

# 目次

<b>1.</b>	<b>コーパスとテキストマイニング</b>	<b>1</b>
1.1	コーパスとテキストマイニングとは	1
1.2	テキストマイニングの準備	2
1.2.1	テキストの収集とクリーニング	2
1.2.2	テキストの処理と加工	2
1.3	テキストの構造化	5
1.4	テキストマイニングの主な方法	6
1.4.1	視覚化	6
1.4.2	重要語句の抽出と関連指標	7
1.4.3	テキストの特徴分析	10
1.4.4	テキストのグループ分け	11
1.4.5	テキストの自動分類	12
1.4.6	テキストの時系列分析	13
1.5	意味解析と辞書	14
	参考文献	14
<b>2.</b>	<b>金融テキストマイニングの紹介</b>	<b>15</b>
2.1	金融テキストマイニングの背景と目的	15
2.2	金融テキストマイニング手法の枠組み	17
2.2.1	問題設定	17
2.2.2	機械学習アルゴリズムの選択	17
2.2.3	データの選択	19
2.2.4	前処理の選択	19
2.2.5	モデルの評価（予測精度の推定）	21
2.3	金融テキストマイニングの研究事例	21
2.3.1	分析目的と利用テキストから見た研究事例の分類	21
2.3.2	ツイッターに現れる心理状態と株価平均	22
2.3.3	リアルタイム配信ニュースと短期市場変動	22
2.3.4	経済リポートと長期市場変動	23

2.4	まとめ	24
	参考文献	24
<b>3.</b>	<b>テキストマイニングを用いた医療専門職のクセの可視化およびその活用</b>	<b>26</b>
3.1	特定健診・特定保健指導制度について	26
3.2	テキストマイニングの適用に至るまで	27
3.3	テキストマイニングの手法の選択について	27
3.4	テキストマイニングの実施環境について	28
3.5	支援内容において注意すべきポイントの可視化	28
3.6	読みにくさを避ける	28
3.6.1	文の長さ・1メッセージあたりの文の数	29
3.6.2	漢字の占める割合	30
3.6.3	専門用語の出現数	31
3.7	単調さを避ける	33
3.7.1	文末表現	33
3.7.2	呼びかけ表現	34
3.7.3	メッセージの非定型度	35
3.8	支援者のグルーピング	37
3.9	まとめおよび展望	38
	参考文献	38
<b>4.</b>	<b>自由記述によるアンケート調査からことばの地域差を探る</b>	<b>40</b>
4.1	自由回答データ解析の目的	40
4.1.1	全国通信調査に基づく調査結果の分析	41
4.1.2	調査結果の分析例1—前置き表現にみられる地域差と世代差—	42
4.1.3	感謝表現にみられる場面差	46
4.2	おわりに	53
	参考文献	54
<b>5.</b>	<b>文章の書き手の特徴情報と書き手の識別</b>	<b>55</b>
5.1	計量的文体論と著者の識別	55
5.1.1	計量的文体論	55
5.1.2	文章の著者の識別	55
5.2	書き手の特徴抽出	56
5.2.1	長さの分布	57
5.2.2	品詞の分布	58
5.2.3	識別語と機能語	59
5.2.4	語彙に関する指標	60
5.2.5	n-gramの分布	61
5.2.6	日本語固有の特徴	61
5.2.7	その他の特徴情報	63

5.3	書き手識別の方法	64
5.3.1	簡単な記述統計	64
5.3.2	推測統計と多変量データ解析	64
5.3.3	データマイニングと機械学習	65
5.4	文献案内	65
	参考文献	66
<b>6.</b>	<b>テキストの多様性をとらえる分類指標 — 体系化の試み —</b>	<b>70</b>
6.1	はじめに	70
6.2	評定実験に基づく分類指標の策定	72
6.2.1	評定実験1	72
6.2.2	分類指標の策定	73
6.3	分類指標の特徴	73
6.3.1	評定実験2	74
6.3.2	言語特徴量	74
6.3.3	分類指標にかかわる言語的特徴の分析	75
6.4	おわりに	80
	参考文献	82
<b>7.</b>	<b>テキストマイニングと付加的な情報の組合せによるニーズ分析支援</b>	<b>83</b>
7.1	研究の課題と方法	83
7.1.1	研究の課題	83
7.1.2	分析の方法	84
7.2	テキストマイニングによる分析	85
7.2.1	分析対象データと分析手順	85
7.2.2	分析結果	86
7.3	統計データによる分析	92
7.3.1	分析対象データの準備	92
7.3.2	分析対象データからの推測	92
7.4	現地事例による分析	93
7.4.1	事例収集	93
7.4.2	二槽式洗濯機を使う理由	95
7.5	インドネシアでは、今後2、3年、二槽式洗濯機ニーズは継続するか	95
7.6	おわりに	95
	参考文献	96
<b>8.</b>	<b>政治テキストの計量分析</b>	<b>97</b>
8.1	政治テキスト分析の概観	97
8.2	政治テキスト分析の実際	99
8.2.1	語彙多様性に関する研究	99
8.2.2	対外認識に関する研究	101

8.3	おわりに	104
	参考文献	104
<b>9.</b>	<b>ブランドマネジメントにおける自由回答データの活用 — 競争環境の変容を把握 するカテゴリー・トレンドマップ分析 —</b>	<b>107</b>
9.1	ブランドマネジメントとブランド知覚把握の必要性	107
9.2	ブランドマネジメントで用いられる分析手法	109
9.2.1	ブランド認知にかかわる分析	109
9.2.2	ブランドイメージにかかわる分析	109
9.2.3	ブランド認知データへの次元縮約系分析の応用と課題	110
9.3	カテゴリー構造変化とカテゴリー・トレンドマップ	111
9.3.1	理論的な二つの背景	112
9.3.2	カテゴリー・トレンドマップの考え方	113
9.4	カテゴリー・トレンドマップの分析例	115
9.4.1	分析対象としたデータについて	115
9.4.2	カテゴリーの抽出とカテゴリー診断に必要な三つの指標	115
9.4.3	カテゴリー・トレンドマップの解釈	116
9.5	課題と応用可能性	117
	参考文献	117
<b>10.</b>	<b>社会調査における計量テキスト分析の手順と実際 — アンケートの自由回答を中 心に</b>	<b>119</b>
10.1	社会調査におけるテキスト型データの利用	119
10.2	大卒の母親はなぜ「なんとなく進学」に賛成するのか	120
10.3	自由回答データの準備	121
10.4	多く出現した主題	123
10.5	自由回答と他の変数の関連	124
10.6	さらに分析を進めるために	126
	参考文献	128
<b>11.</b>	<b>自然言語処理技術と機械学習手法を用いたテキストマイニング</b>	<b>129</b>
11.1	自然言語処理の諸技術	129
11.1.1	形態素・構文解析	129
11.1.2	固有表現の抽出	131
11.1.3	有用な情報を含む文の取り出し	132
11.1.4	教師あり機械学習手法	133
11.2	自然言語処理技術と機械学習手法を用いたテキストマイニングの事例	136
11.2.1	機械学習を用いた論文要約からの重要情報抽出	136
11.2.2	大規模データからの固有表現と数値表現の抽出と可視化	136
11.2.3	対訳コーパスからの日英対訳表現の抽出	138
11.3	むすび	139

参考文献	139
<b>12. 社会心理学研究におけるテキストマイニング</b>	<b>141</b>
12.1 社会心理学におけるテキストデータの分析	141
12.1.1 新聞記事を対象とする分析	141
12.1.2 その他のテキストデータを対象としたテキストマイニング	142
12.2 研究例① 質問紙調査の自由記述回答の分析	143
12.2.1 研究の目的	144
12.2.2 データの取得	144
12.2.3 テキストマイニングとその後の分析	145
12.2.4 テキストマイニングによって見えてくる心理	147
12.3 研究例② ソーシャルメディアのログ分析	148
12.3.1 先行研究	148
12.3.2 データの取得	149
12.3.3 テキストマイニングとその後の分析	150
12.3.4 テキストマイニングによって見えてきた心理	150
12.4 社会心理学におけるテキストマイニングの学術的意義	152
参考文献	153
<b>13. 類似テキストの異同を計る：新約聖書校訂本の比較研究</b>	<b>155</b>
13.1 ギリシャ語新約聖書近代校訂本	155
13.2 校訂本の基本等計量	156
13.3 語彙の特徴度指数からみる校訂本の特徴	157
13.4 TextTiling に基づく校訂本の比較	161
13.5 おわりに	165
参考文献	165
<b>14. 文章のジャンル判別に寄与する指標の研究 — 専門日本語教育への応用 —</b>	<b>166</b>
14.1 はじめに	166
14.2 研究目的と研究方法	167
14.2.1 研究目的	167
14.2.2 研究方法	168
14.3 分析	169
14.3.1 分析例1—複合辞（接続語句と助詞相当句）を指標とした場合—	169
14.3.2 分析例2—複合動詞の後項動詞を指標とした判別分析—	175
14.3.3 分析例3—「手」を含む動詞慣用句と形容詞慣用句を指標とした判別分 析—	177
14.4 おわりに	179
参考文献	180

<b>15. 文章の難易度と語彙の関連性に関する考察 — 学年の違いを特徴づける語彙的要素とは何か —</b>	<b>181</b>
15.1 研究背景と目的	181
15.2 調査方法とデータについて	182
15.2.1 概要	182
15.2.2 調査方法と予想される結果	182
15.2.3 調査データについて	183
15.2.4 統計分析について	184
15.3 結果	184
15.3.1 調査1の結果	184
15.3.2 調査2の結果	185
15.3.3 調査1, 2の統合と検証	187
15.3.4 BCCWJの教科書データの分析結果	188
15.4 考察	190
15.5 最後に	191
参考文献	192

## 付 録

<b>A.1 テキストデータの収集と分析</b>	<b>193</b>
A.1.1 概要	193
A.1.2 TinyWebCrawler	193
A.1.2.1 概要	193
A.1.2.2 インストール	194
A.1.2.3 使い方	194
A.1.3 TinyTweetCrawler	195
A.1.3.1 概要	195
A.1.3.2 インストール	195
A.1.3.3 使い方	195
A.1.3.4 Twitter Search API について	197
A.1.4 TinyTweetTracker	197
A.1.4.1 インストール	197
A.1.4.2 使い方	197
A.1.5 TinyTextMiner	198
A.1.5.1 インストール	198
A.1.5.2 使い方	199
A.1.5.3 Rによる共起グラフの描画	201
A.1.5.4 英文テキスト	201
A.1.6 既知の不具合	202
参考文献	203

<b>A.2</b>	<b>今日から始めるテキストマイニング — 計量テキスト分析の環境「KH Coder」</b>	<b>204</b>
A.2.1	はじめに	204
A.2.2	三つの主要機能	204
A.2.2.1	自動的に取り出した語の統計分析	204
A.2.2.2	コーディング結果の統計分析	206
A.2.2.3	もとのテキストを確認するための検索や閲覧	207
A.2.3	簡単なマウス操作で誰でも手軽に分析を	207
A.2.4	高度なカスタマイズやプラグイン作成も	208
<b>A.3</b>	<b>今日から始めるテキストマイニング・ツール — MTMineR</b>	<b>210</b>
A.3.1	MTMineRについて	210
A.3.1.1	作成の経緯	210
A.3.1.2	ツールの構成	210
A.3.1.3	MTMineRの起動と終了	211
A.3.2	テキストの処理とデータの集計	211
A.3.2.1	集計モデルと集計結果の形式	211
A.3.2.2	平テキスト	212
A.3.2.3	タグ付きテキスト	215
A.3.3	データ解析とマイニング	219
A.3.3.1	データの標準化	219
A.3.3.2	Zipf法則とグラフ作成	220
A.3.3.3	特徴語句の抽出とグラフ作成	221
A.3.3.4	教師なしの解析法	223
A.3.3.5	教師ありの機械学習法	224
<b>A.4</b>	<b>Rでテキストマイニング — RMeCab パッケージを利用する</b>	<b>226</b>
A.4.1	テキストマイニングの準備	226
A.4.1.1	Rについて	226
A.4.1.2	MeCabについて	227
A.4.1.3	RMeCabのインストール	227
A.4.2	RMeCabの使い方	228
A.4.2.1	動作確認：RMeCabC () 関数と RMeCabText () 関数	228
A.4.2.2	RMeCabFreq () 関数と collocate () 関数	229
A.4.2.3	docMatrix () 関数と docMatrix2 () 関数	231
A.4.2.4	Ngram () 関数と docNgram () 関数	233
A.4.2.5	RMeCabDF () 関数と docMatrixDF () 関数, docNgramDF () 関数	235
A.4.2.6	docDF () 関数	236
A.4.3	応用例	238
A.4.3.1	文体の識別	238
A.4.4	まとめ	239

*x* 目次

参考文献 ..... 239

索引 ..... 240