

УДК 004.738.5:004.438R  
ББК 32.971.353

X89

X89 Храмов Д. А.  
Сбор данных в Интернете на языке R. – М.: ДМК Пресс, 2017. – 280 с.: ил.

ISBN 978-5-97060-459-5

Всё, что регистрирует человек и созданные им машины, может считаться данными. Фиксируя новое и переводя архивы в цифровую форму, мы с каждым днём производим всё больше данных. Но гораздо чаще случается так, что данные разбросаны по всемирной сети на многочисленных страницах онлайн-магазинов, заметках в социальных сетях, логах серверов и т. п. Прежде чем начать работать с такими данными, их необходимо собрать и сохранить в пригодном для анализа виде. Решению этих вопросов и посвящена данная книга.

Основной материал книги разделён на две части. В первой части дано краткое введение в R – описание среды разработки, языка и основных пакетов-расширений. Вторая часть посвящена непосредственно сбору данных: работе с открытыми данными, извлечению данных из веб-страниц и из социальных сетей. Также рассмотрены необходимые технические вопросы: протокол HTTP, функции импорта данных различных форматов и регулярные выражения. Завершается рассказ созданием карт на основе собранных данных.

Издание предназначено специалистам по анализу данных, а также программистам, интересующихся сбором данных в Интернете.

УДК 004.738.5:004.438R  
ББК 32.971.353

ISBN 978-5-97060-459-5

© Храмов Д.А., 2016  
© Оформление, издание, ДМК Пресс, 2017

# Содержание

<b>Введение</b> .....	11
Кто и зачем собирает данные.....	11
Почему R? .....	12
Как устроена эта книга .....	13
Обратная связь .....	13
 <b>ЧАСТЬ I. ПРОГРАММИРОВАНИЕ НА R</b> .....	<b>14</b>
 <b>Глава 1. Знакомство с R</b> .....	<b>15</b>
Установка .....	15
Работа в среде RGui .....	17
Справка .....	22
 <b>Глава 2. Скаляры, векторы и матрицы</b> .....	<b>24</b>
Арифметические операции и присваивание .....	24
Имена .....	25
Простые типы данных.....	26
Числа .....	26
Символьный тип.....	28
Логический тип .....	30
Векторы.....	31
Векторизация и логическая индексация.....	36
Матрицы и массивы .....	39
Резюме .....	41
 <b>Глава 3. Списки и таблицы</b> .....	<b>42</b>
Списки .....	42
Таблицы.....	45
Функции, применяемые к составным данным.....	50
apply.....	50
lapply.....	51
sapply.....	52
do.call .....	53
Резюме .....	53
 <b>Глава 4. Управление процессом вычислений</b> .....	<b>54</b>
Циклы .....	54
Цикл со счётчиком .....	54
Цикл с предусловием .....	57
Условные операторы.....	58
Резюме .....	59

<b>Глава 5. Базовая графика</b> .....	60
Функции низкого и высокого уровней.....	60
Глобальные и локальные параметры графиков.....	65
Легенда.....	67
Комбинации графиков .....	67
Графики функций.....	69
Экспорт в файлы .....	70
Резюме и ссылки .....	70
<b>Глава 6. Функции</b> .....	72
Создание функций.....	72
Локальные и глобальные переменные. Области видимости .....	74
Диагностические сообщения.....	76
Функции в качестве аргументов .....	76
Функциональное программирование .....	78
Резюме .....	79
<b>Глава 7. Факторы и даты</b> .....	80
Категориальные данные .....	80
Дата и время .....	83
Резюме .....	86
<b>Глава 8. Пакеты</b> .....	87
Установка и загрузка .....	87
Выбор пакета.....	89
Справка и её разновидности .....	89
Как самому создать пакет R?.....	91
Пакет magrittr: конвейер операций .....	92
<b>Глава 9. Ввод и вывод данных. Работа с файлами</b> .....	94
Рабочий каталог пользователя .....	94
Запись данных в стандартное устройство вывода.....	94
Запись в текстовые файлы .....	95
Таблицы .....	95
Строки.....	97
Матрицы .....	97
Чтение из текстовых файлов.....	97
Элементы данных: scan .....	97
Строки: readLines.....	99
Таблицы .....	100

Работа с данными в бинарном формате .....	101
Управление файлами и каталогами .....	102
Взаимодействие с базами данных .....	103
DBI + RSQLite .....	103
sqldf .....	103
Резюме .....	104
<b>Ссылки к части I .....</b>	<b>105</b>

## **ЧАСТЬ II. СБОР ДАННЫХ..... 106**

### **Глава 10. Открытые данные..... 107**

Что это такое?.....	107
Данные Всемирного банка.....	108
Где взять данные? .....	113
Резюме .....	114

### **Глава 11. Протокол HTTP..... 115**

Основные понятия .....	115
Запрос .....	116
Ответ .....	117
Коды состояния .....	118
Передача параметров .....	119
HTTP в R .....	120
Пакет httr .....	120
Пакет RCurl .....	122
Кириллица и кодирование URL.....	123
Пример: геокодирование с помощью Google Maps Geocoding.....	124
Пример: доступ к API портала открытых данных РФ .....	126
Ссылки .....	129

