Final report for PMLDL Assigment 2

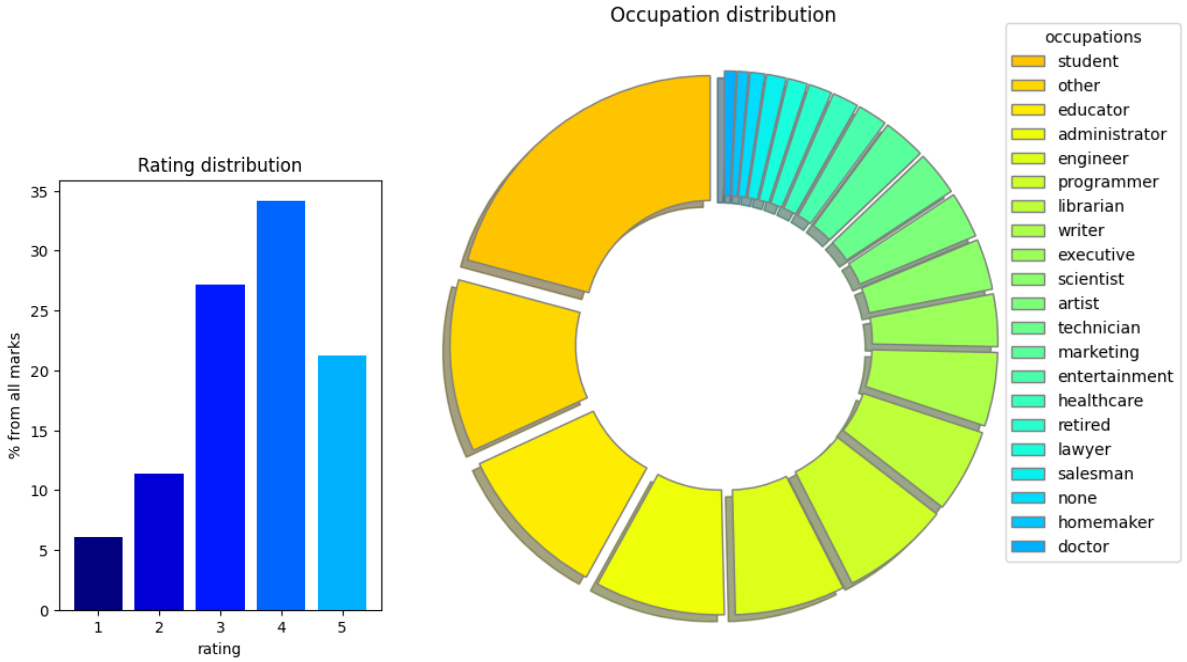
(by Ksenia Shchekina k.shchekina@innopolis.university)

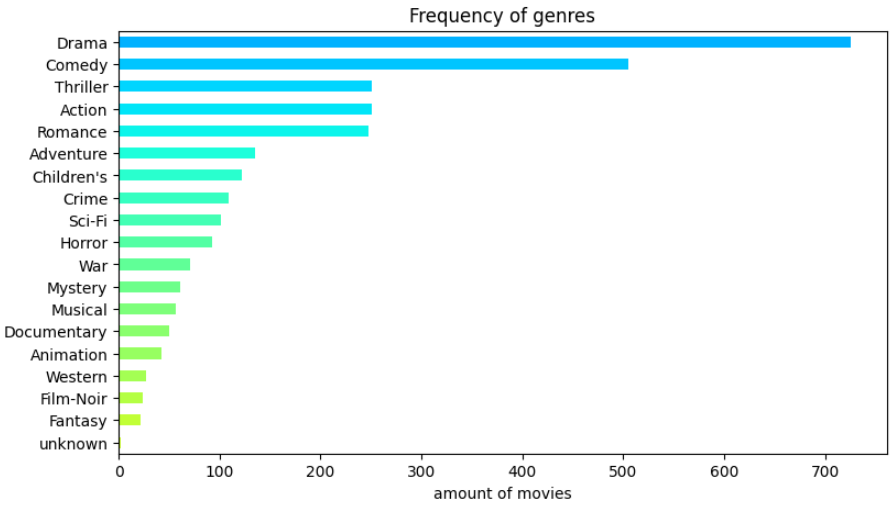
Introduction

To solve this task I need to write a recommendation system for movies. After some exploration I decide to write hybrid system based on two methods: content-based approach and user-user collaborative filtering

Data analysis

In this assignment MovieLens 100K dataset was used. Dataset consists of 10000 ratings to 1682 movies from 943, ratings are ranged from 1 to 5. The user information dataset contains user id, age, gender, occupation, zip code. The movie information dataset contains movie id, movie title, release date, video release date, IMDB URL, and genres. At preprocessing, columns ‘video_release_date’ and ‘IMDB_URL’ were dropped from movie information, because this information cannot affect users' preferences. There are also some statistics from datasets to explore data distribution.





