

目 次

まえがき	iii
第1部 Web とデータの技術入門	1
第1章 導 入	3
1.1 ケーススタディ：危機にある世界遺産	3
1.2 Web のデータの品質について	10
1.3 Web データを配信, 抽出, そして保存する技術	13
1.3.1 Web コンテンツを配信する技術	14
1.3.2 Web ドキュメントから情報抽出する技術	16
1.3.3 データの保存技術	18
1.4 本書の構成	19
第2章 HTML	21
2.1 ブラウザでの表示とソースコード	21
2.2 構文規則	23
2.2.1 タグ, 要素, および属性	23
2.2.2 ツリー構造	24
2.2.3 コメント	26
2.2.4 予約済みの文字と特殊文字	26
2.2.5 ドキュメントのタイプ定義	27
2.2.6 スペースと改行	28
2.3 タグおよび属性	28
2.3.1 アンカータグ<a>	28
2.3.2 メタデータタグ<meta>	29
2.3.3 外部参照タグ<link>	30

2.3.4 強調タグ <code></code> , <code><i></code> , <code></code>	31
2.3.5 段落タグ <code><p></code>	31
2.3.6 見出しタグ <code><h1></code> , <code><h2></code> , <code><h3></code> , …	31
2.3.7 <code></code> , <code></code> , <code><dl></code> タグによるコンテンツのリスト化	32
2.3.8 構造化のためのタグ <code><div></code> , <code></code>	32
2.3.9 <code><form></code> と関連タグ	33
2.3.10 外部スクリプトを指定するタグ <code><script></code>	36
2.3.11 表のタグ <code><table></code> , <code><tr></code> , <code><td></code> , <code><th></code>	37
2.4 構文解析	38
2.4.1 パーサとは	38
2.4.2 ノードの剪定	41
2.4.3 DOM 生成処理による情報抽出	43
第 3 章 XML と JSON	49
3.1 XML ドキュメントの具体例	50
3.2 XML 構文規則	52
3.2.1 要素と属性	52
3.2.2 XML 構造	55
3.2.3 命名規則と特殊文字	57
3.2.4 コメントと文字データ	58
3.2.5 XML 構文の概要	60
3.3 XML ドキュメントが正しく形成され、有効となるのはどのような場合か	61
3.4 XML 拡張機能と技術	63
3.4.1 名前空間	63
3.4.2 XML の拡張	65
3.4.3 例：RSS (Really Simple Syndication)	66
3.4.4 例：SVG (Scalable vector graphics)	69
3.5 XML と R の練習	72
3.5.1 XML の解析	73
3.5.2 XML ドキュメント上の基本的な操作	75
3.5.3 XML からデータフレームやリストへ	79
3.5.4 イベントドリブン型解析	80
3.6 JSON ドキュメントの具体例	83
3.7 JSON の構文規則	84
3.8 JSON と R の練習	87

第 4 章 XPATH	97
4.1 XPath——Web ドキュメント用のクエリ言語	98
4.2 XPath によるノードセットの識別	100
4.2.1 XPath クエリの基本構造	100
4.2.2 ノードの関係性	104
4.2.3 XPath 述語	107
4.3 ノード要素の抽出	115
4.3.1 fun 引数の拡張	116
4.3.2 XML 名前空間	119
4.3.3 XPath に役立つツール	121
第 5 章 HTTP	125
5.1 HTTP の基本	126
5.1.1 Web サーバと短い対話	126
5.1.2 URL 構文	129
5.1.3 HTTP メッセージ	132
5.1.4 リクエストメソッド	133
5.1.5 ステータスコード	134
5.1.6 ヘッダフィールド	135
5.2 HTTP の高度な技術	142
5.2.1 識別	143
5.2.2 認証	148
5.2.3 プロキシ	150
5.3 HTTP 以外のプロトコル	152
5.3.1 セキュアな HTTP (HTTP Secure)	152
5.3.2 FTP	154
5.4 HTTP プロトコルの実際	155
5.4.1 <code>libcurl</code> ライブライ	155
5.4.2 基本的なリクエストメソッド	156
5.4.3 <code>RCurl</code> の低水準関数	160
5.4.4 リクエスト間でのコネクション維持	162
5.4.5 オプション	163
5.4.6 デバッグ	169
5.4.7 エラー処理	174
5.4.8 <code>RCurl</code> と <code>httr</code>	175

第 6 章 AJAX	181
6.1 JavaScript	182
6.1.1 JavaScript を使用する方法	183
6.1.2 DOM 操作	183
6.2 XHR	188
6.2.1 外部 HTML/XML ドキュメントの読み込み	189
6.2.2 JSON の読み込み	192
6.3 Web 開発者ツールで AJAX を調査する	195
6.3.1 Chrome の Web 開発者ツールを使ってみよう	195
6.3.2 Elements パネル	196
6.3.3 Network パネル	196
第 7 章 SQL とリレーションナルデータベース	201
7.1 概要および用語	203
7.2 リレーションナルデータベース	205
7.2.1 データをテーブルに格納する	205
7.2.2 正規化	209
7.2.3 リレーションナルデータベースと DBMS のさらなる機能	213
7.3 SQL: データベースと会話するための言語	216
7.3.1 SQL の総論とその構文および実行例	216
7.3.2 データ制御言語——DCL	218
7.3.3 データ定義言語——DDL	219
7.3.4 データ操作言語——DML	221
7.3.5 句	228
7.3.6 トランザクション制御言語——TCL	231
7.4 データベースの実践	233
7.4.1 データベースを操作するための R パッケージ	233
7.4.2 DBI ベースのパッケージを介した R と SQL の対話	233
7.4.3 RODBC を介した R と SQL の対話	236
第 8 章 正規表現と重要な文字列関数	243
8.1 正規表現	245
8.1.1 マッチした文字列の抽出	245
8.1.2 正規表現の一般化	249
8.1.3 シンプソンズ再び	256

8.2 文字列処理	258
8.2.1 stringr パッケージ	258
8.2.2 いくつかの便利な関数	263
8.3 文字エンコーディング	267
第 2 部 Web スクレイピングとテキストマイニングのためのツールボックス	273
第 9 章 Web からのスクレイピング	275
9.1 収集のシナリオ	277
9.1.1 すでに整形されたデータのダウンロード	277
9.1.2 FTP インデックスからの複数のファイルのダウンロード	282
9.1.3 複数のページにアクセスするために URL を操作する	285
9.1.4 リンクやリスト、表を HTML ドキュメントから取得する便利な関数	290
9.1.5 HTML フォームを処理する	294
9.1.6 HTTP 認証	308
9.1.7 HTTPS による接続	309
9.1.8 Cookies の利用	311
9.1.9 Selenium と Rwebdriver を用いた AJAX による Web ページからのデータスクレイピング	317
9.1.10 API からのデータ収集	326
9.1.11 OAuth による認証	332
9.2 抽出方法	337
9.2.1 正規表現	337
9.2.2 XPath	342
9.2.3 API (Application Programming Interface)	345
9.3 Web スクレイピング：グッドプラクティス	348
9.3.1 Web スクレイピングは合法か？	348
9.3.2 robots.txt とは何か	350
9.3.3 スクレイピングは友好的に	355
9.4 インスピレーションを与えてくれる価値ある情報源	364
第 10 章 統計的テキスト処理	369
10.1 例：英國政府のプレスリリースを分類する	370
10.2 テキストデータの処理	372
10.2.1 tm パッケージによる大規模なテキスト操作	373

10.2.2 単語文書行列の構築	379
10.2.3 データクレンジング	381
10.2.4 スパース性と n-gram	382
10.3 教師あり学習の手法	385
10.3.1 サポートベクターマシン	387
10.3.2 ランダムフォレスト	387
10.3.3 最大エントロピー法	388
10.3.4 RTextTools パッケージ	388
10.3.5 応用：政府によるプレスリリース	388
10.4 教師なし学習の手法	393
10.4.1 Latent Dirichlet allocation と関連するトピックモデル	394
10.4.2 応用：政府のプレスリリース	394
第 11 章 データ分析プロジェクトの管理	403
11.1 ファイルシステムの操作	403
11.2 複数のドキュメントやリンクの処理	404
11.2.1 for ループの使用	405
11.2.2 while ループと制御構造の使用	407
11.2.3 plyr パッケージの使用	409
11.3 スクレイピング処理の構築	410
11.3.1 進捗のフィードバックの実装：メッセージとプログレスバー	414
11.3.2 エラーと例外処理	417
11.4 R スクリプトの定期実行	419
11.4.1 Mac OS と Linux におけるスケジュールタスク	420
11.4.2 Windows プラットフォームにおけるスケジュールタスク	422
第 3 部 事例集	425
第 12 章 アメリカ上院議員間のコラボレーション・ネットワーク	427
12.1 法案に関する情報	428
12.2 上院議員の情報	436
12.3 ネットワーク構造の解析	440
12.3.1 記述統計	441
12.3.2 ネットワーク分析	444
12.4 結論	446

第 13 章 半構造化されたドキュメントから情報を抜き出す	447
13.1 FTP サーバからデータをダウンロードする	448
13.2 半構造化されたテキストデータをパースする	450
13.3 測候所と気温データの可視化	457
第 14 章 Twitter による 2014 年度アカデミー賞予測	463
14.1 Twitter API : 概要	464
14.1.1 REST API	464
14.1.2 ストリーミング API	465
14.1.3 データの収集と準備	466
14.2 Twitter ベースでの 2014 年度アカデミー賞予測	467
14.2.1 データの可視化	467
14.2.2 予測のためのツイートマイニング	468
14.3 結論	473
第 15 章 名字の地理的な分布のマッピング	475
15.1 データ収集戦略の構築	476
15.2 Web サイトの調査	477
15.3 データの検索と情報の抽出	480
15.4 名字のマッピング	484
15.5 プロセスの自動化	487
第 16 章 携帯電話のデータを集める	497
16.1 ページの探索	497
16.1.1 特定のブランドの携帯電話を探す	497
16.1.2 商品情報の抽出	502
16.2 スクレイピングの実施手順	508
16.2.1 他のメーカーに対するデータの検索	508
16.2.2 データクレンジング	510
16.3 グラフィカル分析	511
16.4 データの蓄積	513
16.4.1 一般的な考慮事項	513
16.4.2 データ保存のためのテーブル定義	515
16.4.3 将来的なデータ蓄積のためのテーブル定義	517
16.4.4 データアクセスを便利にするためのビューの定義	518

16.4.5 データを保存するための関数	521
16.4.6 データの保存とその確認	523
第17章 商品レビューのセンチメント分析	525
17.1 イントロダクション	525
17.2 データ収集	526
17.2.1 ファイルのダウンロード	526
17.2.2 情報の抽出	532
17.2.3 データベースの蓄積	537
17.3 データの分析	539
17.3.1 データの準備	540
17.3.2 辞書ベースのセンチメント分析	542
17.3.3 レビューの内容をマイニングする	548
17.4 結論	550
参考文献	553
訳者あとがき	561
事項索引	563
パッケージ索引	568
関数索引	569