
2022 ADL Final Project

TA 蔡尚錡 周敦翔
adl-ta@csie.ntu.edu.tw

Final Group Project (3-4 persons)

Task 1 : Propose your own task

- Research-oriented
- Lecture-related methodology
- **Novelty**

Task 2 : Grand Challenge

- Kaggle competition
- Awards for top teams
- **Performance**

Grand Challenge

Task Description

Hahow : online learning platform



Web3 社群經理的 53 堂課：元宇宙轉職就看	客人自動找到你！Google 地標「我的商	色彩加氛圍，我選 Procreate！	Metashape 3D 掃描 你的相機就是掃描器！
目前無評價	目前無評價	目前無評價	目前無評價
課時 599 分鐘	課時 163 分鐘	課時 507 分鐘	課時 169 分鐘
同學 102 人 NT\$3,800	同學 120 人 NT\$1,890	同學 1683 人 NT\$2,850	同學 45 人 NT\$3,540

Task Description

Target :

- learn the correlation between different courses
- predict the courses the user would buy in the future

Task Description



零基础轉職！網頁設計 HTML、CSS 快速入門

目前無評價

課時 530 分鐘

同學 108 人 **NT\$2,690**

2021/1/20



Python 的 50+ 練習：資料科學學習手冊

★★★★★ 12 則評價

課時 1047 分鐘

同學 469 人 **NT\$3,480**

2021/4/6



Web3 給前端工程師的區塊鏈入門

★★★★★ 4 則評價

課時 163 分鐘

同學 40 人 **NT\$3,599**

2021/5/20

Task Description

model prediction :



Unity 從零開始新手入門：2D 橫向捲軸遊戲

★★★★★ 15 則評價

課時 115 分鐘

同學 700 人 NT\$799

Rank : 1



UX 設計/研究全攻略：給新手的職場通識課

★★★★★ 34 則評價

課時 739 分鐘

同學 1808 人 NT\$3,349

2



Adam 個人理財術：從培養財務認知開始！

★★★★★ 179 則評價

課時 166 分鐘

同學 4486 人 NT\$2,250

3



子麵的影像合成課！創作招式 & 接案流程大公開

預購價 NT\$2,380 剩 30 天

已募資 107%

4

Data Files

- Data: course purchase records from Hahow company
- **course purchase records (2021/1/1 - 2021/12/31)**
 - user_id: 用戶識別 ID
 - course_id: 課程識別 ID, 用空格分隔多項
- unseen : user_id is not observed in train.csv
- seen : user_id is observed in train.csv

不管是seen或unseen都不能用
valid裡面的資料去做training

課程購買記錄 (篩選購買時間 : 2021/1/1 - 2021/12/31)

	train.csv	val_seen.csv	val_unseen.csv	test_seen.csv	test_unseen.csv
Time	2021/01/01 - 2021/08/31	2021/09/01 - 2021/10/31	2021/09/01 - 2021/10/31	2021/11/01 - 2021/12/31	2021/11/01 - 2021/12/31
User count	59737	7748	11622	7205	11097

Data Files

- **courses.csv** - all courses information
 - course_id: 課程識別 ID
 - course_name: 課程名稱
 - course_price: 課程價格
 - teacher_id: 老師識別 ID
 - teacher_intro: 老師簡介
 - groups: 課程分類
 - sub_groups: 課程子分類
 - topics: 課程主題
 - course_published_at_local: 該課程識別 ID 的發行時間
 - description: 課程詳情
 - will_learn: 課程詳情 — 你可以學到
 - required_tools: 課程詳情 — 需要準備的工具 / 軟體
 - recommended_background: 課程詳情 — 需要具備的背景知識
 - target_group: 課程詳情 — 哪些人適合這堂課

Data Files

- `course_chapter_items.csv` - all chapters information
 - `course_id`: 課程識別 ID
 - `chapter_id`: 章節識別 ID
 - `chapter_no`: 章節編號
 - `chapter_name`: 章節名稱
 - `chapter_item_id`: 單元識別 ID
 - `chapter_item_no`: 單元編號
 - `chapter_item_name`: 單元名稱
 - `chapter_item_type`: 單元類型 (課程/作業)
 - `video_length_in_seconds`: 單元影片長度 (秒)

Data Files

- <https://drive.google.com/file/d/1rR7hUqBmi8GtjwPNVSmsJXIf0F1WtRu4/view?usp=sharing>
- **users.csv** - all users information
 - user_id: 用戶識別 ID
 - gender: 用戶性別
 - occupation_titles: 用戶職業類別。為複選, 使用逗號分隔多項
 - interests: 用戶興趣。為複選, 格式為 {分類}_{子分類}, 使用逗號分隔多項
 - recreation_names: 用戶喜好。為複選, 使用逗號分隔多項
- **subgroups.csv** - all subgroup ID and names
- **train_group.csv** or **val_seen_group.csv** or **val_unseen_group.csv**
-users and their corresponding subgroup_ids

