

*Посвящается любви всей моей жизни:
моей жене Саре и моему сыну Питеру,
маленькому супергерою*

Оглавление

Предисловие от издательства	13
Предисловие	14
Благодарности	16
О книге	17
Об авторе	20
Об обложке	21
ЧАСТЬ I. Подготовка к рекомендательным системам	23
Глава 1. Что такое рекомендательная система?	25
1.1. Рекомендации в реальной жизни	25
1.1.1. Рекомендательные системы дома в интернете	27
1.1.2. Длинный хвост	28
1.1.3. Рекомендательная система Netflix	28
1.1.4. Определение рекомендательной системы.....	35
1.2. Таксономия рекомендательных систем	38
1.2.1. Специализация	39
1.2.2. Задача	39
1.2.3. Контекст.....	40
1.2.4. Степень персонализации	40
1.2.5. Чье мнение	42
1.2.6. Конфиденциальность и надежность.....	42
1.2.7. Интерфейс.....	43
1.2.8. Алгоритмы.....	46
1.3. Машинное обучение и Netflix Prize	48
1.4. Интернет-сайт MovieGEEKs.....	49
1.4.1. Оформление и характеристики	51
1.4.2. Архитектура.....	51
1.5. Создание рекомендательной системы	53
Резюме.....	54
Глава 2. Поведение пользователя, и как собирать о нем данные	55
2.1. Как (по моему мнению) Netflix собирает факты, пока вы пользуетесь сервисом	56
2.1.1. Какие факты собирает Netflix.....	58

2.2. Поиск полезных данных о поведении пользователя	60
2.2.1. Как узнать мнение посетителя	61
2.2.2. Что можно узнать по поведению обозревателя в магазине	62
2.2.3. Совершение покупки	67
2.2.4. Пользование товаром	68
2.2.5. Оценки посетителей	69
2.2.6. Знакомство с клиентами по (старому) методу Netflix.....	73
2.3. Идентификация пользователей.....	73
2.4. Получение данных о посетителях из других источников	74
2.5. Сборщик данных.....	74
2.5.1. Создание файлов проекта	76
2.5.2. Модель данных.....	76
2.5.3. Сборщик данных на стороне клиента	77
2.5.4. Интеграция сборщика в MovieGEEKs	78
Регистрация наведения курсора.....	80
Регистрация просмотра подробностей.....	80
Регистрация «сохранения на потом»	80
2.6. Какие пользователи есть в системе, и как их моделировать	81
Резюме.....	84
Глава 3. Мониторинг состояния системы.....	85
3.1. Почему панель аналитики – это круто	86
3.1.1. Ответ на вопрос «Как там дела у сайта?»	86
3.2. Реализация аналитики	88
3.2.1. Веб-аналитика.....	88
3.2.2. Базовые статистические данные	88
3.2.3. Конверсии.....	89
3.2.4. О пути к конверсиям.....	92
3.2.5. Путь конверсии	94
3.3. Архетипы	97
3.4. Панель сайта MovieGEEKs	100
3.4.1. Автоматическая генерация данных в журнале.....	100
3.4.2. Характеристики и дизайн панели аналитики.....	101
3.4.3. Основа панели аналитики.....	101
3.4.4. Архитектура.....	102
Резюме.....	104
Глава 4. Оценки, и как их рассчитывать	105
4.1. Предпочтения элементов пользователями.....	106
4.1.1. Определение оценок.....	106
4.1.2. Матрица пользователь–элемент.....	107
4.2. Явные или неявные оценки	109
4.2.1. Как мы используем доверенные источники для составления рекомендаций	110

