人文社会ビジネス科学学術院 ビジネス科学研究群 2023年度 春C

テキストマイニング
day 4

R05年度 01KA438, 0ADM126

スケジュール

day 1

- 講義 ― テキストマイニング概説 (津田先生)
- 講義 ― 自然言語処理の最新動向

day 2

- 講義 ― テキストマイニングの手順
- 演習 ― テキスト解析 (1)
- 演習 ― データ理解

day 3

- 演習 ― テキスト解析 (2)
- 講義&演習 ― データ分析 (使い方編)

day 4

- TextMining Studio の紹介
- 講義&講義― データ分析 (実践編)

day 5

- 講義&講義― データ分析 (実践編)

KHCoder の解析・分析手法

KHCoder の特徴

- 「単語と単語」、「カテゴリ(外部変数)と単語」の**関係に注目した分析**が得意
  - 特徴的な単語を見つける
    - 特定の文書に特徴的な単語を見つける → TF・IDF
      → その文書に特に頻出するが、他の文書ではそれほどではない

- **特徴的な関係を見つける**
  - 関係性のある単語と単語と見つける → **共起ネットワーク(Jaccard係数)**
    例)「風呂」と「広い」に関係がありそう
  - 関係性の強い単語と外部変数を見つける → **対応分析(カイ2乗値)**
    例)「レジャー」と「風呂」に関係がありそう

KHCoder で使われるデータ表

- 「文書-抽出語」表
  - 【行】ある文中に出現する単語の数を要素とする文ベクトル
  - 【列】全文中に出現する単語の数を要素とする単語ベクトル

「文書-抽出語」頻度表 (文書のクラスター分析)

h5	bun	部屋	ホテル	風呂	温泉	お部屋	スタッ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイキ	家族	場所	トイレ	子供	ベット	コンヒ	良い	美味し	広い	近い	多い	素晴ら	古い	嬉しい	ない	よい	いい	おいし	宿	駅	気	月	人
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	6	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

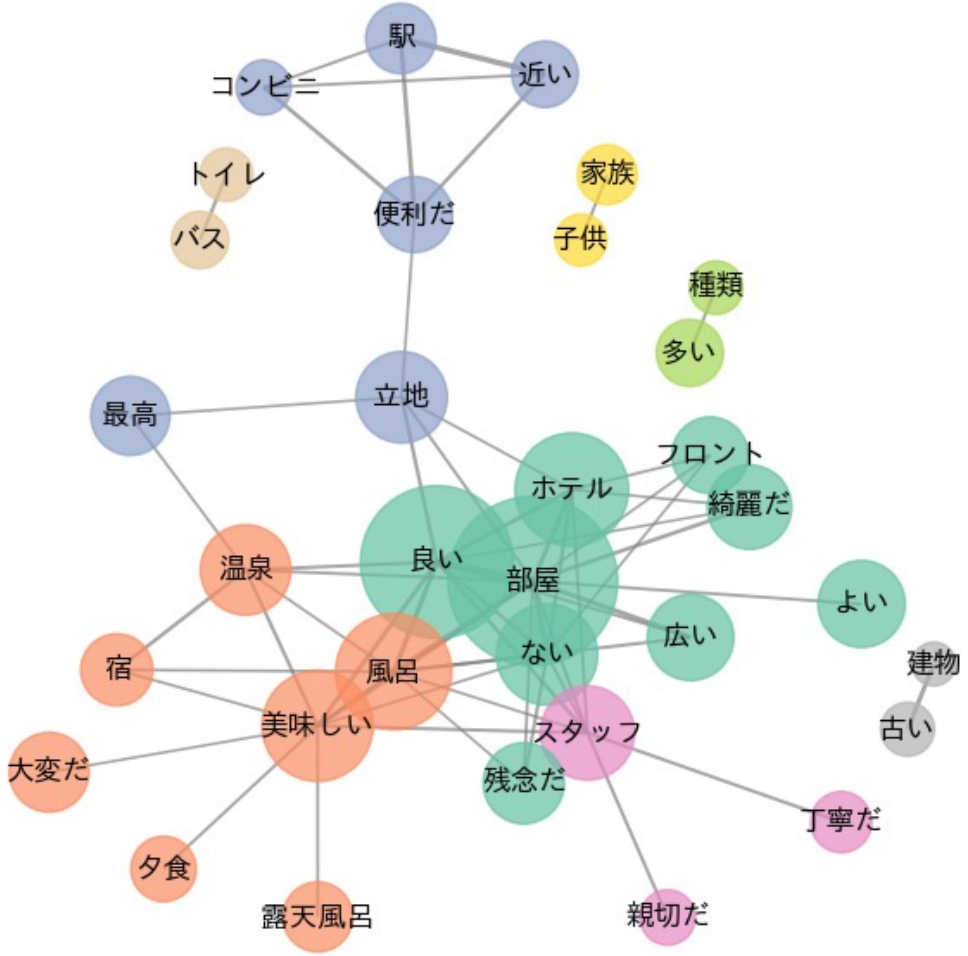
「抽出語-文書」頻度表 (ユークリッド距離、コサイン距離)

h5	1	1	1	1	1	1	1	2	3	3	3	3	4	4	4
bun	1	2	3	4	5	6	7	1	1	2	3	4	1	2	3
id	2	3	4	5	6	7	8	10	12	13	14	15	17	18	19
部屋	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ホテル	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
風呂	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
温泉	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
お部屋	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
スタッフ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
立地	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
フロント	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
最高	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
浴場	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

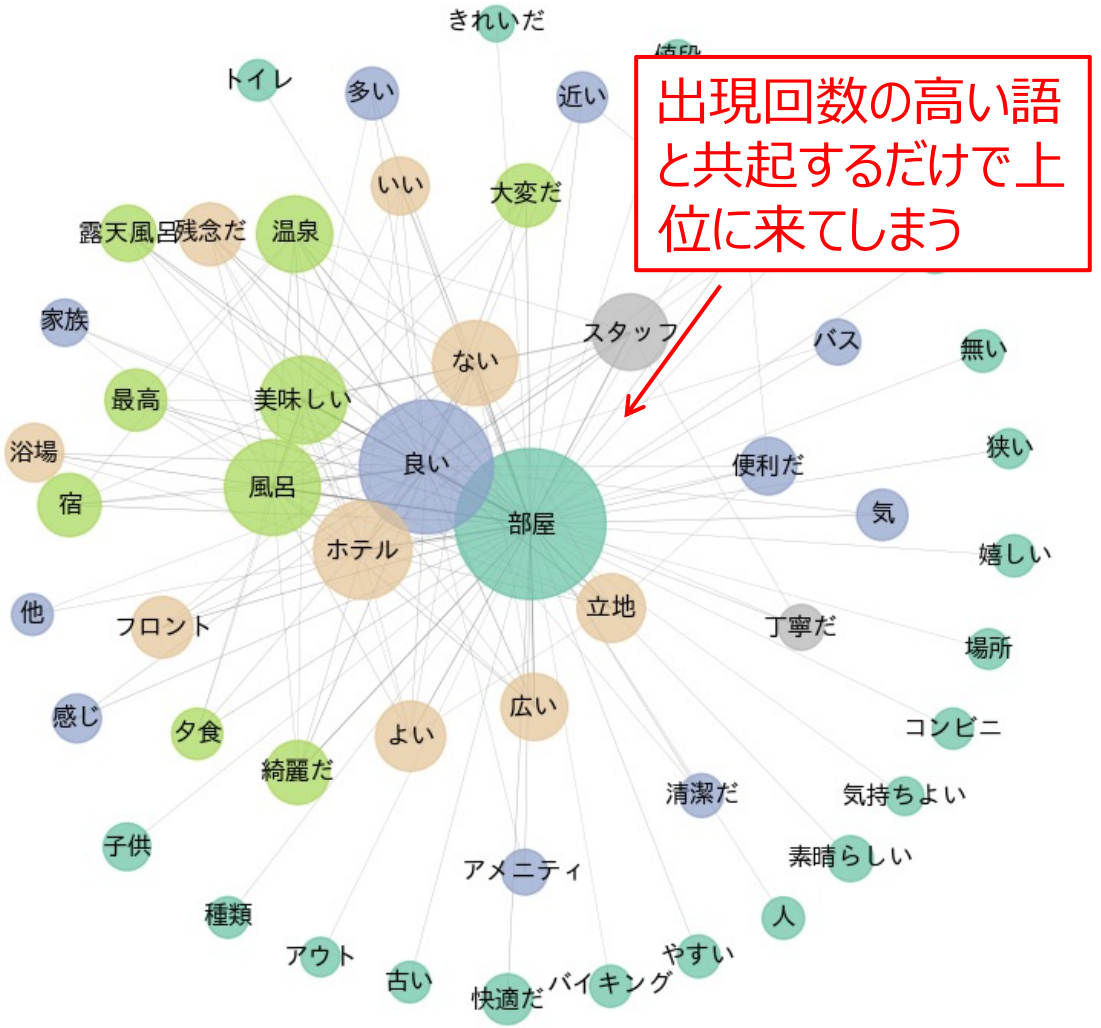
「外部変数-抽出語」クロス集計表 (Jaccard距離、カイ2乗距離)

