

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH



ĐỀ TÀI: HỆ THỐNG KHUYẾN NGHỊ KHÓA HỌC
CHO NỀN TẢNG HỌC TẬP TRỰC TUYẾN

BÁO CÁO PHÂN TÍCH BỘ DỮ LIỆU

MÔN HỌC: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG (CS313)

Nhóm 4

GVHD

ThS. Nguyễn Anh Thư

TP. HO CHI MINH, 6/2024

DANH SÁCH THÀNH VIÊN

STT	Họ và tên	MSSV
1	Đoàn Nhật Sang	21522542
2	Trương Văn Khải	21520274
3	Lê Ngô Minh Đức	21520195
4	Phạm Minh Quốc	22540017
5	Lê Yên Nhi	21522427
6	Hoàng Thị Mỹ Hạnh	21522044
7	Hoàng Tiến Đạt	21520696
8	Lê Minh Quang	21522510

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn cô giáo, ThS. Nguyễn Thị Anh Thư, người đã định hướng, giúp đỡ, trực tiếp hướng dẫn và tận tình chỉ bảo chúng tôi trong suốt quá trình nghiên cứu, xây dựng và hoàn thành đồ án này.

Chúng tôi cũng xin được cảm ơn tới gia đình, những người thân và bạn bè thường xuyên quan tâm, động viên, chia sẻ kinh nghiệm, cung cấp các tài liệu hữu ích trong thời gian học tập, nghiên cứu cũng như trong suốt quá trình thực hiện đồ án.

TP. HCM, ngày 6 tháng 6 năm 2024

ĐÁNH GIÁ CỦA GIẢNG VIÊN HƯỚNG DẪN

TP. HCM, ngày 6 tháng 6 năm 2024

GVHD

Nguyễn Anh Thu

MỤC LỤC

1	TÌM HIỂU DỮ LIỆU.....	1
1.1	Bộ dữ liệu sử dụng	1
1.2	Giới thiệu bộ dữ liệu sử dụng	1
1.3	Mô tả sơ bộ về tập dữ liệu	3
1.3.1	Course resources	4
1.3.2	Student behaviors.....	11
1.3.3	Concepts.....	17
1.3.4	Prerequisites	21
1.4	Nhận xét bộ dữ liệu và dự đoán mục tiêu xử dụng của bộ dữ liệu.....	22
1.4.1	Nhận xét	22
1.4.2	Dự đoán mục tiêu sử dụng bộ dữ liệu	23
2	CHUẨN BỊ DỮ LIỆU	24
2.1	Dữ liệu thực nghiệm.....	24
2.2	Phương pháp tổ chức dữ liệu thực nghiệm.....	24
2.2.1	Data translation	24
2.2.2	EDA, làm sạch dữ liệu	25
2.2.3	Feature selection	40
2.2.4	Data preprocessing	40
2.2.5	Data splitting	42
3	PHÂN TÍCH VÂN ĐÈ.....	43
3.1	Câu hỏi nghiên cứu.....	43
3.2	Kết quả đề tài.....	43

3.3	Khả năng ứng dụng	44
4	TÀI LIỆU THAM KHẢO.....	46

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Từ chuẩn	Điễn giải
KGAT	Knowledge Graph Attention	Cơ chế tập trung cho đồ thị tri thức
KG	Knowledge Graph	Đồ thị tri thức
CSDL	Cơ sở dữ liệu	Cơ sở dữ liệu

DANH MỤC HÌNH ẢNH

Hình 1.1 Hình minh họa trường dữ liệu bình luận của học sinh.....	13
Hình 2.1 Hình minh họa quy trình xử lý dữ liệu của nhóm	24
Hình 2.2 Bảng thống kê mô tả của course.json.....	25
Hình 2.3 Thống kê cơ bản về số từ trong about, name sau khi segmented (len_about_segmented, len_name_segmented)	26
Hình 2.4 Biểu đồ cột thể hiện độ dài văn bản ở trường about, name.....	26
Hình 2.5 Bảng thống kê mô tả của course-field.json	27
Hình 2.6 Histogram thể hiện số lượng fields trong mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải)	27
Hình 2.7 Biểu đồ cột thể hiện số lượng khóa học của các field (bên trái) và bảng thống kê mô tả tương ứng (bên phải).	28
Hình 2.8 Bảng thống kê mô tả của concept.json.....	28
Hình 2.9 Histogram thể hiện số lượng concept của mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải).	29
Hình 2.10 Histogram thể hiện số lượng khóa học của mỗi concept (bên trái) và bảng thống kê mô tả tương ứng (bên phải).	29
Hình 2.11 Bảng thống kê mô tả của school.json.....	30
Hình 2.12 Histogram thể hiện số lượng khóa học của mỗi trường (bên trái) và bảng thống kê mô tả tương ứng (bên phải).	30
Hình 2.13 Bảng thống kê mô tả cho số lượng trường học của mỗi khóa học.....	31
Hình 2.14 Bảng thống kê mô tả của teacher.json.....	31
Hình 2.15 Histogram thể hiện số lượng teachers của mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải).	32

Hình 2.16 Histogram thể hiện số lượng khóa học của mỗi teacher (bên trái) và bảng thống kê mô tả tương ứng (bên phải)	33
Hình 2.17 Bảng thống kê mô tả số lượng giáo viên của một tổ chức.....	33
Hình 2.18 Bảng thống kê mô tả của user.json	34
Hình 2.19 Value counts của trường gender của User.	35
Hình 2.20 Biểu đồ thể hiện số lượng User ứng với từng giới tính.....	35
Hình 2.21 Biểu đồ thể hiện tổng số lượng khóa học đăng ký của mỗi nhóm giới tính	36
Hình 2.22 Histogram thể hiện số lượng videos của mỗi khóa học(trái) và bảng thống kê mô tả tương ứng (phải).	37
Hình 2.23 Bảng thống kê mô tả của video.json	37
Hình 2.24 Thực hiện kiểm định phương sai ANOVA để xem số lượng khóa học đăng kí có bị phụ thuộc vào nhóm giới tính hay không.....	38
Hình 2.25 Một phần của tập luật kết hợp	39
Hình 2.26 Các khóa học thường xuất hiện cùng với khóa học 746997	39
Hình 2.27 Sơ đồ phân rã của quy trình tiền xử lý dữ liệu	41
Hình 3.1 Hình minh họa kết quả sản phẩm website hệ thống khuyến nghị của nhóm	44

DANH MỤC BẢNG

Bảng 1.1 Bảng thống kê mô tả bộ dữ liệu sử dụng	2
Bảng 1.2 Bảng mô tả các trường dữ liệu của course.json	5
Bảng 1.3 Bảng mô tả các trường dữ liệu của video.json	6
Bảng 1.4 Bảng mô tả các trường dữ liệu của problem.json	8
Bảng 1.5 Bảng mô tả các trường dữ liệu của school.json	8
Bảng 1.6 Bảng mô tả các trường dữ liệu của teacher.json	9
Bảng 1.7 Bảng mô tả các trường dữ liệu của course-field.json	10
Bảng 1.8 Bảng mô tả các trường dữ liệu của user.json	12
Bảng 1.9 Bảng mô tả các trường dữ liệu của comment.json	12
Bảng 1.10 Bảng mô tả các trường dữ liệu của reply.json	13
Bảng 1.11 Bảng mô tả các trường dữ liệu của user-video.json	14
Bảng 1.12 Bảng mô tả các trường dữ liệu của user-problem.json	15
Bảng 1.13 Bảng mô tả các trường dữ liệu của user-xiaomu.json	16
Bảng 1.14 Bảng mô tả các trường dữ liệu của concept.json	17
Bảng 1.15 Bảng mô tả các trường dữ liệu của other.json	18
Bảng 1.16 Bảng mô tả các trường dữ liệu của paper.json	20
Bảng 1.17 Bảng mô tả các trường dữ liệu của CS.json	22
Bảng 5.1 Bảng thống kê số lượng của từng thực thể sau khi xử lý dữ liệu	41
Bảng 5.2 Bảng thống kê chi tiết từng loại liên kết sau khi thực hiện N-core filtering	41

1 TÌM HIỂU DỮ LIỆU

1.1 Bộ dữ liệu sử dụng

Sau quá trình khảo sát và tìm hiểu, chúng tôi trước quyết định sử dụng bộ dữ liệu: MOOCCubeX [1]. Đây là bộ dữ liệu ở bài báo: **MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs** tại hội nghị CIKM '21 năm 2021.

1.2 Giới thiệu bộ dữ liệu sử dụng

Bộ dữ liệu được thu thập từ nền tảng XuetangX [2] - Đây là một trong những đối tác của edX. Tuy hệ thống ra mắt vào tháng 10 năm 2013 nhưng đến ngày 31 tháng 5 năm 2021 đã cung cấp hơn 6.000 khóa học, bao gồm các khóa từ Đại học Thanh Hoa, Đại học Bắc Kinh và các khóa học của edX từ MIT, Stanford, UC Berkeley, ... thu hút 4.500.000 người dùng đăng ký. XuetangX cung cấp đa dạng tài nguyên học tập, cho phép người dùng tự do ghi danh vào các khóa học và tham gia vào quá trình học đầy đủ bao gồm học qua video, làm bài tập và tham gia thảo luận. Các dữ liệu này có mối liên hệ chặt chẽ và được quản lý tốt, nên thường được sử dụng làm cơ sở lý tưởng cho MOOCCubeX. Sau đây là bảng thống kê số lượng chi tiết của từng loại tài nguyên:

Tên tài nguyên	Số lượng
Tài nguyên khóa học	3,781 khóa học
Tài nguyên video	59,581 video
Tài nguyên vấn đề	2,454,422 vấn đề
Tài nguyên trường học	429 trường học
Tài nguyên giảng viên	17,018 giảng viên
Tài nguyên Trường học – Lĩnh vực	632 quan hệ

Tài nguyên Khóa học – Trường học	3,983 quan hệ
Tài nguyên Khóa học – Giảng viên	97,192 quan hệ
Tài nguyên phản hồi bình luận	331,011 phản hồi
Tài nguyên User – Xiaomu	108,351 quan hệ
Tài nguyên Course – Comment	10,181,950 quan hệ
Tài nguyên User – Comment	8,422,134 quan hệ
Tài nguyên User – Reply	331011 quan hệ
Tài nguyên Comment - Reply	370,493 quan hệ
Tài nguyên Concepts	637,572 concepts
Tài nguyên Other	210,349 mẫu
Tài nguyên Concept - Other	379,926 quan hệ
Tài nguyên Concept - Paper	5,410,752 quan hệ
Tài nguyên Concept-Problem	33,180 quan hệ
Tài nguyên Concept-Video	624,683 quan hệ
Tài nguyên Concept-Comment	31,074 quan hệ
Tài nguyên CS	492,102 mẫu
Tài nguyên Math	331202 mẫu
Tài nguyên Psy	757,771 mẫu

Bảng 1.1 Bảng thống kê mô tả bộ dữ liệu sử dụng

1.3 Mô tả sơ bộ về tập dữ liệu

Bộ dữ liệu gồm hai phần chính: Tài nguyên khóa học (Course Resource) và Hành vi học sinh (Student Behavior)

Course Resource:

Phần Tài nguyên khóa học của MOOCCubeX bắt đầu bằng việc thu thập dữ liệu khóa học từ XuetangX. Sau khi loại bỏ các khóa học thử nghiệm và khóa học không còn hoạt động, thông tin chi tiết về 4.216 khóa học đã được thu thập. Ở giai đoạn này, tên và mô tả của mỗi khóa học được lưu trữ dưới dạng văn bản, và mỗi khóa học được gán một mã id. Các khóa học trong MOOCs không độc lập với nhau. Một khóa học bao gồm nhiều chương giảng dạy, và một chương thường bao gồm một loạt video và bài tập. Thông tin có cấu trúc như vậy cũng rất quan trọng, do đó, việc thu thập thông tin liên quan đến khóa học, bao gồm giáo trình của khóa học và danh sách tài nguyên bao gồm (video, bài tập, và bình luận) được lưu trữ dưới dạng danh sách. Ngoài ra, thông tin về giáo viên và trường đại học của khóa học, cùng với giới thiệu về họ được thu thập từ web. Loại thông tin này có thể xây dựng các mối liên kết cho các khóa học và hỗ trợ các nhiệm vụ liên quan như phát hiện phong cách giảng dạy.