4.3. Переоценка.....	111
4.4. Что такое неявные оценки?.....	111
4.4.1. Предложения людей	113
4.4.2. Что учитывать при расчете оценок	113
4.5. Расчет неявных оценок.....	116
4.5.1. Просмотр поведенческих данных.....	117
4.6. Как реализовать неявные оценки.....	122
4.6.1. Добавление учета времени	126
4.7. Более редкие элементы имеют большую ценность.....	128
Резюме.....	131
Глава 5. Неперсонализированные рекомендации	132
5.1. Что такое неперсонализированные рекомендации?	133
5.1.1. Что такое реклама?	133
5.1.2. Что делает рекомендация?	134
5.2. Как сделать рекомендации, когда у вас нет данных.....	135
5.2.1. Топ-10: диаграмма элементов	136
5.3. Реализация диаграмм и основы для рекомендатора	138
5.3.1. Компонент рекомендательной системы	138
5.3.2. Код MovieGEEKs на сайте GitHub	139
5.3.3. Рекомендательная система	140
5.3.4. Добавление диаграмм на MovieGEEKs	140
5.3.5. Заставим контент выглядеть более привлекательно	142
5.4. Выборочные рекомендации.....	144
5.4.1. Часто покупаемые элементы, похожие на тот, который вы просматриваете	144
5.4.2. Ассоциативные правила	145
5.4.3. Реализация ассоциативных правил.....	150
5.4.4. Сохранение ассоциативных правил в базе данных.....	154
5.4.5. Запуск калькулятора ассоциаций	155
5.4.6. Использование различных событий для создания ассоциативных правил	157
Резюме.....	158
Глава 6. «Холодные» пользователи и контент.....	159
6.1. Что такое холодный старт?.....	159
6.1.1. Холодные товары	161
6.1.2. Холодный посетитель	161
6.1.3. Серые овцы.....	163
6.1.4. Посмотрим на примеры из реальной жизни	163
6.1.5. Что вы можете сделать с холодным стартом?.....	164
6.2. Отслеживание посетителей.....	165
6.2.1. Анонимные пользователи	165

6.3. Решение проблемы холодного старта алгоритмами	165
6.3.1. Использование ассоциативных правил для создания рекомендаций для холодных пользователей	166
6.3.2. Использование знаний предметной области и бизнес-правил	168
6.3.3. Использование сегментов	168
6.3.4. Использование категорий с целью обойти проблему серых овец и холодных продуктов	170
6.4. Кто не спрашивает, тот не будет знать	172
6.4.1. Когда посетитель уже не новый	173
6.5. Использование ассоциативных правил с целью ускорить показ рекомендаций.....	173
6.5.1. Поиск собранных элементов	174
6.5.2. Получение ассоциативных правил и сортировка в соответствии со значениями уверенности	174
6.5.3. Отображение рекомендаций.....	176
6.5.4. Оценка реализации.....	179
Резюме.....	179
Часть II. Рекомендательные алгоритмы	181
Глава 7. Выявление общих черт у пользователей и контента	183
7.1. Что за сходство?.....	184
7.1.1. Что такое функция подобия?.....	185
7.2. Основные функции подобия	185
7.2.1. Расстояние Жаккара.....	187
7.2.2. Измерение расстояния с помощью L _p -норм.....	189
7.2.3. Коэффициент Отиаи	192
7.2.4. Вычисление сходства с помощью коэффициента корреляции Пирсона	194
7.2.5. Испытание сходства коэффициентом Пирсона	195
7.2.6. Коэффициент корреляции Пирсона на коэффициент Отиаи	198
7.3. Кластеризация k-средних	198
7.3.1. Алгоритм кластеризации k-средних.....	199
7.3.2. Реализация кластеризации k-средних на Python	201
7.4. Реализация вычисления сходства	206
7.4.1. Реализация вычисления сходства на сайте MovieGEEKs	208
7.4.2. Реализация кластеризации на сайте MovieGEEKs	210
Резюме.....	214
Глава 8. Совместная фильтрация в окрестностях	215
8.1. Совместная фильтрация: историческая справка.....	217
8.1.1. Когда начали использовать совместную фильтрацию	217
8.1.2. Взаимопомощь.....	217
8.1.3. Матрица оценок	219

8.1.4. Процедура совместной фильтрации.....	220
8.1.5. Нужно использовать совместную фильтрацию пользователь– пользователь или элемент–элемент?	221
8.1.6. Требования к данным	222
8.2. Расчет рекомендации	222
8.3. Расчет сходства	223
8.4. Алгоритм вычисления сходства элементов с Amazon.....	223
Если проблема повторяется – берегись!.....	227
8.5. Способы выбора окрестности	228
8.6. Поиск правильной окрестности	230
8.7. Методы прогнозирования оценок	230
8.8. Прогнозирование с фильтрацией по элементам.....	232
8.8.1. Вычисление прогнозов.....	233
8.9. Проблема холодного старта	233
8.10. Пара слов о терминах машинного обучения	234
8.11. Совместная фильтрация на сайте MovieGEEKs.....	235
8.11.1. Фильтрация элементов	236
8.12. В чем разница между правилами ассоциации и совместной фильтрацией?.....	242
8.13. Эксперименты с совместной фильтрацией	242
8.14. Преимущества и недостатки совместной фильтрации	244
Резюме	245
Глава 9. Оценка и тестирование рекомендательной системы.....	246
9.1. Бизнесу нужен подъем, перекрестные продажи, рост продаж и конверсии.....	247
9.2. Зачем оценивать?.....	248
Гипотеза.....	249
9.3. Как интерпретировать поведение пользователей	249
9.4. Что измерять.....	249
9.4.1. Понимание вкусов пользователя: сведение к минимуму ошибки предсказания	250
9.4.2. Разнообразие	251
9.4.3. Охват	252
9.4.4. Приятные неожиданности	254
9.5. Перед реализацией рекомендатора.....	255
9.5.2. Регрессионное тестирование	256
9.6. Виды оценки.....	257
9.7. Офлайн-оценка	257
9.7.1. Что делать, если алгоритм не дает рекомендаций	258
9.8. Офлайн-эксперименты	259