	部屋	ホテル	風呂	温泉	お部屋	スタッ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイキ
A_レジャー	2723	1157	2113	1657	1095	1014	531	436	691	518	756	788	504	730	326	501
B_ビジネス	2340	1839	668	85	419	455	812	806	222	383	113	19	280	47	438	135
01_登別	541	251	429	280	168	198	49	123	128	119	77	122	81	141	47	162
02_草津	532	290	493	469	236	173	160	81	157	95	308	102	111	186	129	164
03_箱根	621	250	476	301	283	267	65	89	130	136	133	254	132	172	76	79
04_道後	464	284	216	319	120	118	170	104	79	100	73	56	80	78	58	81
05_湯布院	565	82	499	288	288	258	87	39	197	68	165	254	100	153	16	15
06_札幌	503	351	131	24	77	95	168	161	49	95	20	4	56	4	70	38
07_名古屋	454	377	141	14	80	70	135	164	39	71	31	3	47	13	77	29
08_東京	431	350	106	2	91	98	157	151	41	83	10	3	57	9	81	13
09_大阪	472	350	150	24	91	116	176	183	45	83	25	5	56	9	84	29
10_福岡	480	411	140	21	80	76	176	147	48	51	27	4	64	12	126	26

- Jaccard 係数が上位のエッジを残す

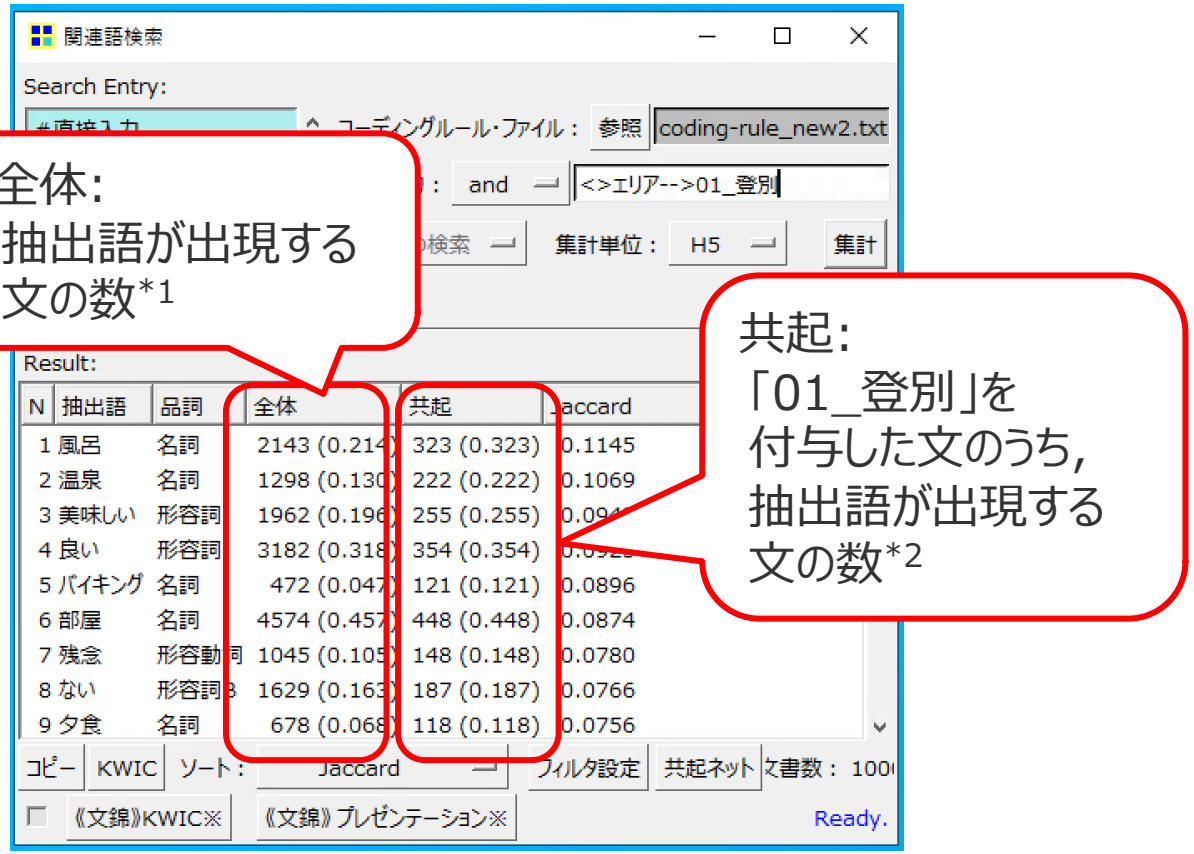


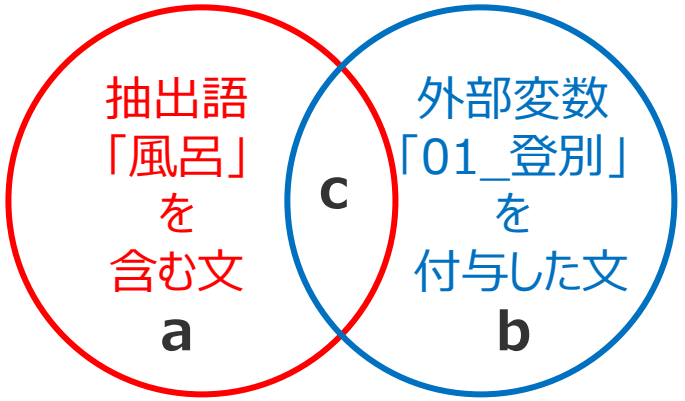
- 共起度の高いエッジを残す



Jaccard 係数

- Jaccard 係数は、共起の強さを測る尺度 (KHCoderで標準的に使用)
  - どちらの語も含まない文書を無視 →言語のようなスパースデータ分析に向いている





$$\text{Jaccard 係数} = \frac{c}{a+b+c}$$

抽出語「部屋」の場合:

c = 323 (“共起”列の値)

