

テキストマイニングの実践

—3日目—

2021/7/9

人文社会ビジネス科学学術院
ビジネス科学研究群

講義スライド

※ manaba にも掲載しています

- <https://github.com/haradatm/lecture/tree/master/gssm-202107>



スケジュール

- 1日目: 6/25(金)
 - 説明 — テキストマイニングの手順
 - 説明 — データをよく知る (Excel)
 - 2日目: 7/2(金)
 - 説明 — テキストマイニングツールの使い方 (KHCoder)
 - 3日目: 7/9(金)
 - 説明 — データ分析の実践 (KHCoder)
 - **実習** — データ分析の実践 (KHCoder)
 - 4日目: 7/16(金)
 - Text Mining Studio 利用体験
 - **実習** — データ分析の実践 (KHCoder)
 - 体育の日: 7/23(金)
 - 5日目: 7/30(金)
 - **発表** — データ分析の実践 (KHCoder)
- ※ お知らせ ※
- 3日目, 4日目後半, 5日目** は, Zoom のブレイクアウトルーム機能を使ったグループワークになります。

(再掲) 前回の課題

- ・コーディングルール「coding-rule.txt」中の「風呂1-2」「風呂4-5」を参考に「総合1-2」「総合4-5」のルールを定義したコーディングルールを作成してください

サンプル:

https://github.com/haradatm/lecture/blob/master/gssm-202107/03-samples/coding-rule_new.txt.zip

- ・前ページで紹介したクロス集計を行い,作成したプロットを **PDF** ファイルで提出してください
- ・形式: PDF, 提出先: manaba, 期限: 次回 7/9 21:00

