同濟大學
TONGJI UNIVERSITY

同濟大學軟件學院
SCHOOL OF SOFTWARE ENGINEERING

School of Software Engineering

Data Analysis and Data Mining

Final Project

Semester 2 2020/2021

31 May, 2021

Final Project of Data Analysis and Data Mining Course

ASSIGNMENT

You can work as at most three-person groups, and the following is the topic of the project you need to complete.

Topic: program recommendation based on TV set-top box

In the final project, you are required to recommend *movies or TV series* to TV set-top box users. Specific requirements of the project are shown below:

a) Extract the features from the provided dataset using any methods as you like.
   Bonus: To extract deep logical connections among television programs, you can introduce additional knowledge graph information into recommendation(An example of the combination of knowledge graph and recommendation system [1]). Both call the API of an off-the-shelf knowledge graph or build your own knowledge graph are allowed.
b) Based on the analysis of the provided datasets, please use three recommendation algorithm(ALS[2], LightGBM[3], NCF[4]) to complete the *movies/TV Series* recommendation to meet the interests of users as much as possible. You can refer to [5] for more tools.
   Bonus: Implement two additional recommendation algorithms from [5].
c) Evaluate the recommendation results. You can use common recommender system metrics(e.g. F1 score, AUC) or other ways to explain your results.

Remarks: The bonus are not required.

REFERENCE

[1]. Wang, Hongwei, et al. "DKN: Deep knowledge-aware network for news recommendation." *Proceedings of the 2018 world wide web conference*. 2018. https://dl.acm.org/doi/pdf/10.1145/3178876.3186175
[2]. Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," in *Computer*, vol. 42, no. 8, pp. 30-37, Aug. 2009. https://ieeexplore.ieee.org/document/5197422
[3]. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf
[4]. Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu & Tat-Seng Chua, Neural Collaborative Filtering, 2017, https://arxiv.org/abs/1708.05031
[5]. https://microsoft-recommenders.readthedocs.io/en/latest/recommender.html

SUBMISSION

Your submission file should be a compressed archive named by Final_StudentIDs.[zip/rar/tar.gz]. The content of your submission should include but not limited to the following part:

1) Source code
2) README file
3) Report document

a) Workflow and functionality of your method.
b) Datasets and how you performed the experiments.
c) Evaluation and analysis of your results.
d) Reference
e) …

Deadline for submission: **17th June**, 2021

ASSESSMENT

A presentation/demonstration is required.

50% of the scores are given from the performance of the methods on the topic: feature extraction(10%), recommendation algorithms(30%), evaluation(10%).

20% are from the documentation/ reports.

The rest 30% are from the presentation. Each presentation should be in 8 minutes. As a team work, you should specify the contributions of each team member in the documentations and presentations.

No plagiarism from other teams!

DATASET

1、 播放记录

ASSET_TYPE 内容编排信息	CHAPTER 第几集	CODE 媒资 ID	CP 内容提供商时长	DURATION
/福建广电/儿童/动画剧集	13	OTHE1000000001983579	GZCTV	792
…	…	…	…	…
ELAPSED 观看断点	**ENDTIME 结束观看时间**	**FOLDERCODE 栏目 ID**	**ITEMCODE 子集 ID**	**NAME 内容名称**
738	01-3 月 -21 09.42.50.000000 上午	MANU00000000002498 53	OTHE000000000203358	猪猪侠之南海日记 1
…	…	…	…	…
OPK 区域码	**PLAYTIME 观看时间**	**PORTAL_VER 互动版本**	**PRICE 价格**	**PROVIDER_ID 内容提供商**
-1	01-3 月 -21 09.36.19.154000 上午	home		华视网聚
…	…	…	…	…
SHOW_TYPE 内容类型	**VIRTUAL_OPK 虚拟区域码**	**WATCHTIME 观看总时长**	**STBID 智能卡号**	
2	-1	391	8595100000000000	
…	…	…	…	

2、 同洲媒资库（影视）

ACTORS_DISPLAY 演员	ASSET_ID 内容 ID	ASSET_TYPE 内容编排信息	CHAPTER 集数	DIRECTOR 导演
乐迪,小爱,酷飞	OTHE1000000001967919	/福建广电/儿童/动漫剧集	12	佚名
…	…	…	…	…
DISPLAY_RUN_TIME 时长	**END_DATE_TIME 版权到期时间**	**FOLDER_ASSET_ID 栏目 ID**	**FOLDER_NAME 栏目名称**	**IS_PACKAGE 是否媒资包**
		MANU0000000000249853	3 至 6 岁幼教	1
…	…	…	…	…
ORDER_NUMBER 排序	**ORIGIN_NAME 区域**	**PREVIEW_ASSET_ID 预览内容 ID**	**PUBLISH_DATE 内容发布日期**	**RUNTIME 总时长**
9999939	中国内地	OTHE1000000001967919	2020/10/2	
…	…	…	…	…
SHOW_TYPE 内容类型	**SUMMAR_MEDIUM 简介**	**TITLE_BRIEF 内容名称**	**YEAR 年代**	**CP 内容提供商**
2	超级飞侠们被中国各种文化深深吸引	超级飞侠看中国	2020	芒果专区
…	…	…	…	