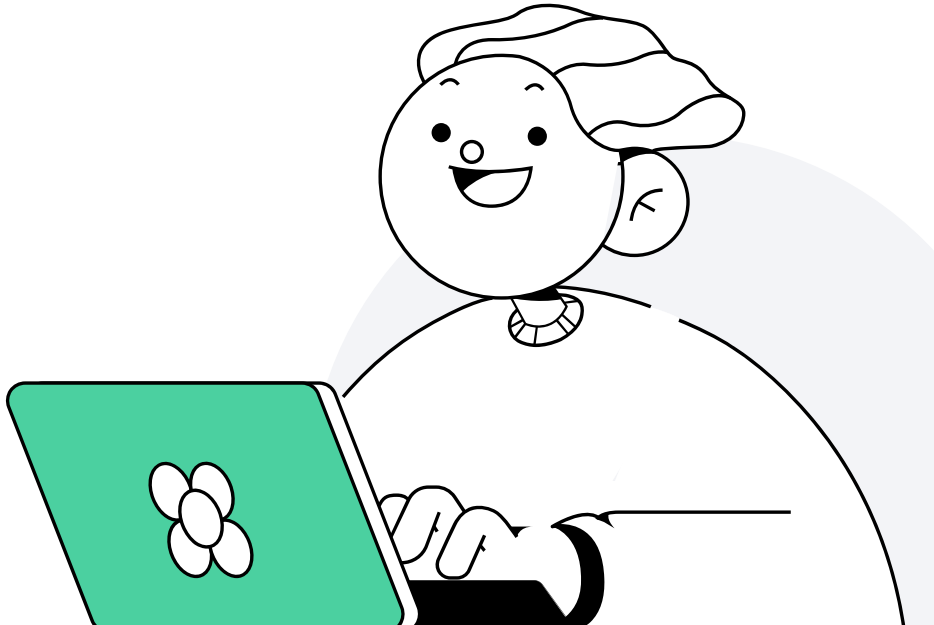
Рекомендации на основе содержания

План занятия

1. Откуда берутся фичи
2. Content-based-модель
3. Рекомендации item-to-item



Извлечение фич

Откуда берутся фичи

1

Ручное
извлечение

2

Парсинг
внешних источников

The Music Genome Project

1. Разработка интернет-радио Pandora

2. Команда экспертов с музыкальным образованием

3. 450 музыкальных фич на каждую звукозапись

Микрожанры онлайн-кинотеатров

1. Команда тегировщиков

2. Десятки страниц правил тегирования

3. Почти 100 тысяч микрожанров:
   - документальные фильмы о чернокожих преступниках
   - страшные фильмы 80-х годов о культах и сектах
   - приключенческие фильмы 30-х годов о шпионах