9.8.1. Подготовка данных для эксперимента	264
9.9. Реализация эксперимента в MovieGEEKs.....	270
9.9.1. Список дел.....	271
9.10. Оценка тестового набора.....	274
9.10.1. Начнем с базовых прогнозов	275
9.10.2. Поиск правильных параметров	277
9.11. Онлайн-оценка.....	278
9.11.1. Контролируемые эксперименты.....	279
9.11.2. А/В-тестирование.....	279
9.12. Непрерывное тестирование с использованием/исследованием	280
9.12.1. Петли обратной связи.....	281
Глава 10. Фильтрация по контенту	283
10.1. Описательный пример	283
10.2. Фильтрация на основе контента	286
10.3. Анализатор контента	288
10.3.1. Выделение признаков для профиля элемента.....	288
10.3.2. Редко встречающиеся данные	290
10.3.3. Преобразование года в сопоставимую функцию.....	290
10.4. Извлечение метаданных из описаний	291
10.4.1. Составление описаний	291
10.5. Поиск важных слов методом TF-IDF.....	295
10.6. Моделирование темы с использованием LDA.....	297
10.6.1. Какими крутилками настраивать LDA?	303
10.7. Поиск подобного контента	306
10.8. Создание профиля пользователя	307
10.8.1. Создание профиля пользователя с помощью модели LDA	307
10.8.2. Создание профиля пользователя с помощью модели TF-IDF	308
10.9. Рекомендации на основе контента на сайте MovieGEEKs	310
10.9.1. Загрузка данных.....	310
10.9.2. Обучение модели	313
10.9.3. Создание профилей элементов.....	314
10.9.4. Создание пользовательских профилей	314
10.9.5. Отображение рекомендаций.....	316
10.10. Оценка рекомендатора на основе контента	317
10.11. Плюсы и минусы фильтрации на основе контента	319
Резюме.....	320
Глава 11. Определение скрытых жанров с помощью матричной факторизации	321
11.1. Иногда чем меньше данных, тем лучше.....	322
11.2. Пример задачи	324

11.3. Немножко линейной алгебры	327
11.3.1. Матрица	327
11.3.2. Что за факторизация?	329
11.4. Выполнение факторизации с использованием SVD.....	331
11.4.1. Добавление новых пользователей путем складывания	336
11.4.2. Как формировать рекомендации с помощью SVD	338
11.4.3. Базисные предикторы	339
11.4.4. Временная динамика	342
11.5. Построение факторизации с помощью Funk SVD	342
11.5.1. Корень средней квадратичной ошибки	343
11.5.2. Градиентный спуск	344
11.5.3. Стохастический градиентный спуск.....	347
11.5.4. Перейдем, наконец, к факторизации	347
11.5.5. Добавление отклонений	349
11.5.6. Как начать и когда остановиться	350
11.6. Генерация рекомендаций с помощью Funk SVD	354
11.7. Реализация Funk SVD на MovieGEEKs	356
11.7.1. Что делать с неподходящими рекомендациями	361
11.7.2. Поддержание актуальности модели	362
11.7.3. Более быстрая реализация.....	363
11.8. Явные данные против неявных данных.....	363
11.9. Оценка	363
11.10. Эксперименты с моделью Funk SVD	365
Резюме.....	367

