC cung cấp số lượng lớn các khóa học đa dạng, nhưng khó khăn lớn đối với người học là tìm kiếm khóa học phù hợp với sở thích và nhu cầu cá nhân. Để giải quyết vấn đề này, hệ thống khuyến nghị khóa học được phát triển nhằm cá nhân hóa trải nghiệm học tập cho từng người dùng dựa trên dữ liệu về hành vi học tập và các đặc điểm cá nhân.

Bài toán đặt ra trong dự án này là: **Làm thế nào để xây dựng một hệ thống khuyến nghị khóa học cá nhân hóa cho từng người học trên nền tảng MOOC?**

#### 1.3.1. Câu hỏi nghiên cứu

- Làm thế nào để dự đoán chính xác các khóa học mà một người dùng có khả năng sẽ đăng ký tiếp theo?
- Làm sao tận dụng các đặc điểm của người dùng như giới tính, độ tuổi, trường học, và lịch sử khóa học để tăng độ chính xác của mô hình khuyến nghị?
- Làm sao đánh giá chất lượng các gợi ý khóa học và xác định mức độ hiệu quả của hệ thống (metric đánh giá là gì) ?



### 1.3.2. Kết quả đề tài

Dự án hướng tới xây dựng một hệ thống khuyến nghị khóa học hiệu quả, dựa trên dữ liệu của người học từ bộ MOOCCubeX. Kết quả mong muốn bao gồm:

- **Xác định yếu tố ảnh hưởng đến việc đăng ký khóa học:** Khám phá các đặc điểm người dùng (giới tính, trường học, năm sinh, các khóa học đã đăng ký...) ảnh hưởng đến hành vi chọn khóa học. Điều này giúp hệ thống có cái nhìn rõ ràng hơn về các yếu tố quan trọng khi gợi ý khóa học.
- **Khả năng khuyến nghị khóa học cá nhân hóa:** Kỳ vọng hệ thống sẽ đưa ra những gợi ý chính xác cho từng người học, dựa trên hành vi đăng ký khóa học trước đây và các yếu tố liên quan. Mục tiêu là hệ thống có thể dự đoán tốt các khóa học mà người dùng có khả năng quan tâm trong tương lai.
- **Định hướng cải thiện trải nghiệm học tập:** Hệ thống khuyến nghị dự kiến sẽ giúp người học tiết kiệm thời gian tìm kiếm, đồng thời cung cấp cho họ trải nghiệm học tập tốt hơn thông qua việc gợi ý các khóa học phù hợp với mục tiêu và sở thích cá nhân.
- **Đánh giá các phương pháp tiếp cận:** Tìm hiểu các mô hình Recommendation System và thử nghiệm với bộ dữ liệu để so sánh độ hiệu quả của các mô hình.

### 1.4. Khả năng ứng dụng

Hệ thống khuyến nghị này có tiềm năng ứng dụng rộng rãi trong các nền tảng học tập trực tuyến. Cụ thể:

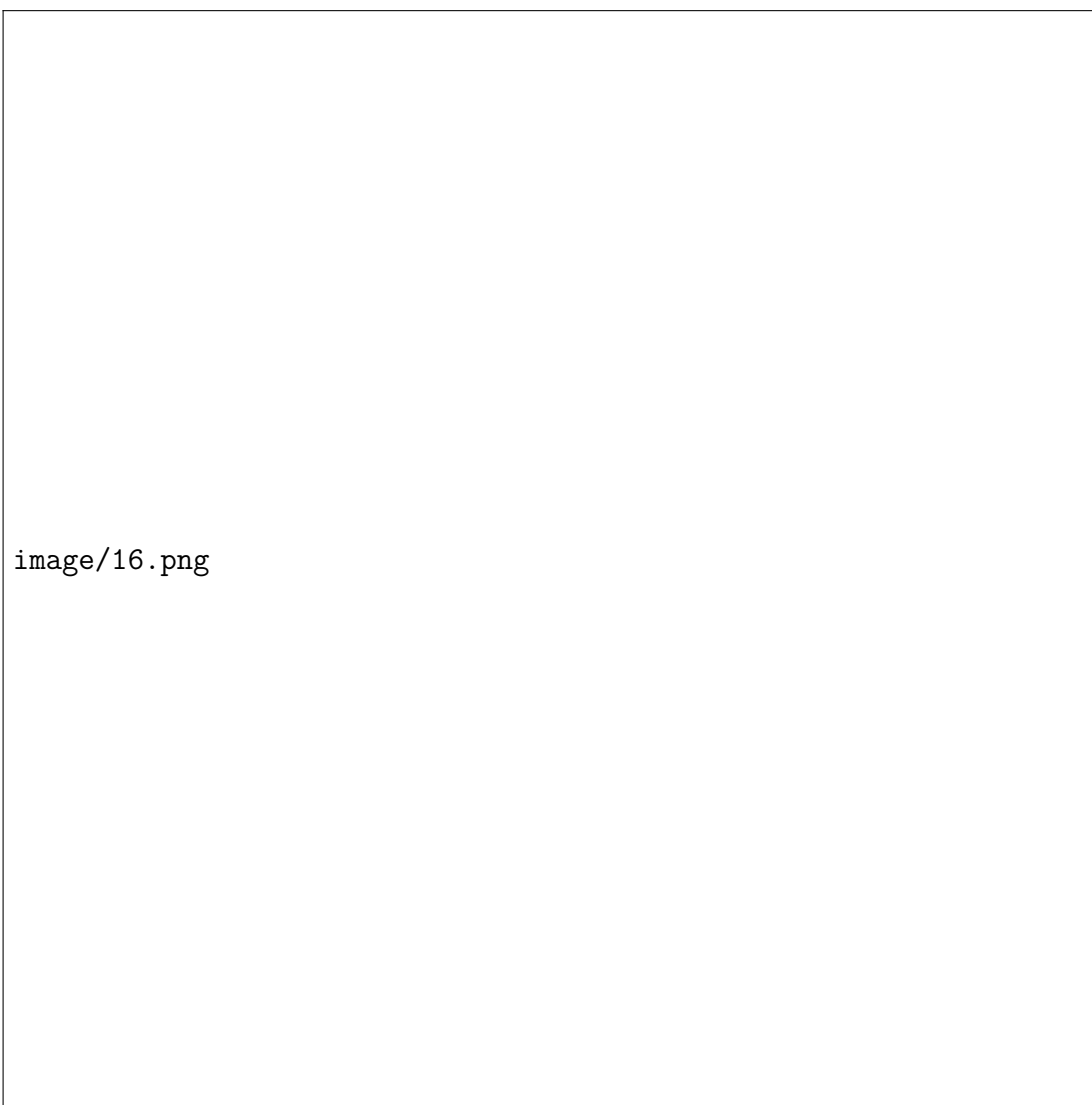
- **Cá nhân hóa trải nghiệm học tập:** Giúp người dùng nhanh chóng tìm được các khóa học phù hợp với mục tiêu học tập và sở thích cá nhân. Tăng cường trải nghiệm người dùng.
- **Thu hút người dùng:** Các gợi ý chính xác và kịp thời có thể dẫn đến tỷ lệ đăng ký khóa học cần thiết cao hơn và cải thiện sự gắn bó của người dùng.



