

Оглавление

| | |
|---|----|
| Об авторе | 13 |
| О научном редакторе | 14 |
| От издательства | 15 |
| Введение | 16 |
| Использование Python для data science | 17 |
| Для кого эта книга? | 17 |
| О чем эта книга? | 18 |
| Глава 1. Базовые знания о данных | 21 |
| Категории данных | 21 |
| Неструктурированные данные | 22 |
| Структурированные данные | 22 |
| Слабоструктурированные данные | 24 |
| Данные временных рядов | 26 |
| Источники данных | 27 |
| API | 28 |
| Веб-страницы | 29 |
| Базы данных | 30 |
| Файлы | 31 |
| Пайплайн обработки данных | 31 |
| Получение | 32 |
| Очистка | 32 |
| Преобразование | 33 |
| Анализ | 34 |
| Хранение | 35 |

| | |
|---|-----------|
| Питонический стиль | 35 |
| Выводы | 36 |
| Глава 2. Структуры данных Python | 37 |
| Списки | 37 |
| Создание списка | 38 |
| Использование общих методов списков | 38 |
| Использование срезов | 40 |
| Использование списка в качестве очереди | 42 |
| Использование списка в качестве стека | 43 |
| Использование списков и стеков для обработки естественного языка | 44 |
| Расширение функциональности с помощью списковых включений .. | 47 |
| Кортежи | 52 |
| Список кортежей | 52 |
| Неизменяемость | 53 |
| Словари | 54 |
| Список словарей | 54 |
| Добавление элементов в словарь с помощью setdefault() | 55 |
| Преобразование JSON в словарь | 57 |
| Множества | 58 |
| Удаление дубликатов из последовательности | 59 |
| Общие операции с множеством | 59 |
| Упражнение № 1: продвинутый анализ тегов фотографий | 60 |
| Выводы | 62 |
| Глава 3. Библиотеки Python для data science | 63 |
| NumPy | 63 |
| Установка NumPy | 64 |
| Создание массива NumPy | 64 |
| Выполнение поэлементных операций | 65 |
| Использование статистических функций NumPy | 66 |
| Упражнение № 2: использование статистических функций numpy .. | 67 |
| pandas | 67 |
| Установка pandas | 67 |
| pandas Series | 68 |

| | |
|--|------------|
| Упражнение № 3: объединение трех серий | 71 |
| pandas DataFrame | 71 |
| Упражнение № 4: использование разных типов join | 79 |
| scikit-learn | 82 |
| Установка scikit-learn | 83 |
| Получение набора образцов | 83 |
| Преобразование загруженного датасета в pandas DataFrame | 84 |
| Разделение набора данных на обучающий и тестовый | 84 |
| Преобразование текста в числовые векторы признаков | 85 |
| Обучение и оценка модели | 86 |
| Создание прогнозов на новых данных | 87 |
| Выводы | 87 |
| Глава 4. Доступ к данным из файлов и API | 88 |
| Импортирование данных с помощью функции open() | 88 |
| Текстовые файлы | 89 |
| Файлы с табличными данными | 91 |
| Упражнение № 5: открытие json-файлов | 93 |
| Двоичные файлы | 94 |
| Экспортирование данных в файл | 94 |
| Доступ к удаленным файлам и API | 96 |
| Как работают HTTP-запросы | 96 |
| Библиотека urllib3 | 97 |
| Библиотека Requests | 100 |
| Упражнение № 6: доступ к api с помощью requests | 101 |
| Перемещение данных в DataFrame и из него | 101 |
| Импортирование вложенных структур JSON | 102 |
| Конвертирование DataFrame в JSON | 103 |
| Упражнение № 7: обработка сложных структур json | 104 |
| Преобразование онлайн-данных в DataFrame с помощью pandas-datereader | 105 |
| Выводы | 106 |
| Глава 5. Работа с базами данных | 107 |
| Реляционные базы данных | 108 |

| | |
|---|-----|
| Понимание инструкций SQL | 109 |
| Начало работы с MySQL | 110 |
| Определение структуры базы данных | 111 |
| Вставка данных в БД | 114 |
| Запрос к базе данных | 116 |
| Упражнение № 8: объединение «один-ко-многим» | 118 |
| Использование инструментов аналитики баз данных | 118 |
| Базы данных NoSQL | 125 |
| Хранилища «ключ — значение» | 125 |
| Документоориентированные базы данных | 128 |
| Упражнение № 9: вставка и запрос нескольких документов | 131 |
| Выводы | 131 |
| Глава 6. Агрегирование данных | 132 |
| Данные для агрегирования | 133 |
| Объединение датафреймов | 135 |
| Группировка и агрегирование данных | 138 |
| Просмотр конкретных агрегированных показателей по MultiIndex .. | 139 |
| Срез диапазона агрегированных значений | 141 |
| Срезы на разных уровнях агрегирования | 142 |
| Добавление общего итога | 143 |
| Добавление промежуточных итогов | 144 |
| Упражнение № 10: исключение из датафрейма строк с итоговой суммой | 146 |
| Выбор всех строк в группе | 146 |
| Выводы | 147 |
| Глава 7. Объединение датасетов | 148 |
| Объединение встроенных структур данных | 149 |
| Объединение списков и кортежей с помощью оператора + | 149 |
| Объединение словарей с помощью оператора ** | 150 |
| Объединение строк из двух структур | 151 |
| Реализация join-объединений списков | 153 |