Evaluation

output_course_format

- you can rank as many **courses_id** as possible

```
user_id,course_id
user_id_1,course_id_1 course_id_2 course_id_3 course_id_4
user_id_2,course_id_1 course_id_2 course_id_3 course_id_4
user_id_3,course_id_1 course_id_2 course_id_3 course_id_4
user_id_4,course_id_1 course_id_2 course_id_3 course_id_4
user_id_5,course_id_1 course_id_2 course_id_3 course_id_4
user_id_6,course_id_1 course_id_2 course_id_3 course_id_4
user_id_7,course_id_1 course_id_2 course_id_3 course_id_4
user_id_8,course_id_1 course_id_2 course_id_3 course_id_4
user_id_9,course_id_1 course_id_2 course_id_3 course_id_4
user_id_10,course_id_1 course_id_2 course_id_3 course_id_4
```

predict and rank the courses a given user will purchase

Evaluation

output_subgroup_format

- you can rank as many **subgroups_id** as possible

```
user_id,subgroup
user_id_1,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_2,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_3,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_4,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_5,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_6,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_7,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_8,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_9,subgroup_1 subgroup_2 subgroup_3 subgroup_4
user_id_10,subgroup_1 subgroup_2 subgroup_3 subgroup_4
```

predict and rank the course subgroup a given user will purchase

Evaluation metrics

Mean Average Precision MAP@50

- It means **mean of AP@k** for all the users
- https://github.com/benhamner/Metrics/blob/master/Python/ml_metrics/average_precision.py

$$\text{AveP} = \frac{\sum_{k=1}^n (P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

$$\text{precision@}k = \frac{tp}{k}$$

1. more correct courses in top-k results
2. higher ranks for correct courses

Challenge Leaderboard

- **Hahow_course challenge -- Seen Domain**
 - <https://www.kaggle.com/competitions/2022-adl-final-hahow-seen-user-course-prediction>
- **Hahow_subgroup challenge -- Seen Domain**
 - <https://www.kaggle.com/competitions/2022-adl-final-hahow-seen-user-topic-prediction>
- **Hahow_course challenge-- Unseen Domain**
 - <https://www.kaggle.com/competitions/2022-adl-final-hahow-unseen-user-course-prediction>
- **Hahow_subgroup challenge-- Unseen Domain**
 - <https://www.kaggle.com/competitions/2022-adl-final-hahow-unseen-user-topic-prediction>

Attention

- **DO NOT CHEAT!**
- **DO NOT TRY TO FIND THE LABELS OF THE TEST SET.**
- You must provide code to reproduce your model predictions
- You can use any other public datasets and models
- Please do not make extra submission personally
- Your team name should be team_{第幾組}_xxxx (xxx可隨便取名不用學號)
- sign the NDA in COOL and we will give you the data permission

Gradings

Grading (35%)

- Oral Presentation (10%)
- Report & Code (12%)
- Performance (10%)
 - Grand challenge: leaderboard performance
 - Choose your own: comparison with current SOTA
- Participation (3%)

Oral Presentation (10%)

- Video recording
 - Duration : 8-10 minutes
 - Note: introduce yourself in the beginning of the video
 - Presenting from all members is highly recommended
- Submission: submit a YouTube Link to Cool by 2023/1/2 23:59
- The grade mainly focuses on model novelty, clarity, and result analysis

Report and Code (12%)

- Wrap-up report

- Content

- Abstract
 - Introduction
 - Related work
 - Approach
 - Experiments
 - Discussion
 - Conclusion
 - Work Distribution

Grand Challenge



Propose Your Own

Task Definition



Related Work



Proposed Method



Experiments



Discussion



Conclusion



- Code w/ README

- Submit to COOL by 2023/01/05 23:59

Performance for Grand Challenge (10%)

- course seen leaderboard (3%)
- subgroup seen leaderboard (3%)
- course unseen leaderboard (2%)
- subgroup unseen leaderboard (2%)
- The score is based on ranking of the test set on 2023/1/2 23:59
 - Top 1~5: 3 points
 - 6~10: 2.5 points
 - 11~30: 2 points
 - 30+: 1.5 points
- 0 point if we can't reproduce your submission!

不一定會完全按照名次硬性給分
如果大家結果都很接近也可能給同樣的分數

Participation (3%)

- Each student ask **at least 1 question** to **at least 3 teams** (1.5%)
 - Videos will be submitted to COOL for everyone access
 - Questions should be asked under the videos at COOL
 - Due on **2023/01/05 23:59**
- Each team should reply **all questions** (1.5%)
 - Due on **2023/01/06 23:59**
 - The team which receives most questions will get additional score
- Note
 - DON'T ask the repeating questions
 - DON'T ask the questions to your own team

Q & A