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

BÀI THỰC HÀNH

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

1. Báo cáo phân tích bộ dữ liệu

1.1. Tìm hiểu dữ liệu

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

1.1.1. 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

1.1.2. Video

-Số lượng: 230,263
-Thông tin: Các video giảng dạy được thu thập từ các khóa học trên nền tảng MOOC. Mỗi vidfeo có các thuộc tính như tiêu đề, thời lượng,

nội dung được giảng dạy, và số lần xem của học viên

1.1.3. 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

1.1.4. 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

1.1.5. 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

1.1.6. 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

1.1.7. 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 tình 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.

1.2. 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:

```python
async def translate(df, batch_start, batch_end):
    tasks = []
    for i in range(batch_start, batch_end):
        tasks.append(async_translate(df.loc[i, COL], i))

    df.loc[batch_start: batch_end - 1, COL] = await asyncio.gather(*tasks)


df = pd.read_json('teacher.json', lines=True)
batch_size = 1000
for i in range(0, len(df), batch_size):
    batch_end = min(len(df), i + batch_size)
    asyncio.run(translate(df, i, batch_end))

df[COL].to_csv(f"translated_{COL}.csv", index=False)
```

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”

```python
course_df = course_df[['id', 'name_trans', 'field', 'prerequisites_trans', 'about_trans', 'resource']]
course_df.head()
```

shape: (5, 6)

id	name_trans	field	prerequisites_trans	about_trans	resource
str	str	list[str]	str	str	list[struct[3]]
'C_584313'	'introduction to "zi zhi tong j…	["history", "chinese language and literature"]	–	"through the teacher's guidance…	[illegible]
'C_584329'	'calculus - limit theory and fu…	["applied economics", "math", … "theoretical economics"]	–	"this course is a basic mathema…	[illegible]
'C_584381'	"photojournalism"	["art", "journalism and communication"]	–	'master basic photography skill…	[illegible]
'C_597208'	"data mining: theory and algori…	["computer science and technology"]	–	"the most interesting theory + …	[illegible]
'C_597225'	"university computer"	[]	–	'university computer courses wi…	[illegible]

-user.json: dịch cột “school”

```
user_df = pd.DataFrame(data_list)
user_df.head()
```
✓ 44.0s Python

	id	name	gender	school	year_of_birth	course_order	enroll_time
0	U_22	我	0.0	None	2015.0	[682129, 2294668]	[2019-10-12 10:28:02, 2020-11-21 14:03:28]
1	U_24	王帅国	1.0	Tsinghua University	6558.0	[597214, 605512, 597211, 597314, 597208, 62950...	[2019-05-20 16:06:48, 2019-05-24 19:34:43, 201...
2	U_25	王帅国	0.0	Tsinghua University	NaN	[1903985]	[2020-08-07 18:59:13]
3	U_53	于歆杰	1.0	Tsinghua University	1973.0	[696679, 1704639, 943255, 1729417, 682164, 177...	[2020-03-01 21:24:30, 2020-03-12 16:17:02, 202...
4	U_54	马昱春	2.0	Tsinghua University	NaN	[682442, 682164, 1748240, 1778890, 1829031, 17...	[2019-10-09 02:17:49, 2019-11-08 00:49:03, 202...

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

```
teacher_df.head()
```

	id	name	name_en	about	job_title	org_name
0	T_1	刘燕妮	Yanni Liu	Graduated from the Philosophy Department of Pe...	lecturer	Tsinghua University
1	T_2	陈怡	Yi Chen	Born in Chongqing in 1945, he graduated from H...	professor	Tsinghua University
2	T_3	程钢	Gang Cheng	Cheng Gang is the course leader of "Introducti...	Associate Professor	Tsinghua University
3	T_4	谢维和	xie wei he	Xie Weihe, PhD, professor, doctoral supervisor...	professor	Tsinghua University
4	T_5	史静寰	Jing-huan Shi	Shi Jinghuan, female, professor and doctoral s...	professor	Tsinghua University

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

```
df = pd.read_json("../translated/concept_translated.json", lines=True)
df.head()
```
✓ 3.1s

	id	name	context
0	K_Nervous system_Histology and Embryology	Nervous system	[]
1	K_TSH cells_Histology and Embryology	TSH cells	['The pituitary gland consists of two parts: t...
2	K_Chromophilic cells_Histology and Embryology	Chromophilic cells	[]
3	K_Growth hormone cells_Histology and Embryology	Growth hormone cells	['Answer: B\n13. Adenohypophysis eosinophils c...
4	K_Limonite_Materials Science and Engineering	Limonite	['\nLimonite is a common iron ore, often forme...