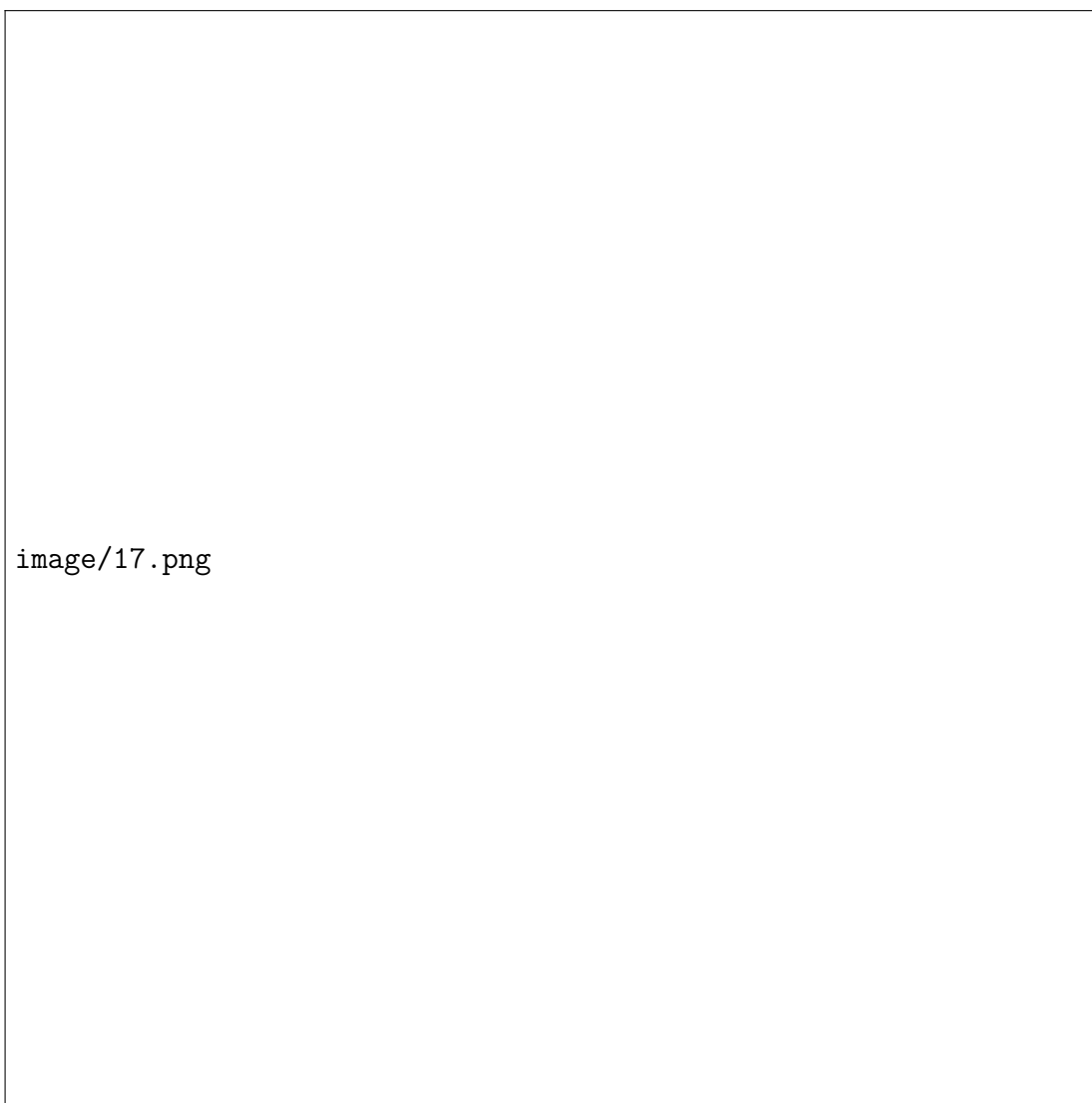


với nền tảng. Các khóa học phù hợp và hấp dẫn có thể giúp giảm tỷ lệ người học từ bỏ giữa chừng, cải thiện tỷ lệ hoàn thành khóa học.

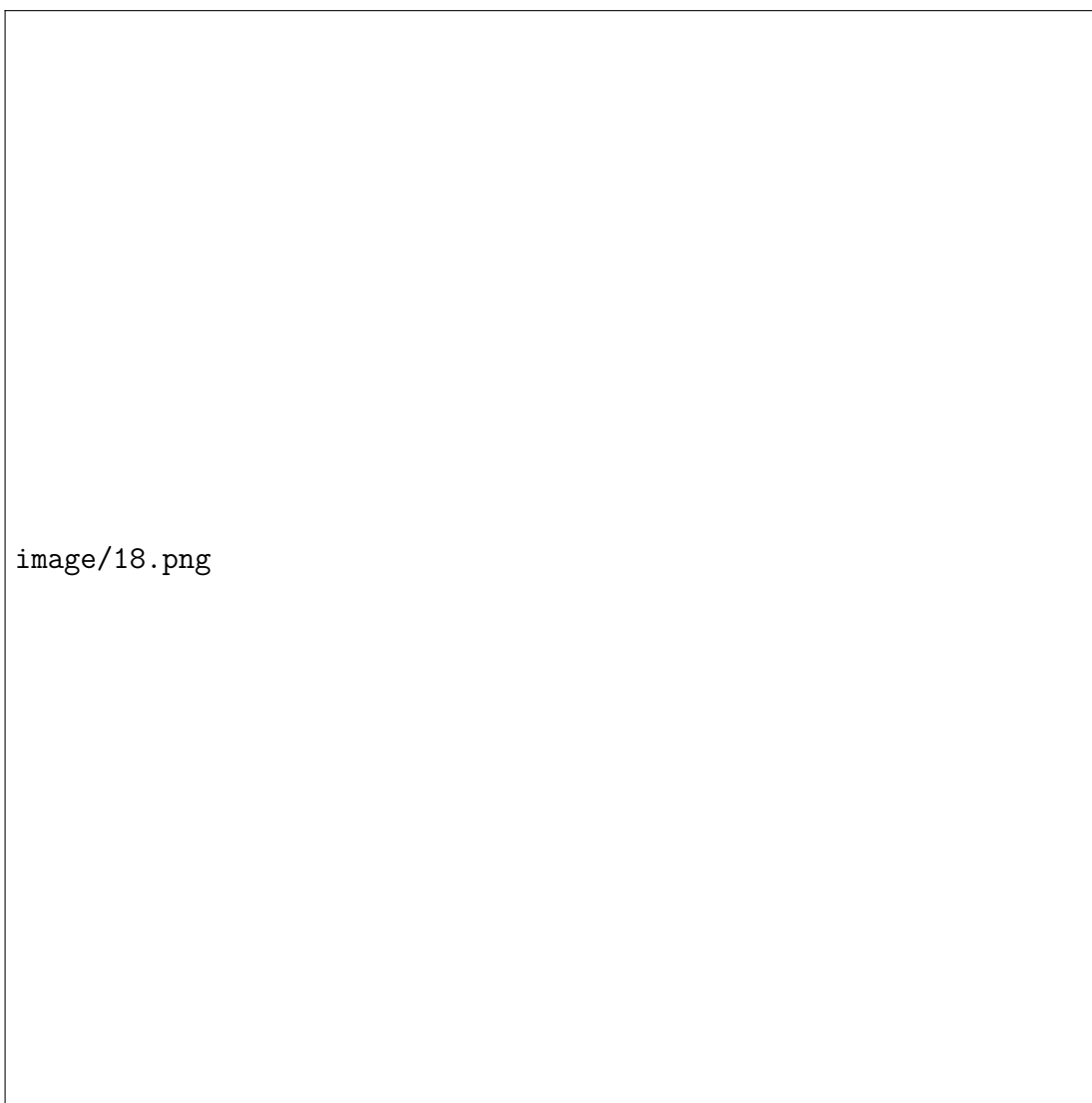
- **Nâng cao hiệu suất học tập của người dùng:** Từ những hành vi học tập của người dùng trong quá khứ, hệ thống sẽ căn cứ vào và tự động đề xuất các khóa học tương thích nhất với khả năng và kỹ năng của người học để tối ưu hóa nhất hiệu suất học tập của người dùng.