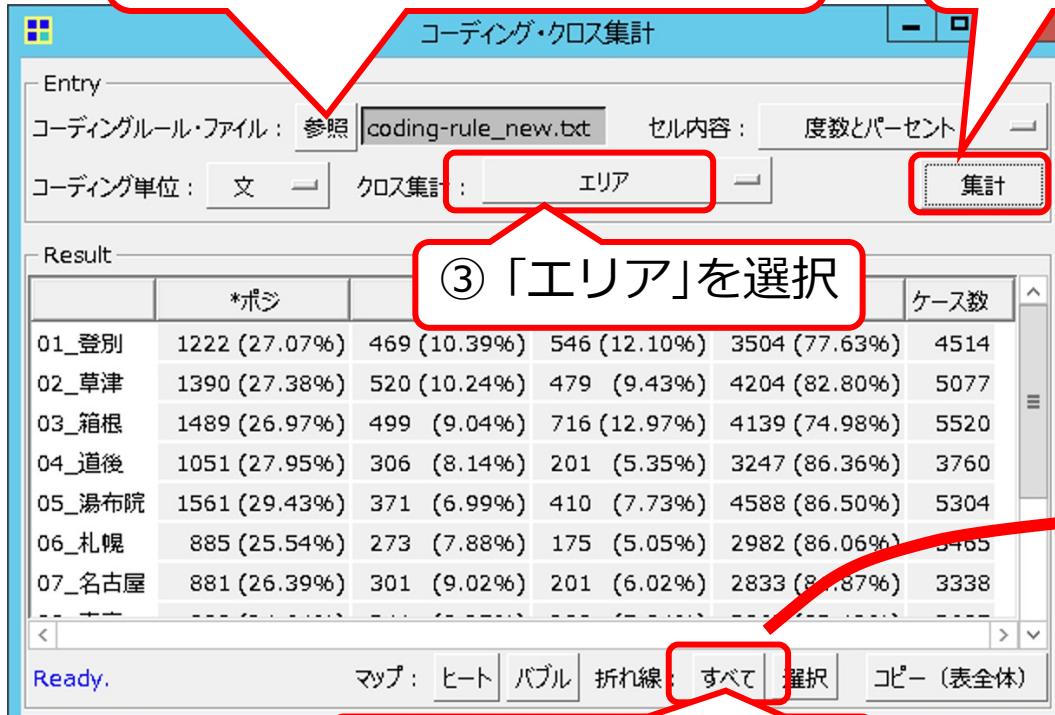
前回の課題 — 正解例

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

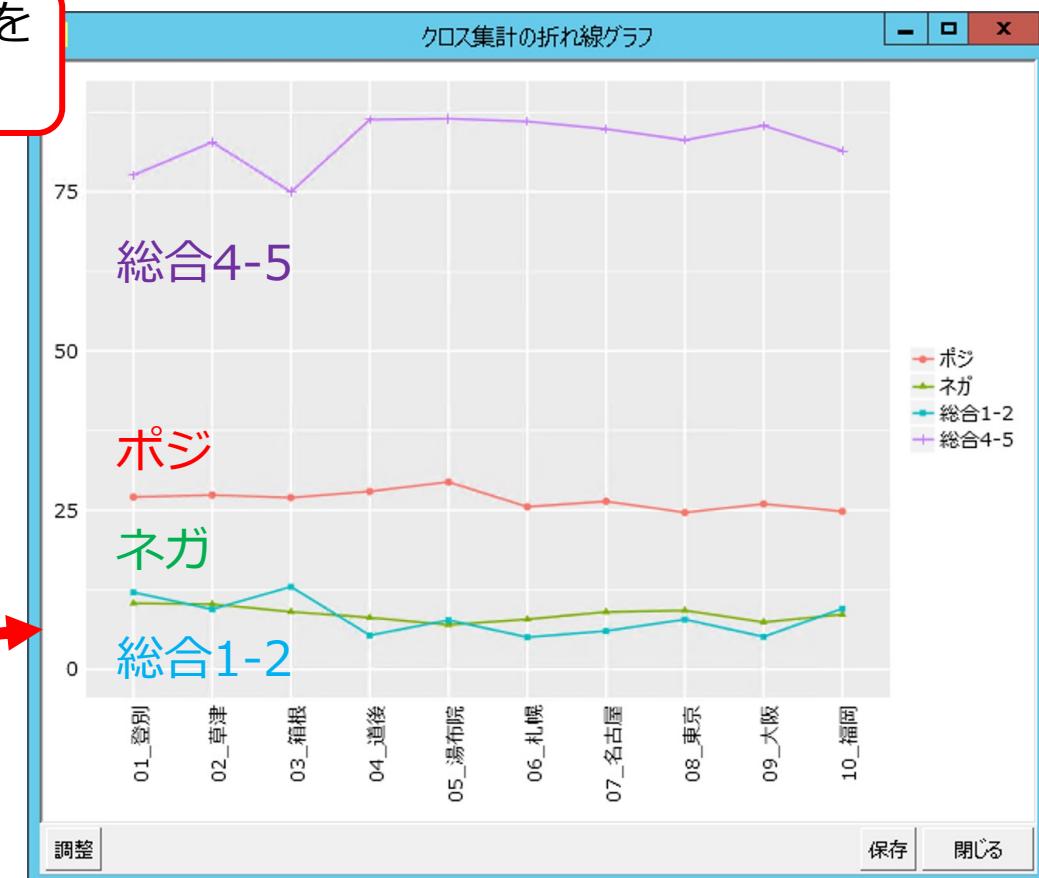
②「参照」をクリックして
「coding-rule_new.txt」を開く

④「集計」を
クリック

③「エリア」を選択



⑤「すべて」をクリック



よくある質問

- Q1. 単語登録したいときは?
- Q2. 表記ゆれ(or 同義語)を統一したいときは?
- Q3. 対応分析の軸って何? (対応分析)
- Q4. Jaccard係数って何? (関連語検索)
- Q5. カイ2乗値で分かるとは何? (クロス集計)

Q1. 単語登録したいときは?

- 目的
 - 複数の単語に分かれる → 1単語として抽出できるようにする
例) 「湯」「畠」の2単語 → 「湯畠」として1単語
- 方法
 - 「前処理の実行」前に「強制出力する語の指定」に追加する
- 手順
 1. メニューから「前処理」「語の取捨選択」を選ぶ
 - 「強制出力する語の指定」欄に抽出したい単語を登録する
 - 「OK」ボタンで画面を閉じる
 2. メニューから「前処理」「前処理の実行」を選ぶ

Q2. 表記ゆれを統一したいときは? (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) 「お湯」 「湯」 の 2単語 → どちらも「お湯」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
 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
- (次ページにつづく)

Q2. 表記ゆれを統一したいときは? (2/2)

- 手順

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

→

編集前	編集後
<pre>1 package z1_edit_words3; 2 use utf8; 3 4 my \$config = { 5 '友達' => 6 [7 '友人', 8 '旧友', 9 '親友', 10 '盟友', 11 '友', 12], 13 '格別' => 14 [15 '特別', 16 '格別', # 通常 17], # の 18 '偶然' => 19 [20 '偶然', # 形容 21], 22 }; 23</pre>	<pre>1 package z1_edit_words3; 2 use utf8; 3 4 my \$config = { 5 'お湯' => 6 [7 '湯', 8], 9 }; 10</pre>

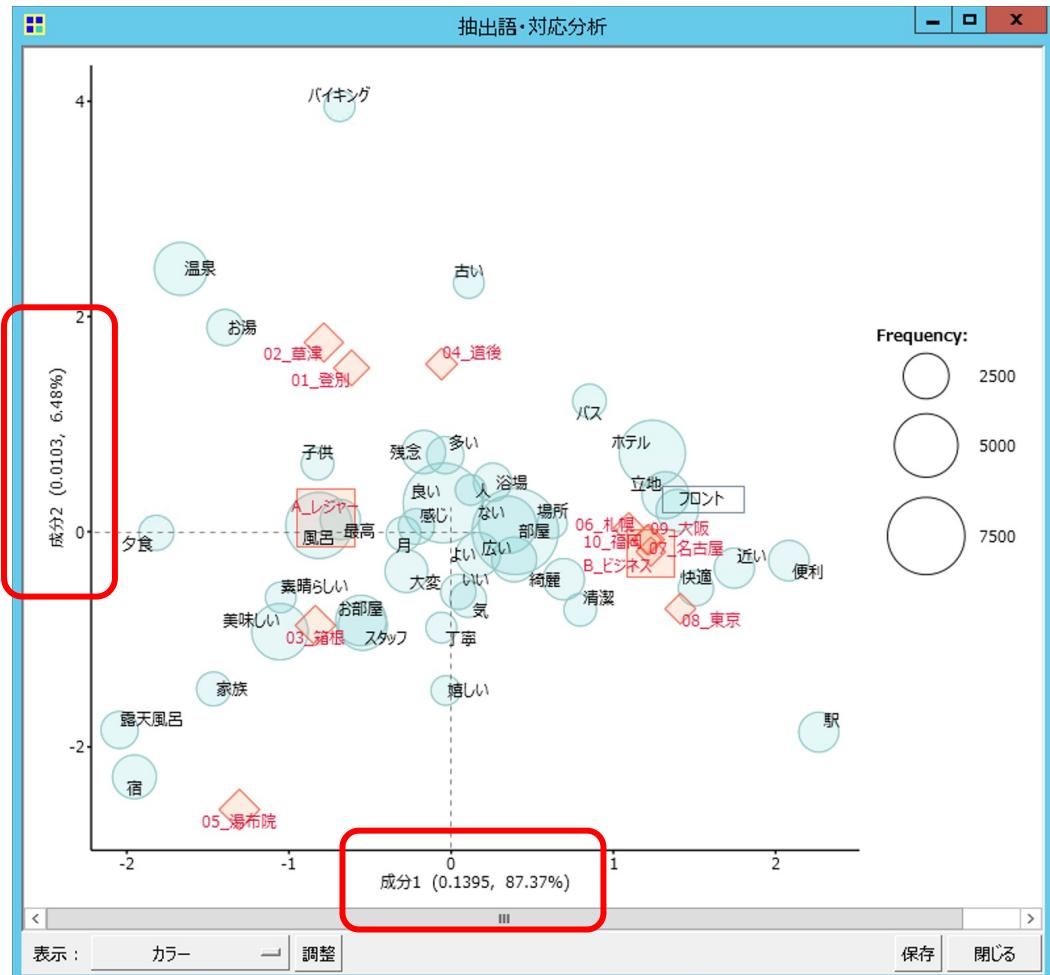
- ↓
3. KH Coder を再起動する
 4. プロジェクトファイルを開く
 5. メニューから「ツール」「プラグイン」「表記ゆれの吸収」を選ぶ
 6. 分析を続ける