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

```
df = pd.read_json("../original_translated/course-field-translated.json", lines=True)
df.head()
```

✓ 0.0s Python

	course_id	course_name	field
0	584313	Introduction to "Zi Zhi Tong Jian"	[Chinese language and literature, History]
1	681932	"Learning by doing" Java programming	[Computer Science and Technology]
2	674962	The spatial art of "Dream of Red Mansions"	[Chinese language and literature]
3	682709	Introduction to the Critique of Pure Reason	[philosophy]
4	682635	Introduction to "Tongwancheng"	[History]

1.2.2. Khám phá dữ liệu

a) Bảng course.json

Ta xem qua bảng course.json:

```
course_df = pd.DataFrame(data_list)
course_df.head()
```

Python

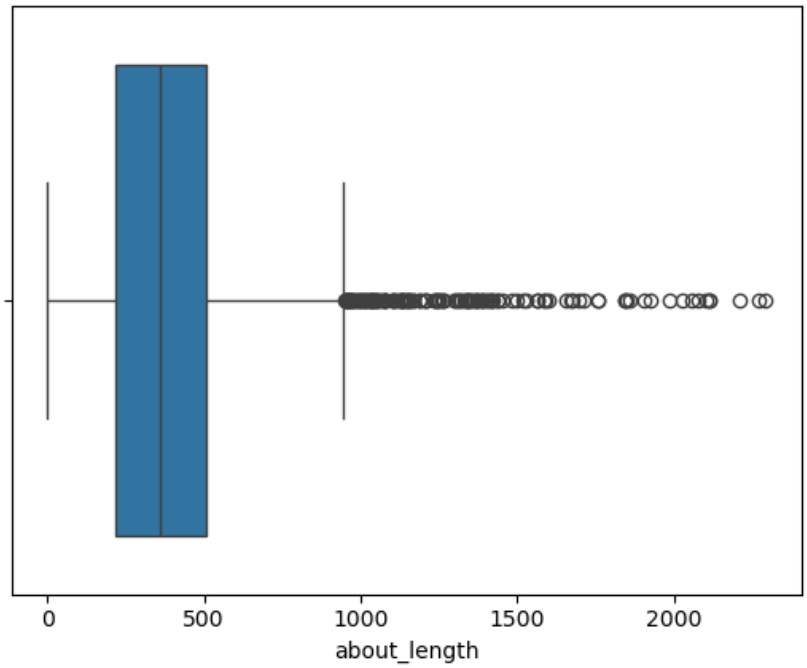
	id	name_trans	field	prerequisites_trans	about_trans	resource
0	C_584313	introduction to "zi zhi tong jian"	[history, chinese language and literature]		through the teacher's guidance, students can g...	[{'titles': ['第一课 导论与三家分晋', '导论', '导论'], 'reso...
1	C_584329	calculus - limit theory and functions of one v...	[applied economics, math, physics, theoretical...		this course is a basic mathematics course in s...	[{'titles': ['序言', '序言', '序言'], 'resource_id':...
2	C_584381	photojournalism	[art, journalism and communication]		master basic photography skills, understand ho...	[{'titles': ['第一章 绪论', '第一讲 引言1', '引言1'], 'res...
3	C_597208	data mining: theory and algorithms	[computer science and technology]		the most interesting theory + the most useful ...	[{'titles': ['走进数据科学：博大精深，美不胜收', '整装待发', 'Vide...
4	C_597225	university computer	[]		university computer courses will be guided by ...	[{'titles': ['第1周： 基于计算机的问题求解', '课程介绍', '开篇'],...

Ta xét độ dài của 3 cột "about", "name_trans" và "resource":

	about_length	name_length	resource_length
count	3781.000000	3781.000000	3781.000000
mean	393.445120	36.942343	71.685533
std	267.904934	21.575065	74.802345
min	0.000000	2.000000	1.000000
25%	217.000000	22.000000	38.000000
50%	361.000000	32.000000	59.000000
75%	509.000000	46.000000	88.000000
max	2293.000000	193.000000	2728.000000

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 tệ 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 tệ 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.

Ta phân tích sâu cột "resource":

```
course_df['resource'][0][0]
```

```
{'titles': ['第一课 导论与三家分晋', '导论', '导论'],
 'resource_id': 'V_849',
 'chapter': '1.1.1'}
```

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 hiển thị cùng một video ccid tại các thời gian bắt đầu khác nhau. Mối 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:

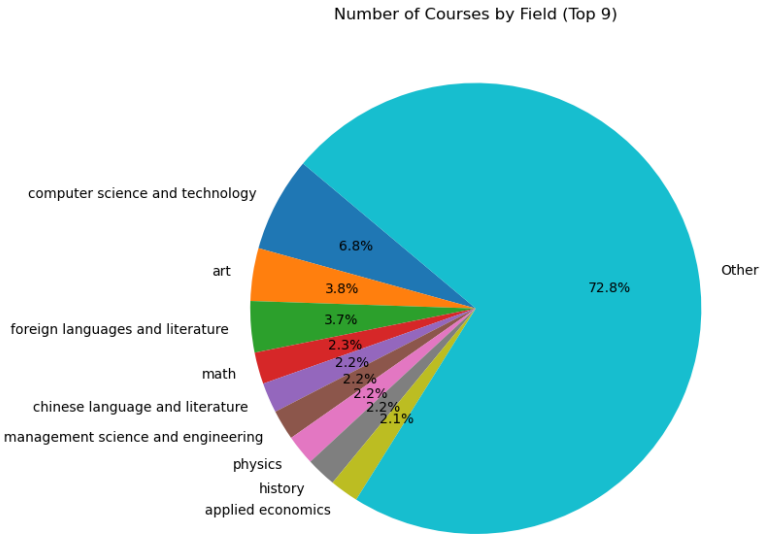
```
videoID = ccid_df['video_id'].unique()

valid_videoID = set(videoID)

non_existent_ids = unique_video_ids - valid_videoID

# Hiển thị kết quả
print(f"Tổng số lượng các video ID không tồn tại: {len(non_existent_ids)}")
print(f"Các video ID không tồn tại: {non_existent_ids}")
```

```
Tổng số lượng các video ID không tồn tại: 2397
Các video ID không tồn tại: {'V_543429', 'V_543378', 'V_543519', 'V_1056006', 'V_374
```

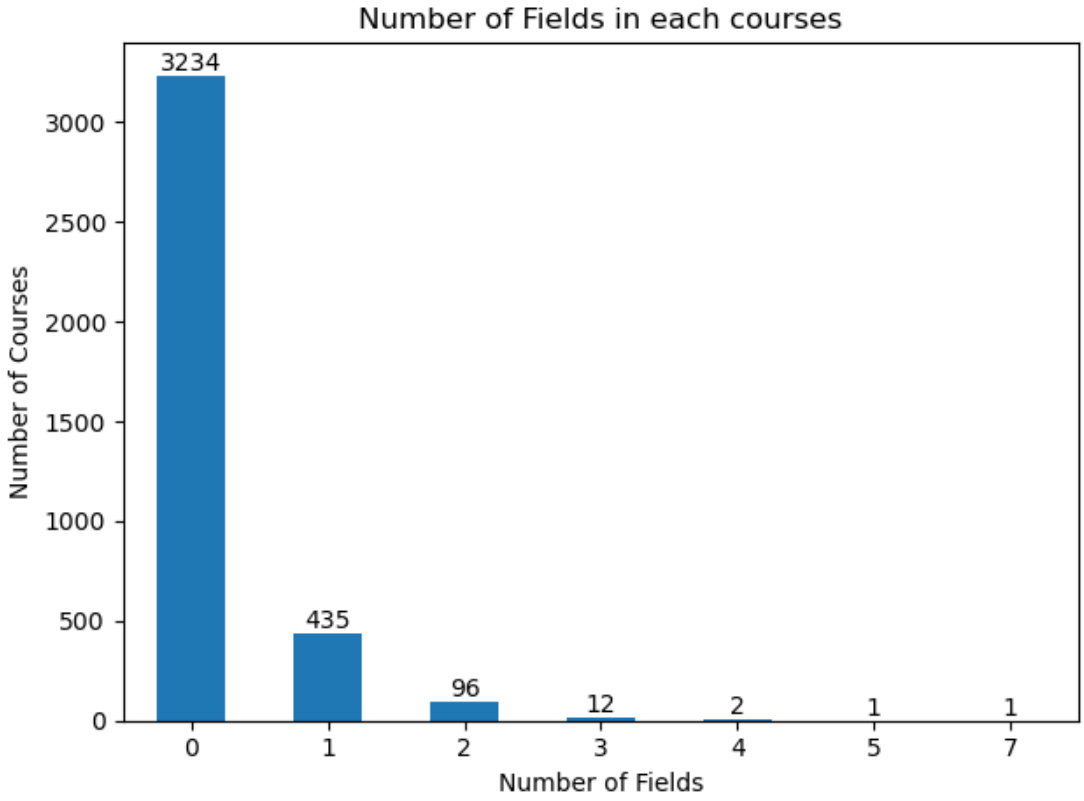
Có 2397 video ID không tồn tại, ta sẽ lọc đi hỗ trợ cho hiển thi thông tin trong tương lai.

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

```
2) Number of courses by field:
field
computer science and technology       63
art                                   35
foreign languages and literature      34
math                                  21
history                               20
                                      ..
marine science                         1
ship and marine engineering            1
army command science                   1
metallurgical engineering              1
basic chinese medicine                 1
Name: count, Length: 81, dtype: int64
```



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"):



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.

1.2.3. Làm sạch dữ liệu

to be continue...

1.2.4. Chuyển đổi dữ liệu

to be continue...

1.3. Phân tích vấn đề

to be continue...

2. Thuyết minh đề tài

2.1. Tên đề tài, thời gian thực hiện, tổng kinh phí

-**Tên đề tài:** Hệ thống khuyến nghị khóa học cho dữ liệu MOOCCubeX
-**Thời gian thực hiện:** 8 tuần
-**Tổng kinh phí dự kiến:** 6.000.000đ (Việt Nam Đồng)

2.2. Nhóm thực hiện:

to be continue...

2.3. Mô tả đề tài

2.3.1. Giới thiệu về bài toán

to be continue...

2.3.2. Ứng dụng

to be continue...

2.3.3. Khó khăn và thách thức

to be continue...

2.3.4. Các dự án liên quan cùng lĩnh vực

to be continue...

2.4. Tổng quan

2.4.1. Ý tưởng và kế hoạch triển khai

to be continue...

2.4.2. Tính cấp thiết

to be continue...

2.4.3. Tính mới

to be continue...

2.5. Mục tiêu đề tài

2.5.1. Mục tiêu về đồ án

to be continue...

2.5.2. Mục tiêu về doanh nghiệp

to be continue...

2.5.3. Mục tiêu về sản phẩm

to be continue...

2.6. Input - Output

to be continue...

2.7. Nội dung bài toán triển khai

2.7.1. Nội dung 1

to be continue...

2.7.2. Nội dung 2

to be continue...

2.7.3. Nội dung 3

to be continue...

2.7.4. Nội dung 4

to be continue...

3. Bộ dữ liệu sau khi tiền xử lý:

4. Content-based Filtering

4.1. Bảng course.json

to be continue...

4.2. Bảng user.json

to be continue...

4.3. Bảng concept.json

to be continue...

4.4. Bảng teacher.json

to be continue...

4.5. Bảng school.json

to be continue...

4.6. Bảng course-field.json

to be continue...