a = 2143 (“全体”列の値) − 323 = 1820

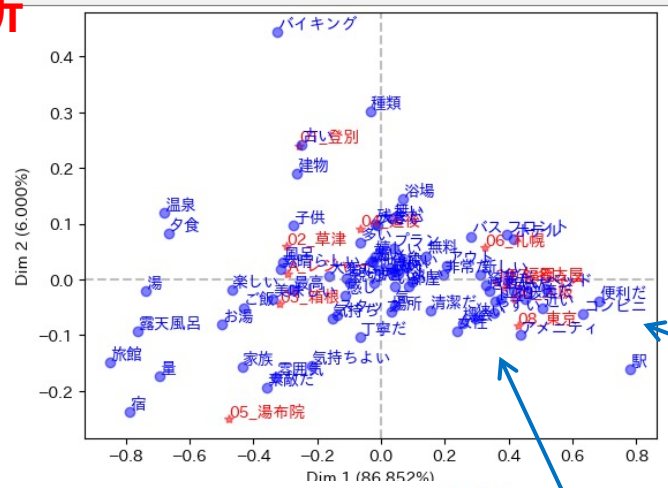
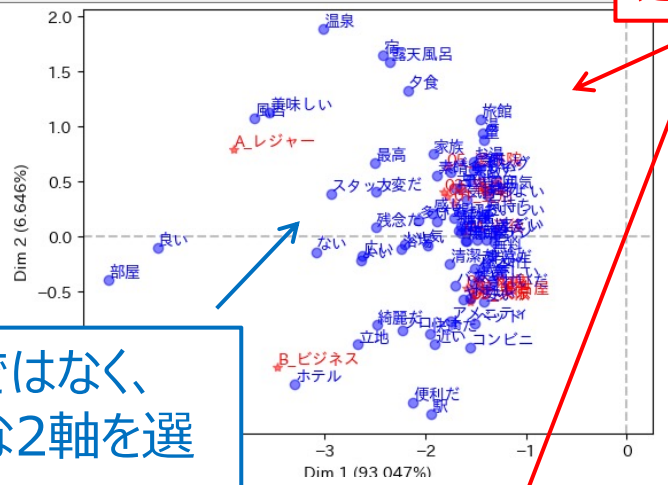
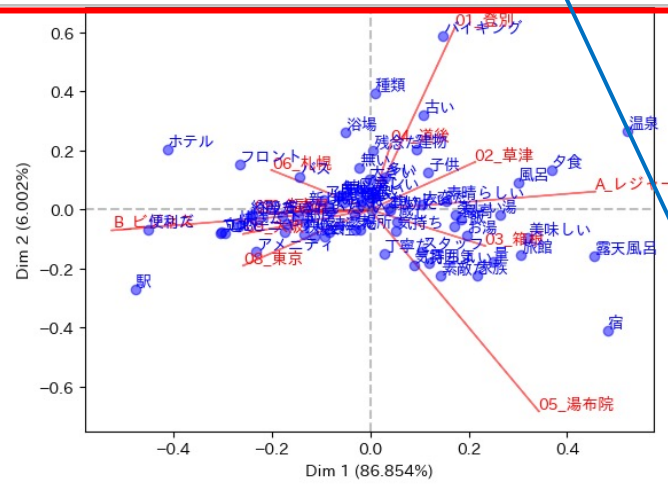
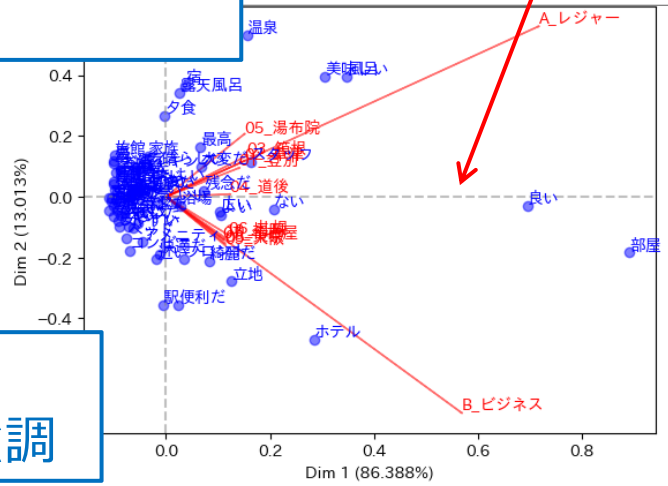
b = (323 / 0.323) − 323 = 677

*1 括弧内はデータ全体に対する割合(前提確率)　*2 括弧内は「10_登別」を付与したデータに対する割合(条件付き確率)

対応分析 (コレスポンデンス分析)

● 対応分析は、カイ2乗値を利用して、関係の強さを強調して表現できる

出現回数の高い語を強調 (原点から遠い位置に布置)

	カイ2乗値	共起頻度
特異値分解		
PCA		

対応分析

- 第2軸の選び方 →特徴的な軸
- 強調するもの →関係の強さ

- 第2軸の選び方 →特徴的な軸
- 強調するもの →出現回数

- 第2軸の選び方 →第1軸と直行
- 強調するもの →関係の強さ

- 第2軸の選び方 →第1軸と直行
- 強調するもの →出現回数

直交性ではなく、特徴的な2軸を選択する

出現頻度より、関係の強さを強調

カイ2乗値

- カイ2乗値は「無関係でない」度合いを測る尺度 → カテゴリと変数間の関連性を測定

$$\text{カイ2乗値} = \frac{\left(\text{観測度数}_{ij} - \text{期待度数}_{ij}\right)^2}{\text{期待度数}_{ij}}$$

「**観測度数**」: カテゴリと変数に従ってクロス集計された度数
「**期待度数**」: 変数が互いに独立している場合に期待される度数
「**観測度数 − 期待度数**」: 実際の度数と独立と期待される度数の差

- カイ2乗値も大きい → カテゴリと変数間の関係が**期待より強い** を示す

クロス集計表 (観測度数)

	A	B	C	D	E	合計
地質学	3	19	39	14	10	85
生物化学	1	2	13	1	12	29
科学	6	25	49	21	29	130
動物学	3	15	41	35	26	120
物理学	10	22	47	9	26	114
工学	3	11	25	15	34	88
微生物学	1	6	14	5	11	37
植物学	0	12	34	17	23	86
統計学	2	5	11	4	7	29
数学	2	11	37	8	20	78
合計	31	128	310	129	198	796

期待度数

	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410	5.996	9.204	37.000
植物学	3.349	13.829	33.492	13.937	21.392	86.000
統計学	1.129	4.663	11.294	4.700	7.214	29.000
数学	3.038	12.543	30.377	12.641	19.402	78.000
合計	31.000	128.000	310.000	129.000	198.000	796.000

観測度数-期待度数

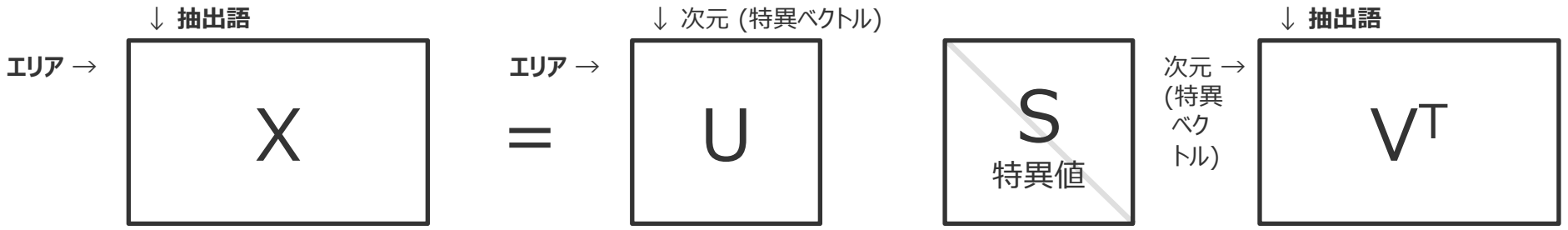
	A	B	C	D	E	合計
地質学	-0.310	5.332	5.897	0.225	-11.143	0.000
生物化学	-0.129	-2.663	1.706	-3.700	4.786	0.000
科学	0.937	4.095	-1.628	-0.068	-3.337	0.000
動物学	-1.673	-4.296	-5.734	15.553	-3.849	0.000
物理学	5.560	3.668	2.603	-9.475	-2.357	0.000
工学	-0.427	-3.151	-9.271	0.739	12.111	0.000
微生物学	-0.441	0.050	-0.410	-0.996	1.796	0.000
植物学	-3.349	-1.829	0.508	3.063	1.608	0.000
統計学	0.871	0.337	-0.294	-0.700	-0.214	0.000
数学	-1.038	-1.543	6.623	-4.641	0.598	0.000
合計	0.000	0.000	0.000	0.000	0.000	0.000

カイ2乗値

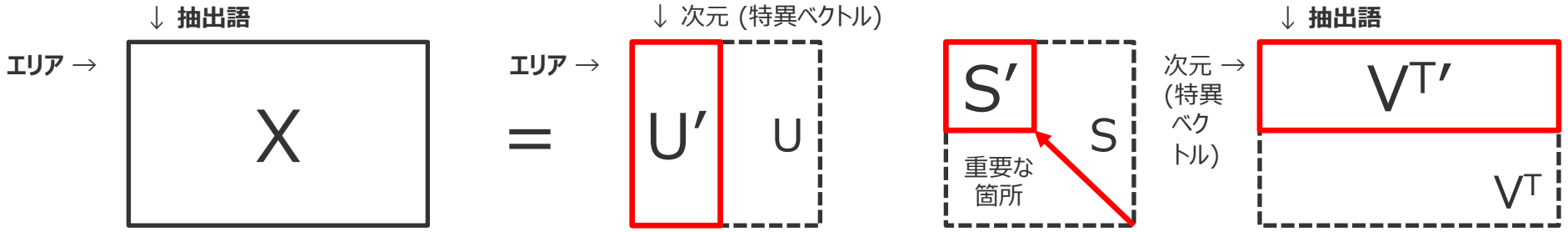
	A	B	C	D	E	合計
地質学	0.029	2.080	1.050	0.004	5.873	9.036
生物化学	0.015	1.521	0.258	2.913	3.176	7.882
科学	0.173	0.802	0.052	0.000	0.344	1.373
動物学	0.599	0.957	0.703	12.438	0.496	15.194
物理学	6.964	0.734	0.153	4.859	0.196	12.906
工学	0.053	0.702	2.508	0.038	6.700	10.001
微生物学	0.135	0.000	0.012	0.166	0.351	0.663
植物学	3.349	0.242	0.008	0.673	0.121	4.393
統計学	0.671	0.024	0.008	0.104	0.006	0.814
数学	0.354	0.190	1.444	1.704	0.018	3.710
合計	12.343	7.252	6.196	22.899	17.282	65.972

