# Задание 2.2. Рекомендательные системы. Библиотека Surprise


Тушканова Ольга Николаевна

Санкт-Петербург

2020

# Входные данные

Набор **ml-100k**

user, item, rating, timestamp

943 users

1682 items

100000 ratings

```
62  498 4   879373848
62  382 3   879375537
28  209 4   881961214
135 23  4   879857765
32  294 3   883709863
90  382 5   891383835
286 208 4   877531942
293 685 3   888905170
216 144 4   880234639
166 328 5   886397722
250 496 4   878090499
271 132 5   885848672
160 174 5   876860807
265 118 4   875320714
198 498 3   884207492
42  96  5   881107178
168 151 5   884288058
110 307 4   886987260
58  144 4   884304936
90  648 4   891384754
271 346 4   885844430
```

# Задание

1. Оценить по метрике **RMSE** с помощью функции **cross_validate** следующие алгоритмы:

- прогнозирование случайного рейтинга на основе распределения всех рейтингов в наборе;
- user-based коллаборативную фильтрацию, метод kNN, **k = 30,** метрика **косинуса;**
- user-based коллаборативную фильтрацию, метод kNN, **k = 30,** метрика **Mean Squared Difference** ;
- user-based коллаборативную фильтрацию, метод kNN, **k = 30, метрика корреляция Пирсона;**
- SVD алгоритм.

2. Для лучшего алгоритма по метрике RMSE рассчитать метрики precision@k and recall@k для k=5 и порога отсечения 3.52, усредненные по всем пользователям.

3. Для заданного пользователя (номер в списке) с помощью лучшего алгоритма по метрике RMSE вывести **топ-5** рекомендаций (те фильмы, для которых у пользователя **нет оценки**) **с названиями, датой выхода и рейтингом**.

# Выходные данные

Пример вывода:

```
User 88
   1240 ('Ghost in the Shell (1995)', '12-Apr-1996') 4.51
   1368 ('Mina Tannenbaum (1994)', '01-Jan-1994’)    4.403
   1449 ('Pather Panchali (1955)', '22-Mar-1996’)    4.338
   1131 ('Safe (1995)', '01-Jan-1995’)               4.214
   922  ('Dead Man (1995)', '10-May-1996’)           4.186
```

# Примечание 1

- Для расчета оценки $\hat{r}_{ui}$ пользователя u и для фильма i в подходе user-based коллаборативной фильтрации, метод kNN, использовать формулу

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v=1}^{k} sim_{vu} \cdot (r_{vi} - \bar{r}_v)}{\sum_{v=1}^{k} |sim_{vu}|}$$

где $\hat{r}_{ui}$ - рассчитываемая оценка,

$\bar{r}_u$ - средняя оценка у пользователя u,

$\bar{r}_v$ - средняя оценка у пользователя v,

$r_{vi}$ - оценка пользователя v для фильма i,

$sim_{vu}$ - значение метрики сходства для пользователей u и v

# Примечание 2

~/.surprise_data/ml-100k/ml-100k/README – описание формата данных

~/.surprise_data/ml-100k/ml-100k/u.item – информация о фильмах

Конфигурация метрики сходства
https://surprise.readthedocs.io/en/stable/prediction_algorithms.html#similarity-measures-configuration

Часто задаваемые вопросы

https://surprise.readthedocs.io/en/stable/FAQ.html#

# Как сдать?

0. Не списывать.
1. Выгрузить на GitHub код своего рекомендательного алгоритма.
2. Текстовый файл с предсказаниями выгрузить на курс в «Сдать задание №2.2. Рекомендательные системы: Библиотека Surprise».
3. Получить мой комментарий.
4. Подойти ко мне на практике и обсудить код.
5. После 03.10 баллы снижаются на 30%.

# Баллы

- 0-10 за все задание

Спасибо за внимание!