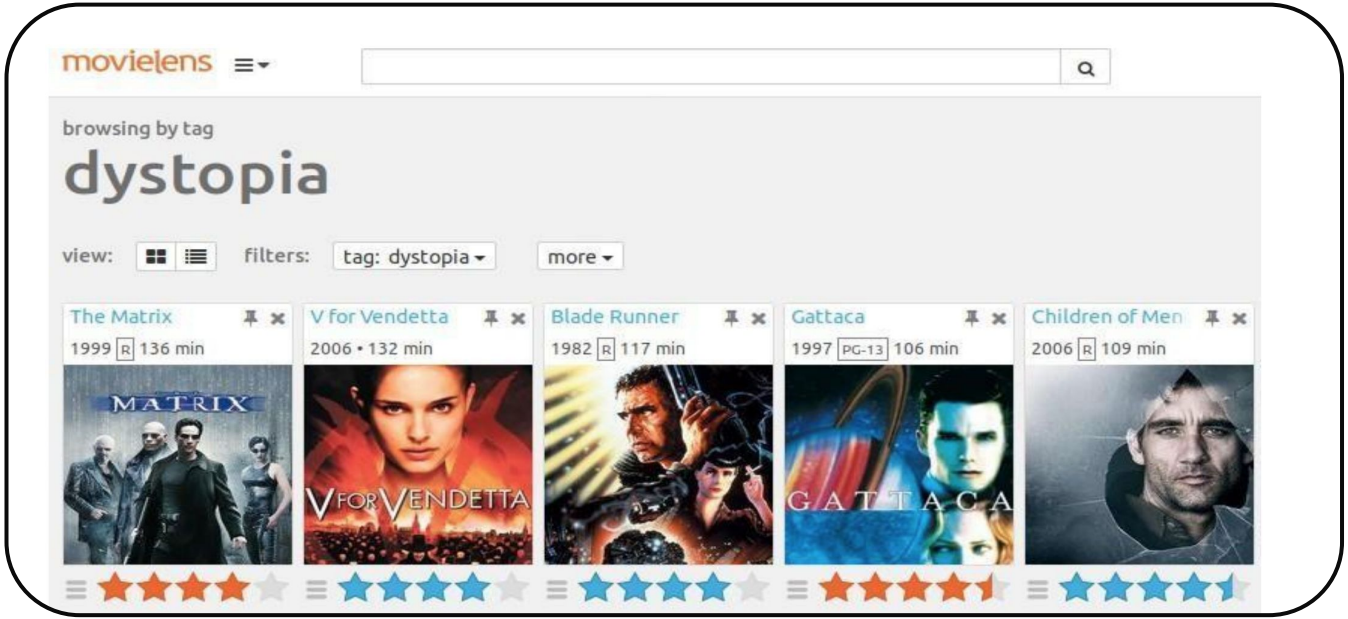
Своя команда

- Разработка правил тегирования
- Найм и обучение экспертов
- Сервисы вроде Толока
- Перекрёстная проверка

MovieLens tags

1. Сотни тысяч пользователей, которые бесплатно размечают данные

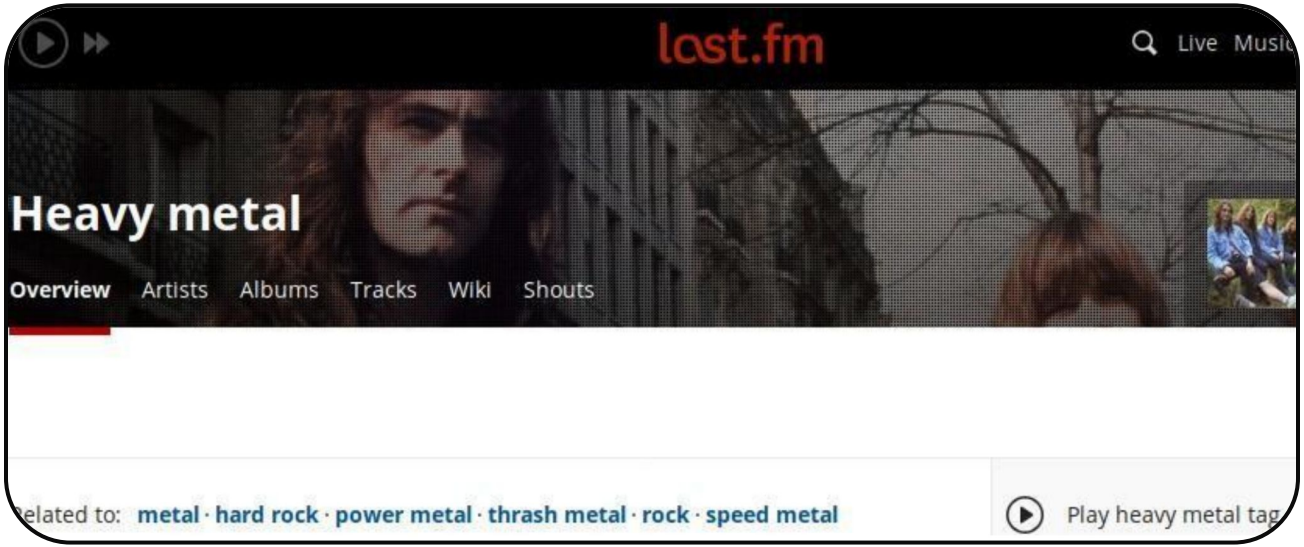
2. Сотни различных тегов



Last.fm tags

1. Десятки миллионов пользователей

2. Сотни тысяч различных тегов



Парсинг внешних данных

- Очень много (сырых) данных
- Бесплатно
- Нужны правила дедубликации и прочее

Практика

Хотим сделать CBRS для фильмов

CBRS (content-based recommender system) - рекомендация на основе содержания.