適用後の例 →

「お湯」と「湯」が
ひとつの単語にまと
まっている

#	抽出語	品詞/活用	頻度
1	お湯	名詞	779
2	湯		426
3	お湯		353

Q3. 対応分析の軸って何?



- KHCoder の対応分析は R の MASS パッケージにある corresp 関数を使用
- 軸ラベルの数値は、固有値および寄与率を示す
- 左図の場合、第2固有値までの累積寄与率は 93.85% で非常に高い
→ 第1,2 固有値に対応する軸のみを分析すればよい
- 寄与率が高い固有値に対応する行や列の得点の大小とその相対関係について分析する

KHCoder の仕組み:

文の出現パターンと単語の出現パターン

【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)

h5	bun	部屋	ホテル	風呂	温泉	お部屋	スタッフ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイク	家族	場所	トイレ	子供	ペット	コンビ	良い
1	1	1	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	
1	3	0	0	0	0	0	0	0	0	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	1	
1	5	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	
1	7	0	0	0	0	0	0	0	0	0	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	
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	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	
3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

(参考) KH Coder で使われるデータ表

「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	スタッフ	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	コンビ	良い	美味しい	広い	近い	多い	素晴らしい	古い	嬉しい	ない	よい	いい	おいしい	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					
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					
1	3	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					
1	4	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					
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					
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					
1	7	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					
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	0	0	0					
3	1	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	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					

「抽出語-文書」頻度表 (対応分析以外)

h5	1	1	1	1	1	1	2	3	3	3	3	4	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 (対応分析)

	部屋	ホテル	風呂	温泉	お部屋	スタッフ	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

(参考) 対応分析で使われる距離尺度

- χ^2 距離でカテゴリー変数間の関連性を測定

- χ^2 は独立性の検定で用いられる指標

$$\chi^2 \text{ 距離} = \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

「観測度数(=実測値)」 カテゴリー変数に従ってクロス集計された度数

「期待度数(=理論値)」 変数が互いに独立している(=無関係の)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(=無関係)と期待される度数の差

- 観測度数と期待度数の差が大きく異なると χ^2 値も大きくなり、
変数間の関係が期待より強い(=無関係でない)ことを示す

(参考) 対応分析のプロット手順

クロス集計表

	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

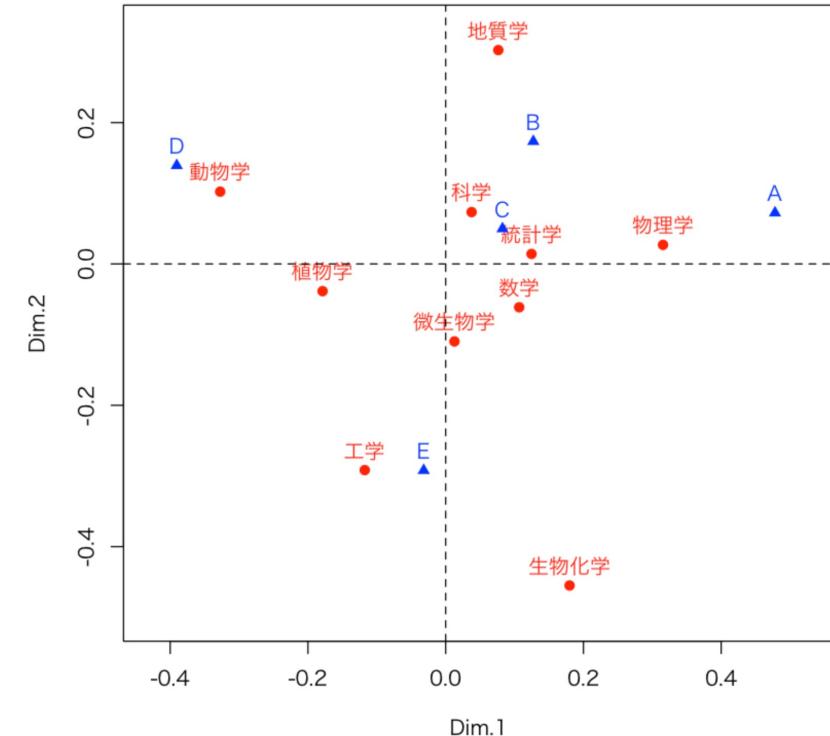
	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410	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

