

Храмов Д. А.

Сбор данных в Интернете на языке R. - М.: ДМ К Пресс, 2017. - 282
е.: ил.

Содержание

Введение	11
Кто и зачем собирает данные.....	11
Почему R?.....	12
Как устроена эта книга.....	13
Обратная связь.....	13
ЧАСТЫ. ПРОГРАММИРОВАНИЕ НА R	14
Глава 1. Знакомство с R	15
Установка.....	15
Работа в среде RGui.....	17
Справка.....	22
Глава 2. Скаляры, векторы и матрицы	24
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
Глава 3. Списки и таблицы	42
Списки.....	42
Таблицы.....	45
Функции, применяемые к составным данным.....	50
apply.....	50
lapply.....	51
sapply.....	52
do.call.....	53
Резюме.....	53
Глава 4. Управление процессом вычислений	54
Циклы.....	54
Цикл со счётчиком.....	54
Цикл с условием.....	57
Условные операторы.....	58
Резюме.....	59

Глава 5. Базовая графика.....	60
Функции низкого и высокого уровней.....	60
Глобальные и локальные параметры графиков.....	65
Легенда	
Комбинации графиков.....	67
Графики функций.....	69
Экспорт в файлы.....	70
	70
Резюме и ссылки	
Глава 6. Функции.....	72
Создание функций	
Локальные и глобальные переменные. Области видимости.....	74
Диагностические сообщения.....	76
Функции в качестве аргументов.....	76
	78
Функциональное программирование	
	79
Резюме	
Глава 7. Факторы и даты.....	80
Категориальные и данные.....	80
Дата и время.....	83
Резюме.....	86
Глава 8. Пакеты.....	87
Установка и загрузка.....	87
Выбор пакета.....	89
Справка и её разновидности.....	89
Как самому создать пакет R?.....	91
Пакет magrittr: конвейер операций.....	92
Глава 9. Ввод и вывод данных. Работа с файлами.....	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
Ссылки к части I	105
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 bR	120
Пакет http	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

Глава 13. Веб-скрапинг.....	140
Используйте структуру данных	140
Элементы HTML и CSS.....	143
div и span	143
Классы и идентификаторы.....	144
Путь к элементу.....	140
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
Глава 14. Пакет rvest.....	159
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
Глава 15. RSelenium: управляем браузером.....	177
Пример: переводе помощью Yandex.Translate.....	179
Пример: динамически генерируемая ссылка на файл.....	180
Selenium и браузеры.....	183
Резюме и ссылки.....	183

Глава 16. PhantomJS и обработка динамических веб-страниц . . .	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
Глава 17. Facebook	192
Протокол авторизации OAuth 2.0	192
Получение маркера доступа пользователя API Graph	193
Доступ к данным с помощью gvest и jsonlite	196
Пакет Rfacebook и создание приложения	198
Глава 18. Сбор информации с помощью API ВКонтакте	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
Глава 19. Использование Twitter API	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

Глава 20. Регулярные выражения.....	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
Глава 21. Создание карт на основе собранных данных.....	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
Ссылки к части II.....	249
Приложение А. Среда разработки RStudio.....	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
Приложение Б. Языки поисковых запросов Google и Яндекс.	262
Почему важно уметь пользоваться ЯИЗ.....	263
Предотвращение перегрузок сервиса.....	263

Приложение В. Введение в HTML и CSS.....	264
Всб-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
Приложение Г. Регулярные выражения.....	274
Предметный указатель.....	276