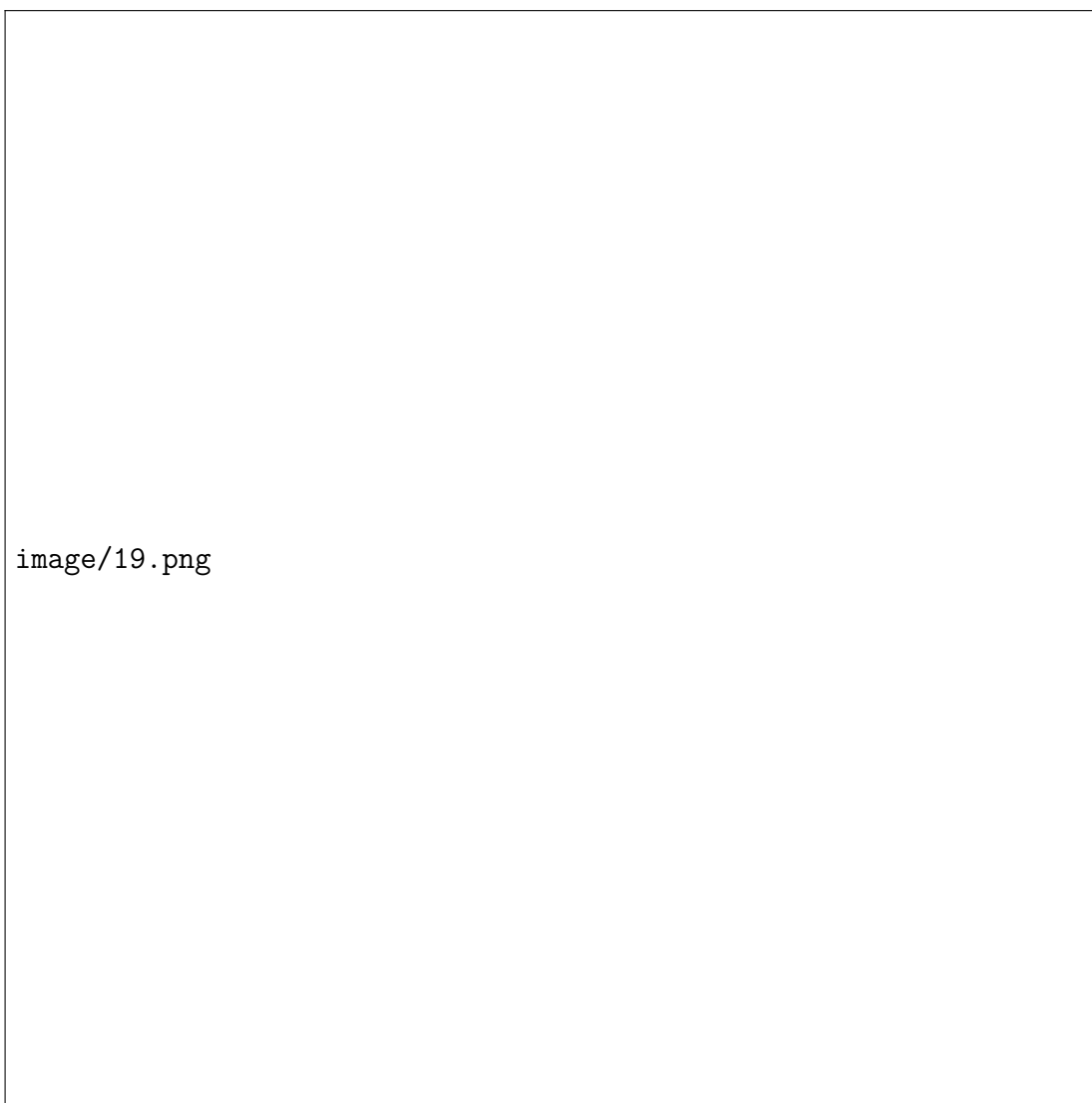
Hình 1: Số lượng users



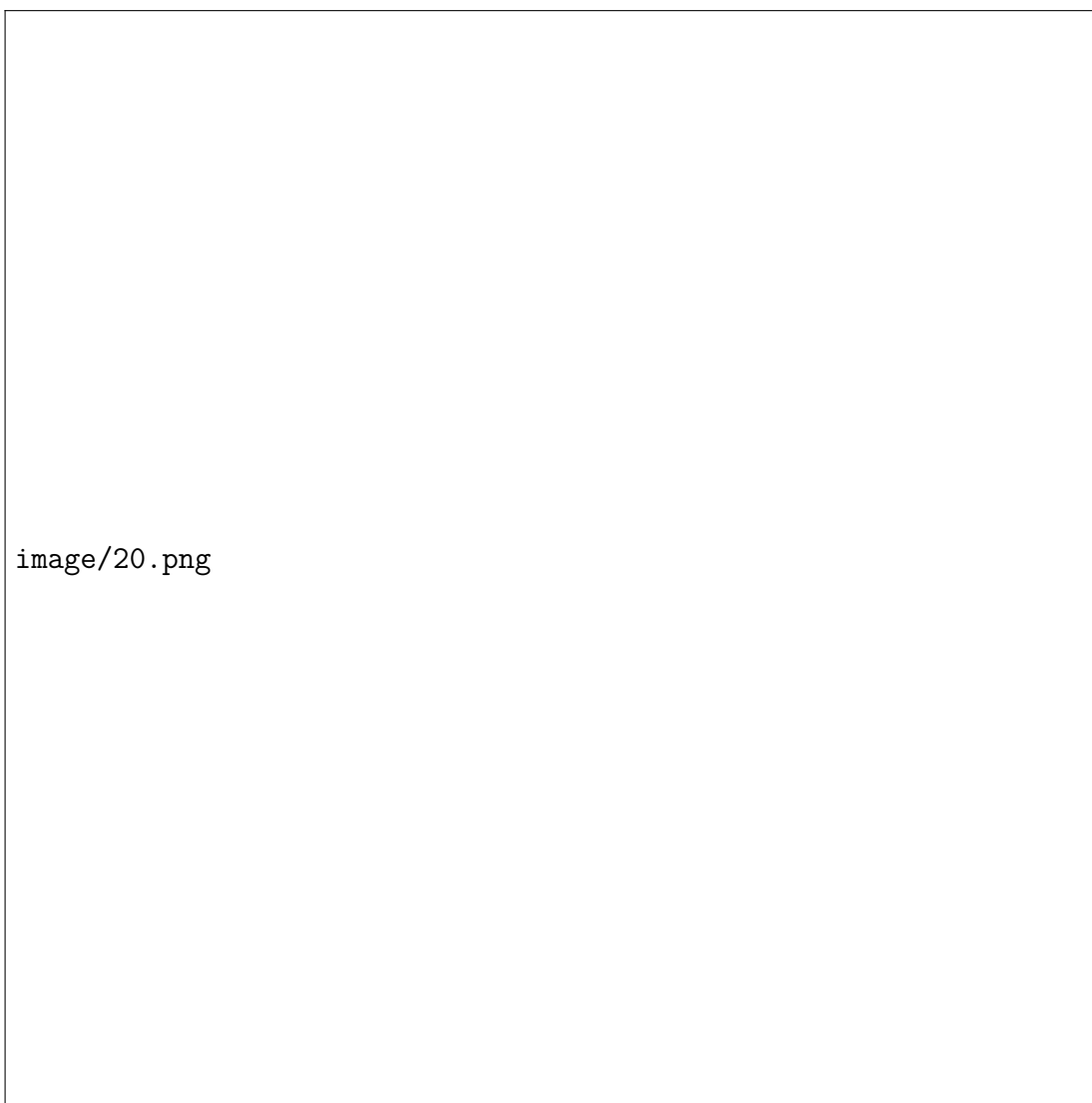
Hình 2: Cột “gender”