Student behaviours:

Ngoài các nguồn tài nguyên tĩnh, các loại hành vi của sinh viên cũng rất quan trọng cho nghiên cứu học tập thích nghi, giúp mô hình hóa ý định học tập của sinh viên ở các cấp độ nhận thức và các hoạt động xã hội. Do đó, tác giả thu thập các bản ghi chi tiết từ XuetangX, bao gồm: hồ sơ sinh viên, hành vi xem video, bài tập và thảo luận. Các hành vi này tự nhiên liên kết với các nguồn tài nguyên của khóa học. Mặc dù đã có giấy phép từ nền tảng, tác giả vẫn cần thực hiện các hoạt động giảm nhẹ cảm như ẩn danh trong quá trình xử lý dữ liệu.

Mô tả chi tiết của các bảng dữ liệu trong MOOCCubeX như các mục sau:

1.3.1 Course resources

1.3.1.1 Course Info (entity)

- Mô tả: Thông tin của khóa học và các tài liệu tương ứng
- Tên file: **entities/course.json**
- Số lượng mẫu: **3781** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
about	Giới thiệu khóa học	string	
id	id của khóa học	string	
field	danh sách các lĩnh vực mà khóa học thuộc về	array của các string	
name	tên của khóa học	string	
prerequisites	mô tả những kiến thức tiên quyết	string	
resource	danh sách các tài nguyên, có thể là một video hoặc một nhóm các bài tập. object gồm các trường: <ul style="list-style-type: none">• resource_id (string): ID của khóa học. Nếu bắt đầu bằng “V_” thì nó là một Video (resource_id được xem như video_id), nếu bắt đầu	list của các object	

	<p>bảng “Ex_” thì nó là một nhóm Exercise (resource_id được xem như exercise_id). Xem thêm lưu ý dưới bảng này.</p> <ul style="list-style-type: none"> • chapter (string): số chapter • titles (list<string>): danh sách các tựa đề, gồm chapter title, video title, v.v. 		
--	---	--	--

Bảng 1.2 Bảng mô tả các trường dữ liệu của course.json

Lưu ý:

- Nhiều video_ids tương ứng với 1 ccid (1 ccid xác định 1 video). Những video_id này tương ứng với phần trình chiếu của cùng 1 ccid tại những thời điểm bắt đầu khác nhau, ví dụ (spring 2018 / fall 2020). Sự tương ứng giữa video_id và ccid có thể tìm thấy ở mục 1.3.1.10.
- Mỗi tập hợp exercises (exercise) tương ứng với nhiều câu hỏi (problem). Sự tương ứng giữa chúng có thể tìm thấy ở mục 1.3.1.9.

1.3.1.2 Video (entity)

- Mô tả: Tên video và chú thích. Nội dung, khóa học, chương và thứ tự của video có thể tìm thấy trong file course.json
- Tên file: **entities/video.json**
- Số lượng mẫu: **59581** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
ccid	id duy nhất của video	string	

name	tên của khóa học	string	
start	thời gian bắt đầu mỗi câu phụ đề của video	list<float>	[0, +vc)
end	thời điểm mỗi câu của phụ đề video kết thúc	list<float>	[0, +vc)
text	nội dung phụ đề từng câu trong video	list<string>	

Bảng 1.3 Bảng mô tả các trường dữ liệu của video.json

1.3.1.3 Problem (entity)

- Mô tả: câu hỏi (problem) thực hành của nhóm bài tập. Lưu ý: Mỗi nhóm bài tập sẽ tương ứng với nhiều câu hỏi (problem).
- Tên file: **entities/problem.json**
- Số lượng mẫu: **2454422** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
problem_id	id của vấn đề, bắt đầu bằng Pm_	int	
excercise_id	id của bài tập, bắt đầu bằng Ex_	string	

language	ngôn ngữ mô tả của vấn đề	string	{‘English’, ‘Chinese’}
title	tựa của bài tập	string	
content	mô tả vấn đề	string	
option	lựa chọn của vấn đề	object chứa các cặp key, value. Key có kiểu string, là ký hiệu A, B, C, ... của lựa chọn. Value có kiểu string, là mô tả của lựa chọn	Key nhận các giá trị như: “A”, “B”, “C”, “D”, ...
answer	câu trả lời cho câu hỏi	string	
score	điểm cho câu hỏi	float	
type	loại câu hỏi	int	
typetext	loại câu hỏi	string	
location	chương của vấn đề	string	

context_id	các leaf_id liên quan đến vấn đề	list<int>	
------------	--	-----------	--

Bảng 1.4 Bảng mô tả các trường dữ liệu của problem.json.

1.3.1.4 School (entity)

- Mô tả: thông tin của trường học
- Tên file: **entities/school.json**
- Số lượng mẫu: **429** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID trường, bắt đầu với s_	string	
name	Tên tiếng Trung của trường	string	
name_en	Tên tiếng Anh của trường	string	
sign	Tên viết tắt của tên tiếng Anh của trường	string	
about	Giới thiệu	string	
motto	Châm ngôn của trường	string	

Bảng 1.5 Bảng mô tả các trường dữ liệu của school.json

1.3.1.5 Teacher (entity)

- Mô tả: thông tin của giáo viên
- Tên file: **entities/teacher.json**

- Số lượng mẫu: **17018** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID của giáo viên, bắt đầu với T_	string	
name	Tên tiếng Trung của giáo viên	string	
name_en	Tên tiếng Anh của giáo viên	string	
about	Hồ sơ của giáo viên	string	
job_title	chức danh công việc	string	
org_name	tổ chức liên kết	string	

Bảng 1.6 Bảng mô tả các trường dữ liệu của teacher.json

1.3.1.6 Course - Field (relation)

- Mô tả: Các lĩnh vực mà khóa học thuộc về. Các lĩnh vực của các khóa học được đánh dấu thủ công được lấy từ 88 lĩnh vực trong “Danh mục ngành, chuyên ngành cấp bằng tiến sĩ, thạc sĩ và đào tạo sau đại học” do Bộ Giáo dục ban hành năm 1997. Mỗi khóa học có thể thuộc nhiều lĩnh vực.
- Tên file: **relations/course-field.json**
- Số lượng mẫu: **632** mẫu

Trường	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 đánh nhãn thủ công	list<str>, mỗi string đại diện cho một lĩnh vực	

Bảng 1.7 Bảng mô tả các trường dữ liệu của course-field.json

1.3.1.7 Course - School (relation)

- Mô tả: Trường mà khóa học tương ứng được dạy.
- Định dạng: {course ID}\t{school ID}
- Tên file: **relations/course-school.txt**
- Số lượng mẫu: **3983** mẫu

1.3.1.8 Course - Teacher (relation)

- Giảng viên của khóa học.
- Định dạng: {course ID}\t{teacher ID}
- Tên file: **relations/course-teacher.txt**
- Số lượng mẫu: **97192** mẫu

1.3.1.9 Exercise - Problem (relation)

- Mô tả: Tập hợp các vấn đề (câu hỏi) chứa trong bài tập.
- Định dạng: {exercise ID}\t{question ID}
- Mỗi dòng trong bộ Exercise là một bộ gồm bài tập (exercise) tương ứng với câu hỏi (problem). Ví dụ: Ex_143 Pm_1
- Tên file: **relations/exercise-problem.txt**
- Số lượng mẫu: **6252830** mẫu

1.3.1.10 Video ID - CCID (relation)

- Mô tả: Video và ccid tương ứng của nó.
- Định dạng: {Video ID}\t{ccid}

- Tên file: **relations/video_id-ccid.txt**
- Số lượng mẫu: **2798892** mẫu

1.3.2 Student behaviors.

1.3.2.1 Student profile (entity)

- Mô tả: Thông tin của học sinh (user)
- Tên file: **entities/user.json**
- Số lượng mẫu: **3330294** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Id người dùng, bắt đầu bằng “U_”	string	
name	Tên người dùng	string	
gender	Giới tính	int	
school	Tên trường	string	
year_of_birth	Năm sinh	int	
course_order	Các mã khóa học đã chọn	list<int>	
enroll_time	Thời gian đăng ký tương ứng với từng khoá học	list<DateTime>. DateTime có định dạng “YYYY-MM-DD HH:MM:SS”	

Bảng 1.8 Bảng mô tả các trường dữ liệu của user.json

1.3.2.2 Comment (entity)

- Mô tả: Thông tin các bình luận của các học sinh (user) đăng tải lên
- Tên file: **entities/comment.json**
- Dung lượng file: 2.1GB

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Comment ID, bắt đầu bằng “Cm_”	String	
user_id	ID của người dùng đã bình luận, bắt đầu bằng “U_”	Int	
text	Nội dung bình luận	String	
create_time	Thời gian bình luận	DateTime, có định dạng “YYYY-MM-DD HH:MM:SS”	
resource_id	ID của tài nguyên (như video, exercise) mà user bình luận	String	Có thể nhận giá trị null

Bảng 1.9 Bảng mô tả các trường dữ liệu của comment.json

Lưu ý: Thực tế, trường resource_id không được nhắc đến trong file user-en.md trên github của nhóm tác giả nhưng dữ liệu thực lại có thêm trường này.

```
[{"id": "Cm_187", "user_id": 11731, "text": "资源统建", "resource_id": "V_454874", "create_time": "2019-09-05 17:12:29"}, {"id": "Cm_188", "user_id": 11731, "text": "资质统建", "resource_id": "V_454874", "create_time": "2019-09-05 17:12:24"}, {"id": "Cm_190", "user_id": 11731, "text": "资质统建", "resource_id": "V_454874", "create_time": "2019-09-05 17:12:29"}, {"id": "Cm_192", "user_id": 11731, "text": "资质统建", "resource_id": "V_454874", "create_time": "2019-09-05 17:12:29"}]
```

Hình 1.1 Hình minh họa trường dữ liệu bình luận của học sinh

1.3.2.3 Reply (entity)

- Mô tả: Thông tin của phần trả lời bình luận (reply) của học sinh (user)
- Tên file: **entities/reply.json**
- Số lượng mẫu: **331011** mẫu

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

Bảng 1.10 Bảng mô tả các trường dữ liệu của reply.json

1.3.2.4 User-video (relation)

- Mô tả: Tốc độ và các bước nhảy thời gian của người dùng khi xem video.

