The Hong Kong Polytechnic University

COMP4434 Big Data Analytics
Individual Project

Introduction

PolyTube is an online media platform which provides online teleplay services. For teleplays already in service, the platform will record user ratings and give an average rating for each teleplay on the web page. For recently published teleplays without ratings, how to predict them is essential for the investment policy of PolyTube. On the other hand, to attract users watching more teleplays on the platform, providing accurate and personalized recommendation services is also meaningful for revenue increase. In this project, you are required to help PolyTube improve their system with the knowledge you learned in Big Data Analytics.

Dataset

This dataset is composed of two parts:

1. *Teleplay.csv* contains history information about a total of 10,204 teleplays and their average ratings given by the users.
   - teleplay_id – a unique id identifying a teleplay.
   - name – the full name of the teleplay.
   - genre – the genre of the teleplay, separate with a comma if it belongs to multiple genres.
   - length – short (<20 min), medium (20~40 min), long (>40 min).
   - episodes – how many episodes in the teleplay.
   - rating – the average rating out of 10 for the teleplay.
   - members – the number of registered users that have rated the teleplay.
2. *Rating.csv* contains the latest rating of 73516 new users who have watched these teleplays.
   - user_id – a unique id identifying a new user.
   - teleplay_id – the id of a teleplay that the user has watched.
   - rating – rating out of 10 the user assigns (-1 if the user watched it but didn't assign a rating).

Task

This project contains two tasks:

1. Design prediction models to predict the rating of recently published teleplays.
2. Design recommendation systems to provide personalized recommendation services.

Submission Format

1. For task 1, fill in the blank space of *New_Teleplay.csv,* And rename it as *your_student_ID_task1.csv*. (e.g., *190XXXXXXR_task1.csv*)
2. For task 2, predict user 53698's personalized rating of all teleplays. Save one file with the following format and name it as *your_student ID_task2.csv*. (e.g., *190XXXXXXR_task2.csv*)

Teleplay_id	Predicted rating
1	
2	
…	

3. The source code, readme file, report, and above two files need to be packed into a zip file named *your_student_ID.zip* and then submitted to Blackboard. (e.g., *190XXXXXXR.zip*)

Mid Report

A mid report should include following information of this project, e.g.

- Introduction
- Data preprocessing/analytics of Task 1
- Model design and implementation of Task 1
- Preliminary result of Task 1
- Future plan
- Reference

Final Report

A final report should include following information of this project, e.g.

- Introduction
- Data preprocessing/analytics
- Model design and implementation
- Performance evaluation and discussions
- Summary and future work
- Reference

The Hong Kong Polytechnic University

Project Grading

The project is 40% of the total subject assessment. There are 4 sets of deliverables throughout the semester, e.g.

- Mid report (5%)
- Final report (8%)
- Project presentation (7%)
  - ✓ Presentation slides in 15-20 pages
  - ✓ Record a presentation video in 10-15 minutes
- Code (20%)
  - ✓ Work independently
  - ✓ You are required to implement using Python
  - ✓ Only using the Third-party related packages will lose the point for coding

Grading Criteria for Project Report

We will grade your report based on following 3 aspects:

1. Integration of course content:
   - You are encouraged to apply the knowledge you learned in the course as much as possible.
2. Diversity of your methodology:
   - You are encouraged to use more than one algorithm for each task. Creative comparisons, critical thinking, and valuable discussions will be very impressive in your project report.
3. Performance of your models:
   - Your prediction model will be compared with the true value and the grading is based on the Root Mean Square Error (RMSE). Lower the RMSE as much as possible.

Timeline:

Submit deadline:

- Mid report: **23:59 15 March 2021**
- Final report: **23:59 2 May 2021**