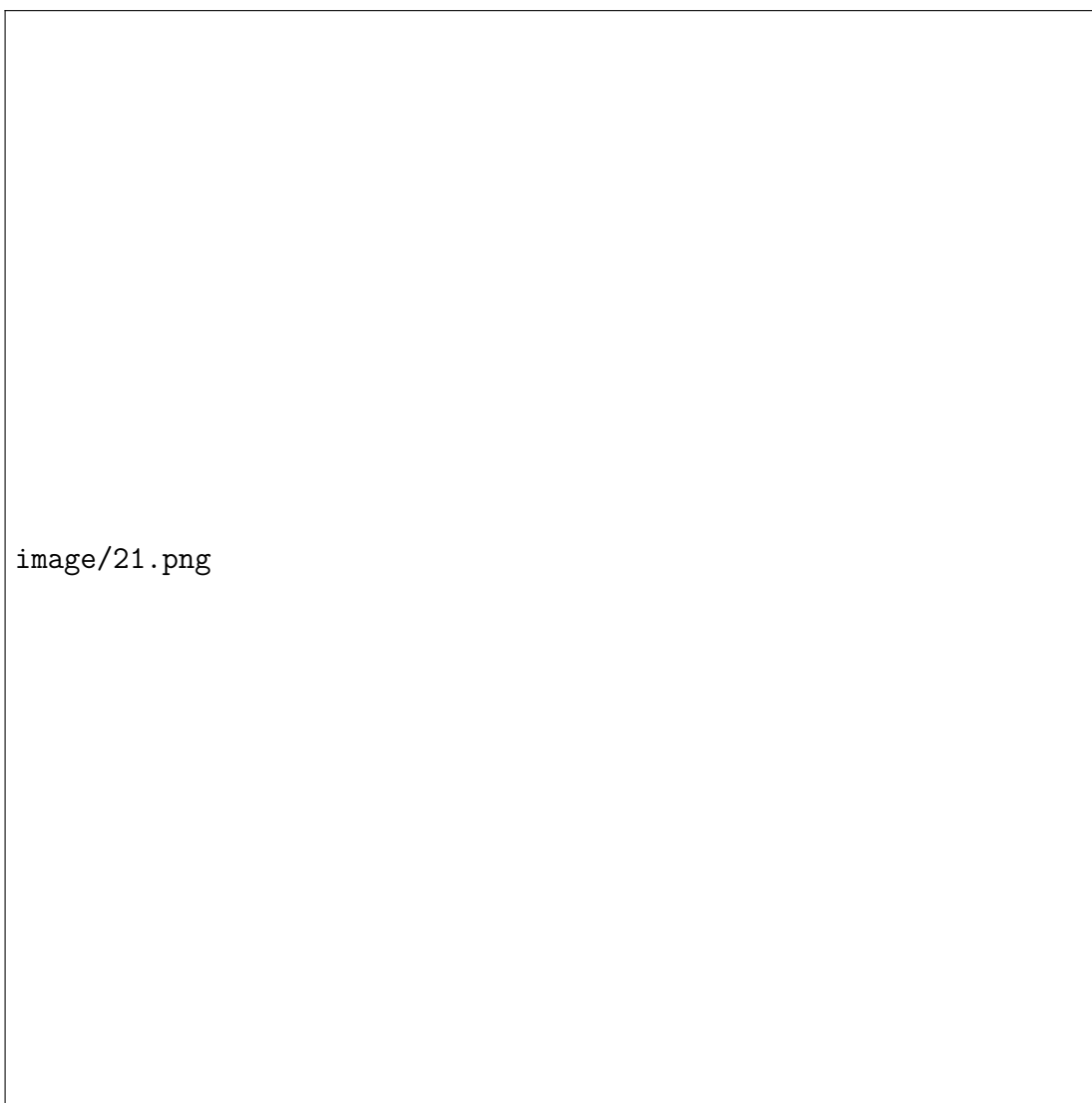
Hình 3: Phân bố các các giá trị trong cột “gender”:



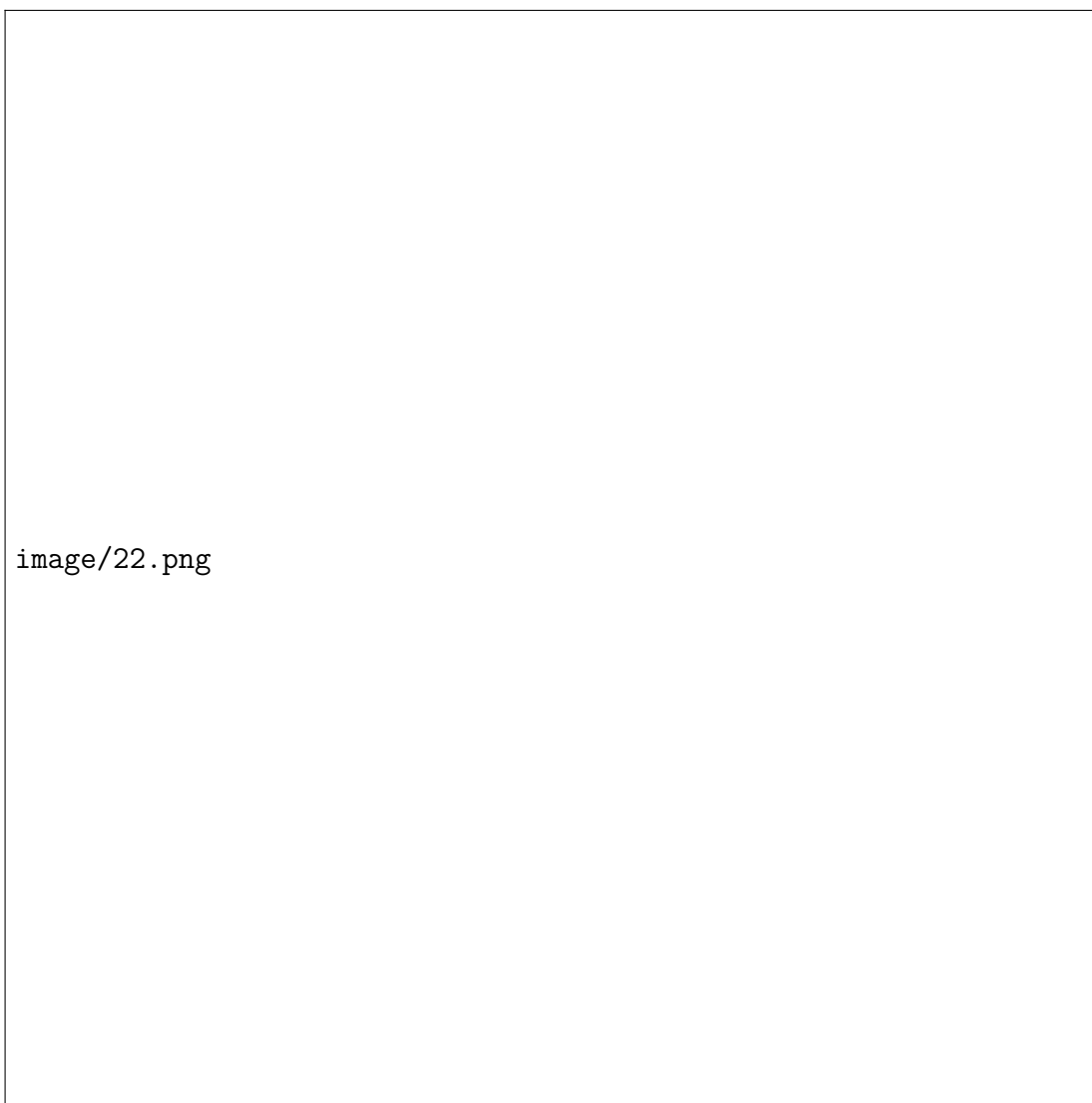
Hình 4: Thông tin cột “school”