特異値分解

- 特異値分解　$X = USV^T$



- S の特異値が小さいものを削る

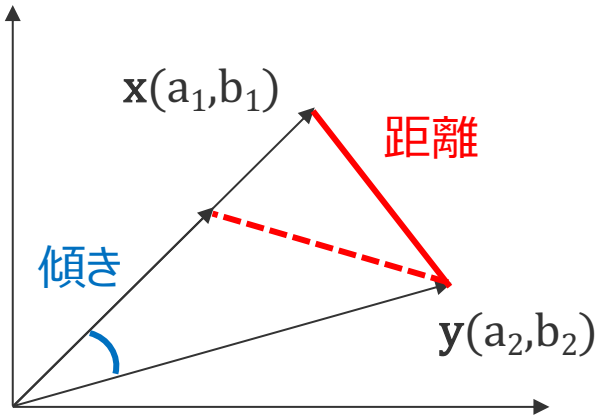


その他の尺度

- **出現パターンが似てる** を測る ＝ ユークリッド距離、コサイン距離
  - 1つひとつの文が長く、各文中での語の出現回数の大小が重要なケースに向く
    (語が1回出現したか、10回出現したかを区別したい)

ユークリッド距離	コサイン距離
サイズ(出現回数の大小)の差までも見る場合向き	傾きが似ているかどうかだけを見る場合向き
$d(\boldsymbol{x},\mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$	$d(\boldsymbol{x},\mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2 \sum {y_i}^2}}$

※ $\mathbf{x}, \mathbf{y}$ はそれぞれの単語ベクトル (単語の出現パターン)



テキスト分析 (実践編)

(再掲) 数値評価で違いを見るのは難しい

【再掲】⑧-a 数値評価の平均 (エリア別x数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	**4.22**	**4.28**	**4.11**	**4.01**	**4.29**	**4.26**	**4.28**
01_登別	4.03	4.27	3.95	3.88	4.31	4.08	4.10
02_草津	4.19	4.28	4.03	3.92	4.31	4.15	4.25
03_箱根	4.22	4.15	4.12	3.97	4.22	4.28	4.23
04_道後	4.16	4.41	4.10	4.00	4.09	4.21	4.26
05_湯布院	4.52	4.28	4.36				4.55
B_ビジネス	**4.00**	**4.34**	**4.10**				**4.19**
06_札幌	3.99	4.37	4.09				4.20
07_名古屋	3.98	4.26	4.06	3.92	3.82		4.16
08_東京	3.97	4.34	4.11	3.91	3.73	3.99	4.14
09_大阪	4.06	4.34	4.14	3.96	3.86	4.12	4.24
10_福岡	4.01	4.40				4.02	4.18

- ユーザーの 8割が 4～5 の評価, 1～2をつけない →本音が見えない
- 同じ点数でもテキストを見れば差異があるかも
- すべての項目に回答する → どこに注目しているかよくわからない

【再掲】⑧-b 数値評価の平均 (カテゴリ別x数値評

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.22	4.28	4.11	4.01	4.29	4.26	4.28
B_ビジネス	4.00	4.34	4.10	3.92	3.82	4.06	4.19

実践的な分析

- 実践1: カテゴリーやエリアごとの**ユーザーの注目ポイント**を押さえる
- 実践2: カテゴリーやエリアごとの**ユーザーの注目ポイント**の**評価の違い**を見つける
- 実践3: 高評価のエリアに倣って、低評価のエリアを**改善するプランを提案**する

  → 注意: プロットによる可視化 と 宿泊客の生の声(原文) を使って解釈する

例) 実践3のまとめ方

対象エリア	エリアX の評価ポイント	エリアY の課題	エリアYの改善案
エリアX:　XXX エリアY:　XXX	• 風呂が広い 　根拠原文: … • …	• エアコンが臭い 　根拠原文: … • …	• … • …

実践1 ― ユーザーの注目ポイントを押さえる

- カテゴリーやエリアごとでの注目する観点の違いを確認する
  - カテゴリー「レジャー」と「ビジネス」を比較する
  - カテゴリー「レジャー」(or「ビジネス」) の 5エリアを比較する
- 手順:
  - カテゴリーやエリアごとの特徴語の違いから,宿泊客が注目する観点を調べる

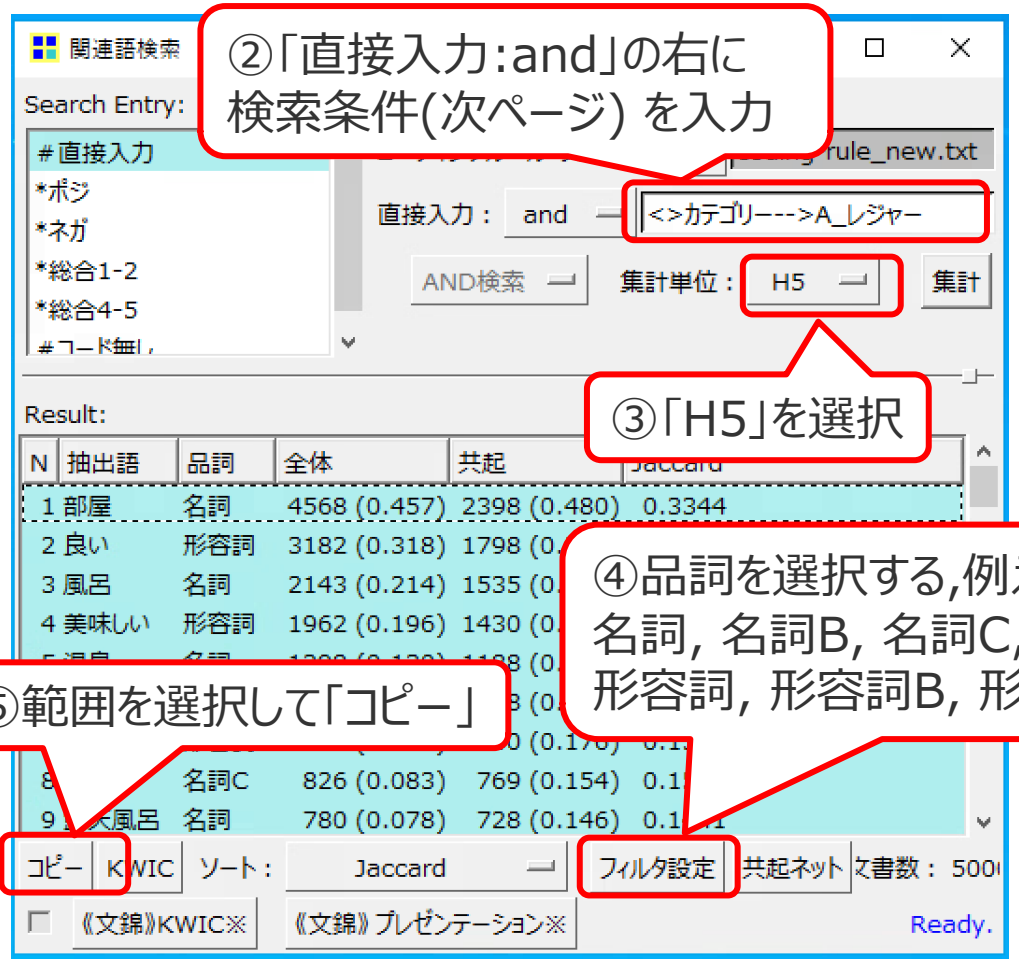
「ツール」→「抽出語」→「関連語検索」→「＃直接入力[and]“<>カテゴリー-->A_レジャー”」「集計単位:H5」→「フィルタ設定」→「品詞＝名詞, 形容動詞, 未知語,タグ,形容詞,名詞B, 形容詞B,名詞C」を選択→「集計」→結果を選択し「コピー」

「ツール」→「抽出語」→「関連語検索」→「＃直接入力[and]“<>エリア-->01_登別”」「集計単位:H5」→「フィルタ設定」→「品詞＝名詞, 形容動詞, 未知語,タグ,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→結果を選択し「コピー」

実践1 ― ユーザーの注目ポイントを押さえる

● カテゴリーやエリアごとの特徴語を抽出する

①メニューから「ツール」「抽出後」「関連語検索」を選ぶ



②「直接入力:and」の右に
検索条件(次ページ) を入力

③「H5」を選択

④品詞を選択する,例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B, 形容動詞, タグ

⑤範囲を選択して「コピー」

⑥EXCEL にペースト

	A	B	C	D	E	F
1	1	部屋	名詞	4568 (0.457)	2398 (0.480)	0.3344
2	2	良い	形容詞	3182 (0.318)	1798 (0.360)	0.2816
3	3	風呂	名詞	2143 (0.214)	1535 (0.307)	0.2737
4	4	美味しい	形容詞	1962 (0.196)	1430 (0.286)	0.2585
5	5	温泉	名詞	1298 (0.130)	1188 (0.238)	0.2325
6	6	スタッフ	名詞	1411 (0.141)	888 (0.178)	0.1608
7	7	ない	形容詞B	1629 (0.163)	880 (0.176)	0.1531
8	8	宿	名詞C	826 (0.083)	769 (0.154)	0.1521
		天風呂	名詞	780 (0.078)	728 (0.146)	0.1441
		高	名詞	992 (0.099)	698 (0.140)	0.1318
		変	形容動詞	1007 (0.101)	652 (0.130)	0.1218
		食	名詞	678 (0.068)	611 (0.122)	0.1206
		い	形容詞	1186 (0.119)	631 (0.126)	0.1136
14	14	残念	形容動詞	1045 (0.105)	609 (0.122)	0.112
15	15	よい	形容詞B	892 (0.089)	495 (0.099)	0.0917

実践1 ― ユーザーの注目ポイントを押さえる

- **直接入力: [and]** の右側に入力する条件式

レジャー:

<>カテゴリー-->A_レジャー

<>エリア-->01_登別
<>エリア-->02_草津
<>エリア-->03_箱根
<>エリア-->04_道後
<>エリア-->05_湯布院

ビジネス:

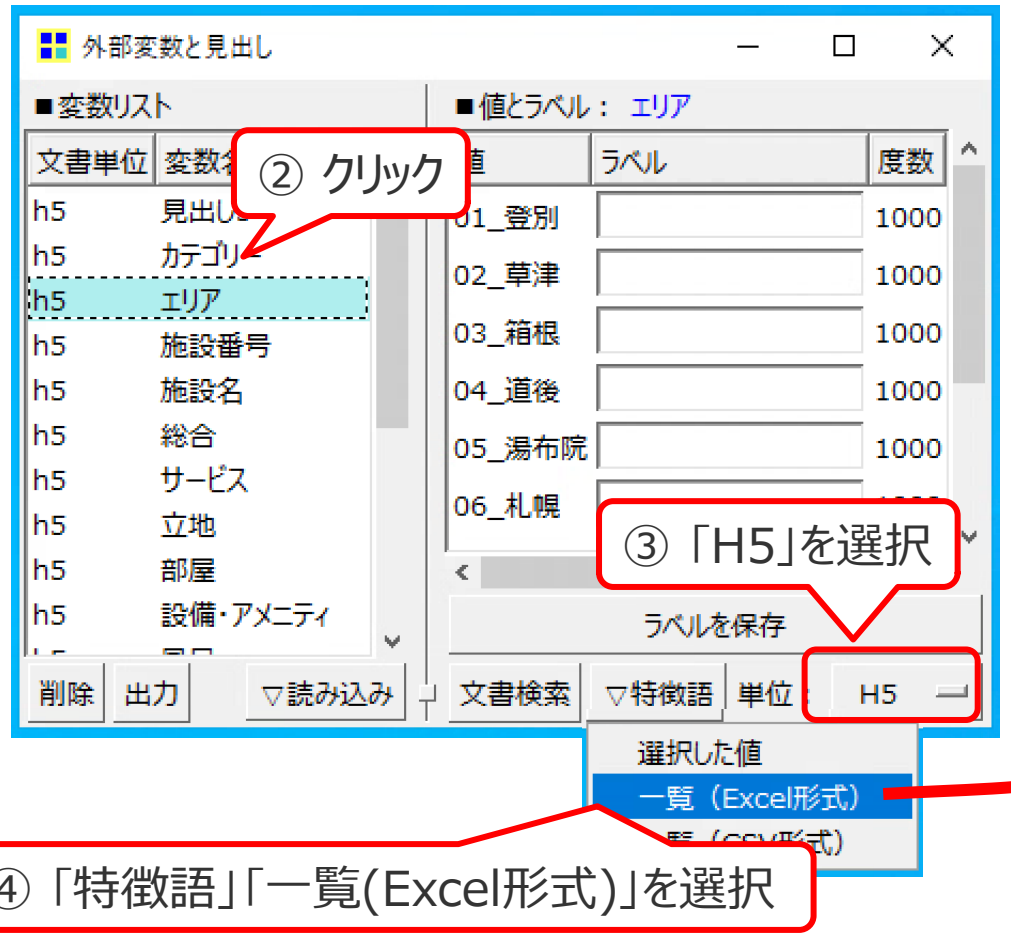
<>カテゴリー-->B_ビジネス

<>エリア-->06_札幌
<>エリア-->07_名古屋
<>エリア-->08_東京
<>エリア-->09_大阪
<>エリア-->10_福岡

実践1 ― ユーザーの注目ポイントを押さえる

● カテゴリーやエリアごとの特徴語を抽出する (一括でEXCELに出力)

①メニューから「ツール」「外部変数と見出し」を開く



④「特徴語」「一覧(Excel形式)」を選択

	A	B	C	D	E	F	G	H	I	J	K
1											
2	01_登別			02_草津			03_箱根			04_道後	
3	食事	.134		湯畑	.327		食事	.159		道後	.128
4	風呂	.115		草津	.171		美味しい	.136		温泉	.109
5	温泉	.107		温泉	.136		露天風呂	.134		松山	.098
6	美味しい	.094		食事	.129		風呂	.116		朝食	.092
7	良い	.093		風呂	.126		部屋	.109		立地	.082
8	満足	.090		宿	.120		良い	.106		最高	.066
9	バイキング	.090		美味しい	.102		温泉	.102		広い	.063
10	宿泊	.088		良い	.100		思う	.100		浴場	.059
11	思う	.082		部屋	.096		料理	.100		駐車	.057
12	料理	.082		思う	.096		宿	.097		フロント	.057
13	05_湯布院			06_札幌			07_名古屋			08_東京	
14	宿	.180		札幌	.158		名古屋	.150		駅	.102
15	食事	.159		朝食	.099		朝食	.097		利用	.087
16	美味しい	.144		ホテル	.092		ホテル	.086		ホテル	.086
17	露天風呂	.135		利用	.085		利用	.083		便利	.078
18	風呂	.127		立地	.077		便利	.072		立地	.077
19	温泉	.124		便利	.077		駅	.070		東京	.072
20	料理	.123		綺麗	.071		綺麗	.069		近い	.071
21	最高	.114		浴場	.070		フロント	.066		朝食	.070
22	スタッフ	.110		対応	.066		立地	.065		綺麗	.064
23	満足	.107		フロント	.065		近い	.059		快適	.063
24	09_大阪			10_福岡							
25	大阪	.111		博多	.126						
26	ホテル	.108		ホテル	.090						
27	利用	.097		便利	.087						
28	駅	.096		利用	.085						
29	便利	.080		立地							
30	立地	.074		朝食							
31	綺麗	.072		福岡							
32	フロント	.067		近く							
33	快適	.064		駅							
34	広い	.064		フロン							

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

実践1 ― ユーザーの注目ポイントを押さえる

- 数値評価ではすべての項目に回答
→ レジャーとビジネスでは注目する項目にかなり偏りがありそう

● 特徴語の抽出結果の例

A_レジャー	
部屋	.334
良い	.282
風呂	.274
美味しい	.259
温泉	.233
スタッフ	.161
ない	.153
宿	.152
露天風呂	.144
最高	.132

数値評価指標

風呂
部屋
食事
サービス
設備
~~立地~~

01_登別		02_草津		03_箱根		04_道後		05_湯布院	
風呂	.115	湯畑	.327	美味しい	.136	温泉	.109	宿	.180
温泉	.107	温泉	.136	露天風呂	.134	立地	.082	美味しい	.144
美味しい	.094	風呂	.126	風呂	.116	最高	.066	露天風呂	.135
良い	.093	宿	.120	部屋	.109	広い	.063	風呂	.127
バイキング	.090	美味しい	.102	良い	.106	浴場	.059	温泉	.124
残念	.078	良い	.100	温泉	.102	よい	.058	最高	.114
ない	.077	部屋	.096	宿	.097	フロント	.057	スタッフ	.110
夕食	.076	最高	.090	スタッフ	.096	大変	.057	家族	.104
種類	.075	夕食	.085	夕食	.095	夕食	.055	部屋	.098
露天風呂	.074	ない	.074	ない	.083	便利	.055	良い	.097

B_ビジネス	
ホテル	.223
立地	.147
便利	.134
駅	.124
綺麗	.123
フロント	.107
近い	.091
快適	.090
アメニティ	.072
コンビニ	.069

数値評価指標

~~風呂~~
部屋
~~食事~~
サービス
設備
立地

06_札幌		07_名古屋		08_東京		09_大阪		10_福岡	
ホテル	.092	ホテル	.086	駅	.102	ホテル	.108	ホテル	.090
立地	.077	便利	.072	ホテル	.086	駅	.096	便利	.087
便利	.077	駅	.070	便利	.078	便利	.080	立地	.082
綺麗	.071	綺麗	.069	立地	.077	立地	.074	駅	.063
浴場	.070	フロント	.066	近い	.071	綺麗	.072	フロント	.063
フロント	.065	立地	.065	綺麗	.064	フロント	.067	綺麗	.060
広い	.063	近い	.059	快適	.063	快適	.064	トイレ	.053
快適	.056	アメニティ	.056	コンビニ	.059	広い	.064	コンビニ	.053
駅	.056	快適	.055	フロント	.055	近い	.064	よい	.051
ベッド	.055	コンビニ	.051	アメニティ	.052	アメニティ	.054	快適	.051

Tips: 「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=カテゴリー」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→「フィルタ設定」→「品詞=名詞,形容動詞,未知語,タグ,形容詞,名詞B,形容詞B,名詞C」を選択→「▽特徴語」→「一覧(EXCEL形式)」で連続実行

(参考) 表記ゆれを吸収する方法 (1/2)

目的: 同じ意味の単語を同一視する別の単語として扱わない

例) 「部屋」「お部屋」の 2単語 → どちらも「部屋」としてカウント

方法: 「表記揺れを吸収」プラグインを利用する

手順: (出典 https://github.com/ko-ichi-h/khcoder/issues/101)

1. プラグインをダウンロードし、解凍して plugin_jp 配下へコピー

ダウンロードURL	https://github.com/ko-ichi-h/khcoder/files/4809463/z1_edit_words3.zip
解凍後ファイル名	z1_edit_words3.zip → z1_edit_words3.pm
配置後のパス	**khcoder3¥plugin_jp¥**z1_edit_words3.pm

(参考) 表記ゆれを吸収する方法 (2/2)

手順 (続き):

2. プラグインファイル「**z1_edit_words3.pm**」を編集する

編集前

```
package z1_edit_words3;
use utf8;

my $config = {
    '友達' =>
        [
            '友人',
            '旧友',
            '親友',
            '盟友',
            '友',
        ],
    '格別' =>
        [
            '特別',
            '格別', # 通常
        ],          # の
    '偶然' =>
        [
            '偶然', # 形容
        ],
};
```

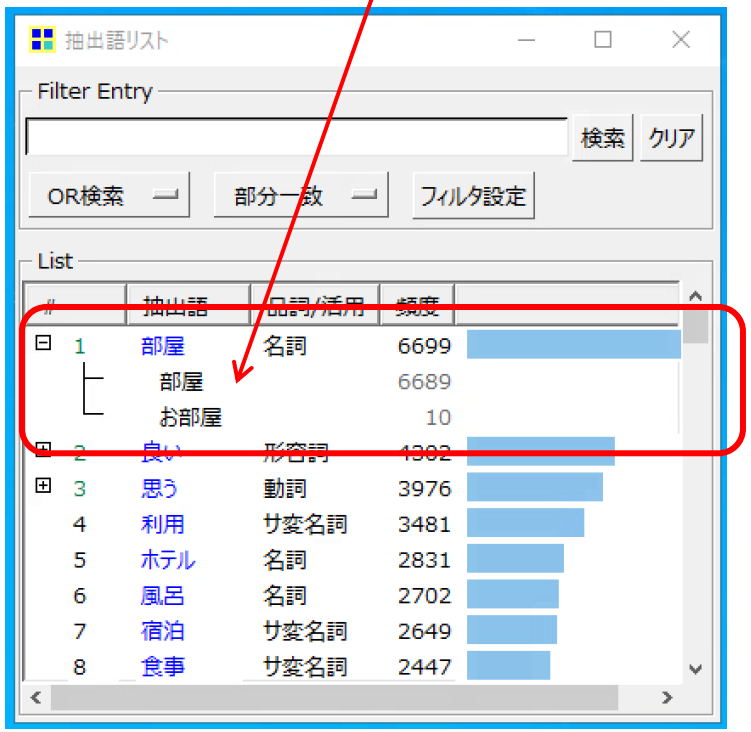
→

編集後

```
package z1_edit_words3;
use utf8;

my $config = {
    '部屋' =>
        [
            'お部屋',
        ],
};
```

適用後の例:
「部屋」と「お部屋」がひとつの単語にまとまっている



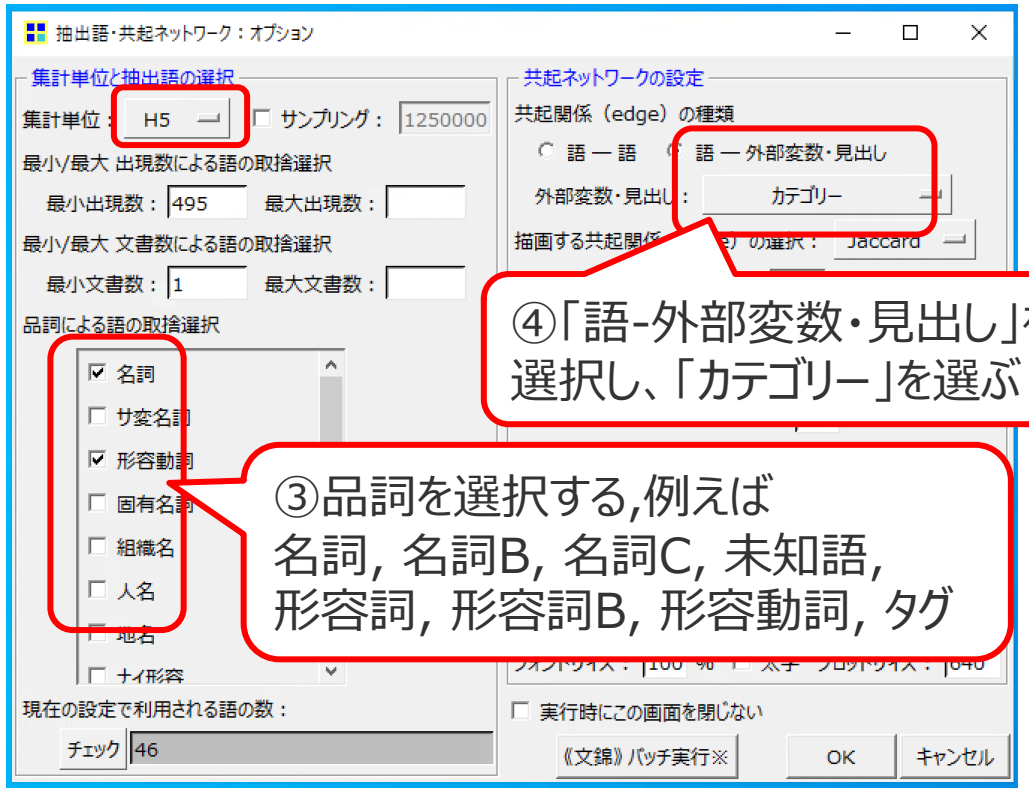
3. KH Coder を再起動する
4. プロジェクトファイルを開く
5. メニューから「ツール」「プラグイン」「表記ゆれの吸収」を選ぶ