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

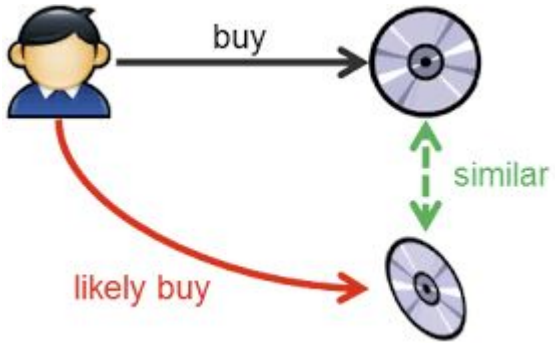
Что делать

- Получите гистограмму количества тегов на фильм и пользователя
- Получите график количества тегов по месяцам
- Получите гистограмму количества жанров на фильм

Content-based-модель

Основные принципы

- У объекта должно быть какое-то признаковое описание (жанры фильмов)



Основные принципы

- Фичи — свойства объекта (пользователя)
- Простой вариант: один пользователь — одна модель
- Иначе: один объект (товар/услуга) - одна модель
- Иначе: одна модель на все, а каждый элемент данных это пара пользователь-объект
- Целевая переменная — релевантность пользователю

Один пользователь - одна модель

1. Должна быть предельно простая модель (линейная)
2. Коэффициенты модели — профиль пользователя
3. Важна L1-регуляризация («совсем не нравится»)
4. Не учитывает общие паттерны поведения пользователей
5. Должны быть богатые по содержанию объекты

genre	**tag**	**year**	...	**buy**
horror	Action| Adventure	2006	...	1
...	...	...	...	...
love	Comedy| Romance	2012	...	0

Один объект - одна модель

1. Когда мало объектов
2. Описание пользователей
3. Для каждого нового продукта новая модель

age	city	cite	...	buy
18	Moscow	2	...	1
...	...	...	...	...
36	Kirov	6	...	0

Одна модель на все

1. Учитывает общие закономерности (больше обобщает)
2. Требует больше параметров/сложность
3. Ограничения в применимости на большом объеме данных

genre	**tag**	**year**	...	**age**	**city**	**cite**	...	**buy**
horror	Action| Adventure	2006	...	18	Moscow	2	...	1
...	...	...	...	...	...	...	...	...
love	Comedy| Romance	2012	...	36	Kirov	6	...	0

Рекомендации item-to-item

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

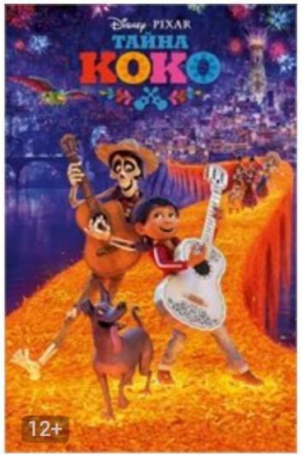
Как продолжить (с МЛ)

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

genre	**tag**	**year**	...	**genre**	**tag**	**year**	...	**target**
horror	Action| Adventure	2006	...	horror	Action	2009	...	1
...	...	...	...	...	...	...	...	...
love	Comedy| Romance	2012	...	horror	Adventure	2009	...	0

С фильмом «Гарри Поттер и философский камень» также смотрят

Приключения Паддингтона 2



Тайна Коко

Фантастические твари и где они обитают

Гадкий я 3

Фердинанд

С фильмом «Гарри Поттер и философский камень» также смотрят

К чему рекомендовали	Что рекомендовали	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
- Регрессия вместо классификации

Какие ещё можно брать фичи

- Свойства пользователя
- Контекст: время, место, устройство и т. д.
- Расстояние между тем, к чему рекомендуются, и тем, что рекомендуют
- Всё что угодно :)

Почему все любят item-to-item

- Обладаете знаниями о своих товарах/услугах
- Большой простор для экспериментов
- Полезно и понятно бизнесу

Практика

Рекомендации к фильму

Гипотеза: рекомендации похожих фильмов увеличат время сессии и конверсию в просмотр.

Что делать

- TF-IDF на тегах, жанрах
- Найдите ближайших соседей любимого фильма
- Проделайте то же для других расстояний

Итоги

- Сделали EDA для датасета MovieLens
- Поняли, как можно сформировать признаки на основе представленных данных
- Построили простую item-to-item рекомендательную систему на основе метода ближайших соседей

Рекомендации на основе содержания