Hình 5: Số lượng trường học trong bảng

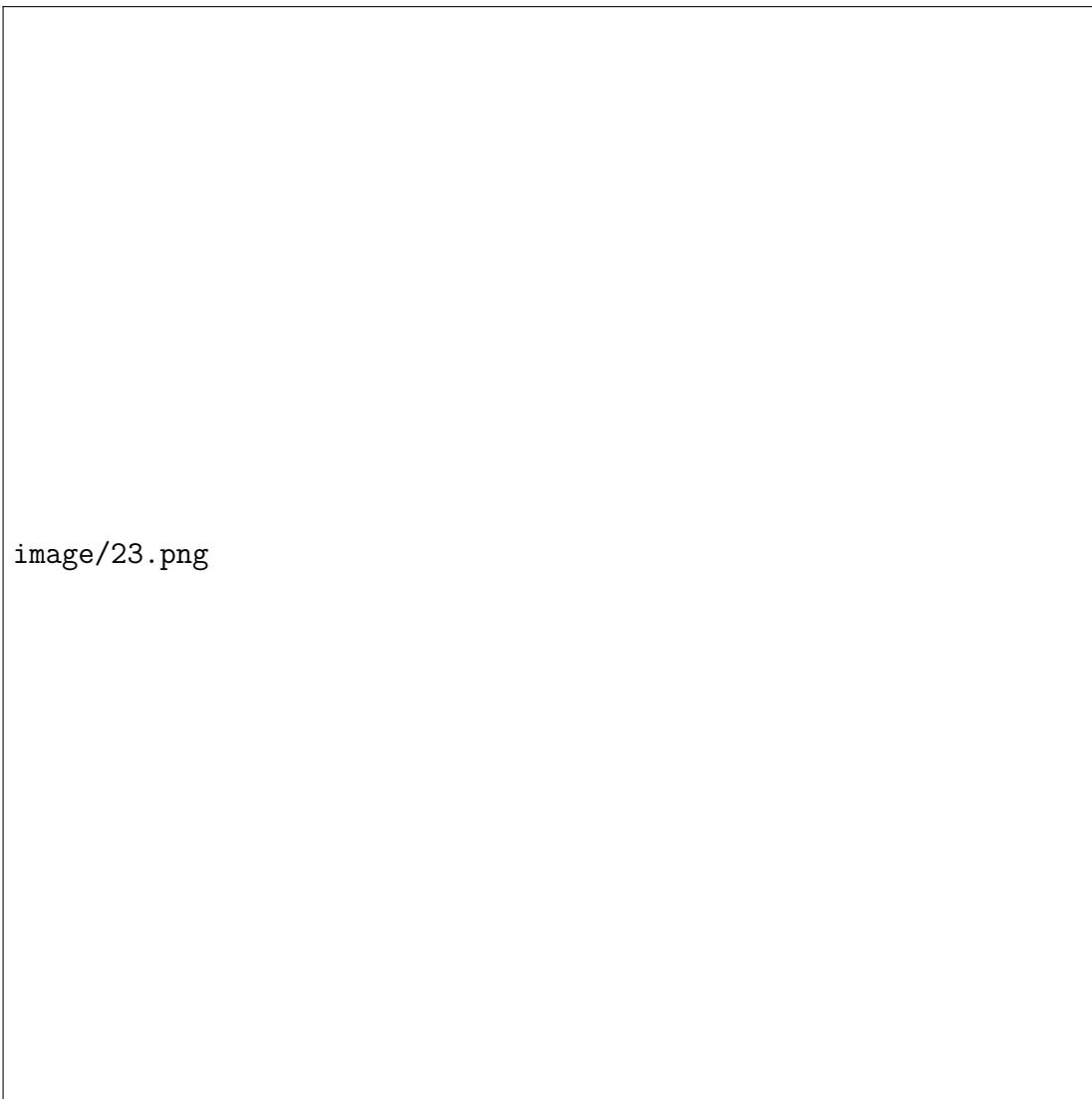


Hình 6: Kiểm tra thông tin tổng quan sau cùng

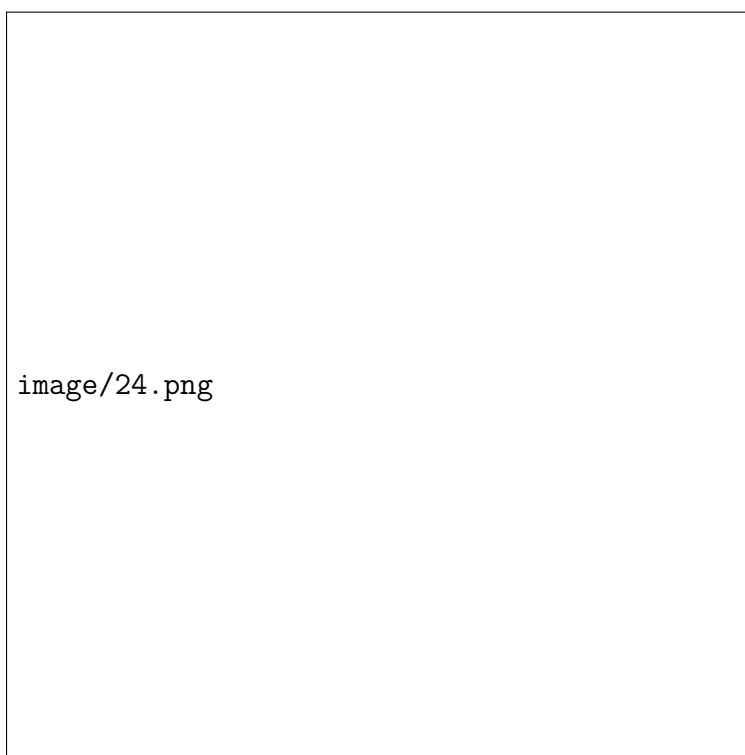


Hình 7: Số lượng sample (users) có trong bảng và số lượng users thuộc về mỗi trường học

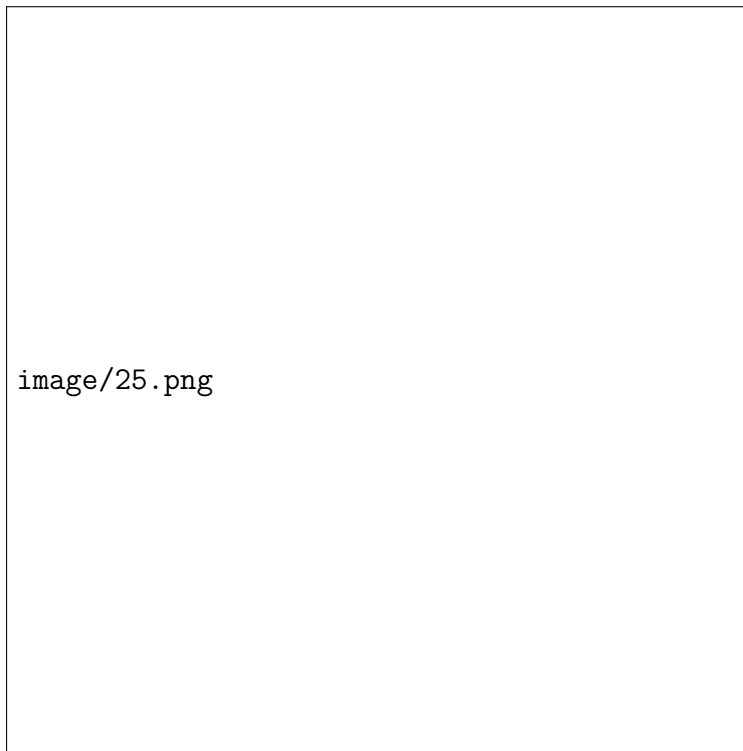




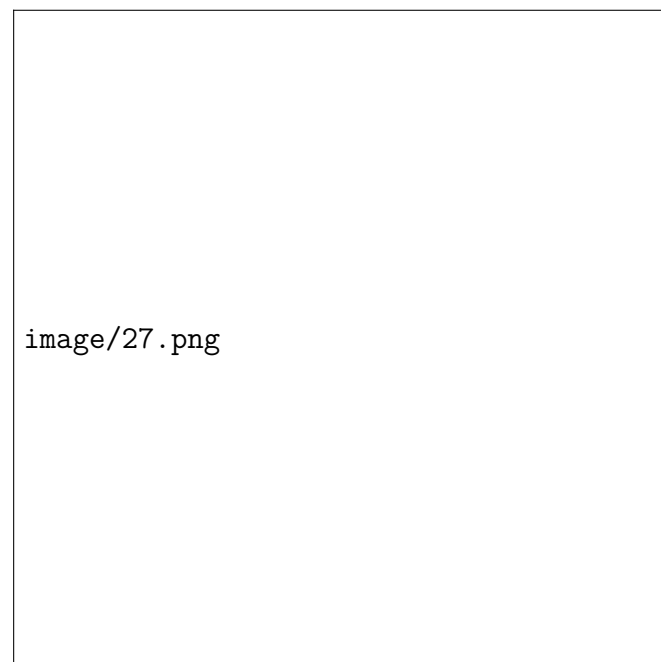
Hình 8: Tạo một cột “school\_length” để phân tích độ dài mỗi sample của cột



Hình 9: Trực quan hóa độ dài của sample cột “school”



Hình 10: Trực quan hóa phân bố các giá trị của cột “gender”





