

РЕКОМЕНДАЦИИ НА ОСНОВЕ СОДЕРЖАНИЯ



Артур Сапрыкин

—

ПЛАН ЗАНЯТИЯ

ПЛАН ЗАНЯТИЯ

1

откуда
берутся фичи

2

item-to-user
content-based
filtering

3

item-to-item
рекомендации



ЦЕЛИ ЗАНЯТИЯ

ЦЕЛИ ЗАНЯТИЯ

ПОСЛЕ ЗАНЯТИЯ СМОЖЕТЕ

1

feature
engineering
категориальных
признаков

2

Строить item
to item RS

3

Строить item to
user RS

—

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ

ЦЕЛИ ЗАНЯТИЯ

ОТКУДА БЕРУТСЯ ФИЧИ

1

ручное
извлечение

2

парсинг
внешних
источников

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ



THE MUSIC GENOME PROJECT

1. Замкнутая разработка интернет-радио Pandora
2. Команда экспертов с музыкальным образованием
3. 450 уникальных фич на каждую звукозапись

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ

МИКРО-ЖАНРЫ NETFLIX

1. Распределённая команда тегировщиков
2. Десятки страниц правил тегирования
3. Почти **сто тысяч** микро-жанров:
 - документальные фильмы о чернокожих преступниках
 - страшные фильмы 80-х годов о культах и сектах
 - приключенческие фильмы 30-х годов о шпионах


РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ


СВОЯ КОМАНДА

- разработка правил тегирования
- найм и обучение экспертов
- сервисы вроде **Mechanical Turk** или **Толока**
- перекрёстная проверка
- голосование или арбитраж

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ





MOVIELENS TAGS



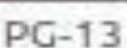











movielens 



browsing by tag

dystopia

view:   filters: tag: dystopia  more 

| The Matrix | V for Vendetta | Blade Runner | Gattaca | Children of Men |
|--|--|--|--|--|
| 1999  136 min | 2006 • 132 min | 1982  117 min | 1997  106 min | 2006  109 min |
|  |  |  |  |  |
|  |  |  |  |  |

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ

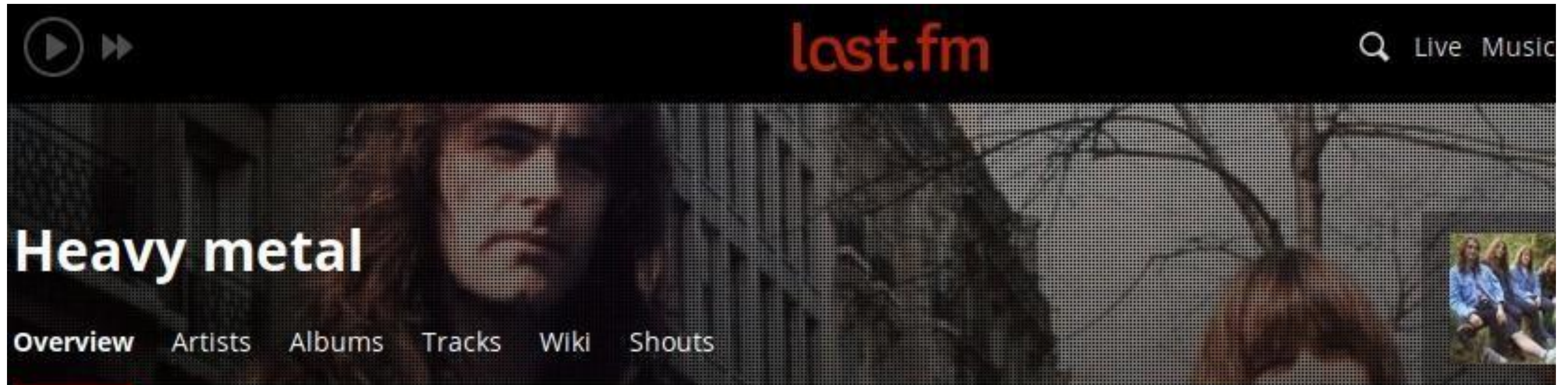
MOVIELENS TAGS

1. сотни тысяч
пользователей

2. сотни различных тегов

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ

LAST.FM TAGS



Related to: [metal](#) · [hard rock](#) · [power metal](#) · [thrash metal](#) · [rock](#) · [speed metal](#)