- Tên file: **relations/user-video.json**
- Dung lượng: **3 GB**

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
user_id	ID của user, bắt đầu bằng “U_”	string	
seq	Mảng, trình tự người dùng xem video, mỗi đối tượng trong mảng là trình tự thời gian người dùng xem một video nhất định, 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 1.11 Bảng mô tả các trường dữ liệu của user-video.json.

Lưu ý: ví dụ về 1 phần tử trong seq.

```
{'video_id': 'V_1395639', 'segment': [{ 'start_point': 100.0, 'end_point': 106.25, 'speed': 1.25, 'local_start_time': 1588438980}, { 'start_point': 180.0, 'end_point': 186.25, 'speed': 1.25, 'local_start_time': 1588439045} ]}
```

1.3.2.5 User-problem

- Mô tả: Người dùng làm bài tập nào.
- Tên file: **relations/user-problem.json**
- Dung lượng: **21 GB**

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
log_id	ID của bản ghi câu hỏi của người dùng, kết hợp với khóa duy nhất của user_id và problem_id	string	
user_id	ID người dùng, bắt đầu bằng U_	string	
problem_id	ID vấn đề, bắt đầu bằng Pm_	string	
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, có định dạng “YYYY-MM-DD HH:MM:SS”	

Bảng 1.12 Bảng mô tả các trường dữ liệu của user-problem.json

1.3.2.6 User-xiaomu

- Mô tả: Tương tác của người dùng với Xiaomu (bot QA của XuetangX).
- Tên file: **relations/user-xiaomu.json**
- Số lượng mẫu: **108351** mẫu

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

Bảng 1.13 Bảng mô tả các trường dữ liệu của user-xiaomu.json

1.3.2.7 Course-comment

- Mô tả: Bình luận của người dùng liên quan đến khóa học.
- Định dạng: {course ID}\t{review ID}.
- Tên file: **relations/course-comment.txt**
- Số lượng mẫu: **10181950** mẫu

1.3.2.8 User-comment

- Mô tả: Bình luận của Người dùng.
- Định dạng: {User ID}\t{Comment ID}.
- Tên file: **relations/user-comment.txt**
- Số lượng mẫu: **8422134** mẫu

1.3.2.9 User-reply

- Mô tả: Phản hồi Bình luận của Người dùng.
- Định dạng: {User ID}\t{Reply ID}.
- Tên file: **relations/user-reply.txt**
- Số lượng mẫu: **331011** mẫu

1.3.2.10 Comment-reply

- Mô tả: Phản hồi bình luận liên quan đến khái niệm.

- Định dạng là {Concept ID}\t{Reply ID}.
- Tên file: **relations/comment-reply.txt**
- Số lượng mẫu: **370493** mẫu

1.3.3 Concepts

1.3.3.1 Concept

- Mô tả: Các khái niệm của khóa học
- Tên file: **entities/concept.json**
- Số mẫu dữ liệu: **637572** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Mã khái niệm, thể hiện theo format K_{tên khái niệm}_{trường khái niệm}	string	
name	Tên khái niệm (mặc định giống với tên khái niệm trong mã khái niệm)	string	
context	Nội dung của khái niệm được xuất hiện trong “ <u>part of Wiki/Baidu Encyclopedia, Zhihu Q&A</u> ”. Nội dung này bắt buộc có sẵn tên khái niệm	string	

Bảng 1.14 Bảng mô tả các trường dữ liệu của concept.json

1.3.3.2 Other

- Mô tả: Các tài liệu liên quan được thu thập bên ngoài những khoá học
- Tên file: **entities/other.json**
- Số mẫu dữ liệu: **210349** mẫu

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Mã dữ liệu	string	
concept	khái niệm của thông tin của tài liệu thu thập được	string	
type	Nguồn dữ liệu, bao gồm [“zhihu”, “baike”, “wiki”]	string	[“zhihu”, “baike”, “wiki”]
content	Nội dung của tài liệu	string	

Bảng 1.15 Bảng mô tả các trường dữ liệu của other.json

1.3.3.3 Paper

- Mô tả: Những bài báo khoa học liên quan
- Tên file: **entities/paper.json**
- Dung lượng: **6.8 GB**

Trường	Nội dung	Kiểu dữ liệu	Miền giá trị
id	Mã bài báo	string	
concept	Mã khái niệm	string	

abstract	Phần giới thiệu (abstract) của bài báo	string	[“zhihu”, ”baike”, “wiki”]
author	Nội dung của tài liệu	string	
lang	Ngôn ngữ của bài báo	string	“en”: Tiếng Anh “zh”: Tiếng Trung
pages	Số lượng trang	Integer	
num_citation	Số lượng citation (trích dẫn từ bài báo khác) tính trong năm 2020	Integer	
score	Điểm số tương đồng giữa bài báo và khái niệm. Điểm càng cao thì càng liên quan nhiều đến khái niệm	Float	
sourcetype	Nguồn của bài báo (hiện tại tất cả là publication)	String	
title	Tên bài báo	String	
venue	Diễn đàn bài báo được đăng tải	String	

urls	các đường link dẫn đến bài báo	List <String>	
year	năm xuất bản	Year	

Bảng 1.16 Bảng mô tả các trường dữ liệu của paper.json

1.3.3.4 Concept-Other

- Mô tả: Khái niệm liên quan với tài nguyên ngoài môn học
- Định dạng: {concept ID}\t{resource ID}
- Tên file: **relations/concept-other.txt**
- Số lượng mẫu: **379926** mẫu

1.3.3.5 Concept-Paper

- Khái niệm liên quan với các bài báo ngoài môn học
- Định dạng: {concept ID}\t{paper ID}
- Tên file: **relations/concept-paper.txt**
- Số lượng mẫu: **5410752** mẫu

1.3.3.6 Concept-Problem

- Mô tả: Khái niệm liên quan với các vấn đề
- Định dạng: {Concept ID}\t{Question ID}
- Tên file: **relations/concept-problem.txt**
- Số lượng mẫu: **33180** mẫu

1.3.3.7 Concept-Video

- Mô tả: Khái niệm liên quan đến video
- Định dạng: {concept ID}\t{ccid}
- Tên file: **relations/concept-video.txt**
- Số lượng mẫu: **624683** mẫu

1.3.3.8 Concept-Comment

- Mô tả: Khái niệm liên quan đến bình luận
- Định dạng: {concept ID}\t{review ID}
- Tên file: **relations/concept-comment.txt**
- Số lượng mẫu: **31074** mẫu

1.3.4 Prerequisites

1.3.4.1 CS.json

- Mô tả: Dự đoán và chú thích của con người về các tiên điều kiện của môn Khoa học Máy tính.
- Tên file: **prerequisites/cs.json**
- Số lượng mẫu: **492102** mẫu

Trường	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, 1 có nghĩa là có, 0 có nghĩa là không.	int	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 1.17 Bảng mô tả các trường dữ liệu của CS.json

1.3.4.2 Math.json

- Mô tả: Chú thích và dự đoán các khái niệm trong lĩnh vực toán học, theo định dạng giống CS.json.
- Tên file: **prerequisites/math.json**
- Số lượng mẫu: **331202** mẫu

1.3.4.3 Psy.json

- Mô tả: Chú thích và dự đoán các khái niệm trong lĩnh vực tâm lý học, theo định dạng giống CS.json.
- Tên file: prerequisites/psy.json
- Số lượng mẫu: **757771** mẫu

1.4 Nhận xét bộ dữ liệu và dự đoán mục tiêu sử dụng của bộ dữ liệu

1.4.1 Nhận xét

Sau khi đọc và khảo sát sơ lược bộ dữ liệu, chúng tôi rút ra được một số nhận xét sau:

- Bộ dữ liệu MOOCCubeX là một bộ dữ liệu có mức độ đa dạng cao và bao gồm các thông tin, tài nguyên về giáo dục cũng như một số thông tin có liên quan khác đến học sinh về việc học.
- Đây là một bộ dữ liệu có kích thước lớn, do đó từ bộ dữ liệu này có thể hỗ trợ việc khám phá dữ liệu phục vụ cho các mục đích hỗ trợ học tập với các phương pháp tiếp cận học máy, học sâu, ...
- Nhìn chung, đây là một bộ dữ liệu không đồng nhất nhưng được tổ chức một cách bài bản, linh hoạt với mức độ chi tiết rất cao. Điều này giúp cho việc sử dụng tài nguyên có thể linh hoạt với nhiều mục đích sử dụng khác nhau, đồng

thời việc tìm kiếm dữ liệu cũng như thiết lập các mô hình để khai phá dữ liệu cũng dễ dàng hơn.

1.4.2 Dự đoán mục tiêu sử dụng bộ dữ liệu

Vì đây là một bộ dữ liệu có liên quan đến chủ đề học tập nên chúng tôi dự định sẽ khai thác các bài toán có liên quan đến lĩnh vực “Cố vấn học tập thông minh tại các trường đại học” hay “Cố vấn học tập thông minh cho các nền tảng học tập trực tuyến”. Sau khi tiến hành khảo sát cũng như thảo luận nhóm, chúng tôi đưa ra hai bài toán có thể sẽ giải quyết sau:

- Bài toán 1: Hệ thống khuyến nghị hay cố vấn cho người học đăng ký các môn học, khóa học theo các định hướng chuyên ngành dựa trên hành vi học tập của người học.
- Bài toán 2: Dự đoán chất lượng giảng dạy của khóa học.

Ngoài ra, trong quá trình tìm hiểu của cả nhóm sau này, chúng tôi sẽ cố gắng kết hợp sử dụng các bộ dữ liệu khác cũng như các loại bài toán khác để tiến hành xây dựng hệ thống phù hợp hơn với môi trường học tập ở Việt Nam nói chung, cũng như cho các sinh viên đang học tập tại trường Đại học Công Nghệ Thông Tin nói riêng.

