

Задание для итоговой работы

В качестве датасета используем данные условного онлайн-кинотеатра. У нас есть информация о пользователях, фильмах, а так же оценках, которые пользователи поставили тому или иному фильму.

В рамках работы хотим провести исследование текущей ситуации и решить бизнес-кейс с рекомендациями фильмов пользователям (очевидно, что не все пользователи смотрели все фильмы и нам нужно каким-то образом рекомендовать пользователю, что ему посмотреть следующим)

Датасет доступен тут - <https://grouplens.org/datasets/movielens/100k/>

Описание тут - <http://files.grouplens.org/datasets/movielens/ml-100k-README.txt>

Работа состоит из 3-х частей:

- Практика Google Sheets
- Практика Python
- Теоретическая часть

За каждую часть можно получить зачет или не зачет. Для зачета по работе достаточно получить зачет за две из трех частей.

САМИ ЗАДАНИЯ РАСПОЛОЖЕНЫ НА СЛЕДУЮЩИХ СТРАНИЦАХ ЭТОГО ДОКУМЕНТА.

Практика Google Sheets

Скачайте датасет MovieLens 100K к себе на компьютер

<https://grouplens.org/datasets/movielens/100k/>

В данной части мы выступаем в роли bi-аналитика и хотим представить заказчику общую информацию по фильмам, пользователям, а так же построить топ активных пользователей за последние 3 месяца для мотивации.

Для того, чтобы загрузить данные в Google Sheets надо переименовать файл u.user в файл с именем u.user.csv. В качестве разделителей там используется символ |

Номер	Задание	Баллы
1	Построить гистограмму пользователей по возрасту	10
2	Построить 2 графика, показывающих распределение людей по профессиям в зависимости от их пола	10

Аналогично хотим посмотреть на данные по фильмам. Для того, чтобы загрузить данные в Google Sheets надо переименовать файл u.item в файл с именем u.item.csv. В качестве разделителей там используется символ |

Номер	Задание	Баллы
3	Построить график количества фильмов по жанрам	10
4	Построить график количества фильмов по годам	10

Наконец, мы хотим найти самых активных пользователей нашего портала для мотивирования. Для того, чтобы загрузить данные в Google Sheets надо переименовать файл u.data в файл с именем u.data.csv. В качестве разделителей там используется символ табуляции

Номер	Задание	Баллы
5	Построить график количества оценок по	10

	месяцам и годам (преобразование timestamp в дату см тут https://stackoverflow.com/questions/45227380/convert-unix-epoch-time-to-date-in-google-sheets)	
6	Выявить top-5 самых активных пользователей (больше всего оценок) за последние 3 месяца	10

Итого - максимум 60 баллов. Для зачета надо набрать минимум 40

Практика Python

В данном разделе мы выступим в роли data scientist и попытаемся построить простую модель для рекомендации фильмов пользователям.

Номер	Задание	Баллы
1	Загрузите в колаб файлы по оценкам (ratings) и фильмам (movies) и создайте на их основе pandas-датафреймы	10

Сформировав общий топ фильмов в прошлой практике, мы хотим сделать шаг вперед и начать советовать пользователю те фильмы, которые могли бы быть для него наиболее интересны. Наша цель - научиться предсказывать оценку фильма пользователем. Для тестирования модели найдем пользователя, который поставил больше всего оценок

Номер	Задание	Баллы
2	Средствами Pandas, используя dataframe ratings, найдите id пользователя, поставившего больше всего оценок	10

Отберем фильмы, которые оценил данный пользователь

Номер	Задание	Баллы
3	Оставьте в датафрейме ratings только те фильмы, который оценил данный пользователь	10

Для построения модели нам нужны признаки. В качестве таковых будем использовать:

- Год выхода
- Жанры
- Общее количество оценок
- Суммарную оценку

Номер	Задание	Баллы
4	Добавьте к датафрейму из задания 3 столбцы: <ul style="list-style-type: none">• По жанрам. Каждый столбец - это жанр. Единицу записываем, если фильм принадлежит данному жанру и 0 - если нет	10

	<ul style="list-style-type: none"> • столбцы с общим количеством оценок от всех пользователей на фильм и суммарной оценкой от всех пользователей 	
--	---	--

Теперь все готово и можно строить модель

Номер	Задание	Баллы
5	Сформируйте X_train, X_test, y_train, y_test	10
6	Возьмите модель линейной регрессии (или любую другую для задачи регрессии) и обучите ее на фильмах	10
7	Оцените качество модели на X_test, y_test при помощи метрик для задачи регрессии	10

Вторая часть практики Python связана со Spark'ом

8	Загрузить данные в spark	10
9	Средствами спарка вывести среднюю оценку для каждого фильма	20
10	Посчитайте средствами спарка среднюю оценку для каждого жанра	20
11	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

Итого - максимум 140 баллов. Для зачета необходимо набрать 100

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 бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу	10
2	Описать основные имеющиеся данные и источники их поступления	20
3	Описать основные сущности в хранилище данных (схема звезда) и процесс заливки данных	20
4	Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке	10
5	Придумать Data-проект, который должен улучшить показатели Вашего бизнеса и расписать его по Crisp-DM	20
6	Описать требуемые роли в команде по работе с данными на этапах 4 и 5	30
	Итого	110

Итого - максимум 110. Для зачета необходимо набрать 90