カイ二乗距離

	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



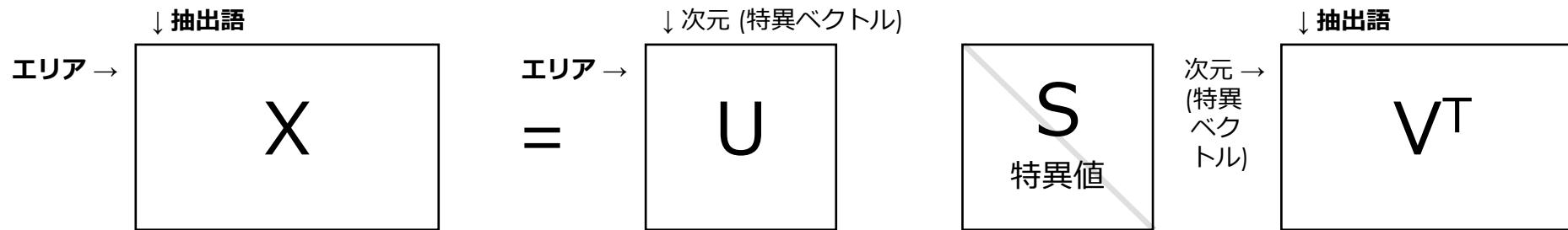
特異値分解してプロット

固有値 = {0.0391, 0.0304, 0.0109, 0.0025, 0}

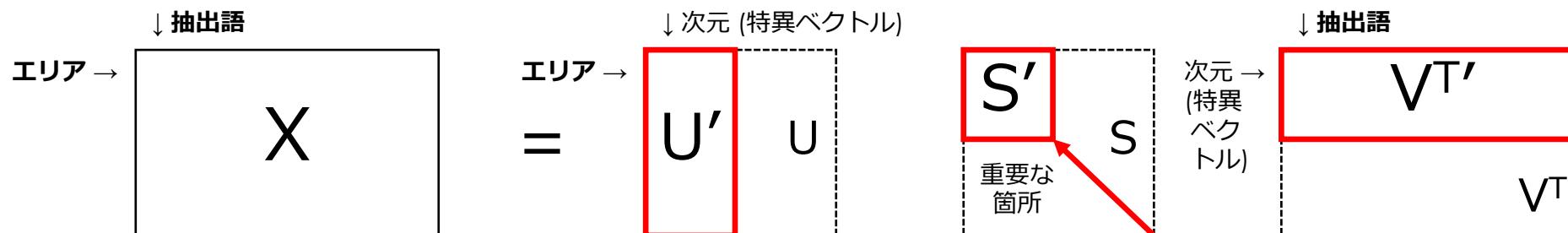
寄与率 = {47.2%, 36.66%, 13.11%, 3.03%, 0%}

(参考) 特異値分解とは

- 特異値分解 $X = USV^T$

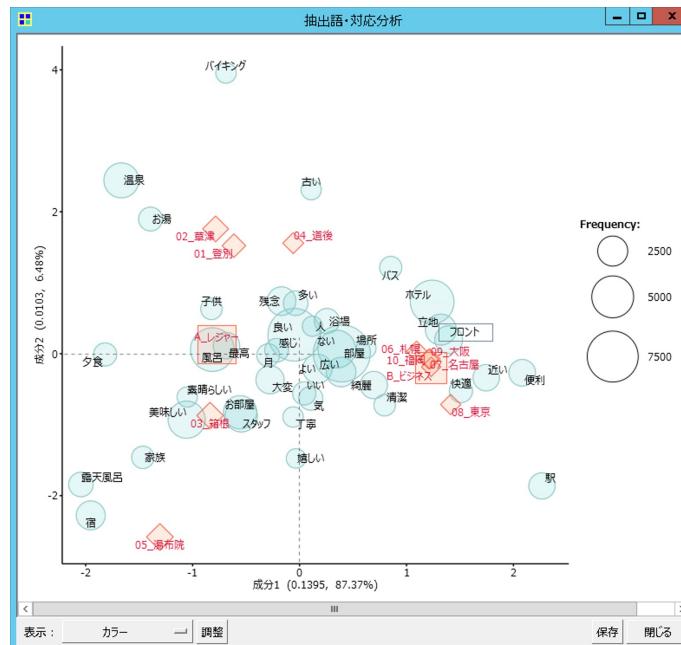


- S の特異値が小さいものを削る

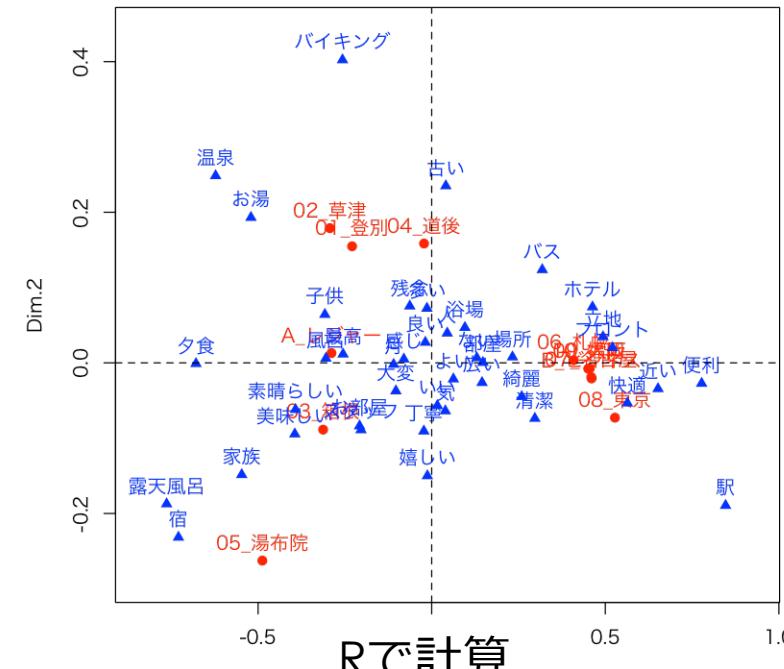


(参考) Rによる計算過程の解説

- https://github.com/haradatm/lecture/blob/master/gssm-202007/03-samples/practice-5_sample.ipynb



