

# Оглавление

<b>Предисловие</b> .....	13
<b>Введение</b> .....	14
Что вы найдете в этой книге .....	15
Использование примеров исходного кода .....	15
Благодарности .....	15
<b>Глава 1. Анализ больших данных</b> .....	17
Основные проблемы науки о данных .....	19
Знакомство с Apache Spark .....	21
Об этой книге .....	23
<b>Глава 2. Введение в анализ данных с помощью Scala и Spark</b> ...	25
Scala для исследователей данных .....	26
Модель программирования Spark .....	27
Сопоставление записей .....	28
Начинаем работу: командная оболочка Spark и SparkContext .....	29
Доставка клиенту данных из кластера .....	34
Отправка кода с клиента на кластер .....	38
Структурирование данных с помощью кортежей и case-классов .....	38

Агрегации . . . . .	43
Создание гистограмм . . . . .	44
Сводные статистические данные для непрерывных переменных. . . . .	45
Создание многоцветного кода для вычисления сводных статистических данных. . . . .	46
Простой выбор и оценка переменных. . . . .	51
Куда двигаться дальше . . . . .	53
<b>Глава 3. Рекомендация музыки и набор данных сервиса AudioScrobbler . . . . .</b>	<b>54</b>
Набор данных . . . . .	55
Рекомендации на основе метода чередующихся наименьших квадратов . . . . .	56
Подготовка данных . . . . .	59
Создание первой модели . . . . .	61
Выборочная проверка модели . . . . .	64
Оценка качества рекомендаций . . . . .	65
Вычисление AUC . . . . .	67
Выбор гиперпараметров . . . . .	69
Подготовка рекомендаций . . . . .	71
Куда двигаться дальше . . . . .	72
<b>Глава 4. Прогнозирование лесного покрова с использованием деревьев принятия решений . . . . .</b>	<b>74</b>
Быстрый переход к регрессии . . . . .	75
Векторы и признаки. . . . .	75
Примеры обучения . . . . .	76
Деревья и леса принятия решений. . . . .	77

Набор данных Covtype . . . . .	80
Подготовка данных . . . . .	81
Первое дерево принятия решений . . . . .	82
Гиперпараметры деревьев принятия решений . . . . .	86
Настройка деревьев принятия решений . . . . .	88
Возвращаемся к категориальным признакам . . . . .	90
Случайные леса принятия решений . . . . .	92
Выполнение прогнозов . . . . .	94
Куда двигаться дальше . . . . .	95

## **Глава 5.** Обнаружение аномалий сетевого трафика

с помощью кластеризации методом k-средних . . . . .	97
Обнаружение аномалий . . . . .	98
Кластеризация методом k-средних . . . . .	98
Сетевая атака . . . . .	99
Набор данных кубка KDD-1999 . . . . .	100
Первая попытка кластеризации . . . . .	101
Выбор k . . . . .	103
Визуализация в R . . . . .	106
Нормирование признаков . . . . .	108
Категориальные переменные . . . . .	110
Использование меток с энтропией в качестве меры неоднородности . . . . .	111
Кластеризация в действии . . . . .	113
Куда двигаться дальше . . . . .	114

## **Глава 6.** Описание «Википедии»

с помощью латентно-семантического анализа . . . . .	116
Матрица «терм — документ» . . . . .	118
Получение данных . . . . .	119

Синтаксический разбор и подготовка данных . . . . .	119
Лемматизация . . . . .	121
Вычисление TF-IDF . . . . .	122
Сингулярное разложение. . . . .	125
Поиск важных концептов. . . . .	127
Выполнение запросов и оценок с помощью низкоразмерного представления. . . . .	130
Релевантность «терм — терм». . . . .	131
Релевантность «документ — документ» . . . . .	133
Релевантность «терм — документ» . . . . .	134
Запросы с несколькими термами . . . . .	135
Куда двигаться дальше . . . . .	137
<b>Глава 7. Анализ сетей совместной встречаемости с помощью GraphX. . . . .</b>	<b>138</b>
Индекс цитирования MEDLINE: сетевой анализ . . . . .	139
Получение данных. . . . .	141
Разбор XML-документов с помощью XML-библиотеки языка Scala . . . . .	143
Анализ основных тем MeSH и их взаимосвязей. . . . .	144
Построение сети совместной встречаемости с помощью GraphX . . . . .	147
Понимание структуры сетей. . . . .	150
Фильтрация зашумленных ребер . . . . .	155
Сети типа «мир тесен». . . . .	159
Вычисление средней длины пути с помощью Pregel . . . . .	161
Куда двигаться дальше . . . . .	167
<b>Глава 8. Анализ геопространственных и временных данных на примере поездок нью-йоркских такси . . . . .</b>	<b>168</b>
Получение данных. . . . .	169
Работа с временными и геопространственными данными в Spark . . . . .	170

Временные данные, JodaTime и NScalaTime . . . . .	171
Геопространственные данные, геометрическое API Esri и Spray . . . . .	172
Подготовка данных о поездках нью-йоркских такси . . . . .	177
Сеансирование в Spark . . . . .	185
Куда двигаться дальше . . . . .	189

<b>Глава 9. Оценка финансовых рисков с помощью моделирования по методу Монте-Карло . . . . .</b>	<b>191</b>
Терминология . . . . .	192
Методы вычисления VaR . . . . .	193
Наша модель . . . . .	194
Получение данных . . . . .	195
Предварительная обработка . . . . .	196
Выборка . . . . .	201
Выполнение испытаний . . . . .	204
Визуализация распределения доходов . . . . .	208
Оценка результатов . . . . .	209
Куда двигаться дальше . . . . .	211

<b>Глава 10. Анализ геномных данных и проект BDG . . . . .</b>	<b>213</b>
Разделяем хранение и моделирование . . . . .	214
Получение и обработка геномных данных с помощью ADAM CLI . . . . .	216
Прогнозирование факторов транскрипции участков связывания на основе данных ENCODE . . . . .	224
Запросы генотипов из проекта «1000 геномов» . . . . .	231
Куда двигаться дальше . . . . .	232

<b>Глава 11. Анализ нейровизуальных данных с помощью PySpark и Thunder</b> . . . . .	234
Обзор PySpark . . . . .	235
Обзор и установка библиотеки Thunder . . . . .	238
Загрузка данных с помощью Thunder . . . . .	239
Категоризация типов нейронов с помощью Thunder . . . . .	247
Куда двигаться дальше . . . . .	252
<b>Приложение А. Spark: копнем поглубже</b> . . . . .	253
Сериализация . . . . .	255
Сумматоры . . . . .	255
Spark и технологический процесс исследования данных . . . . .	256
Форматы файлов . . . . .	258
Подпроекты Spark . . . . .	260
<b>Приложение Б. Новый API конвейеров библиотеки MLlib</b> . . . . .	263
Выходим за пределы простого моделирования . . . . .	263
API конвейеров . . . . .	264
Пошаговый разбор примера классификации текста . . . . .	266