2 CHUẨN BỊ DỮ LIỆU

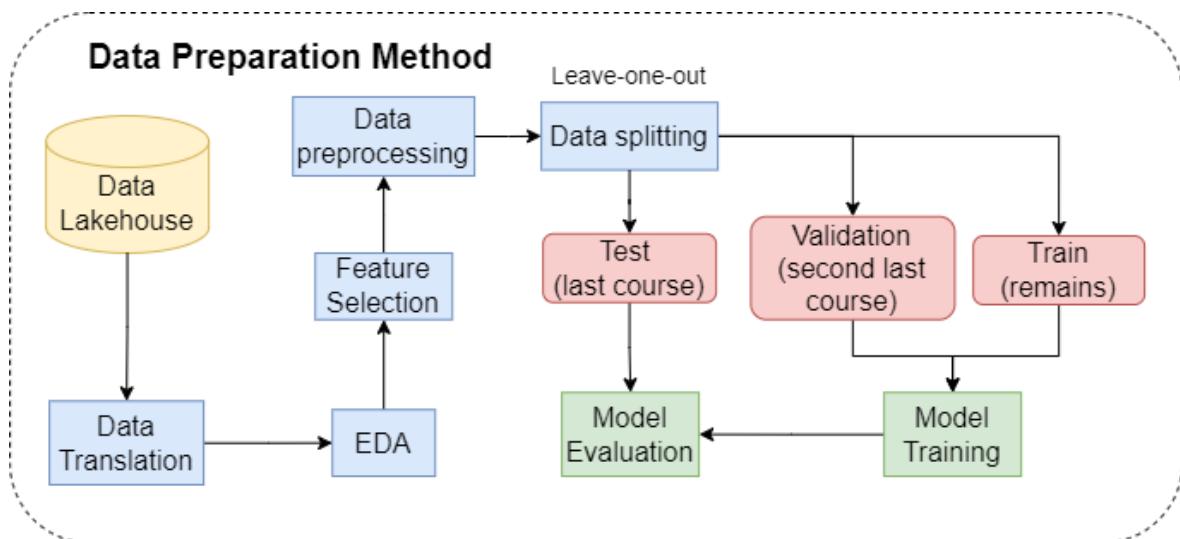
2.1 Dữ liệu thực nghiệm

Từ bộ dữ liệu MOOCCubeX [1], nhóm đã chọn lọc các files sau để xây dựng hệ thống gợi ý khóa học và ứng dụng web:

- Entities: course.json, teacher.json, school.json, concept.json, user.json, video.json.
- Relations: course-field.json, course-school.txt, course-teacher.txt, course-concept.txt, user-video.json, vid_ccid.txt.

2.2 Phương pháp tổ chức dữ liệu thực nghiệm

Các thao tác chuẩn bị dữ liệu được thực hiện theo trình tự sau:



Hình 2.1 Hình minh họa quy trình xử lý dữ liệu của nhóm

2.2.1 Data translation

Để dịch dữ liệu từ tiếng Trung sang tiếng Việt, nhóm sử dụng Googletrans [3], một thư viện python miễn phí và không giới hạn, triển khai API Google Translate. Thư viện này sử dụng API Google Translate Ajax để thực hiện lệnh gọi đến các phương thức như phát hiện và dịch. Nhưng do có một số trường có lượng dữ liệu cần dịch lớn nên thời gian dịch rất lâu, và trong quá trình dịch cũng xảy ra hiện tượng bị mất kết

nội. Bên cạnh đó, việc dịch một số trường có lượng dữ liệu lớn cũng không cần thiết để huấn luyện mô hình nên nhóm chỉ dịch một số trường có dữ liệu nhỏ để hiển thị thông tin trên ứng dụng web. Dưới đây là các trường mà nhóm đã dịch:

- course.json: name, prerequisites, fields
- user.json: không dịch, nhưng ở cột name, nhóm đã sử dụng bộ sinh tên tiếng Việt dựa trên giới tính được cung cấp bởi [4]. Việc này giúp thuận tiện cho việc mô phỏng khả năng tìm kiếm bằng username của ứng dụng web.
- school.json: name, about, motto
- teacher.json: name, job_title, org_name, about

2.2.2 EDA, làm sạch dữ liệu

Để hiểu rõ hơn về tập dữ liệu và xác định các mẫu, xu hướng tiềm ẩn, nhóm đã thực hiện phân tích dữ liệu khám phá

2.2.2.1 Thống kê mô tả, trực quan hóa dữ liệu, xử lý dữ liệu

2.2.2.1.1 Course

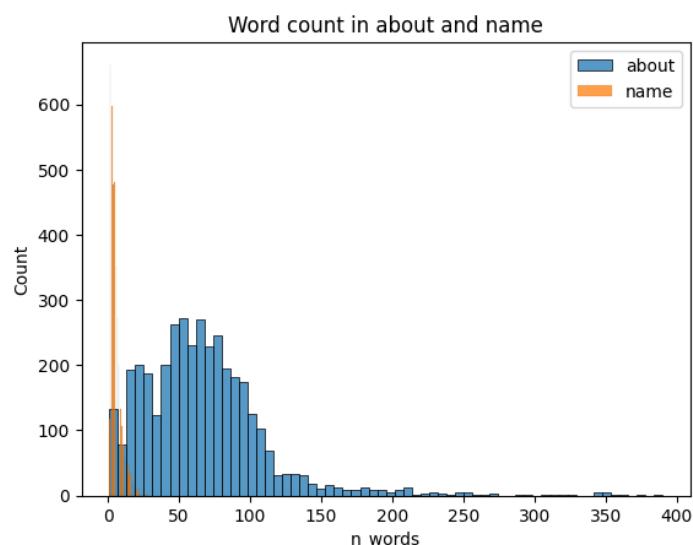
describe	id	name	field	prerequisites	about	resource	name_trans	about_trans
str	str	str	str	str	str	str	str	str
"count"	"3781"		"3781" "3781"	"3779"		"3779" "3781"	"3781"	"3781"
"null_count"	"0"		"0" "0"	"2"		"2" "0"	"0"	"0"

Hình 2.2 Bảng thống kê mô tả của course.json

course.json có 3781 hàng với các trường thông tin như sau: case_id, name (đã được dịch), field (đã được dịch), prerequisites, about (đã được dịch), resource. Tất cả các trường đều có rất ít null. Nhận thấy bên cạnh sử dụng tfidf để tạo đặc trưng trên trường about, name, ta vẫn có thể sử dụng PhoBERT [5] để embed tạo đặc trưng. Do PhoBERT được huấn luyện trên dữ liệu ở mức độ từ, nên ta sẽ phải gôm các tiếng lại thành từ bằng rdrsegmenter của VNCoreNLP [6]. Sau đây là một số thống kê cơ bản:

describe	about_segmented	name_segmented	len_about_segmented	len_name_segmented
	str	str	f64	f64
"count"	"3781"	"3781"	3781.0	3781.0
"null_count"	"0"	"0"	0.0	0.0
"mean"	null	null	66.694261	5.525522
"std"	null	null	44.815107	3.791343
"min"	""	"" diễn_dàn gia...	1.0	1.0
"25%"	null	null	39.0	3.0
"50%"	null	null	62.0	5.0
"75%"	null	null	87.0	7.0
"max"	"🔥 " tâm_lý_học... "中医基础理论俄文版 (co b...		390.0	28.0

Hình 2.3 Thống kê cơ bản về số từ trong about, name sau khi segmented (len_about_segmented, len_name_segmented)



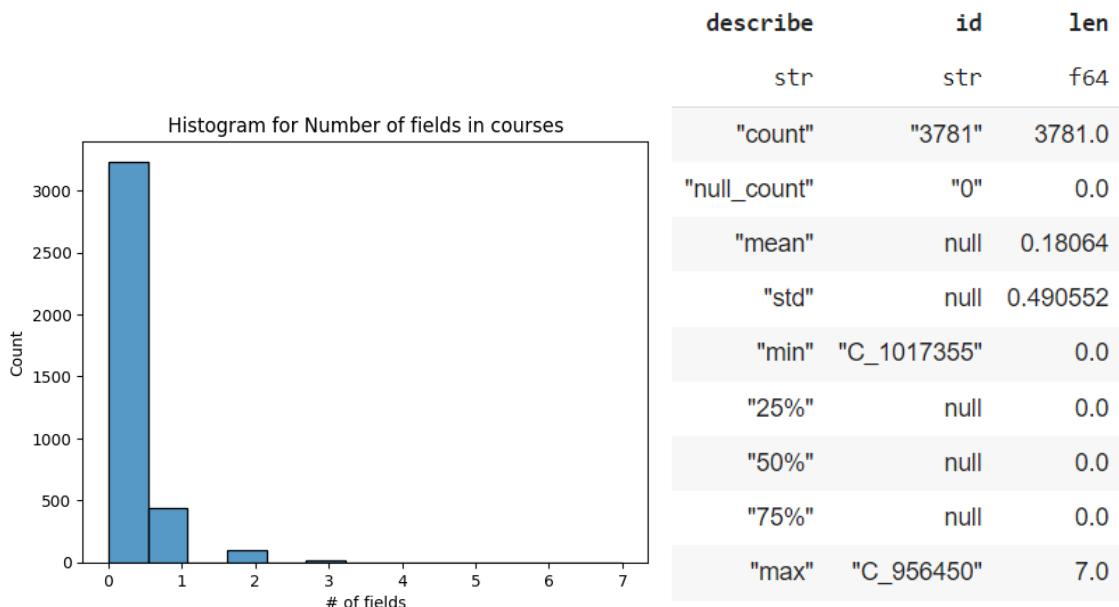
Hình 2.4 Biểu đồ cột thê hiện độ dài văn bản ở trường about, name.

Từ bảng trên, ta có độ dài ngắn nhất trong trường about, name là 1; độ dài lớn nhất trong about, name lần lượt là 390 và 28; độ dài trung bình trong about, name lần lượt là 67 và 6. Từ hình vẽ, ta thấy được độ dài văn bản trong trường about, name lần lượt có dạng phân phối chuẩn và long tail. Tiếp đến, ta sẽ xét trường field và file course-field.json.

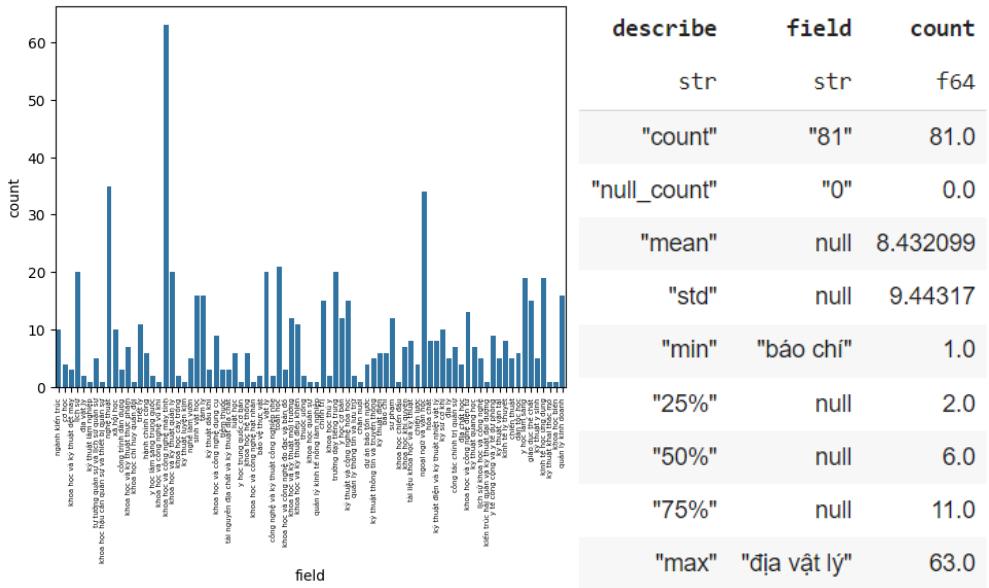
describe	course_id	course_name	field	course_name_trans
str	str	str	str	str
"count"	"632"	"632"	"632"	"632"
"null_count"	"0"	"0"	"0"	"0"

Hình 2.5 Bảng thống kê mô tả của course-field.json

Trường field của course có nội dung tương tự như file course-field.json. Ta sử dụng phép hợp để gồm 2 thông tin này lại với nhau. Ngoài ra, ta sẽ lọc bỏ những khóa học của file course-field.json mà không tồn tại trong course.json. Sau đó, ta trực quan hóa thông tin về field của khóa học như hình sau:



Hình 2.6 Histogram thể hiện số lượng fields trong mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải)



Hình 2.7 Biểu đồ cột thể hiện số lượng khóa học của các field (bên trái) và bảng thống kê mô tả tương ứng (bên phải).