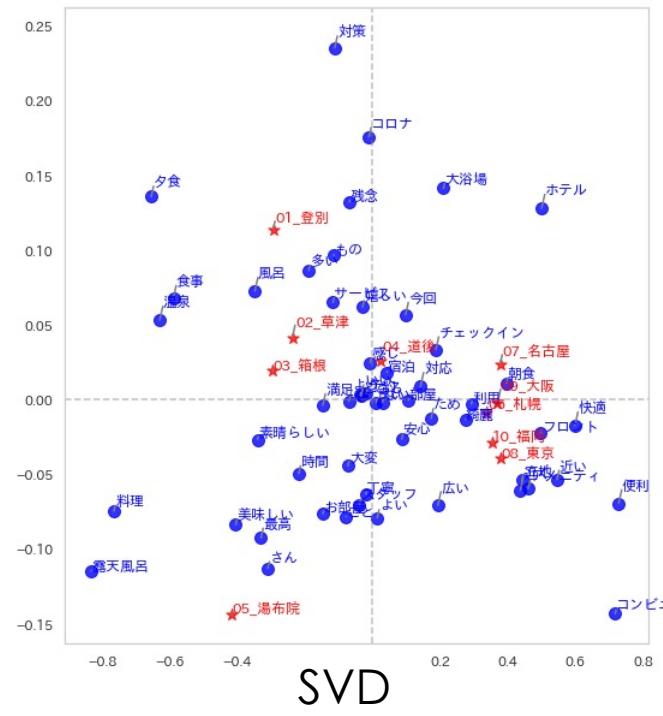
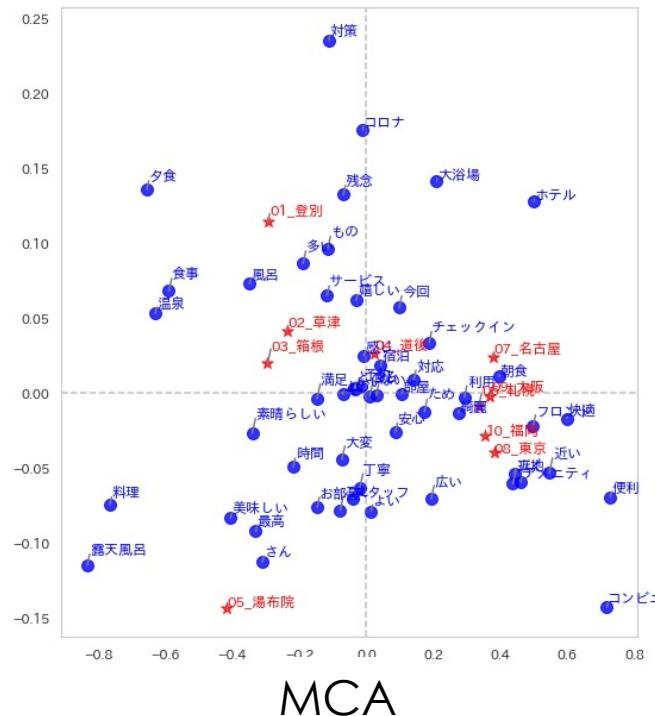
KHCoder



固有値:
{ 0.1395, 0.0103, ... }
寄与率:
{ 87.37%, 6.48%, ... }

(参考) Pythonによる計算過程の解説

- <https://github.com/haradatm/lecture/tree/master/gssm-202107/05-colab> → **corresp_example.ipynb**



固有値:
 $\{ 0.1048, 0.0046, \dots \}$
 寄与率:
 $\{ 32.37\%, 6.8\%, \dots \}$

(参考) 引用した文献

(対応分析)

- [1] 中山慶一郎. “<研究ノート> 対応分析によるデータ解析.” 関西学院大学社会学部紀要 108 (2009): 133-145.
- [2] 金明哲. Rによるデータサイエンス: データ解析の基礎から最新手法まで. 森北出版, 2007. (P.85 「7.2 対応分析」)
- [3] 使用したRのコード. https://github.com/haradatm/lecture/blob/master/gssm-202107/03-samples/practice-4_sample.ipynb

Q3. Jaccard係数って何？

- ①メニューから「ツール」「外部変数と見出し」「リスト」を開く

The screenshot shows two windows side-by-side. The left window is titled '外部変数と見出し' (External Variables and Headings) and lists various area names (e.g., 見出, カテゴリ, エリア, 施設番号, 施設名, 総合, サービス, 立地, 部屋, 設備・アメニティ) with their corresponding codes (e.g., h5, 02_草津, 03_箱根, 04_道後, 05_湯布院, 06_札幌). A red box labeled ② クリック surrounds the 'エリア' row. Another red box labeled ③ 「文」を選択 surrounds the '単位:' dropdown set to '文'. A third red box labeled ④ 「特徴語」「一覧(Excel形式)」を選択 surrounds the '一覧 (Excel形式)' button at the bottom. The right window is titled '関連語検索' (Related Word Search) and displays a search entry for '#直接入力' (Direct Input) with the query 'and <>エリア-->10_福岡'. It shows a result table with columns: N, 抽出語, 品詞, 全体, 共起, Jaccard. The results are listed as follows:

N	抽出語	品詞	全体	共起	Jaccard
1	部屋	名詞	4256 (0.105)	400 (0.123)	0.0563
2	ホテル	名詞	2536 (0.062)	272 (0.084)	0.0494
3	立地	名詞	1329 (0.033)	175 (0.054)	0.0398
4	フロント	名詞	1035 (0.025)	153 (0.047)	0.0371
5	近い	形容詞	931 (0.023)	135 (0.042)	0.0334
6	ない	形容詞B	1891 (0.047)	159 (0.049)	0.0319
7	駅	名詞C	930 (0.023)	128 (0.039)	0.0316
8	便利	形容動詞	974 (0.024)	125 (0.038)	0.0305
9	快適	形容動詞	728 (0.018)	95 (0.029)	0.0245

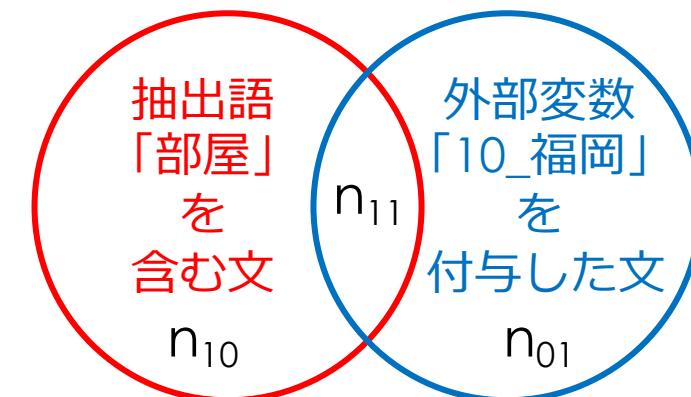
A blue box labeled '各エリアの特徴語を Jaccard 係数の降順で表示' (Display characteristic words of each area in descending order of Jaccard coefficient) is overlaid on the right window.

Jaccard 系数 — 関連の強い語が分かる



全体:
抽出語が出現する
文の数*1

共起:
「10_福岡」を
付与した文のうち,
抽出語が出現する
文の数*2



$$\text{Jaccard 系数 } J^S = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

抽出語「部屋」の場合:

$$n_{11} = 400 \text{ ("共起"列の値)}$$

$$n_{10} = 4256 \text{ ("全体"列の値)} - 400 = 3856$$

$$n_{01} = (400 / 0.123) - 400 = 2852$$

*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_福岡」を付与したデータに対する割合(条件付き確率)

「条件付き確率が同等ないし低下する語も表示」とは

The screenshot shows two windows of the 'Related Word Search' application. The main window ('関連語検索') displays search results for the query '#直接入力'. The results table has columns: N, 抽出語, 品詞, 全体, 共起, Jaccard. Red boxes highlight the '前提確率' (Probability) and '条件付き確率' (Conditional Probability) columns. The '条件付き確率' column shows values like (0.105), (0.062), (0.033), etc. A red arrow points from the '条件付き確率' column to the 'フィルタ設定' (Filter Settings) button in the main window's toolbar. The 'Filter Settings' window ('関連語検索・フィルタ設定') contains several filter options. A red dashed box highlights the checkbox '条件付き確率が同等ないし低下する語も表示' (Display words with conditional probability equal to or lower than the baseline). This checkbox is currently unchecked.

【注意】

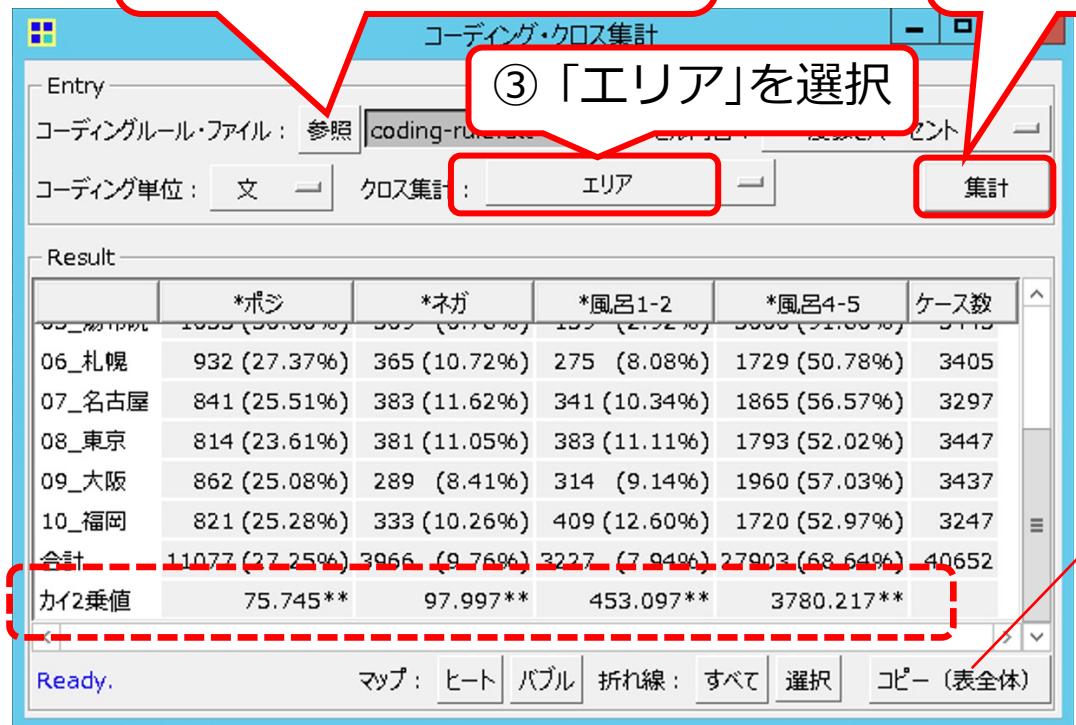
- デフォルトでは「前提確率」より「条件付き確率」が高くなっている語はリストアップされない
- データ全体における出現確率と同等以下の確率でしか出現していない語は、「関連の強い」「特徴的な語」ではないという考え方
- ただし「フィルタ設定」ボタンをクリックして、「条件付き確率が同等ないし低下する語も表示」にチェックを入れると条件付き確率の方が低い語も表示できる

Q4. カイ2乗値で分かることは?