### **Глава 12. Импорт данных ..... 130**

Чтение файлов.....	130
Скачивание.....	131
Excel .....	132
JSON.....	133
Пример: какой из JSON-пакетов самый популярный?.....	133
Google Spreadsheets .....	136
Архивы .....	137
Завершающий штрих: проверка типа данных.....	138
Ссылки .....	139

<b>Глава 13. Веб-скрапинг</b>	<b>140</b>
Используйте структуру данных	140
Элементы HTML и CSS	143
div и span	143
Классы и идентификаторы	144
Путь к элементу	146
XPath	146
CSS	149
Как найти путь к элементу при помощи браузера	150
Проверка и упрощение пути. Консоль разработчика	153
Резюме	155
Лирическое отступление: построение графов	155
Ссылки	157
Поиск в Интернете	157
HTML и CSS:	158
XPath	158
<b>Глава 14. Пакет rvest</b>	<b>159</b>
Пакеты для веб-скрапинга	159
Получение и обработка HTML-документа	160
Поиск элемента	162
Разбор элемента	164
Пример: получаем ссылку и скачиваем файл	165
Таблицы	166
Пример: извлечение таблицы из Википедии	166
Пример: разбор страницы сериала «Светлячок»	167
Пример: извлечение данных об инвестиционных фондах	169
Работа с формами. Сессии	171
Пример: аутентификация на форуме	173
Функции навигации	174
Работа с кодировками	175
Заключительные замечания и ссылки	175
<b>Глава 15. RSelenium: управляем браузером</b>	<b>177</b>
Пример: перевод с помощью Yandex.Translate	179
Пример: динамически генерируемая ссылка на файл	180
Selenium и браузеры	183
Резюме и ссылки	183

<b>Глава 16. PhantomJS и обработка динамических веб-страниц</b> .....	185
Динамические страницы: описание проблемы .....	185
Установка .....	186
Запуск .....	186
Пример: рендеринг веб-страницы .....	187
Сохранение веб-страницы в файл .....	188
Резюме и ссылки .....	190
<b>Глава 17. Facebook</b> .....	192
Протокол авторизации OAuth 2.0 .....	192
Получение маркера доступа пользователя API Graph .....	193
Доступ к данным с помощью rvest и jsonlite .....	196
Пакет Rfacebook и создание приложения .....	198
<b>Глава 18. Сбор информации с помощью API ВКонтакте</b> .....	204
Создание приложения .....	204
Регистрация приложения .....	204
Получение кода доступа .....	206
Получение данных .....	207
Реализация в R .....	208
Построение графа связей .....	210
Получение другой информации из сети .....	212
Поиск пользователя .....	213
Ограничения .....	214
<b>Глава 19. Использование Twitter API</b> .....	215
Получение доступа к Twitter API .....	215
Подключение к Twitter из R .....	215
Поиск и сохранение его результатов в базе данных .....	217
Фильтрация результатов поиска .....	218
Построение облака слов .....	219
Данные для анализа .....	220
Лексический корпус и терм-документная матрица .....	220
Ключевые слова и их частоты .....	221
Облако слов .....	221
Ограничения Search API .....	223
Streaming API .....	223
Ссылки .....	223

<b>Глава 20. Регулярные выражения</b>	225
Символы и метасимволы	225
Квантификаторы	227
Положение образца внутри строки	228
Операторы	229
«Жадность» и «лень» квантификаторов	230
Классы символов	232
Заключительные замечания	233
Ссылки	234
<b>Глава 21. Создание карт на основе собранных данных</b>	235
Интерактивные карты в leaflet	235
Переходим к созданию карты	239
Извлечение адресов и названий магазинов	240
Геокодирование	242
Отображение на карте	243
Работа с шейп-файлами	244
Ссылки	247
<b>Ссылки к части II</b>	249
<b>Приложение А. Среда разработки RStudio</b>	250
Создание скрипта	251
Автодополнение имён объектов	252
Выполнение	252
Рабочее пространство	253
История команд	254
Сохранение файлов	256
Кодировки файлов	256
Управление файлами в рабочем каталоге	257
Управление пакетами	257
Поиск и замена	258
Автоматическое создание функций	259
Комментирование	260
Переход к определению функции	260
Ссылки	261
<b>Приложение Б. Языки поисковых запросов Google и Яндекс</b>	262
Почему важно уметь пользоваться ЯПЗ	263
Предотвращение перегрузок сервиса	263

<b>Приложение В. Введение в HTML и CSS</b> .....	264
Веб-страница.....	264
Гиперссылки .....	266
Шрифт .....	267
Цвет.....	268
Стиль.....	268
Выравнивание .....	270
Рисунки.....	270
Списки .....	271
Маркированные.....	271
Нумерованные .....	271
Вложенные.....	272
Таблицы.....	272
Ссылки .....	273
<b>Приложение Г. Регулярные выражения</b> .....	274
<b>Предметный указатель</b> .....	276