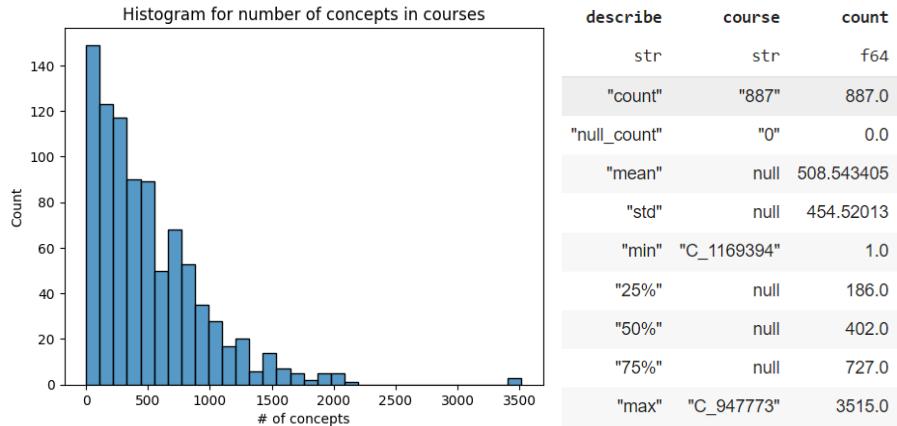
Biểu đồ cho thấy, có rất nhiều khóa học không có field nào, và có nhiều field có số lượng khóa học ít hơn 5. Điều này cho thấy, feature field của khóa học sẽ có rất ít đóng góp.

2.2.2.1.2 Concept

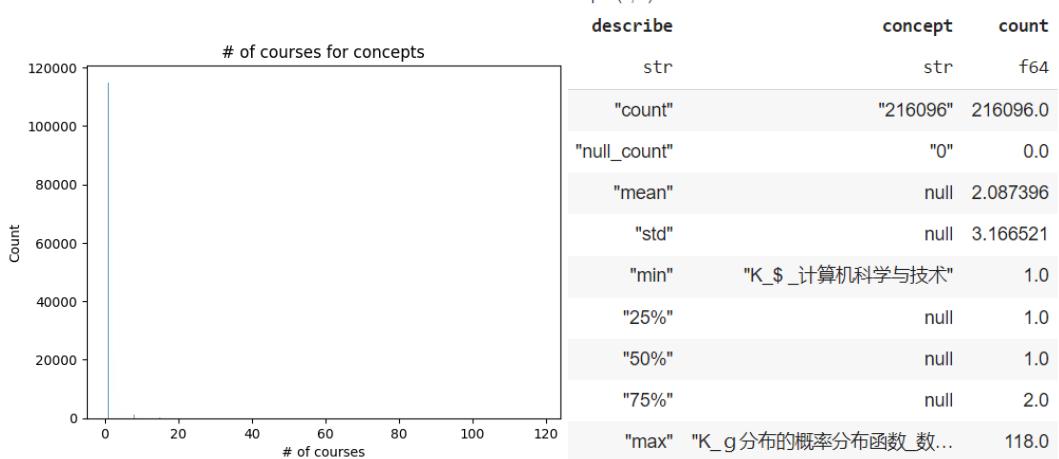
describe	id	name	context
str	str	str	str
"count"	"637572"	"637572"	"637572"
"null_count"	"0"	"0"	"0"

Hình 2.8 Bảng thống kê mô tả của concept.json

concept.json có 637572 hàng và các trường thông tin như id, name, context. Các trường thông tin này tuy không có null nhưng không có nhiều ý nghĩa, nên ta sẽ chỉ quan tâm đến liên kết giữa khóa học và concept (concept-course.txt với 451078 hàng). Do ta không quan tâm đến thông tin của concept nên sẽ không lọc bỏ những liên kết của các concept không hợp lệ. Sau đây là một số thông tin được trực quan hóa:



Hình 2.9 Histogram thể hiện số lượng concept của mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải).



Hình 2.10 Histogram thể hiện số lượng khóa học của mỗi concept (bên trái) và bảng thống kê mô tả tương ứng (bên phải).

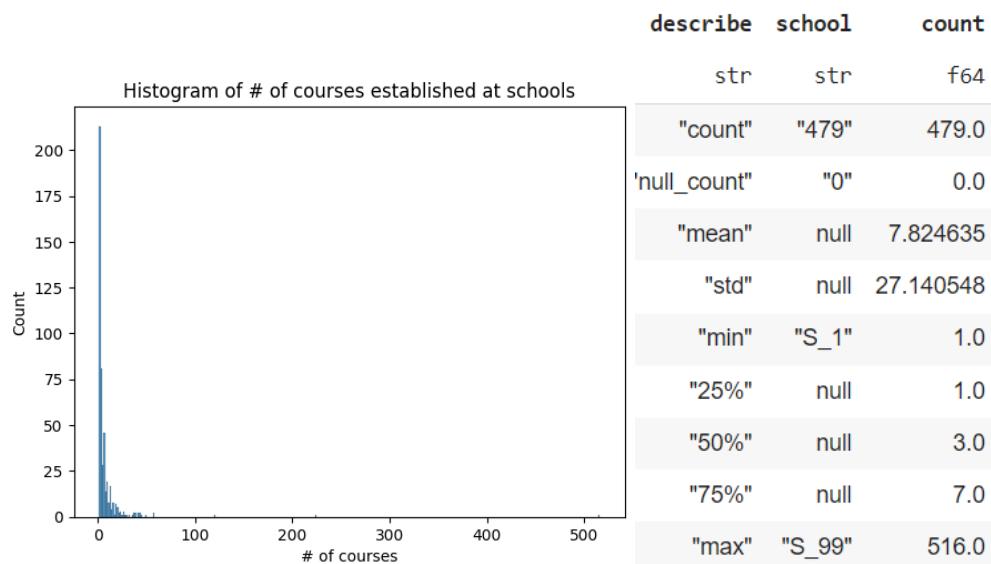
Số lượng concepts của khóa học có phân phối long tail. Số lượng concept nhiều nhất trong khóa học là 3515, ít nhất là 0. Phần lớn concepts có số lượng khóa học là 1. Số lượng khóa học ít nhất của 1 concept là 1, lớn nhất là 118. Số lượng concepts với số lượng khóa học ≥ 5 là 14016, điều này cho thấy, đây có thể là đặc trưng đóng góp nhiều vào dự đoán của mô hình.

2.2.2.1.3 School

describe	id	name	name_en	sign	about	motto	about_trans	motto_trans
str	str	str	str	str	str	str	str	str
"count"	"429"	"429"	"429"	"429"	"429"	"429"	"429"	"429"
"null_count"	"0"	"0"	"0"	"0"	"0"	"0"	"0"	"0"

Hình 2.11 Bảng thống kê mô tả của school.json

school.json gồm 429 hàng với các trường thông tin như: id, name (đã được dịch), name_en, sign, about (đã được dịch), motto (đã được dịch). Tuy các trường thông tin này không có null nhưng lại không có quá nhiều ý nghĩa để ta khai thác. Vì vậy, ta sẽ chỉ quan tâm đến liên kết giữa khóa học và school (course-school.txt). Trong file course-school.txt có 3748 hàng. Tương tự concept, ta sẽ không lọc bỏ những liên kết của các school không hợp lệ. Một số thông tin được trực quan hóa như sau:



Hình 2.12 Histogram thể hiện số lượng khóa học của mỗi trường (bên trái) và bảng thống kê mô tả tương ứng (bên phải).

describe	course	count
str	str	f64
"count"	"3717"	3717.0
"null_count"	"0"	0.0
"mean"	null	1.00834
"std"	null	0.090955
"min"	"C_1017355"	1.0
"25%"	null	1.0
"50%"	null	1.0
"75%"	null	1.0
"max"	"C_956450"	2.0

Hình 2.13 Bảng thống kê mô tả cho số lượng trường học của mỗi khóa học.

Số lượng khóa học của một trường theo phân phối long tail. Phần lớn school có số lượng khóa học ≤ 7 (Phân vị 75%). Có một số school đặc biệt tổ chức từ 100 khóa học trở lên. Số lượng khóa học ít nhất của một trường là 1, nhiều nhất là 516. Có 185 schools tổ chức ít nhất 5 khóa học.

Ngoài ra, nhóm cũng thống kê xem có bao nhiêu trường cùng tổ chức 1 khóa học. Kết quả cho thấy phần lớn khóa học chỉ được tổ chức bởi 1 school (Phân vị 75%). Số schools ít nhất của một khóa học là 1 (không xét trường hợp khóa học không có school), nhiều nhất là 2. Điều này là khác biệt so với trường field, concept vì 1 khóa học có nhiều field, concept. Tuy nhiên sự khác biệt này không ảnh hưởng nhiều.