Глава 12. С каждого по способностям – реализуем гибридный алгоритм рекомендательной системы.....	368
12.1. Сложности мира гибридов	369
12.2. Монолитные рекомендаторы.....	370
12.2.1. Смешивание функций контента с поведенческими данными для улучшения алгоритмов на основе совместной фильтрации	371
12.3. Смешанный гибридный рекомендатор	372
12.4. Ансамбль.....	372
12.4.1. Переключаемый ансамбль рекомендаторов.....	374
12.4.2. Взвешенная ансамбль рекомендаторов	375
12.4.3. Линейная регрессия	376
12.5. Признако-взвешенное линейное сочетание (FWLS)	377
12.5.1. Представляем веса в виде функций.....	378
12.5.2. Алгоритм.....	380
12.6. Реализация	387
Резюме.....	396

Глава 13. Ранжирование и обучение ранжированию	397
13.1. Обучение ранжированию на примере Foursquare	398
13.2. Переранжирование	402
13.3. Еще раз – что такое обучение ранжированию?	403
13.3.1. Три типа алгоритмов LTR	403
13.4. Байесовское персонализированное ранжирование	405
13.4.1. Ранжирование с BPR	407
13.4.2. Магия математики (продвинутое колдовство)	409
13.4.3. Алгоритм BPR	412
13.4.4. BPR с матричной факторизацией	413
13.5. Реализация BPR	413
13.5.1. Генерация рекомендаций	419
13.6. Оценка	421
13.7. Эксперименты с BPR	423
Резюме	424
Глава 14. Будущее рекомендательных систем	425
14.1. Вся книга в паре предложений	426
14.2. Темы для дальнейшего изучения	429
14.2.1. Дальнейшее чтение	429
14.2.2. Алгоритмы	430
14.2.3. Контекст	430
14.2.4. Взаимодействие «Человек–Машина»	431
14.2.5. Выбор подходящей архитектуры	431
14.3. Что ждет рекомендательные системы в будущем?	432
14.4. Послесловие	436
Предметный указатель	438

Предисловие от издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв прямо на нашем сайте www.dmkpress.com, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в тексте или в коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу dmkpress@gmail.com, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или веб-сайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты dmkpress@gmail.com со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Предисловие

Когда в 2003 году я окончил университет, были опасения, что в Европе программисты будут не востребованы, поскольку все разработки будут проводиться в странах с гораздо более низким уровнем зарплат. К счастью, по множеству причин эти опасения не сбылись. Как я предполагаю, одной из главных проблем было то, что компании недооценивали следующую проблему: разработчики не понимали культурных реалий тех мест, где будут применяться их разработки. Разрабатывалось необходимое программное обеспечение, но функциональные возможности не соответствовали ожиданиям.

В настоящее время аналогичная проблема возникла в сфере машинного обучения и анализа данных. Разница лишь в том, что у ее истоков не низкие зарплаты, а программное обеспечение как услуга (SaaS), система, в которую вы загружаете данные и которая делает за вас всю остальную работу.

Меня, как и всех остальных, беспокоит тот факт, что машины не понимают домены и людей. Машины еще не настолько умны, чтобы можно было исключить человека из задачи. Развитие идет быстрыми темпами, но я рискну предположить, что любой, кто читает эту книгу, сможет работать с рекомендательными системами вплоть до завершения своей трудовой деятельности.

Как я попал в эту сферу? Я работал разработчиком ПО в Италии и собирался переезжать в Англию, где мне хотелось получить работу поинтереснее простого управления данными в базе. К счастью, на меня вышел замечательный сотрудник кадрового агентства RedRock Consulting Ltd. Агентство связало меня с компанией, занимающейся разработкой рекомендательных систем, куда меня взяли для работы над базовым программным обеспечением. И все – я с головой ушел в машинное обучение (мне было невероятно интересно). Помимо работы над рекомендательными системами, я начал изучать информацию об интернете и читал множество книг по этой и схожим тематикам.

Сейчас и чихнуть нельзя без того, чтобы как минимум десять человек не попытались поведать вам какие-то знания из области машинного обучения. Мне очень удивительно видеть одностраничные или одночасовые учебные курсы, авторы которых заявляют, что научат вас всему, что нужно знать о машинном обучении. С таким же успехом я могу написать руководство, как стать летчиком-истребителем:

«Взлетайте и управляйте полетом с помощью штурвала. Если вам нужно стрелять, нажмите кнопку. Прежде чем у вас закончится топливо, необходимо совершить посадку...»