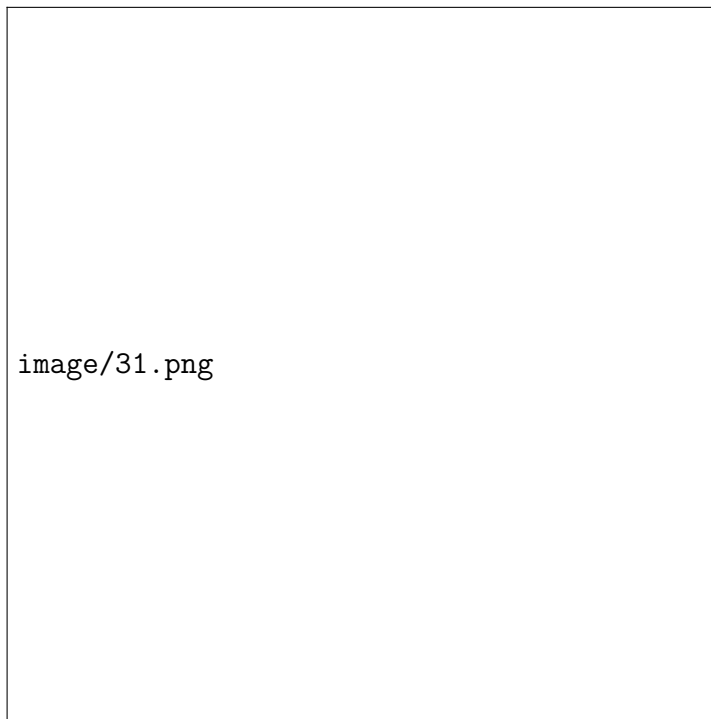
image/28.png



image/29.png



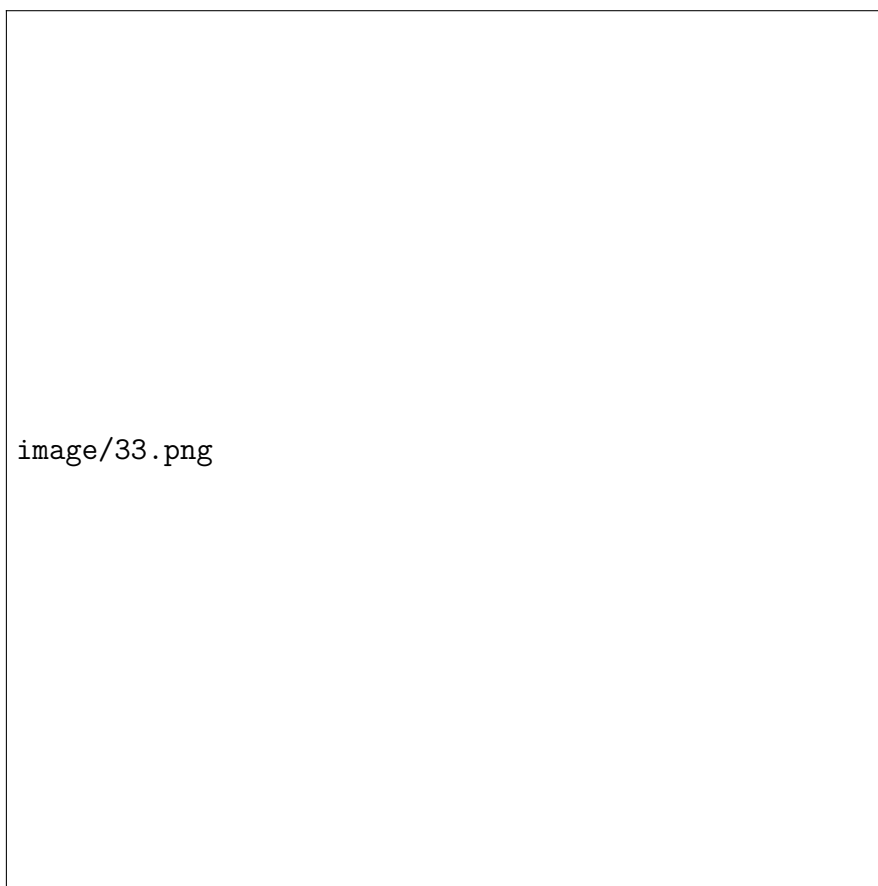
image/30.png





Hình 11: Histogram thể hiện số lượng khóa học của mỗi teacher và bảng thống kê mô tả tương ứng





Hình 12: Histogram thể hiện số lượng teacher của mỗi khóa học và bảng thống kê mô tả tương ứng