Play heavy metal tag

РУЧНОЕ ИЗВЛЕЧЕНИЕ ФИЧ

LAST.FM TAGS

1. десятки миллионов
пользователей

2. сотни тысяч различных тегов

Из своей практики

- Контекст покупки товара - магазин
- Всего магазинов в пилотном проекте ± 50
- Хотим извлечь внешние признаки (которых нет в качестве данных ритейлера)
- Много внешних статических данных о локации (google, yandex, 2gis, cian, etc.)

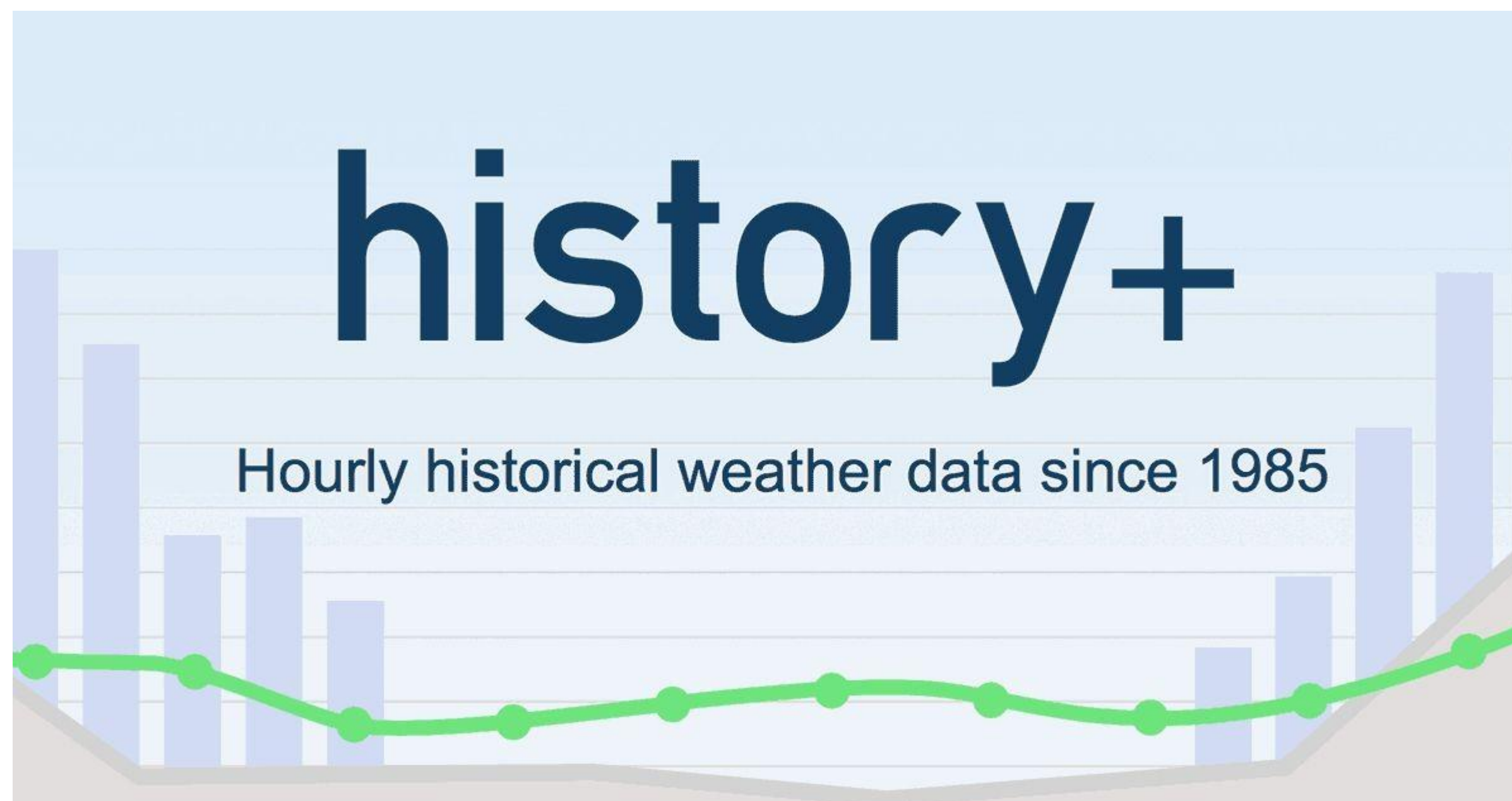
ПАРСИНГ

ПАРСИНГ ВНЕШНИХ ДАННЫХ

- очень много *(сырых)* данных
- бесплатно
- нужны правила дедубликации и прочее
- достоверность

ПАРСИНГ

Из своей практики



ПАРСИНГ

Из чужой практики



DOUBLE DATA



—

ПРАКТИКА

Хотим сделать CBRS для фильмов

Предварительно нужно посмотреть на распределения/статистики имеющихся фич. Знаем про TF-IDF, хотим посмотреть, как его лучше использовать

Что делать?

1. Получите гистограмму количества тегов на фильм/пользователя
2. Получите график количества тегов по месяцам
3. Получите гистограмму количества жанров на фильм

Сколько есть времени?

10 минут

ОСНОВНАЯ CONTENT-BASED МОДЕЛЬ

ОСНОВНАЯ CONTENT-BASED МОДЕЛЬ

ОСНОВНЫЕ ПРИНЦИПЫ

фичи — свойства объекта

один пользователь — одна модель

целевая переменная — релевантность пользователю

ОСНОВНАЯ CONTENT-BASED МОДЕЛЬ

НА УРОВНЕ ПОЛЬЗОВАТЕЛЯ