Model Implementation

Model for this task is a hand-written hybrid model based on two methods: content-based approach and user-user collaborative filtering, implemented with references to lecture[1].

Model Advantages and Disadvantages

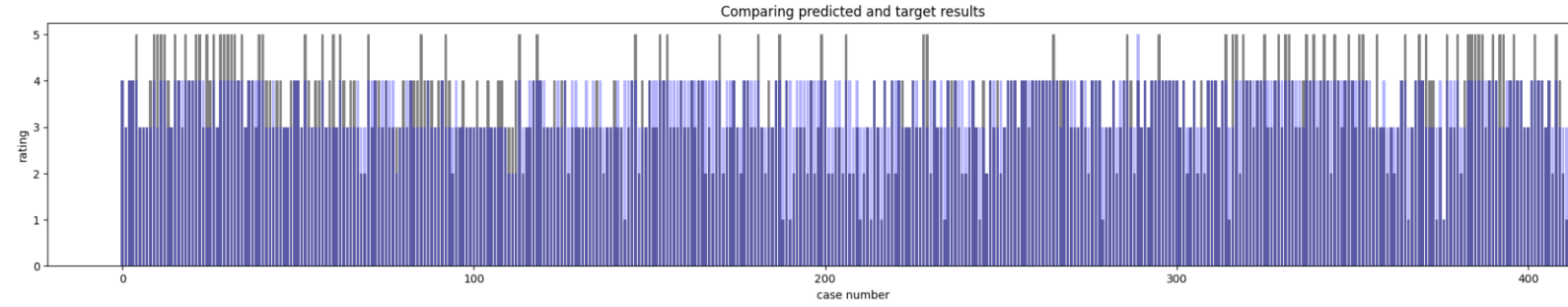
Separately, methods have some advantages and disadvantages. But hybrid system combines all advantages and have only one main disadvantage: the problem with new users.

Training Process & Evaluation

Model consists of two implemented methods. Collaborative filtre approach doesn't need a training process, this method just makes computation on the dataset. Size of the test dataset for this approach was 417 cases (~5 % of the whole dataset). Another approach is the implementation of a content-based method that includes a decision tree for each user.

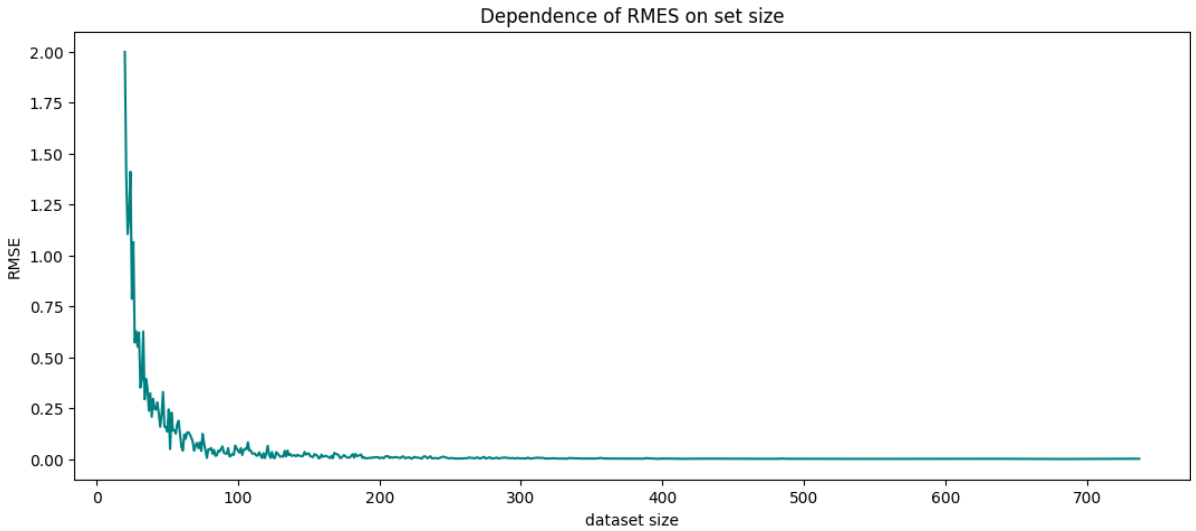
The RMSE was chosen as an evaluation metric.
Result of collaborative filtre method' testing:

```
RMSE = 0.9157067802780935
```



Result of content based method' testing:

```
average RMSE = 0.08249232980342763
```



Results

Main result of work is a hybrid recommendation system model implemented in a file `model.py`. Also there are evaluation graphics of this model that provide educational value. As shown in the evaluation the recommendation system has better results in mode ‘content_b’, which performs a content_based approach, but to recover disadvantages of this approach it is recommended to use a ‘hybrid’ method.

References

[1] file:///C:/Users/Admin/OneDrive/Desktop/PMLDL.F2023.Week11.pdf