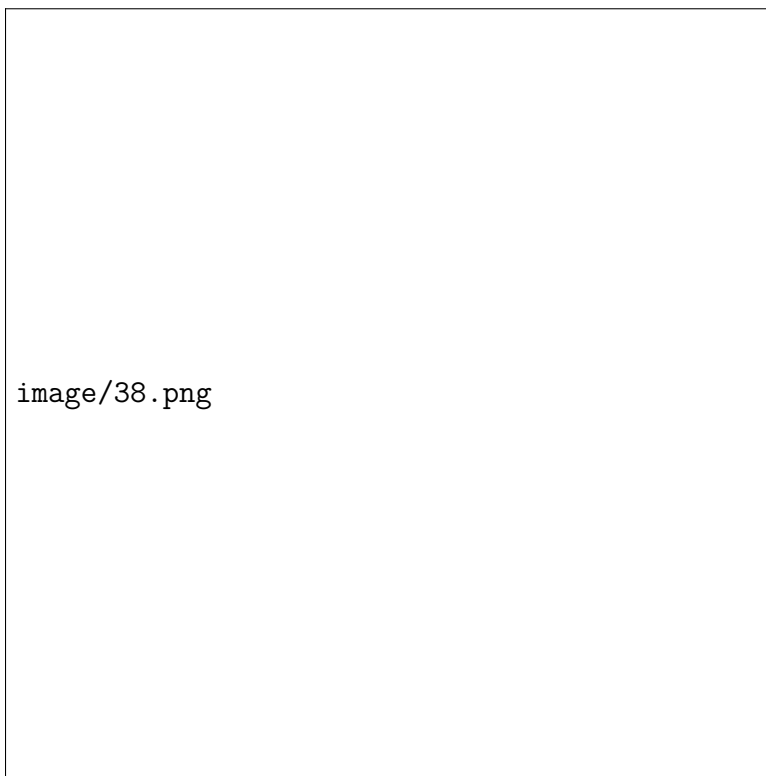
image/34.png

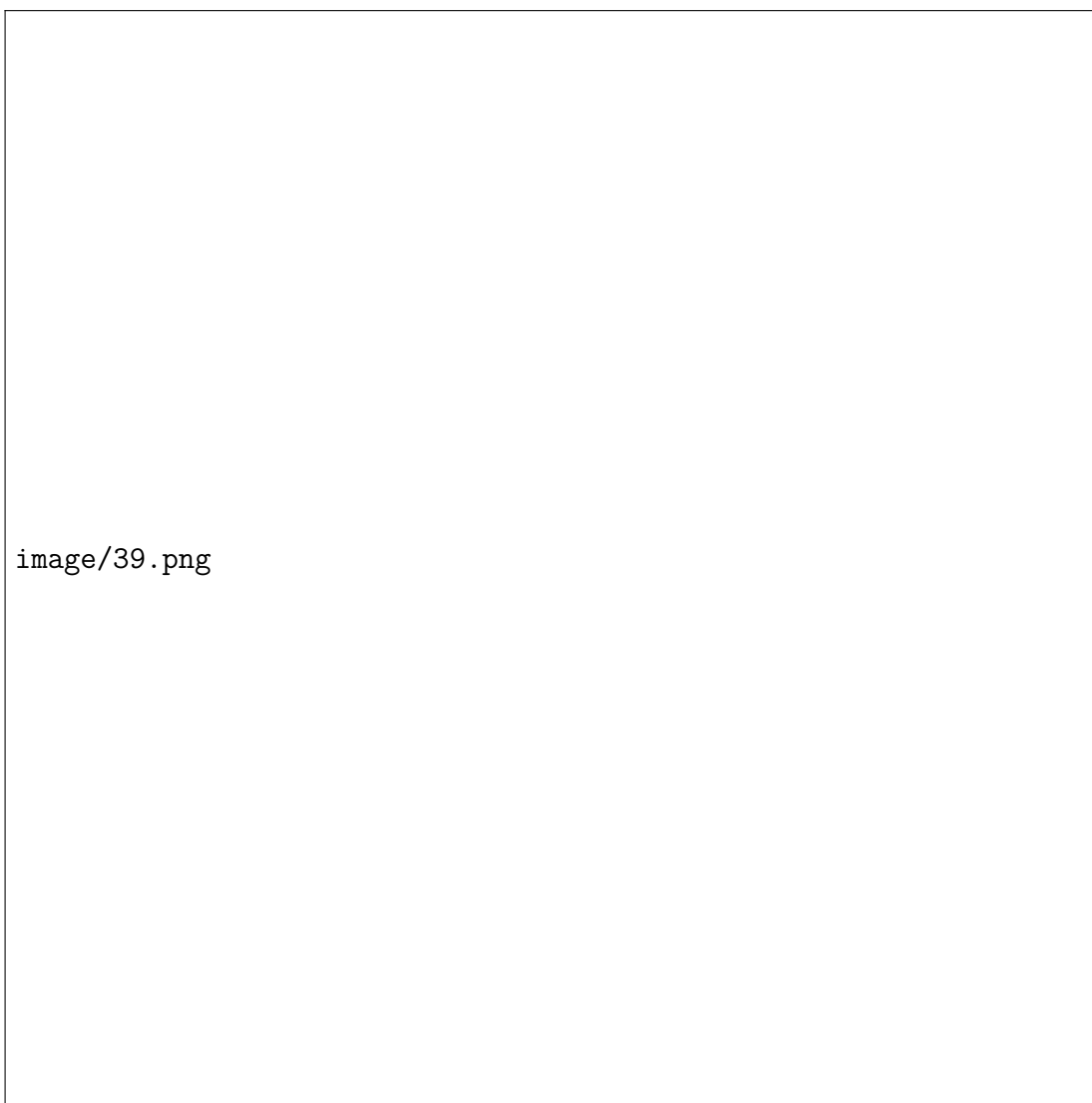
image/35.png



image/36.png

image/37.png





Hình 13: Tổng số lượng khóa học và tổng số lượng các lĩnh vực khác nhau

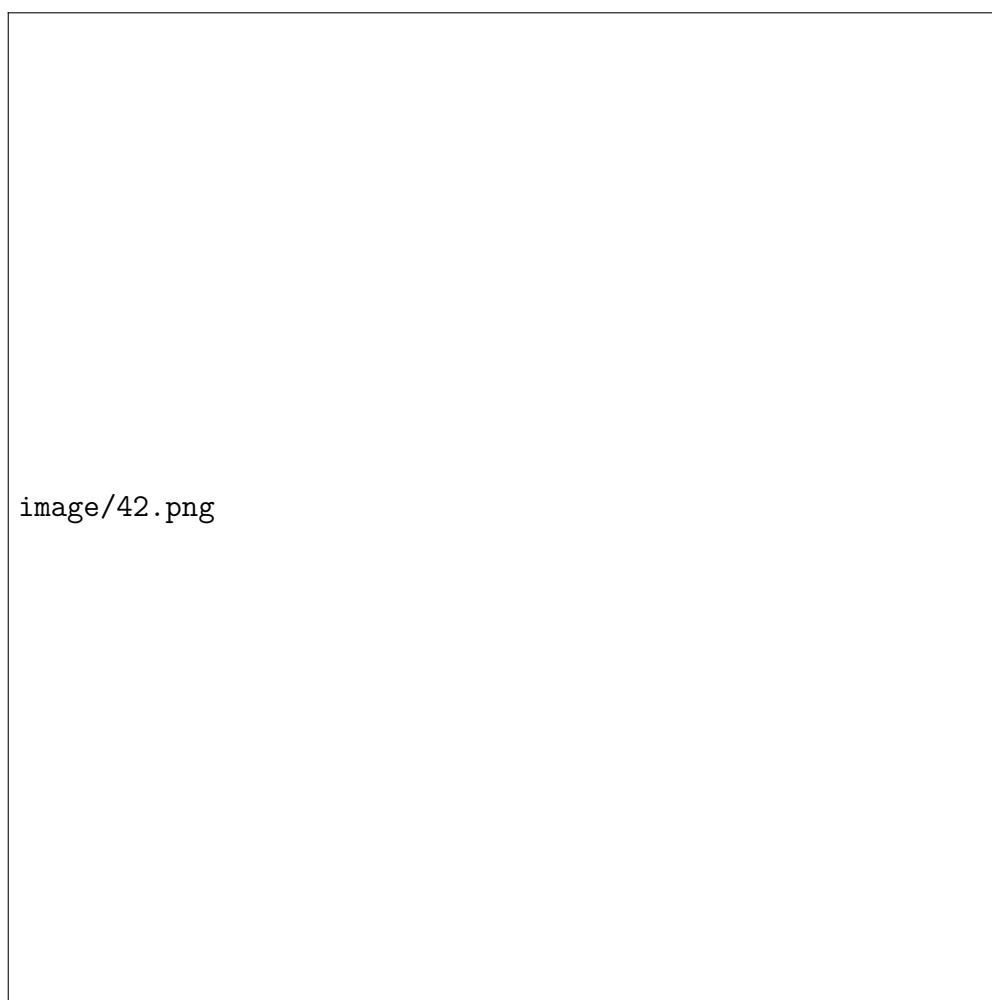


Hình 14: Phân bố số lượng khóa học theo từng lĩnh vực



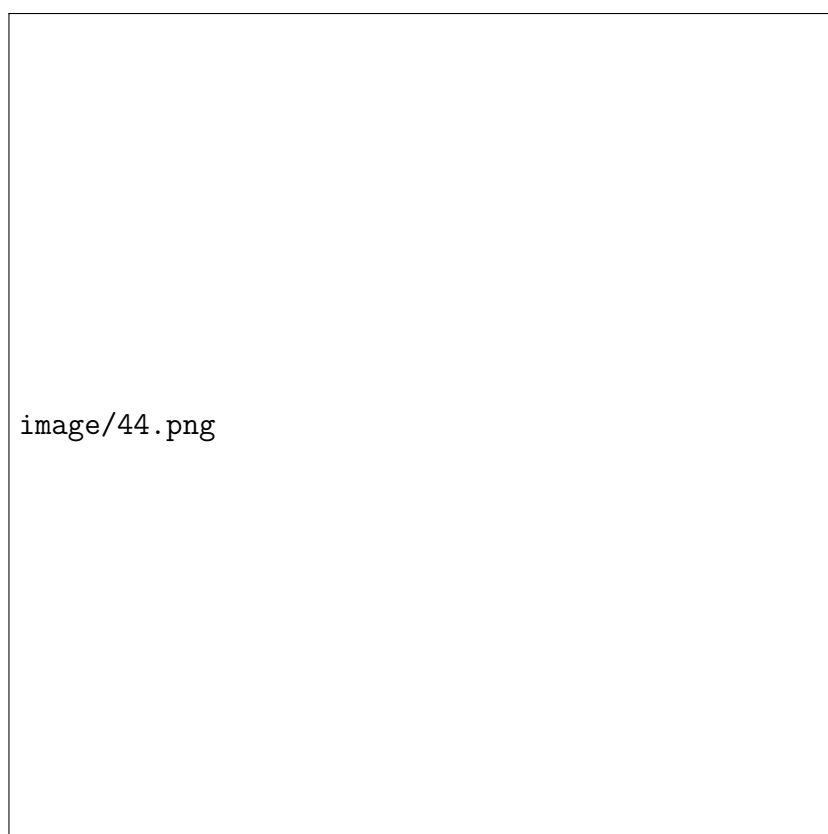
image/41.png

Hình 15: Phân bố độ dài tên khóa học



Hình 16: Biểu đồ thanh thể hiện sự phân bố số lượng khóa học theo từng lĩnh vực





Hình 17: Biểu đồ phân phối cho độ dài tên khóa học



image/26.png



image/45.png



image/46.png



image/47.png



image/48.png



image/49.png



image/50.png

image/51.png





image/52.png

image/53.png