2.2.2.1.4 Teacher

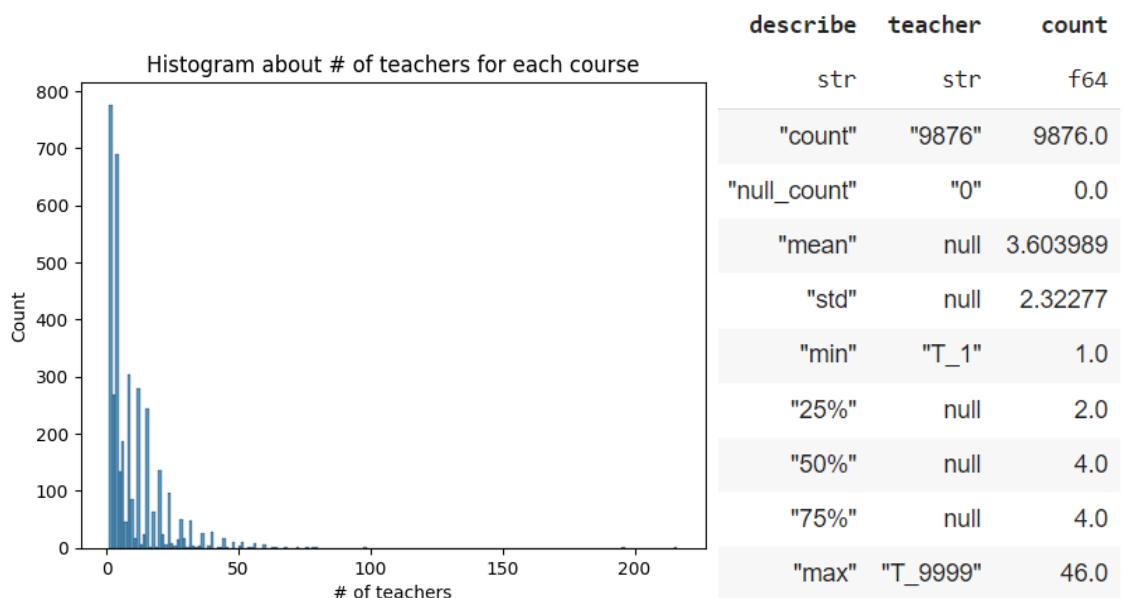
describe	id	name	name_en	about	job_title	org_name	about_trans
str	str	str	str	str	str	str	str
"count"	"17018"	"17018"	"9525"	"17018"	"14768"	"17018"	"13892"
"null_count"	"0"	"0"	"7493"	"0"	"2250"	"0"	"3126"

Hình 2.14 Bảng thống kê mô tả của teacher.json

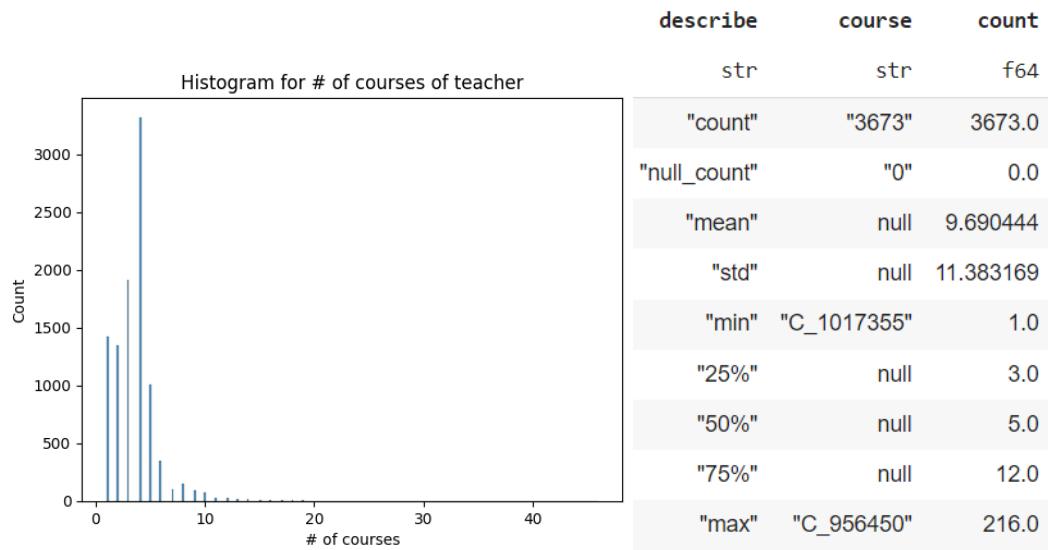
teacher.json gồm 17018 hàng và các trường thông tin như id, name, name_en, about (đã được dịch), job_title, org_name. Trong đó, trường job_title là chức danh của teacher, có thể có ích trong việc thiết kế feature. Thực chất, job_title nên được xem như là văn bản thay vì categorical bởi vì job_title không chỉ bao gồm chức vụ, học vị, mà còn có một số thông tin khác như tên công ty, phòng khoa... VD: "giáo sư, trường

kinh tế và quản lý, đại học thanh hóa", "phó bí thư, phó tổng bí thư đảng ủy viện kê toán công chứng trung quốc". Ta có thể sử dụng một mô hình bên ngoài (ChatGPT [7], Gemini [8]) để trích xuất các thông tin về chức vị, tổ chức, sau đó xem các biến này là categorical. Bên cạnh đó, trường org_name, tổ chức của teacher, cũng có thể ảnh hưởng nhiều đến số lượng user tham gia vào khóa học.

Về liên kết giữa khóa học và teacher, file course-teacher.txt chứa 97192 hàng. Trong tương lai, nhóm sẽ dùng một số trường thông tin của teacher nếu có thể. Vì vậy, ta cần lọc bỏ các liên kết có khóa học hoặc teacher không tồn tại. Sau khi lọc bỏ, số hàng còn lại là 35593. Các thông tin được trực quan hóa như sau:



Hình 2.15 Histogram thể hiện số lượng teachers của mỗi khóa học (bên trái) và bảng thống kê mô tả tương ứng (bên phải).



Hình 2.16 Histogram thể hiện số lượng khóa học của mỗi teacher (bên trái) và bảng thống kê mô tả tương ứng (bên phải).

Phần lớn teachers có số lượng khóa học < 5 . Có 1883 teachers với số lượng khóa học ≥ 5 . Phần lớn khóa học có số teachers ≤ 12 (Phân vị 75%).

Về org_name của teacher, số lượng giáo viên trong một tổ chức như sau:

describe	count
str	f64
"count"	724.0
"null_count"	0.0
"mean"	13.640884
"std"	35.463706
"min"	1.0
"25%"	1.0
"50%"	4.0
"75%"	14.0
"max"	678.0

Hình 2.17 Bảng thống kê mô tả số lượng giáo viên của một tổ chức.

2.2.2.1.5 User

describe	id	name	gender	school	year_of_birth	course_order	enroll_time
str	str	str	f64	str	str	str	str
"count"	"3330294"	"3330240"	3.33024e6	"1128399"	"0"	"3330294"	"3330294"
"null_count"	"0"	"54"	54.0	"2201895"	"3330294"	"0"	"0"
"mean"	null	null	0.945575	null	null	null	null
"std"	null	null	0.83211	null	null	null	null
"min"	"U_10000"	""	0.0	"Queen's Univ..."	null	null	null
"25%"	null	null	0.0	null	null	null	null
"50%"	null	null	1.0	null	null	null	null
"75%"	null	null	2.0	null	null	null	null
"max"	"U_999999"	"□"	232.0	"🔧 工程大学"	null	null	null

Hình 2.18 Bảng thông kê mô tả của user.json

user.json gồm 3330294 hàng với các trường thông tin như id, name (được khởi tạo), gender, school, year_of_birth, course_order, enroll_time. Trường school có tới 66.12% là null, ngoài ra có một số ký tự đặc biệt như: "🔧 工程大学"; "Queen's Univ...", có lẽ trường này không được kiểm tra trước khi đưa vào CSDL của XuetangX. Trường year of birth gần như toàn bộ là null nên không đem lại ý nghĩa. Các khóa học trong course_order sẽ tương ứng với enroll_time và chúng được sắp xếp theo thời điểm tăng ký, nên sẽ thuận tiện cho việc chia tập train, val, test theo thời gian. Tiếp theo, nhóm sẽ lọc bỏ những khóa học không hợp lệ (không tồn tại trong course.json)

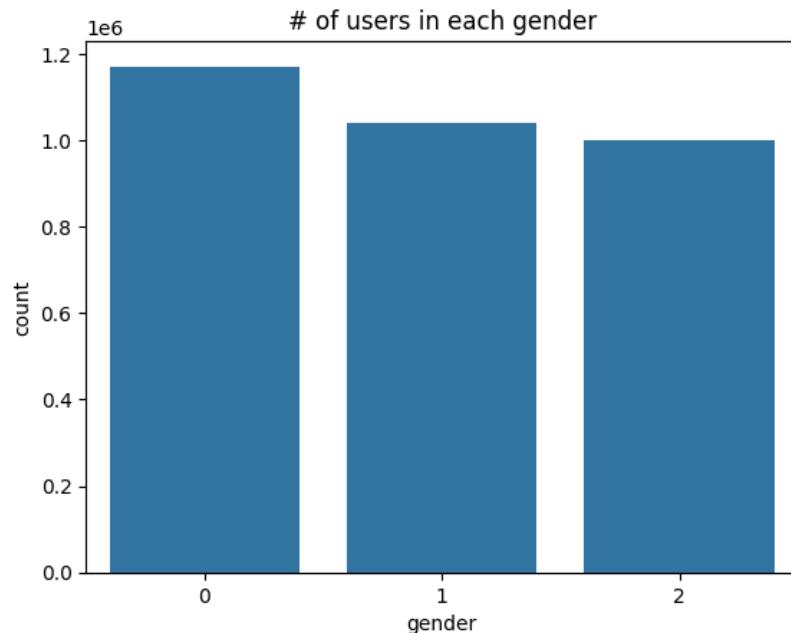
Sau đó, ta xét trường gender. Trường này chỉ có $\frac{54}{3330294} * 100 = 0.0016\%$ giá trị null nên ta có thể sử dụng để làm feature.

gender	count
i64	u32
232	1
2	1040449
3	1
1	1067858
0	1221931
null	54

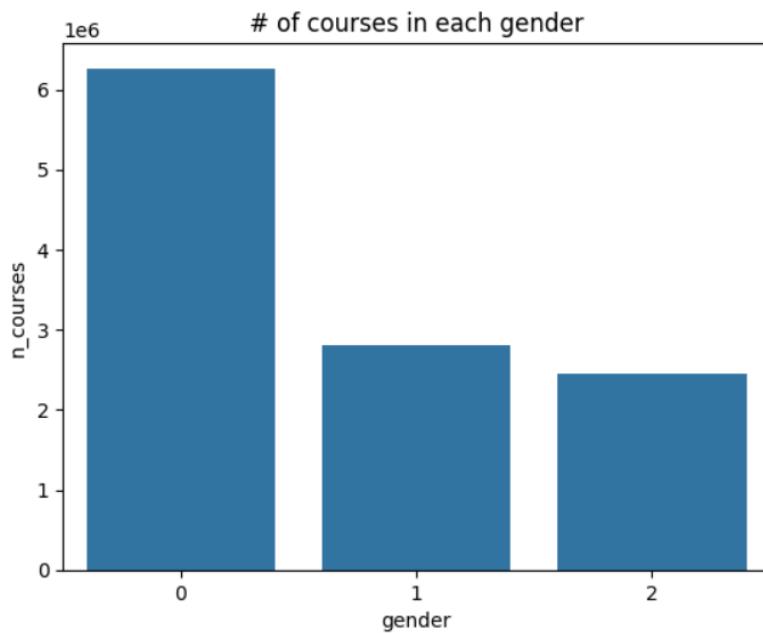
Hình 2.19 Value counts của trường gender của User.

