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. Feature Engineering

Feature 1:

- Sử dụng trường name, about, field
- Vectorize các trường
- Tính độ tương đồng giữa các khóa học bằng cosine

Feature 2:

- Sử dụng thêm trường school, concept được nối từ các relations
- Vectorize các trường
- Tính độ tương đồng giữa các khóa họ bằng cosine

1.2. Training

in progress...

1.3. Evaluation

in progress...

2. Thuyết minh đề tài

2.1. Feature Engineering

Sử dụng các cặp user_id và course_id để tạo ra 1 bảng chứa các tương tác giữa người dùng và khóa học như sau:

user_id	course_id
U_24	C_55110
U_24	C_55231
...	...

2.2. Training

Đối với mỗi người dùng, loại bỏ khóa học cuối cùng mà người đó đã học, sau đó với mỗi khóa học chúng ta cặp theo một khóa chưa học hay còn gọi là negative sampling

2.3. Evaluation

Với các khóa học được để lại ở bước trước, cặp chúng vưới 99 khóa học khác mà người dùng đó chưa học, sau đó dùng mô hình đánh giá các khóa học trên rồi tính độ đo Recall@10 và NDCG@10

3. Bộ dữ liệu sau khi tiền xử lý:

3.1. Feature Engineering

Ta sử dụng các dữ liệu từ các bảng sau để tạo ra các feature cho mô hình Factorization Machine:

- User info: Sử dụng trường `id`, `gender`.
- Course info: Sử dụng trường `id`, `school` (course-school), `teacher` (course-teacher), `concept` (course-concept).
- User-course interactions: Sử dụng trường `id`, `course_order`

3.2. Training

3.3. Evaluation

Models	Recall@K		NDCG@K	
	1	10	1	10
FM	0.1408	0.7601	0.1408	0.4309