Подобное руководство, возможно, станет для вас неплохой отправной точкой – я начинал точно так же. Но не пытайтесь себя обмануть: для понимания машинного обучения необходимы комплексные знания. Добавьте сюда человеческий фактор, который всегда еще больше все усложняет.

Возвращаясь к моей истории: я работал над рекомендательными системами и получал от этого удовольствие, а затем сменил работу. На новой должно-

сти я должен был продолжить работу над рекомендательными системами, но этот проект был отложен. Тогда я стал переживать, что больше мне не представится возможность работать с рекомендательными системами, и именно в этот момент издательство Manning предложило мне написать о них книгу. Конечно же, я с радостью взялся за это задание. Как только контракт был подписан, проект по рекомендательным системам все-таки начался. Работая над этой книгой, я существенно обогатил свои знания, и надеюсь, что и вам она принесет пользу.

Задача этой книги – познакомить вас с рекомендательными системами: не только с алгоритмами, но и со всей экосистемой. Алгоритмы не отличаются особой сложностью, но, чтобы их понимать и применять, необходимо понимать пользователей, для которых система будет предлагать рекомендации. Содержание книги менялось в процессе ее написания, поскольку я пытался включить в нее как можно больше информации. Надеюсь, что, прочитав эту книгу, вы получите все необходимые знания, чтобы начать работать с рекомендательными системами, и у вас появится база для дальнейшего обучения.

Благодарности

Я хочу отметить и поблагодарить две группы людей: тех, кто активно работал над книгой, и тех, кому приходилось терпеть и поддерживать меня в течение последних трех лет, пока я работал, не обращая внимания ни на что вокруг.

Хотя на обложке «Рекомендательных систем на практике» стоит мое имя, эта книга не появилась бы на свет без активного участия замечательных специалистов из издательства Manning. Я хочу отдельно поблагодарить редактора-консультанта Хелен Стергиус (Helen Stergius) за постоянную помощь и содействие. Она и другие сотрудники превратили мой несколько несвязный текст в руководство, обучающее людей создавать рекомендательные системы.

Также я хочу поблагодарить технических редакторов Фуркана Камаси (Furkan Kamasi) и Валентина Креттаза (Valentin Crettaz), а также всех рецензентов, которые нашли время прочитать первые варианты книги и помогли мне отточить текст, в их числе: Адхир Рамджиаван (Adhir Ramjiavan), Александр Мильцев (Alexander Myltsev), Элвин Радж (Alvin Raj), Амит Ламба (Amit Lamba), Эндрю Колльер (Andrew Collier), Фазел Кешткар (Fazel Keshtkar), Джаред Дункан (Jared Duncan), Яромир Немец (Jaromir Nemes), Мартин Бир (Martin Beer), Маюр Патил (Mayur Patil), Майк Далримпл (Mike Dalrymple), Норин Дертинджер (Noreen Dertinger), Оливье Дукаттеу (Olivier Ducatteeuw), Питер Хэмптон (Peter Hampton), Симеон Лейзерзон (Simeon Leyzerzon), Серен Линд Кристиансен (Søren Lind Kristiansen), Стивен Парр (Steven Parr), Тобиас Бергер (Tobias Bürger), Тобиас Гетрост (Tobias Getrost) и Випул Гупта (Vipul Gupta).

Работая над книгой, я обращался ко множеству различных библиотек, систем и баз данных, и очень благодарен всем сообществам, которые мне помогли. Также я благодарен сообществам разработчиков ПО с открытым исходным кодом, которые создали ряд инструментов, избавив нас от необходимости делать все с нуля.

И в первую очередь я благодарю свою жену, сына и тещу, остальных родственников, а также близких друзей за поддержку, любовь и, самое главное, терпение. Им было нелегко жить с членом семьи, который все время норовит ускользнуть и засесть за написание книги, пока вся семья переезжает в новый дом и лицезреет, как наши дома в Италии разрушаются до основания в результате землетрясений. Не говоря уже о том, что при этом новоиспеченный писатель взялся одновременно за две новые работы. Спасибо вам, и я обещаю, что хотя бы пару лет не буду начинать новые проекты. Люблю вас всех!

О книге

Испытываете ли вы зависть, когда видите рекомендуемые товары Amazon или когда Netflix попадает в точку с рекомендациями подписчикам? Тогда у вас появился шанс пополнить свой арсенал подобными умениями. Читая эту книгу, вы получите представление о том, что такое рекомендательные системы и каково их практическое применение. Чтобы рекомендательная система работала, необходимо, чтобы одновременно выполнялось несколько процессов. Нужно понимать, как собирать данные о пользователях и как их интерпретировать, а также нужно владеть различными алгоритмами рекомендательных систем, чтобы для каждой ситуации можно было выбрать наиболее подходящий из них. И самое главное, нужно уметь чувствовать, хорошо ли функционирует рекомендательная система. Все это и многое другое вы найдете в данной книге.