| | |
|--|------------|
| Конкатенация массивов NumPy | 156 |
| Упражнение № 11: добавление новых строк/столбцов в массив numpy | 158 |
| Объединение структур данных pandas | 158 |
| Конкатенация датафреймов | 158 |
| Удаление столбцов/строк из датафрейма | 161 |
| Join-объединение двух датафреймов | 164 |
| Выводы | 169 |
| Глава 8. Визуализация | 170 |
| Распространенные способы визуализации | 170 |
| Линейные диаграммы | 170 |
| Столчатые диаграммы | 172 |
| Круговые диаграммы | 173 |
| Гистограммы | 173 |
| Построение графиков с помощью Matplotlib | 174 |
| Установка Matplotlib | 175 |
| Использование matplotlib.pyplot | 175 |
| Работа с объектами Figure и Axes | 177 |
| Создание гистограммы с помощью subplots() | 178 |
| Упражнение № 12: объединение интервалов в сегмент other (другое) | 182 |
| Совместимость Matplotlib с другими библиотеками | 182 |
| Построение графиков для данных pandas | 182 |
| Отображение данных геолокации с помощью Cartopy | 184 |
| Упражнение № 13: составление карты с помощью cartopy и matplotlib | 189 |
| Выводы | 190 |
| Глава 9. Анализ данных о местоположении | 191 |
| Получение данных о местоположении | 191 |
| Преобразование стандартного вида адреса в геокоординаты | 192 |
| Получение геокоординат движущегося объекта | 193 |
| Анализ пространственных данных с помощью geopy и Shapely | 196 |
| Поиск ближайшего объекта | 197 |
| Поиск объектов в определенной области | 200 |

| | |
|---|-----|
| Упражнение № 14: определение двух и более многоугольников | 201 |
| Объединение двух подходов | 202 |
| Упражнение № 15: совершенствование алгоритма подбора машины | 204 |
| Объединение пространственных и непространственных данных | 204 |
| Получение непространственных характеристик | 204 |
| Упражнение № 16: фильтрация данных с помощью спискового включения | 206 |
| Объединение датасетов с пространственными и непространственными данными | 207 |
| Выводы | 208 |
| Глава 10. Анализ данных временных рядов | 209 |
| Регулярные и нерегулярные временные ряды | 209 |
| Общие методы анализа временных рядов | 211 |
| Вычисление процентных изменений | 213 |
| Вычисление скользящего окна | 215 |
| Вычисление процентного изменения скользящего среднего | 216 |
| Многомерные временные ряды | 216 |
| Обработка многомерных временных рядов | 218 |
| Анализ зависимости между переменными | 219 |
| Упражнение № 17: введение дополнительных метрик для анализа зависимостей | 222 |
| Выводы | 224 |
| Глава 11. Получение инсайтов из данных | 225 |
| Ассоциативные правила | 226 |
| Поддержка | 227 |
| Доверие | 227 |
| Лифт | 228 |
| Алгоритм Apriori | 228 |
| Создание датасета с транзакциями | 229 |
| Определение часто встречающихся наборов | 231 |
| Генерирование ассоциативных правил | 233 |
| Визуализация ассоциативных правил | 234 |

| | |
|--|-----|
| Получение полезных инсайтов из ассоциативных правил | 238 |
| Генерирование рекомендаций | 238 |
| Планирование скидок на основе ассоциативных правил | 239 |
| Упражнение № 18: извлечение данных о реальных транзакциях . . . | 242 |
| Выводы | 246 |
| Глава 12. Машинное обучение для анализа данных | 247 |
| Почему машинное обучение? | 247 |
| Типы машинного обучения | 248 |
| Обучение с учителем | 248 |
| Обучение без учителя | 250 |
| Как работает машинное обучение | 250 |
| Данные для обучения | 250 |
| Статистическая модель | 252 |
| Неизвестные данные | 252 |
| Пример анализа тональности: классификация отзывов о товарах | 253 |
| Получение отзывов о товарах | 253 |
| Очистка данных | 255 |
| Разделение и преобразование данных | 257 |
| Обучение модели | 260 |
| Оценка модели | 260 |
| Упражнение № 19: расширение набора примеров | 263 |
| Прогнозирование тенденций фондового рынка | 264 |
| Получение данных | 265 |
| Извлечение признаков из непрерывных данных | 266 |
| Генерирование выходной переменной | 267 |
| Обучение и оценка модели | 268 |
| Упражнение № 20: экспериментируем с различными акциями и новыми метриками | 269 |
| Выводы | 270 |