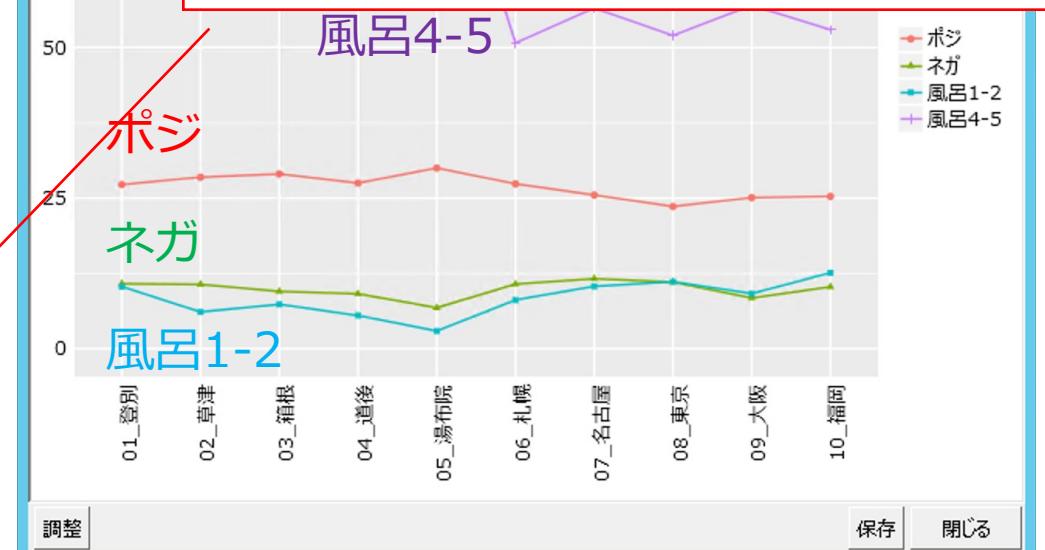
①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

②「参照」をクリックして
「coding-rule.txt」を開く

④「集計」を
クリック



- χ² 値の欄に表示されるアスタリスク「*」の数は、1% 水準で有意な場合は2つ、5% 水準で有意な場合は1つ
- 左の例では、(右のプロットのスケールでは見づらいが)すべてのコードについて出現割合に 1% 水準で有意な変化 があったと見ることができる



(復習) テキストマイニングの手順

・データをよく知る

- ・データ件数や構成比を集計 → データを理解する
 - ・旅行目的別の人気エリアは?
 - ・同伴者別の人気エリアは?
 - ・数値評価による人気エリアの差異は?

・テーマを設定する

- ・解決すべき課題を決める → 分析目的を明確にする
 - ・数値評価が低い原因は?
 - ・高評価の施設に学ぶ改善点は?

・データ分析に取り組む

- ・これら課題を解決するために、テキスト分析を実施

(再掲) 数値評価で違いを見る

の / 売上 / 、
 • 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

数値評価の平均 (エリア別)

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	4.24	4.28	4.12	4.05	4.32	4.24	4.29
01_登別	4.13	4.24	3.97	3.96	4.38	4.10	4.18
02_草津	4.21	4.32	4.02	3.97	4.34	4.13	4.26
03_箱根	4.17	4.15	4.11	3.97	4.18	4.22	4.17
04_道後	4.18	4.38	4.14	4.03	4.03	4.30	4.30
05_湯布院	4.52	4.32	4.39	4.03	3.91	4.52	4.52
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.28	4.28
06_札幌	4.16	4.38	4.22	4.10	3.95	4.07	4.33
07_名古屋	4.10	4.31	4.16	3.98	3.89	3.96	4.25
08_東京	4.01	4.37	4.11	3.99	3.97	4.02	4.22
09_大阪	4.12	4.43	4.11	4.03	3.91	4.07	4.36
10_福岡	4.07	4.39	4.11	4.03	3.91	4.09	4.25

数値評価の平均 (レジャー, ビジネス別)

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	4.24	4.28	4.12	4.05	4.32	4.24	4.29
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.04	4.28