1. Коэффициенты модели - профиль пользователя
2. Важна L_1 -регуляризация (“совсем не нравится”)
3. Мало данных

ПРЕИМУЩЕСТВА

1. Нет проблемы холодного старта объектов
2. Можно что-то рекомендовать даже пользователю с одним действием
3. Работает для любого типа объектов

НЕДОСТАТКИ

1. Холодный старт пользователя
2. Требуется распределенного обучения
3. Не учитывает общие паттерны поведения пользователей

ОСНОВНАЯ CONTENT-BASED МОДЕЛЬ



КОГДА ХОРОШО ИСПОЛЬЗОВАТЬ

1. Богатые по содержанию объекты
2. Мало взаимодействий пользователь-объект

—

ПРАКТИКА

Рекомендации к фильму

Холодный старт. Гипотеза: рекомендации похожих фильмов увеличат время сессии / конверсию в просмотр

Что делать?

1. TF-IDF на тегах/жанрах
2. Найдите ближайших соседей любимого фильма
3. Найдите ближайших соседей от одного из ближайших соседей
4. Прodelайте то же для других расстояний

Сколько есть времени?

20 минут



ITEM-TO-ITEM РЕКОМЕНДАЦИИ

КАК НАЧАТЬ

1. Получить векторные представления объектов
2. Выбрать какую-нибудь метрику (*формулу расстояния*)
3. Найти матрицу расстояний между объектами
4. Рекомендовать к выбранному объекту его ближайших соседей

А КАК ЖЕ МАШИННОЕ ОБУЧЕНИЕ?

1. Нужен функционал качества
2. Нужен алгоритм оптимизации функционала качества
3. Нет данных — нет машинного обучения