Theo Hình 2.19, trường gender có 2 giá trị nhiều là 232, 3 và có rất ít hàng có gender là null, vì vậy ta sẽ bỏ đi những hàng chứa các giá trị này.

Số lượng users trong từng giới tính khá tương đồng (Hình 2.20). Số lượng khóa học của nhóm giới tính thứ 0 nhiều vượt trội so với 2 giới tính còn lại (Hình 2.21).



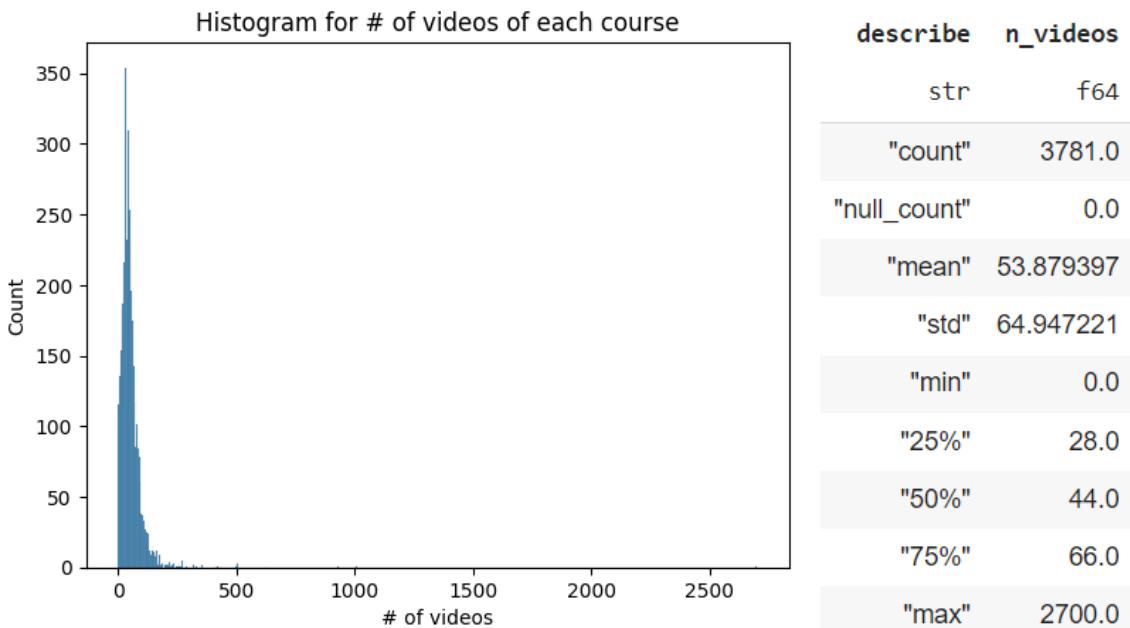
Hình 2.20 Biểu đồ thể hiện số lượng User ứng với từng giới tính.



Hình 2.21 Biểu đồ thể hiện tổng số lượng khóa học đăng ký của mỗi nhóm giới tính

2.2.2.1.6 Video

Ở đây, ta sẽ tập trung vào danh sách các video của khóa học trong trường resource của course.json. Do trường resource chứa 2 loại tài nguyên là video và exercise, nên ta cần tách video ra để dễ dàng thống kê. Sau đây là một số thống kê:



Hình 2.22 Histogram thể hiện số lượng videos của mỗi khóa học(trái) và bảng thống kê mô tả tương ứng (phải).

Số videos nhiều nhất trong 1 khóa học là 2700. Số videos trung bình trong 1 khóa học là 53.88. Có một số khóa học không có video nào. Do videos có thể chỉ là thông tin bổ trợ cho bài toán, nên nhóm sẽ không xóa đi các khóa học không có videos nào. Tuy nhiên, các khóa học có thể chứa các video không tồn tại, nên chúng ta sẽ xét tiếp đến video.json và vid_ccid.txt.

describe	ccid	name	start	end	text
str	str	str	str	str	str
"count"	"59581"	"59581"	"59581"	"59581"	"59581"
"null_count"	"0"	"0"	"0"	"0"	"0"

Hình 2.23 Bảng thống kê mô tả của video.json

video.json chứa 59581 hàng với các trường: ccid, name, start, end, text. Trong đó, trường start, end có thể được sử dụng để tính tổng thời lượng của từng video. Lưu ý, ccid khác với video_id. Hiểu 1 cách đơn giản, khi ccid được trình chiếu tại một môn học nào đó thì nó mới là video_id. Do đó, 1 video_id sẽ tương ứng với 1 ccid, và 1 ccid sẽ tương ứng với nhiều video_id, liên kết của chúng được lưu trữ trong vid_ccid.txt.

vid_ccid.txt chứa 2798892 hàng. Ta tiến hành lọc bỏ các liên kết không hợp lệ (video_id không tồn tại trong resource của course.json). Kết quả thống kê cho thấy, trong vid_ccid.txt, chỉ có 7% liên kết có ý nghĩa (video_id tồn tại trong resource của course.json); có tới 63% ccid không tồn tại trong video.json (xét ccid unique).

2.2.2.2 Phân tích thống kê

Nhóm thực hiện kiểm định ANOVA để so sánh trung bình số lượng khóa học đăng ký giữa các nhóm giới tính.

Bộ dữ liệu bao gồm ba loại giới tính khác nhau được đánh theo số thứ tự 0, 1, 2. Sau đó nhóm gọi hàm `scipy.stats.f_oneway(g0, g1, g2)` để thực hiện kiểm định ANOVA, kết quả trả về giá trị thống kê f và p-value.

```
[ ] g_nc = user_df.select('gender', 'course_order') \
    .with_columns(pl.col('course_order').list.len()) \
    .rename({'course_order': 'n_courses'}) \n\n
g0 = g_nc.filter(pl.col('gender') == 0).select('n_courses')
g1 = g_nc.filter(pl.col('gender') == 1).select('n_courses')
g2 = g_nc.filter(pl.col('gender') == 2).select('n_courses')\n\n
f, p = scipy.stats.f_oneway(g0, g1, g2)
print(f'Thống kê f : {f}')
print(f'p-value : {p}')
```

➡ Thống kê f : [26606.91252138]
p-value : [0.]

Hình 2.24 Thực hiện kiểm định phương sai ANOVA để xem số lượng khóa học đăng kí có bị phụ thuộc vào nhóm giới tính hay không.

Với việc p-value < 0.05, ta có thể kết luận rằng: “Có sự khác nhau với số lượng khóa học đăng kí giữa 3 nhóm giới tính”.

2.2.2.3 Khai phá tri thức

Tại mục 5.2.2.1, nhóm đã thấy được bộ dữ liệu có số lượng user rất nhiều, tuy nhiên số lượng user đăng kí các khóa học lại khá ít (Hơn 75% user có số khóa học đăng kí

≤ 2). Vì vậy, nhóm sẽ lọc bỏ và chỉ lấy số lượng user đã đăng ký từ 5 khóa học trở lên.

Sau đó, nhóm thực hiện một số bước khai phá dữ liệu và tìm ra được các quy tắc kết hợp từ dữ liệu các khóa học đã được đăng ký bởi người dùng, sau đó dựa vào những quy tắc đã tìm được để dự đoán các khóa học mà một học viên có thể quan tâm. Các bước khai phá tri thức được thực hiện cụ thể như sau:

- Sử dụng thuật toán FP-Growth [9] để tìm được tập các khóa học phổ biến với min_support là 0.01.
- Sau đó sử dụng hàm association_rules từ thư viện mlxtend.frequent_patterns để tạo ra được các quy tắc kết hợp từ tập các khóa học phổ biến.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(948410)	(696700)	0.062076	0.103303	0.010022	0.161453	1.562902	0.003610	1.069346	0.384002
1	(696700)	(948410)	0.103303	0.062076	0.010022	0.097018	1.562902	0.003610	1.038697	0.401657
2	(948410)	(697791)	0.062076	0.139220	0.011482	0.184968	1.328602	0.002840	1.056130	0.263698
3	(697791)	(948410)	0.139220	0.062076	0.011482	0.082474	1.328602	0.002840	1.022232	0.287331
4	(948410)	(629559)	0.062076	0.119984	0.014332	0.230875	1.924216	0.006884	1.144179	0.512097
...
1771	(916828)	(735123)	0.077565	0.033061	0.010670	0.137569	4.161018	0.008106	1.121178	0.823553
1772	(758208)	(679390)	0.031045	0.078778	0.010767	0.346821	4.402512	0.008321	1.410367	0.797619
1773	(679390)	(758208)	0.078778	0.031045	0.010767	0.136674	4.402512	0.008321	1.122352	0.838948
1774	(758208)	(916828)	0.031045	0.077565	0.010183	0.328013	4.228900	0.007775	1.372698	0.787995
1775	(916828)	(758208)	0.077565	0.031045	0.010183	0.131285	4.228900	0.007775	1.115389	0.827735

1776 rows x 10 columns

Hình 2.25 Một phần của tập luật kết hợp

Từ tập quy tắc này, nhóm xây dựng một function nhỏ có input là khóa học, tập luật; output là các khóa học thường đi kèm. Ví dụ minh họa như sau:

```
# Giả định `new_student_courses` là một danh sách các khóa học mà một học viên mới muốn đăng ký
new_student_courses = [746997]

# Dự đoán course_order cho học viên mới
predicted_order = predict_course_order(new_student_courses, rules)

print("Recommended course order:", predicted_order)
```

⇒ Recommended course order: {697018, 782555, 696679}

Hình 2.26 Các khóa học thường xuất hiện cùng với khóa học 746997

2.2.2.4 Làm sạch dữ liệu

Đầu tiên, chúng tôi tiến hành xử lý dữ liệu trùng lặp ở các bảng course, course-chool, teacher, course-teacher, course-concept, user. Kết quả phát hiện được số mẫu dữ liệu bị trùng lặp ở từng bảng như sau:

- course: Số dòng bị trùng lặp là 0, Số ID bị trùng lặp là 0.
- course-chool: Số dòng bị trùng lặp là 0.
- teacher: Số dòng bị trùng lặp là 0, Số ID bị trùng lặp là 0.
- course – teacher: Số dòng bị trùng lặp là 22281.
- course-concept: Số dòng bị trùng lặp là 0
- user: Số dòng bị trùng lặp là 0, Số ID bị trùng lặp là 0.

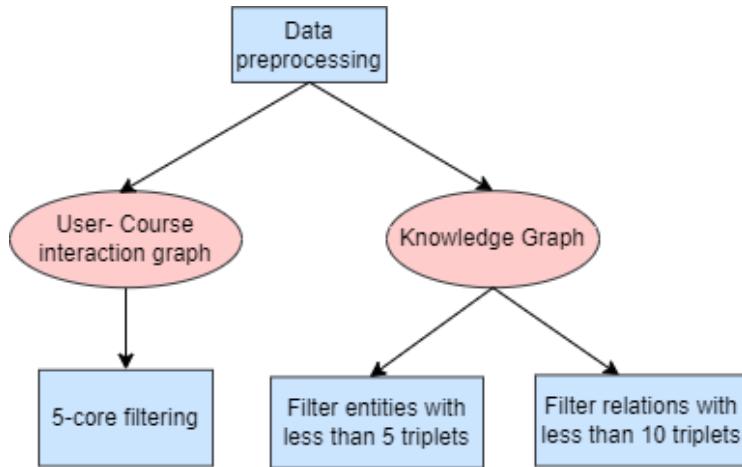
Đối với các mẫu dữ liệu bị trùng lặp, nhóm tiến hành xóa bỏ các mẫu giống nhau này và giữ lại một mẫu duy nhất.