Кто должен прочитать эту книгу

Книга «Рекомендательные системы на практике» в первую очередь предназначена для разработчиков, которые хотели бы создать рекомендательную систему. В книге собраны практические советы, и материал изложен простым, доступным языком. Есть и вычисления, и статистика, но всегда обязательно присутствуют цифры и программный код. Новоиспеченные специалисты по обработке и анализу данных также почерпнут много полезного из данной книги: получают базовые представления об алгоритмах рекомендательных систем и инфраструктуре, необходимой для их запуска и функционирования. Менеджерам эта книга будет интересна в качестве общего пособия по рекомендательным системам: что это такое и как это можно применить на практике.

Чтобы получить от книги максимум пользы, нужно разбираться в языках программирования, таких как Python и Java, понимать SQL-запросы и обладать базовыми знаниями высшей математики и статистики. Цифры и листинги кодов, которые приводятся для наглядности, дают лишь самое общее представление о теме.

Структура книги

Книга делится на две части: первая посвящена инфраструктуре рекомендательных систем а вторая – алгоритмам.

Из части I вы узнаете, как после добавления рекомендательной системы в приложение получать данные и применять их:

- глава 1 представляет собой общий обзор рекомендательных систем и их ключевых элементов. Она знакомит читателя с рекомендательными системами в целом и базовыми принципами их работы;

- глава 2 говорит о том, как научиться понимать пользователей и их поведение, а также перечисляет способы сбора данных о пользователях;
- глава 3 посвящена веб-аналитике и рассказывает, как можно создать сводную информационную панель для отслеживания данных о своих рекомендательных системах;
- глава 4 говорит о том, как на основе данных о поведении пользователей составлять рейтинги;
- глава 5 рассматривает общие рекомендации (без учета индивидуальных предпочтений пользователя);
- глава 6 описывает проблемы, связанные с новыми пользователями и товарами, и предлагает простые решения.

Часть II рассказывает об алгоритмах рекомендательных систем, а также о том, как на основе собранных системой данных выстроить рекомендации для пользователя:

- глава 7 говорит о формулах для определения степени сходства между различными пользователями или контентом, например фильмами;
- глава 8 рассказывает о составлении персональных рекомендаций путем совместной фильтрации;
- глава 9 приводит методы оценки рекомендательной системы вне интернета и описывает способы составления рекомендаций онлайн;
- глава 10 знакомит вас с фильтрацией по контенту, которая позволяет найти сходные черты в контенте с помощью различных типов алгоритмов, таких как латентное размещение Дирихле и TF-IDF;
- глава 11 возвращается к обсуждению совместной фильтрации, о которой говорилось в главе 8, но сосредотачивается на методах снижения размерности;
- глава 12 рассказывает, как можно совместить различные типы рекомендательных систем;
- глава 13 говорит об алгоритмах ранжирования и способах оценивания рекомендаций;
- глава 14 подводит общий итог, дает задел на будущее, рассказывает, какие вопросы необходимо изучить и какие книги нужно еще прочитать, а также рассуждает об алгоритмах и контексте.

Книга построена таким образом, что лучше ее читать от начала до конца, поскольку во многих местах есть отсылки к предыдущим главам, но вполне допустимо читать и отдельные главы.

Загрузки

Программный код для запуска демонстрационного сайта под названием MovieGEEKs можно загрузить с сайта издательства по ссылке www.manning.com/books/practical-recommendersystems, а также с сервиса Github.com по ссылке [mng.bz/04K5](https://github.com/mng/bz/04K5). Сайт создан с помощью фреймворка Django. У нас будет два набора данных: один генерируется автоматически, а второй нужно загрузить из базы MovieTweetings. Все инструкции по установке можно найти на сайте GitHub.

Формат кода

В книге содержится много примеров программного кода – как в виде пронумерованных листингов, так и в составе обычного текста. И в том, и в другом случае исходный код выделен шрифтом определенной ширины, чтобы его было легко отличить от обычного текста. Иногда код также выделен жирным шрифтом, чтобы акцентировать внимание читателя на измененных фрагментах, например при добавлении новой функции в строку кода, приведенную ранее в той же главе.