ITEM-TO-ITEM РЕКОМЕНДАЦИИ


КАК ПРОДОЛЖИТЬ

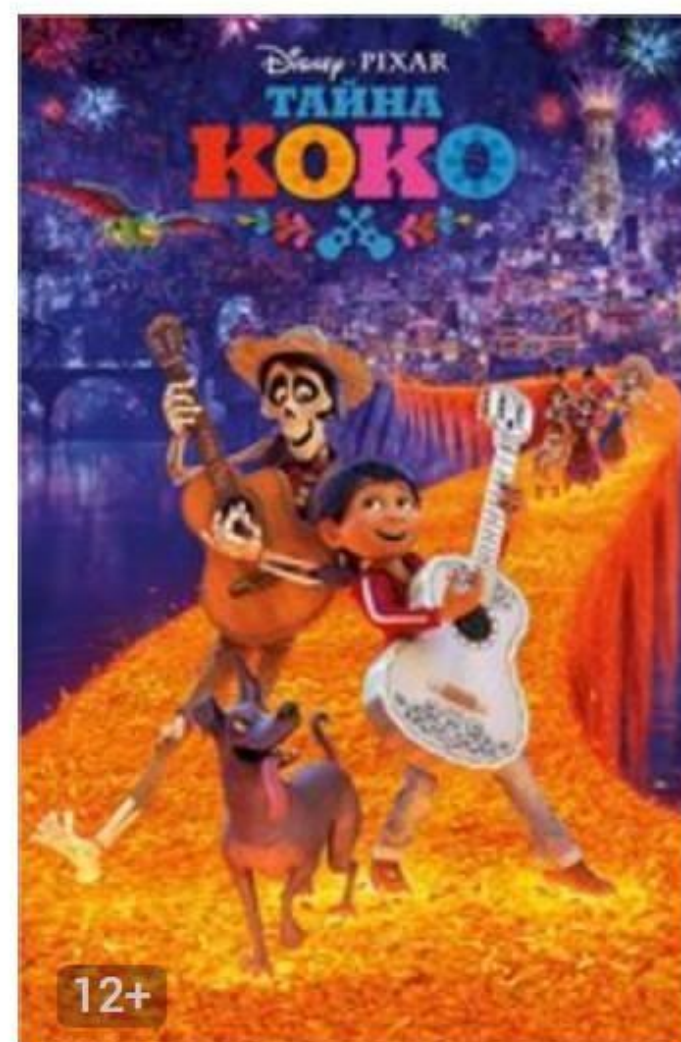
1. Взять фичи объекта, *к которому* рекомендуют
2. Добавить фичи объекта, *который* рекомендуют
3. Целевая переменная — 1/0, было ли целевое действие
4. Построить модель бинарной классификации


ІТЕМ-ТО-ІТЕМ РЕКОМЕНДАЦИИ

С ФИЛЬМОМ “ГАРРИ ПОТТЕР И ФИЛОСОФСКИЙ КАМЕНЬ” ТАКЖЕ СМОТРЯТ




 Приключения
Паддингтона 2




 Тайна Коко



 Фантастические твари
и где они обитают



 Гадкий я 3



 Фердинанд

ITEM-TO-ITEM РЕКОМЕНДАЦИИ

| К чему рекомендовали | Что рекомендовали | Был клик |
|----------------------|-------------------------|----------|
| Гарри Поттер | Приключения Паддингтона | 0 |
| Гарри Поттер | Тайна Коко | 0 |
| Гарри Поттер | Фантастические твари | 1 |
| Гарри Поттер | Гадкий я | 0 |
| Гарри Поттер | Фердинанд | 0 |

ЧТО ЕЩЁ МОЖНО СДЕЛАТЬ

- негативное сэмплирование
- размытие выдачи
- регрессия вместо классификации
- классификация на вероятностях (*logloss*)

КАКИЕ ЕЩЁ МОЖНО БРАТЬ ФИЧИ

- свойства пользователя
- контекст (время, место, устройство...)
- расстояния между тем, к чему рекомендуются и тем, что рекомендуют
- всё, что угодно:)

ПОЧЕМУ ВСЕ ЛЮБЯТ ITEM-TO-ITEM?

- большой простор для экспериментов
- легко внедрять и интегрировать с другими сервисами
- полезно и понятно бизнесу

ДОМАШНЯЯ РАБОТА

КАК ПОЛУЧИТЬ ЗАЧЁТ

1. Использовать dataset [MovieLens](#)
2. Построить рекомендации (*регрессия, предсказываем оценку*)
на фичах:
 - TF-IDF на тегах и жанрах
 - Средние оценки (+ median, variance, etc.) пользователя и фильма
3. Оценить RMSE на тестовой выборке

—

ВОПРОСЫ