

目 次

第 1 章	テキストアナリティクス	1
1.1	テキストアナリティクスとは	1
1.2	テキストアナリティクスの諸相	2
1.2.1	テキストアナリティクスの由来	2
1.2.2	計量文体学	3
1.2.3	計量言語学とコーパス言語学	4
1.2.4	情報・社会科学	6
1.3	テキストアナリティクスの手順	7
第 2 章	テキストアナリシスのための前処理	9
2.1	電子化とテキストの収集	9
2.2	テキストのクリーニングと正規表現	10
2.2.1	テキストエディタ	12
2.2.2	正規表現	14
2.3	プログラミング言語	16
2.4	テキストの処理	17
2.4.1	形態素解析	18
2.4.2	構文解析	21
2.5	要素・項目の集計	24
2.5.1	n -gram 統計モデル	24
2.5.2	特徴ベクトル	25
第 3 章	テキストデータの視覚化	29
3.1	棒グラフと折れ線グラフ	29
3.2	ワードクラウド	31

3.3	格子グラフ	33
3.4	ネットワークプロット	34
3.4.1	ネットワークの統計量	35
3.4.2	コミュニティ分析	38
3.5	テキストにおけるネットワーク分析	41
第4章	法則と語句の重みおよび特徴語句抽出	47
4.1	ジップの法則	47
4.2	語彙の豊かさ	49
4.2.1	延べ語数と異なり語数を用いた指標	50
4.2.2	頻度スペクトルを用いた指標	52
4.3	語句の重み	53
4.3.1	プーリアン重み付け	54
4.3.2	頻度重み付け	54
4.3.3	TF-IDF 重み付け	55
4.3.4	エントロピー重み付け	56
4.3.5	相互情報量による共起頻度の重み付け	56
4.4	特徴語句の抽出	61
4.4.1	カイ二乗統計量	61
4.4.2	外的基準の利用	64
第5章	テキストの特徴分析	65
5.1	特徴分析のデータの形式	65
5.2	特異値分解	66
5.3	主成分分析	67
5.3.1	主成分と寄与率・累積寄与率	67
5.3.2	主成分得点	68
5.3.3	主成分分析の例	68
5.4	対応分析	74
5.4.1	固有値分解と対応分析	75

5.4.2	対応分析の例	75
5.5	潜在意味解析	76
5.6	確率潜在意味解析	78
5.6.1	pLSA とは	78
5.6.2	pLSA の分析例	79
5.7	トピックモデル LDA	84
5.7.1	LDA とは	85
5.7.2	LDA の分析例	86
5.7.3	トピックモデル	89
5.7.4	トピックの数について	89
5.8	NMF 分析	92
5.8.1	基本アルゴリズム	92
5.8.2	NMF 分析の例	96
5.9	その他の方法	99
第 6 章 テキストのクラスター分析		101
6.1	類似度と非類似度	101
6.1.1	量的データの類似度	101
6.1.2	名義尺度の類似度	103
6.1.3	多値名義尺度	105
6.2	非類似度と距離	106
6.2.1	量的データの距離	107
6.2.2	相対頻度データの距離	108
6.3	階層的クラスタリング	109
6.3.1	階層的クラスタリングのプロセス	110
6.3.2	階層的クラスタリングの流れ	110
6.3.3	階層的クラスタリングの方法	111
6.4	クラスターのヒートマップ	114
6.5	非階層的クラスタリング	116
6.6	クラスターの数の決定方法	118

第7章 テキストの分類と判別分析	123
7.1 分類と判別分析	123
7.1.1 線形判別分析	124
7.1.2 ベイズ判別分析	126
7.1.3 ロジスティック判別分析	127
7.1.4 k 近傍法	128
7.2 サポートベクターマシン	129
7.2.1 サポートベクターマシンの基本定式	129
7.2.2 カーネル法	132
7.3 ツリーモデル	133
7.4 アンサンブル学習	138
7.4.1 ブースティング	138
7.4.2 ランダムフォレスト	139
7.5 ニューラルネットワーク	142
7.5.1 ニューラルネットワークとは	142
7.5.2 階層ニューラルネットワーク	145
7.6 モデルと結果の評価	147
7.6.1 交差確認法	147
7.6.2 分類結果の評価指標	148
7.7 いくつかの分類器の比較	152
7.7.1 スпамメール	153
7.7.2 文章の著者の識別	158
7.8 統合的分析	160
7.8.1 統合的分類アルゴリズム	161
7.8.2 用いるコーパスとデータセット	161
7.8.3 書き手の特徴データ	163
7.8.4 用いる分類器	167
7.8.5 評価方法	168
7.8.6 分類器ごとの判別結果	169
7.8.7 統合的判別の結果	172

第 8 章 テキストデータによる予測と要因分析	175
8.1 テキストの経時的分析	175
8.2 重回帰分析	176
8.2.1 重回帰分析の定式	176
8.2.2 変数の選択	177
8.2.3 執筆時期の推定	177
8.3 正則化回帰モデル	180
8.3.1 ridge 回帰モデル	181
8.3.2 lasso 回帰モデル	181
8.3.3 elastic net 回帰モデル	183
8.3.4 正則化回帰モデルによる執筆時期の推定	184
8.4 その他の回帰分析	188
8.4.1 サポートベクター回帰	189
8.4.2 回帰木とランダムフォレスト	190
8.4.3 いくつかの回帰分析の結果の比較	192
参考文献	197
索引	207