Во многих случаях мы внесли коррективы в первоначальный формат исходного кода, например добавили переносы строк и изменили отступы, чтобы максимально продуктивно использовать пространство книжной страницы. В редких случаях даже этого было недостаточно, и тогда мы вынуждены были применять символы, указывающие на продолжение строки (⇒). Кроме того, из листингов часто приходилось убирать комментарии к исходному коду (когда код описывается в тексте). Многие листинги сопровождаются аннотациями к коду, в которых отражаются важные моменты.

Форум для читателей книги

При покупке «Рекомендательных систем на практике» читатель получает доступ к закрытому интернет-форуму издательства Manning Publications, где можно оставить комментарии о книге, задать технические вопросы и получить помощь от автора и других пользователей. Для входа на форум перейдите по ссылке forums.manning.com/forums/practical-recommendersystems. Подробнее познакомиться с форумами издательства Manning и правилами поведения на них можно на странице forums.manning.com/forums/about.

Издательство Manning стремится предоставить читателям площадку, где читатели могут общаться друг с другом или с автором книги. Издательство не берет на себя ответственности за обязательное привлечение на форум автора, чье участие в обсуждениях на форуме исключительно добровольное (и неоплачиваемое). Советуем задавать автору интересные вопросы, чтобы вовлечь его в общение! Форум и архивы старых обсуждений будут доступны на сайте издательства, пока книгу не снимут с печати.

Об авторе



Ким Фальк – специалист по обработке и анализу данных, обладающий большим опытом создания приложений, управляемых данными. Рекомендательные системы и машинное обучение в целом вызывают его живейший интерес. Он разрабатывал рекомендательные системы, предлагающие конечным пользователям фильмы и рекламу, и даже помогал юристам находить материалы по судебным прецедентам. Он занимается большими данными и машинным обучением с 2010 года. Ким часто говорит и пишет о рекомендатель-

ных системах. Его веб-страница: **kimfalk.org**.

Когда Ким не занят тем, что обучает машины следить за людьми, он отличный муж, отец и владелец курцхаара.

Об обложке

Иллюстрация на обложке «Рекомендательных систем на практике» – это работа под названием «*Amazone d’Afrique*», т. е. «Африканская амазонка». Этот рисунок Жака Грассе де Сен-Совера (1757–1810) является частью серии изображений национальных костюмов различных стран под названием «*Costumes de Différents*», опубликованной во Франции в 1797 году. Каждая работа тщательно прорисована и раскрашена вручную.

Богатое разнообразие рисунков в серии Жака Грассе де Сен-Совера напоминает нам о том, насколько сильно различалась культура городов и регионов мира всего 200 лет назад. Люди, жившие изолированно друг от друга, разговаривали на разных диалектах и языках. На городских улицах или в сельской местности по одежде человека можно было легко определить, чем он занимается и где живет. С тех пор наша манера одеваться сильно изменилась, а региональное разнообразие, столь богатое в те времена, стерлось. Теперь трудно различить жителей разных континентов, а тем более разных стран, регионов и городов. Похоже, что мы пожертвовали культурным многообразием ради того, чтобы достичь большего разнообразия в личной жизни каждого отдельного человека, которая, безусловно, теперь более многогранна, динамична и технологична.

Во времена, когда трудно отличить одну компьютерную книгу от другой, издательство Manning привносит в компьютерную индустрию элемент новизны и свежести, создавая обложки, воспроизводящие богатое региональное разнообразие, которое было присуще миру два века назад и запечатлено на рисунках Жака Грассе де Сен-Совера.

ЧАСТЬ I

Подготовка к рекомендательным системам

Окружающая среда – это все, что не является мной.
Альберт Эйнштейн

Применение рекомендательных систем и, по сути, большинства методов машинного обучения в условиях промышленной эксплуатации подразумевает не только применения наиболее подходящего алгоритма, но и требует понимания пользователей и сферы деятельности.

Главы 1–6, т. е. часть I «Рекомендательных систем на практике», познакомят вас с экосистемой и инфраструктурой рекомендательных систем. Вы научитесь собирать данные и применять их, добавив рекомендательную систему в приложение. Вы узнаете, чем отличается рекомендация от рекламы и персональная рекомендация от неперсональной. Вы также узнаете, как собирать данные для создания собственной рекомендательной системы.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru