

Оглавление

Предисловие	11
Благодарности	12
О книге	14
Структура книги	14
Для кого написана эта книга	15
Условные обозначения и загружаемые файлы	15
Об авторах	16
От издательства	17
Глава 1. Data science в мире больших данных	18
1.1. Область применения data science и больших данных и их преимущества . . .	19
1.2. Грани данных	21
1.2.1. Структурированные данные	21
1.2.2. Неструктурированные данные	22
1.2.3. Данные на естественном языке.	22
1.2.4. Машинные данные	23
1.2.5. Графовые, или сетевые, данные	24
1.2.6. Аудио, видео и графика	25
1.2.7. Поточковые данные	26
1.3. Процесс data science	26
1.3.1. Назначение цели исследования	27
1.3.2. Сбор данных.	27
1.3.3. Подготовка данных.	27
1.3.4. Исследование данных.	27

1.3.5. Моделирование данных или построение модели	27
1.3.6. Отображение и автоматизация	28
1.4. Экосистема больших данных и data science	28
1.4.1. Распределенные файловые системы	30
1.4.2. Инфраструктура распределенного программирования.	30
1.4.3. Инфраструктура интеграции данных.	31
1.4.4. Инфраструктуры машинного обучения	31
1.4.5. Базы данных NoSQL	32
1.4.6. Инструменты планирования	33
1.4.7. Инструменты сравнительного анализа.	33
1.4.8. Развертывание системы	33
1.4.9. Программирование служб.	34
1.4.10. Безопасность	34
1.5. Вводный пример использования Hadoop	34
1.6. Итоги	40

Глава 2. Процесс data science 42

2.1. Обзор процесса data science	42
2.1.1. Не будьте рабом процесса	45
2.2. Этап 1: Определение целей исследования и создание проектного задания	46
2.2.1. Выделите время на то, чтобы разобраться в целях и контексте исследования	46
2.2.2. Создайте проектное задание.	47
2.3. Этап 2: Сбор данных	47
2.3.1. Начните с данных, хранимых в компании	48
2.3.2. Не бойтесь покупок во внешних источниках	49
2.3.3. Проверьте качество данных сейчас, чтобы предотвратить проблемы в будущем	50
2.4. Этап 3: Очистка, интеграция и преобразование данных	50
2.4.1. Очистка данных	51
2.4.2. Исправляйте ошибки как можно раньше	58
2.4.3. Комбинирование данных из разных источников	59
2.4.4. Преобразование данных	62
2.5. Этап 4: Исследовательский анализ данных	66
2.6. Этап 5: Построение моделей	70
2.6.1. Выбор модели и переменных	71
2.6.2. Выполнение модели	72

2.6.3. Диагностика и сравнение моделей	77
2.7. Этап 6: Представление результатов и построение приложений на их основе	78
Итоги	79

Глава 3. Машинное обучение 81

3.1. Что такое машинное обучение, и почему оно важно для вас?	82
3.1.1. Применение машинного обучения в data science	83
3.1.2. Применение машинного обучения в процессе data science.	84
3.1.3. Инструменты Python, используемые в машинном обучении	85
3.2. Процесс моделирования	87
3.2.1. Создание новых показателей и выбор модели	88
3.2.2. Тренировка модели	89
3.2.3. Проверка адекватности модели	90
3.2.4. Прогнозирование новых наблюдений	91
3.3. Типы машинного обучения	92
3.3.1. Контролируемое обучение	92
3.3.2. Неконтролируемое обучение	100
3.4. Частично контролируемое обучение	111
3.5. Итоги	112

Глава 4. Работа с большими данными на одном компьютере 114

4.1. Проблемы при работе с большими объемами данных	115
4.2. Общие методы обработки больших объемов данных	116
4.2.1. Правильный выбор алгоритма.	117
4.2.2. Правильный выбор структуры данных.	126
4.2.3. Правильный выбор инструментов	128
4.3. Общие рекомендации для программистов при работе с большими наборами данных	131
4.3.1. Не повторяйте уже выполненную работу	131
4.3.2. Используйте все возможности оборудования	132
4.3.3. Экономьте вычислительные ресурсы.	133
4.4. Пример 1: Прогнозирование вредоносных URL-адресов	134
4.4.1. Этап 1: Определение цели исследования	134
4.4.2. Этап 2: Сбор данных URL	135
4.4.3. Этап 4: Исследование данных.	136
4.4.4. Этап 5: Построение модели	137

4.5. Пример 2: Построение рекомендательной системы внутри базы данных	139
4.5.1. Необходимые инструменты и методы	139
4.5.2. Этап 1: Вопрос исследования	142
4.5.3. Этап 3: Подготовка данных	142
4.5.4. Этап 5: Построение модели	147
4.5.5. Этап 6: Отображение и автоматизация	148
4.6. Итоги	150

Глава 5. Первые шаги в области больших данных 151

5.1. Распределение хранения и обработки данных в инфраструктурах	152
5.1.1. Hadoop: инфраструктура для хранения и обработки больших объемов данных	152
5.1.2. Spark: замена MapReduce с повышенной производительностью	156
5.2. Учебный пример: Оценка риска при кредитовании	157
5.2.1. Этап 1: Цель исследования	159
5.2.2. Этап 2: Сбор данных	160
5.2.3. Этап 3: Подготовка данных	164
5.2.4. Этап 4: Исследование данных и Этап 6: построение отчета	169
5.3. Итоги	182

Глава 6. Присоединяйтесь к движению NoSQL 183

6.1. Введение в NoSQL	186
6.1.1. ACID: базовые принципы реляционных баз данных	186
6.1.2. Теорема CAP: проблема баз данных, распределенных по многим узлам	187
6.1.3. Принципы BASE баз данных NoSQL	190
6.1.4. Типы баз данных NoSQL	192
6.2. Учебный пример: Диагностика болезней	199
6.2.1. Этап 1: Назначение цели исследования	201
6.2.2. Этапы 2 и 3: Сбор и подготовка данных	202
6.2.3. Этап 4: Исследование данных	211
6.2.4. Этап 3 (снова): Подготовка данных для профилирования болезни	220
6.2.5. Этап 4 (повторно): Исследование данных для профилирования болезни	223
6.2.6. Этап 6: Отображение и автоматизация	224
6.3. Итоги	226

Глава 7. Графовые базы данных 227

7.1. Связанные данные и графовые базы данных	227
7.1.1. Когда и почему используются графовые базы данных?	231
7.2. Neo4j: графовая база данных	234
7.2.1. Cypher: язык запросов к графам	235
7.3. Пример использования связанных данных: рекомендательная система	242
7.3.1. Этап 1: Определение цели исследования	242
7.3.2. Этап 2: Сбор данных.	244
7.3.3. Этап 3: Подготовка данных.	245
7.3.4. Этап 4: Исследование данных.	248
7.3.5. Этап 5: Моделирование данных	251
7.3.6. Этап 6: Отображение	254
7.4. Итоги	255

Глава 8. Глубокий анализ текста 257

8.1. Глубокий анализ текста в реальном мире	259
8.2. Методы глубокого анализа текста	263
8.2.1. Набор слов.	264
8.2.2. Выделение основы и лемматизация	266
8.2.3. Классификатор на базе дерева принятия решений	267
8.3. Учебный пример: классификация сообщений Reddit	269
8.3.1. NLTK.	270
8.3.2. Обзор процесса data science и этап 1: назначение цели исследования	272
8.3.3. Этап 2: Сбор данных.	273
8.3.4. Этап 3: Подготовка данных.	277
8.3.5. Этап 4: Исследование данных.	280
8.3.6. Этап 3 (повторно): Подготовка данных (адаптированная)	283
8.3.7. Этап 5: Анализ данных	287
8.3.8. Этап 6: Отображение и автоматизация	291
8.4. Итоги	293

Глава 9. Визуализация данных для конечного пользователя 295

9.1. Способы визуализации данных	296
9.2. Crossfilter, библиотека MapReduce для JavaScript	300
9.2.1. Подготовка необходимых компонентов	300
9.2.2. Использование Crossfilter для фильтрации набора данных.	305

9.3. Создание информационной панели с использованием dc.js	309
9.4. Средства разработки	315
9.5. Итоги	317
Приложение А. Настройка Elasticsearch	319
A.1. Установка в Linux	319
A.2. Установка в Windows	321
Приложение Б. Установка Neo4j	325
Б.1. Установка в Linux	325
Б.2. Установка в Windows	326
Приложение В. Установка сервера MySQL	328
В.1. Установка в Windows	328
В.2. Установка в Linux	330
Приложение Г. Установка Anaconda в виртуальной среде	332
Г.1. Установка в Linux	332
Г.2. Установка в Windows	332
Г.3. Настройка среды	333