実践1 ― ユーザーの注目ポイントを押さえる

● 共起ネットワークを使う (カテゴリー)

①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

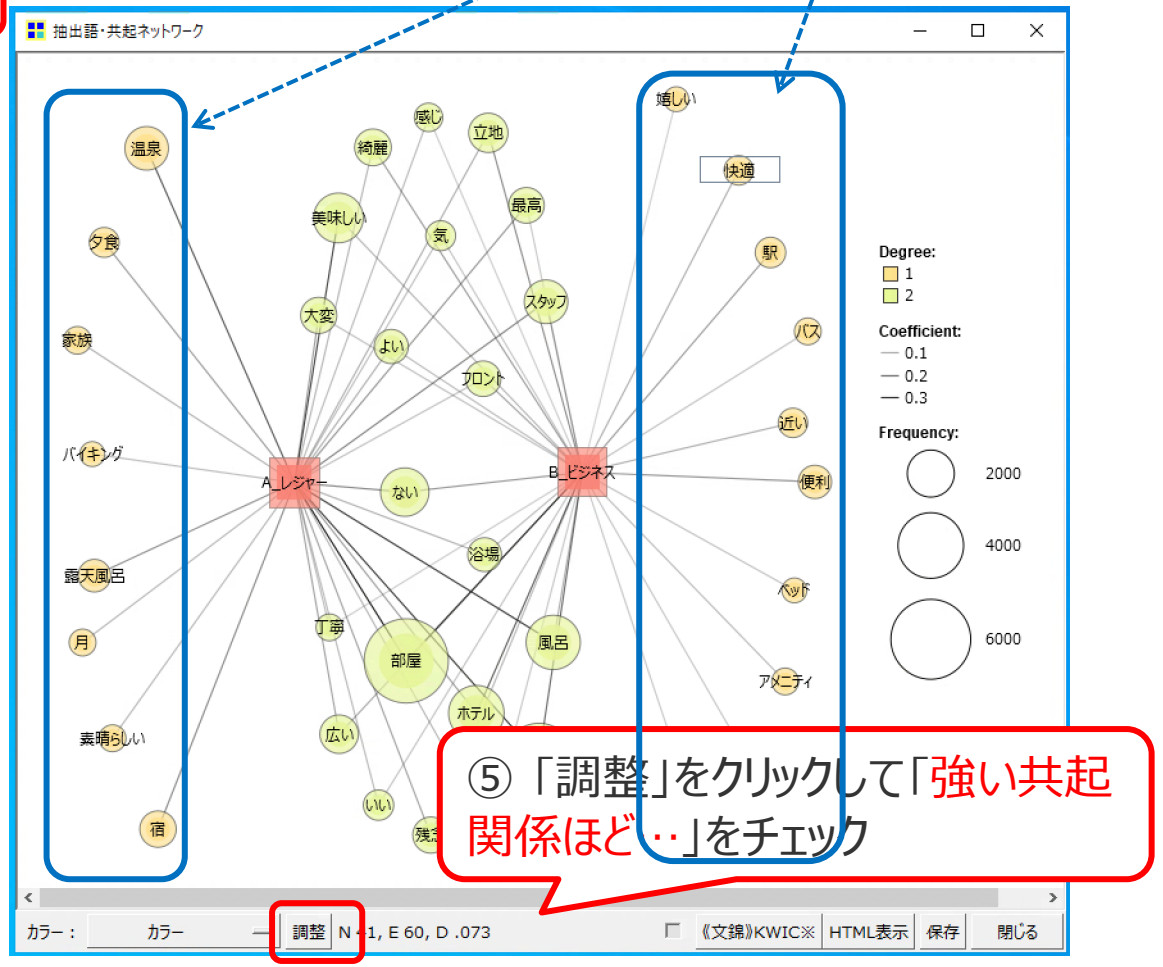
②「集計単位」として「H5」を選んで「OK」をクリック



④「語-外部変数・見出し」を選択し、「カテゴリー」を選ぶ

③品詞を選択する,例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B, 形容動詞, タグ

レジャーとビジネスで注目している観点が異なる



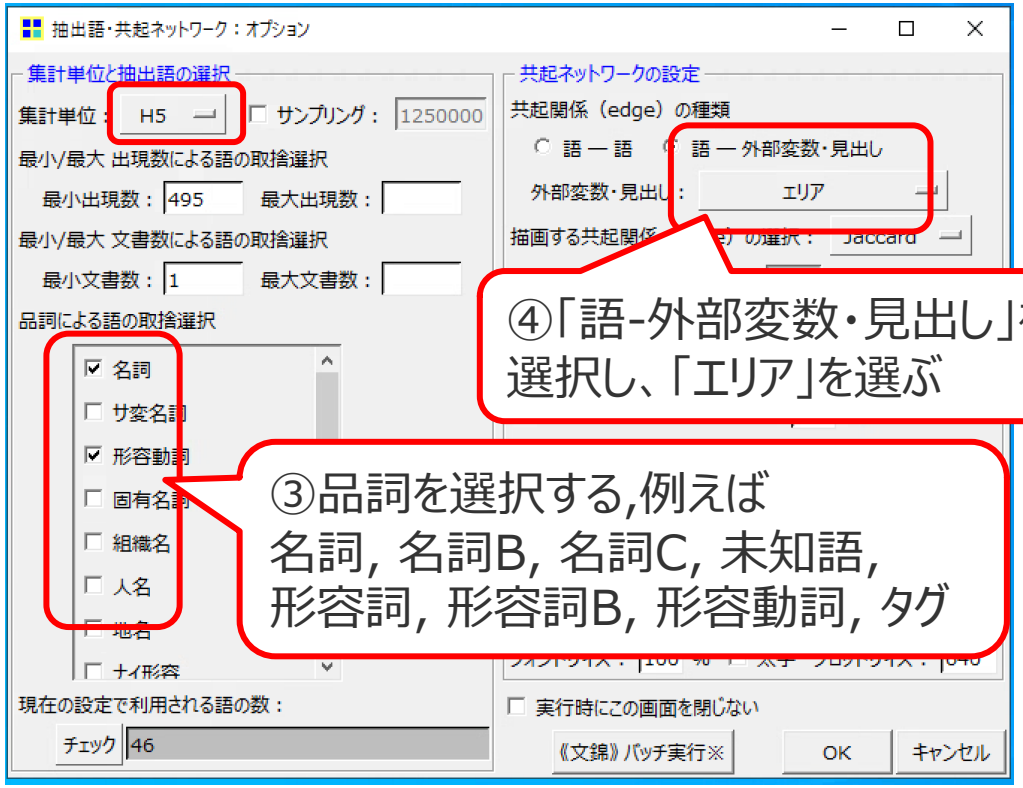
⑤「調整」をクリックして「強い共起関係ほど‥」をチェック

実践1 ― ユーザーの注目ポイントを押さえる

● 共起ネットワークを使う (エリア)

