# Практическая часть

Работаем с набором данных MovieLens 100K Dataset https://grouplens.org/datasets/movielens/100k/

Вам нужны только файлы u.data, u.genre, u.info, u.item и u.occupation. Это текстовые файлы, которые можно загрузить в pandas при помощи read\_csv с правильным набором параметров

|  |  |  |
| --- | --- | --- |
| Номер | Задание | Баллы |
| 1 | Загрузить данные в pandas | 10 |
| 2 | Посчитайте среднюю оценку, которую поставил каждый пользователь | 20 |
| 3 | Отсортируйте фильмы по количеству оценок | 20 |
| 4 | Средствами pandasa соедините данные по оценкам с данными по пользователям (u.user) | 30 |
| 5 | Построить модель, которая предсказывает оценку для фильма и пользователя (воспользоваться RandomForestRegressor, тк мы решаем задачу регрессии) | 50 |
| 6 | Оценить качество регрессора на отложенной (train\_test\_split) выборке по одной из метрик на выбор - RMSE, RMAE, R^2 | 20 |
| 7 | Загрузить данные в spark | 10 |
| 8 | Средствами спарка вывести среднюю оценку для каждого фильма | 20 |
| 9 | В спарке получить 2 датафрейма с 5-ю самыми популярными и самыми непопулярными фильмами (по количеству оценок, либо по самой оценке - на Ваш выбор) | 20 |
| 10 | Средствами спарка соедините информацию по фильмам и жанрам (u.genre) | 30 |
| 11 | Посчитайте средствами спарка среднюю оценку для каждого жанра | 40 |
|  | **Итого максимум баллов** | 270 |

# Теоретическая часть

Вы - главный по данным в среднем по объему просмотров интернет-кинотеатре. Ваша задача разработать стратегию внедрения хранилища данных и работы с большими данными в этой компании. Задания:

|  |  |  |
| --- | --- | --- |
| Номер | Задание | Баллы |
| 1 | Описать основные имеющиеся данные и источники их поступления | 20 |
| 2 | Сформулировать цель проекта и ключевые метрики, на которые мы будем влиять | 30 |
| 3 | Описать основные сущности в хранилище данных | 30 |
| 4 | Описать основные проверки на качество данных (10-20 штук) | 20 |
| 5 | Описать основные бизнес-отчеты, которые мы сможем построить из имеющегося хранилища данных | 20 |
| 6 | Описать основные DS-модели, которые мы сможем построить на основе имеющихся данных | 30 |
| 7 | Описать источники внешних данных и какой value они смогут дать | 30 |
|  | Итого | 180 |

# Оценки

|  |  |  |
| --- | --- | --- |
| Оценка | Мин баллы за практику | Мин баллы за теорию |
| Зачет | 140 | 140 |
| Зачет с отличием | 210 | 170 |