2.2.3 Feature selection

Trong phần này, nhóm sẽ tìm các thuộc tính có thể của khóa học để đưa vào Knowledge graph. Có 4 thuộc tính của khóa học được chọn: school, teacher, concept, field. Bên cạnh đó, nhóm đã thử kiểm tra xem liệu có thể sử dụng mức độ hoàn thành khóa học dựa trên thời gian xem video của từng user hay không. Mức độ này được tính theo công thức $\frac{\text{tổng thời gian xem các video của 1 khóa học của 1 người dùng}}{\text{tổng thời gian video của khóa học đó}}$. Tuy nhiên, lượng thông tin này rất ít. Để có được thông tin về tổng thời gian video của 1 khóa học, ta cần ánh xạ video_id sang ccid (nhiều video_id sẽ tương ứng với 1 ccid), sau đó lấy thông tin về thời gian của từng video. Nhưng đánh giá cho thấy, trong video_id-ccid.txt có đến 63% ccid không tồn tại trong video.json (file chứa thông tin thời gian video). Vì vậy, nhóm sẽ không sử dụng thông tin thời gian này để tạo đặc trưng.

2.2.4 Data preprocessing

Nhóm sẽ lọc bỏ đối tượng có liên kết ít để đảm bảo đồ thị collaborative knowledge graph không bị thừa thớt, đảm bảo chất lượng của bộ dữ liệu huấn luyện, đánh giá.



Hình 2.27 Sơ đồ phân rã của quy trình tiền xử lý dữ liệu

Đối với user-course bipartite (interaction) graph, nhóm sử dụng 5-core filtering, nghĩa là sẽ lọc bỏ những user đăng ký ít hơn 5 khóa học và những khóa học được đăng ký bởi ít hơn 5 user. Sau khi lọc, số lượng liên kết trong đồ thị đã giảm từ 11523022 xuống 7470942; số lượng user, khóa học còn lại lần lượt là 373351, 3118. Nhưng do không đủ tài nguyên tính toán, nhóm sẽ chỉ giữ lại 100000 user ngẫu nhiên rồi lọc lại như trên. Kết quả cuối cùng như bảng sau:

	Số lượng
User-course interactions	1992150
Users	99969
Courses	2831

Bảng 2.1 Bảng thống kê số lượng của từng thực thể sau khi xử lý dữ liệu

Đối với Knowledge graph, nhóm lọc bỏ những entity có số lượng triplet (course, relation, entity) ít hơn 5 và những relation có số lượng triplet ít hơn 10. Kết quả cuối như bảng sau:

Relation	# of triplets	# of unique entities
course.school	2296	144
course.concept	63680	7162
course.teacher	262	40
course.field	471	41
Tổng	66709	7387

Bảng 2.2 Bảng thống kê chi tiết từng loại liên kết sau khi thực hiện N-core filtering

Thực chất, trong paper KGAT [10], để đảm bảo chất lượng dữ liệu huấn luyện, nhóm nghiên cứu đã sử dụng 10-core filtering, lọc bỏ entity ít hơn 10 triplets, lọc bỏ relation ít hơn 50 triplets. Nhưng do paper KGAT thực hiện gợi ý các sản phẩm như sách, bài hát,... có thời gian hoàn thành ngắn hơn nhiều so với việc hoàn thành một khóa học, nên việc giảm tiêu chí để lọc là điều cần thiết.

2.2.5 Data splitting

Nhóm chia dữ liệu theo chiến lược leave-one-out. Với mỗi user, nhóm giữ khóa học cuối cùng làm test, khóa học kế cuối làm val, các khóa học còn lại làm train.

3 PHÂN TÍCH VẤN ĐỀ

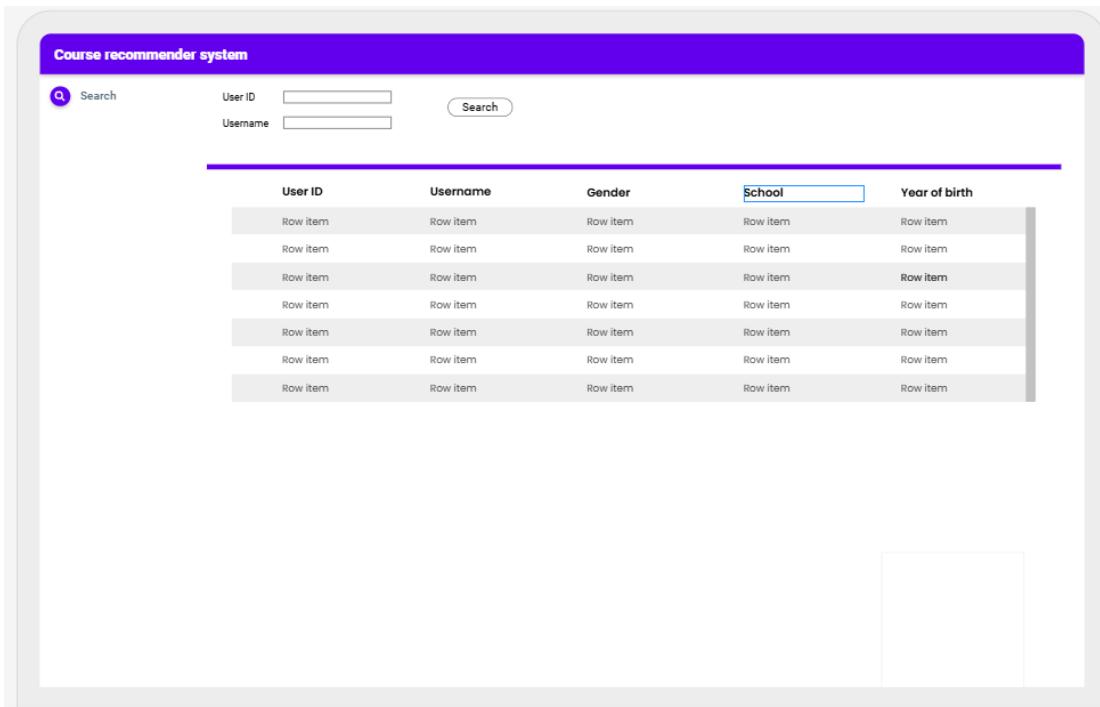
3.1 Câu hỏi nghiên cứu

Trong thời đại công nghệ số kéo theo sự biến đổi ngày càng tiến bộ của mạng Internet, hình thức học trực tuyến càng được chú ý với sự phổ biến và tính tiện lợi của chúng. Tuy nhiên, với hàng ngàn, thậm chí là hàng trăm ngàn các khóa học được tổ chức trên các nền tảng học tập trực tuyến khác nhau thì người dùng sẽ gặp rất nhiều khó khăn trong việc lựa chọn được những khóa học phù hợp. Từ thực trạng này, nhóm đề xuất một hệ thống khuyến nghị các khóa học có thể hoạt động tốt trên các nền tảng học tập trực tuyến.

Với những ý tưởng trên, nhóm tự hỏi rằng: Liệu có thể áp dụng có kĩ thuật học máy học, học sâu để xây dựng một hệ thống để xuất đủ tốt để phục vụ mục đích khuyến nghị các khóa học phục vụ người dùng dựa trên các nền tảng học tập trực tuyến?

3.2 Kết quả đề tài

Với những mong muốn và câu hỏi đặt ra, kết quả của đề tài mà nhóm chúng tôi mong muốn sẽ tạo ra được một hệ thống để xuất khóa học với giao diện website phục vụ tương tác người dùng.



Hình 3.1 Hình minh họa kết quả sản phẩm website hệ thống khuyến nghị của nhóm

3.3 Khả năng ứng dụng

Từ những lý do, câu hỏi đề tài và kết quả mong muốn thu được sau khi hoàn thiện đề tài, chúng tôi nhận thấy tính ứng dụng thực tiễn của đề tài là vô cùng lớn, có thể bao gồm:

- Cá nhân hóa trải nghiệm học tập của người dùng:
 - o Tối ưu hóa lộ trình học tập: Hệ thống đề xuất các khóa học phù hợp với nhu cầu, sở thích, và mục tiêu của từng người học, giúp họ chọn được những khóa học phù hợp nhất.
 - o Gợi ý dựa trên hành vi học tập: Dựa trên dữ liệu hành vi như các khóa học đã hoàn thành, thời gian học, và thành tích, hệ thống có thể đề xuất các khóa học tiếp theo một cách chính xác.
- Nâng cao hiệu suất học tập của người học:
 - o 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.

- Tăng cường trải nghiệm và sự hài lòng của người học:
 - o Bằng cách đề xuất các khóa học thú vị và phù hợp, hệ thống giữ người học gắn bó hơn với nền tảng, tăng thời gian và mức độ tham gia.
 - o 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.

4 TÀI LIỆU THAM KHẢO

- [1] Yu, J.; Wang, Y.; Zhong, Q.; Luo, G.; Mao, Y.; Sun, K.; Feng, W.; Xu, W.; Cao, S.; Zeng, K.; et al., “MOOCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs,” trong *In Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.
- [2] “XuetangX: Online Courses from Top Universities,” Tsinghua University, [Trực tuyến]. Available: <https://www.xuetangx.com/global>. [Đã truy cập 28th May 2024].
- [3] S. H. Han, “PyPI,” Python Software Foundation, 14th June 2020. [Trực tuyến]. Available: <https://pypi.org/project/googletrans/>. [Đã truy cập 28th May 2024].
- [4] L. Long, “vietnamese-fullname-generator,” 19th March 2021. [Trực tuyến]. Available: <https://github.com/lhlong/vietnamese-fullname-generator/commits/main/>. [Đã truy cập 28th May 2024].
- [5] Dat Quoc Nguyen, Anh Tuan Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” trong *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.
- [6] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson, “VnCoreNLP: A Vietnamese Natural Language Processing Toolkit,” trong *Proceedings of NAACL: Demonstrations*, 2018.
- [7] “ChatGPT,” OpenAI, [Trực tuyến]. Available: <https://chatgpt.com/>. [Đã truy cập 28th May 2024].

- [8] “Gemini,” Google, [Trực tuyến]. Available: <https://gemini.google.com/app>. [Đã truy cập 28th May 2024].
- [9] Han, J., Pei, J., Yin, Y. et al, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” trong *Data Mining and Knowledge Discovery* 8, 2004.
- [10] Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua, “KGAT: Knowledge graph attention network for recommendation,” trong *KDD 2019*, 2019.