①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

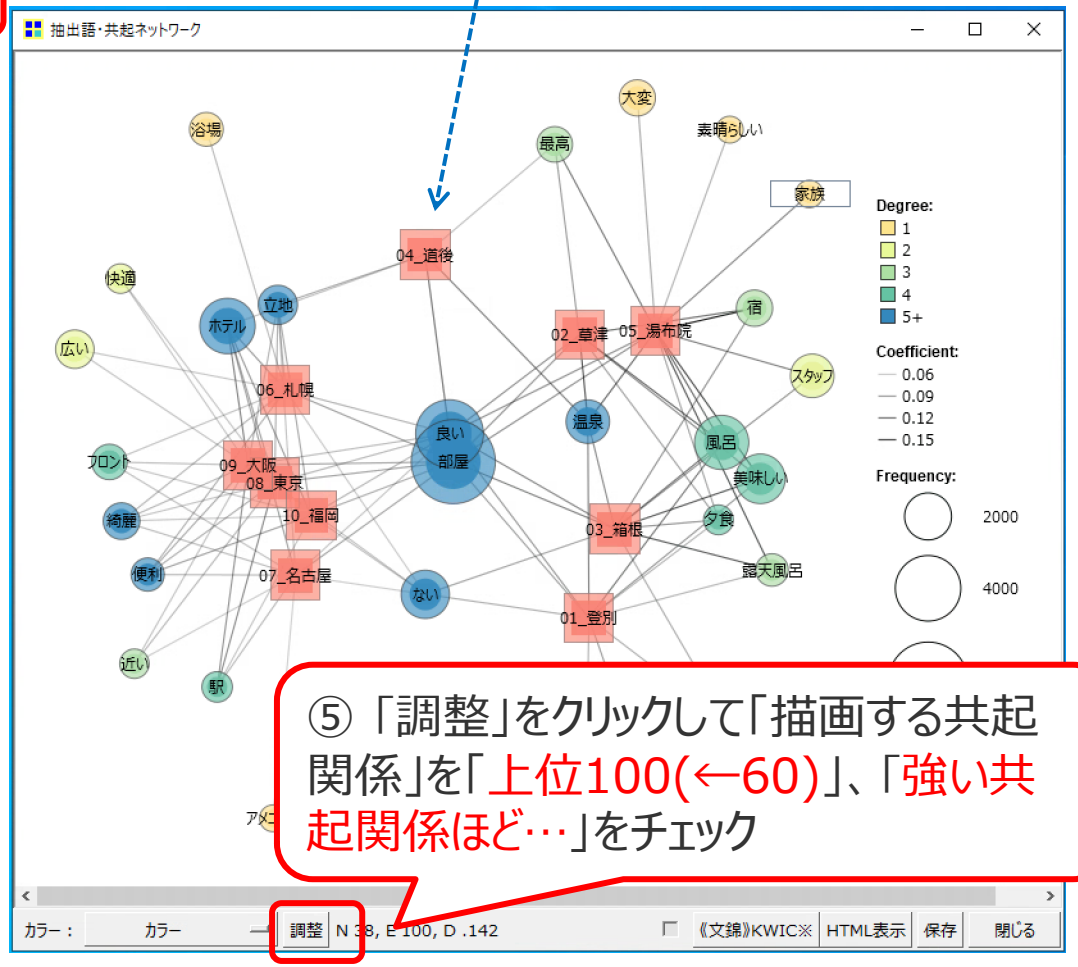
②「集計単位」として「H5」を選んで「OK」をクリック



③品詞を選択する,例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B, 形容動詞, タグ

④「語-外部変数・見出し」を選択し、「エリア」を選ぶ

- 特徴語抽出と似た傾向が確認できる
- 道後はレジャーとビジネスの中間的な位置付け

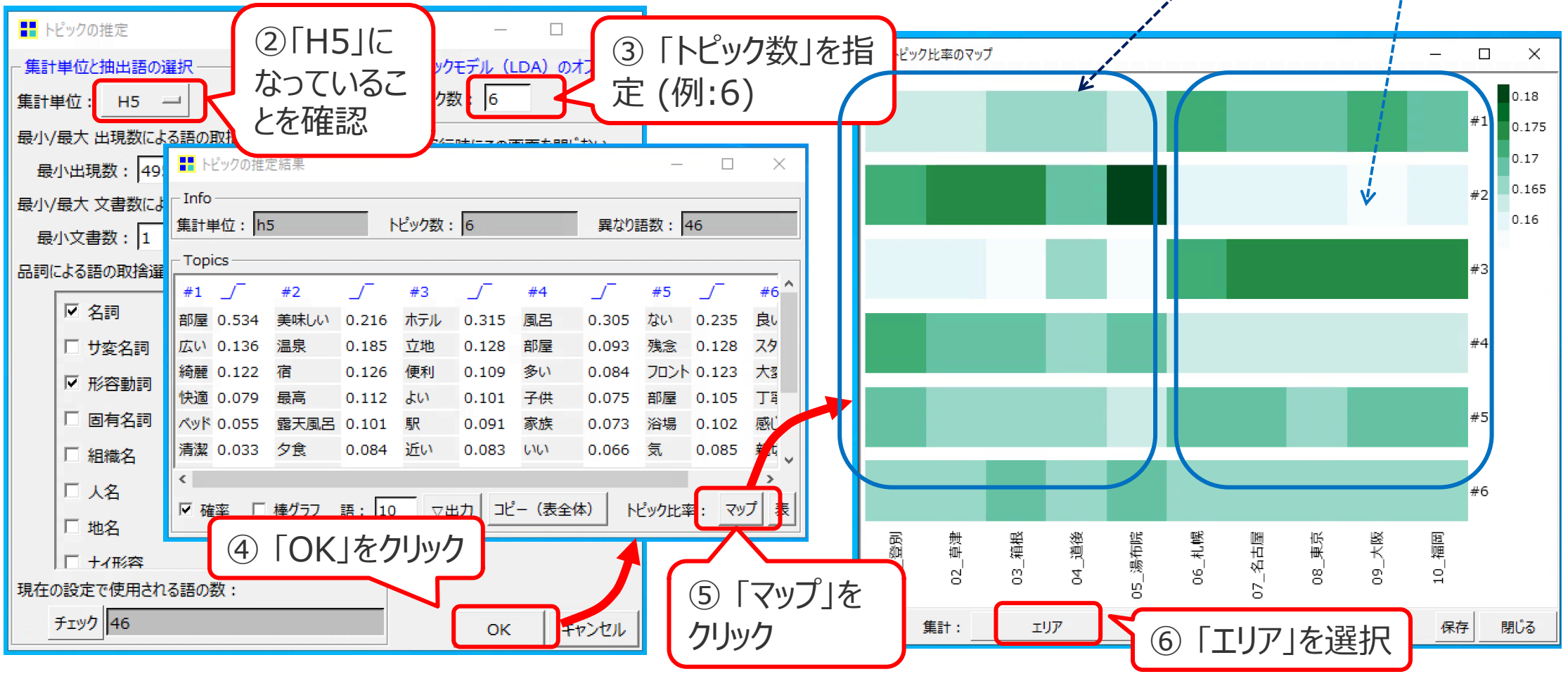


⑤「調整」をクリックして「描画する共起関係」を「上位100(←60)」、「強い共起関係ほど…」をチェック

実践1 ― ユーザーの注目ポイントを押さえる

● トピックモデルを使う (エリア)

①メニューから「ツール」「文書」「トピックモデル」「トピックの推定」を選ぶ



day 4 – レポート課題

- 以下の課題 A と B 両方について PDF ファイルで提出 してください

A. 演習用のデータを用いて、以下の2つのネットワーク図を作成し、キャプチャーを撮るとともに、2つのネットワーク図の違いを観察し、違いについて考察してください

1. KHCoder の共起ネットワーク図
2. TextMining Studio のことばネットワーク図

B. 今後、テキストマイニングを使ってご自身で分析したいデータやテーマ(目的)を挙げてください (1件以上)

※ 何らかの事情で A. の2つ以上のツールが試せない場合、本日の講義の感想を文章で記述してください

レポート形式	提出先	期限
PDF	manaba	次回～18:20

Q&A

参考資料

- KH Coder
  - 樋口耕一. 社会調査のための計量テキスト分析―内容分析の継承と発展を目指して【第2版】KH Coder オフィシャルブック. ナカニシヤ出版, 2020.
  - 樋口耕一. テキスト型データの計量的分析 ―2つのアプローチの峻別と統合―. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.
  - 牛澤賢二. やってみよう テキストマイニング ―自由回答アンケートの分析に挑戦!. 朝倉書店, 2019
  - 樋口耕一. 動かして学ぶ! はじめてのテキストマイニング: フリー・ソフトウェアを用いた自由記述の計量テキスト分析 KH Coder オフィシャルブック II.ナカニシヤ出版, 2022.
- Windows環境によるデータ収集方法の参考
  - テキストマイニングソフトを利用した新未来洞察手法の研究. 第10分科会, (財)市場創造研究会. [発表スライド]
- R を使った参考書
  - 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
  - 石田基広. "RMeCab によるテキスト解析. R によるテキストマイニング入門." 森北出版, 2008, 51-82.
- 他のツールを使った参考書
  - 那須川哲哉. "テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法." 東京電機大学出版局, 2006.
  - 上田隆穂, 黒岩祥太, 戸谷圭子. "テキストマイニングによるマーケティング調査." 講談社, 2005.
- 統計解析を中心とした参考書
  - 前田忠彦; 山崎誠. 言語研究のための統計入門. くろしお出版株式会社, 東京, 2013.