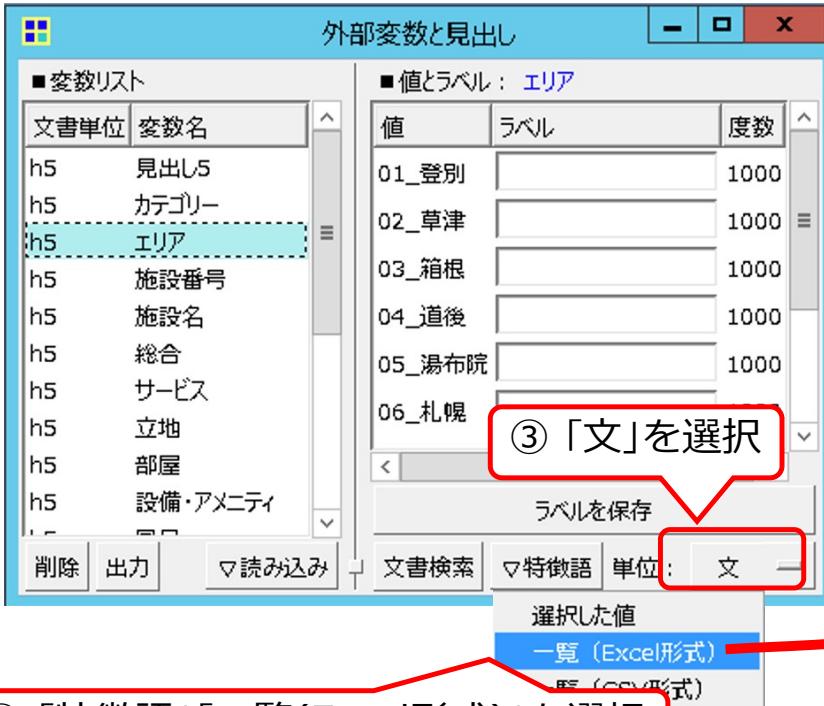
実践的な分析 — 投稿者の関心事を知る

- 宿泊客は、どの項目に注目しているか？
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 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
 - テキスト中の特徴語を集計

「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=エリア」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→再度「▽特徴語」→「一覧(EXCEL形式)」を実行
 - エリアによって特徴語がどう異なるかを比較
 - 注目する項目の違いを考察する

使い方 — 外部変数(エリア)を利用する

①メニューから「ツール」「外部変数と見出し」を開く



A	B	C	D	E	F	G	H	I	J	K
1	01_登別			02_草津			03_箱根			04_道後
2	食事	.068	湯畑	.066	食事	.071	部屋	.053		
3	風呂	.058	風呂	.064	思う	.063	利用	.051		
4	思う	.056	食事	.061	良い	.058	温泉	.048		
5	宿泊	.049	良い	.058	風呂	.052	宿泊	.045		
6	温泉	.045	温泉	.058	美味しい	.049	朝食	.042		
7	美味しい	.043	草津	.052	露天風呂	.043	ホテル	.041		
8	満足	.042	美味しい	.043	満足	.041	風呂	.037		
9	残念	.034	満足	.042	お部屋	.039	美味しい	.035		
10	行く	.034	宿	.038	温泉	.039	道後	.034		
11	料理	.034	行く	.037	宿	.037	立地	.028		
12	05_湯布院			06_札幌			07_名古屋			08_東京
13	食事	.069	利用	.059	ホテル	.058	利用	.060		
14	宿	.065	朝食	.059	利用	.057	部屋	.056		
15	美味しい	.062	ホテル	.055	部屋	.056	ホテル	.054		
16	良い	.057	部屋	.054	名古屋	.053	駅	.033		
17	風呂	.056	思う	.051	朝食	.050	フロント	.033		
18	温泉	.046	札幌	.050	良い	.045	朝食	.032		
19	料理	.045	良い	.048	綺麗	.031	便利	.031		
20	露天風呂	.044	宿泊	.045	対応	.031	立地	.031		
21	満足	.044	立地	.033	快適	.029	快適	.029		
22	お部屋	.041	対応	.031	駅	.029	広い	.027		
23	09_大阪			10_福岡			各エリアの特徴語を10件ずつ 一覧 (数値は Jaccard係数)			
24	ホテル	.064	ホテル							
25	利用	.059	利用							
26	部屋	.056	部屋							
27	思う	.047	思う							
28	朝食	.042	朝食							

実践的な分析 — 特徴語の集計例

A_レジヤー	数値評価指標
良い	.088
風呂	.075
美味しい	.062
温泉	.060
お部屋	.043
宿	.041
スタッフ	.038
露天風呂	.037
最高	.031
大変	.031

B_ビジネス	数値評価指標
部屋	.104
ホテル	.092
ない	.047
立地	.039
フロント	.036
綺麗	.036
広い	.035
便利	.035
駅	.034
快適	.032

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
風呂	.058	湯畑	.066	良い	.058	部屋	.053	宿	.065
温泉	.045	風呂	.064	風呂	.052	温泉	.048	美味しい	.062
美味しい	.043	良い	.058	美味しい	.049	ホテル	.041	良い	.057
残念	.034	温泉	.058	露天風呂	.043	風呂	.037	風呂	.056
お部屋	.032	美味しい	.043	ない	.040	美味しい	.035	温泉	.046
最高	.030	宿	.038	お部屋	.039	立地	.028	露天風呂	.044
バイキング	.030	ない	.035	温泉	.039	広い	.024	お部屋	.041
露天風呂	.029	湯	.030	宿	.037	残念	.024	最高	.037
大変	.028	最高	.029	スタッフ	.036	フロント	.023	スタッフ	.036
夕食	.027	立地	.026	残念	.029	夕食	.021	大変	.031

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
ホテル	.055	ホテル	.058	部屋	.056	ホテル	.064	ホテル	.062
部屋	.054	部屋	.056	ホテル	.054	部屋	.056	部屋	.054
良い	.048	良い	.045	ない	.035	駅	.035	ない	.035
ない	.034	ない	.032	駅	.033	便利	.033	立地	.034
立地	.033	綺麗	.031	フロント	.033	綺麗	.030	便利	.032
広い	.030	快適	.029	便利	.031	フロント	.028	フロント	.031
スタッフ	.029	駅	.029	立地	.031	立地	.028	広い	.030
フロント	.029	立地	.028	快適	.029	快適	.027	綺麗	.029
快適	.028	フロント	.027	広い	.027	広い	.027	駅	.028
便利	.027	便利	.024	綺麗	.026	近い	.024	よい	.024

操作: 「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

使い方 – トピックモデル

NEW

- 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



レジャーとビジネスで注目している観点
(=トピック)が異なる

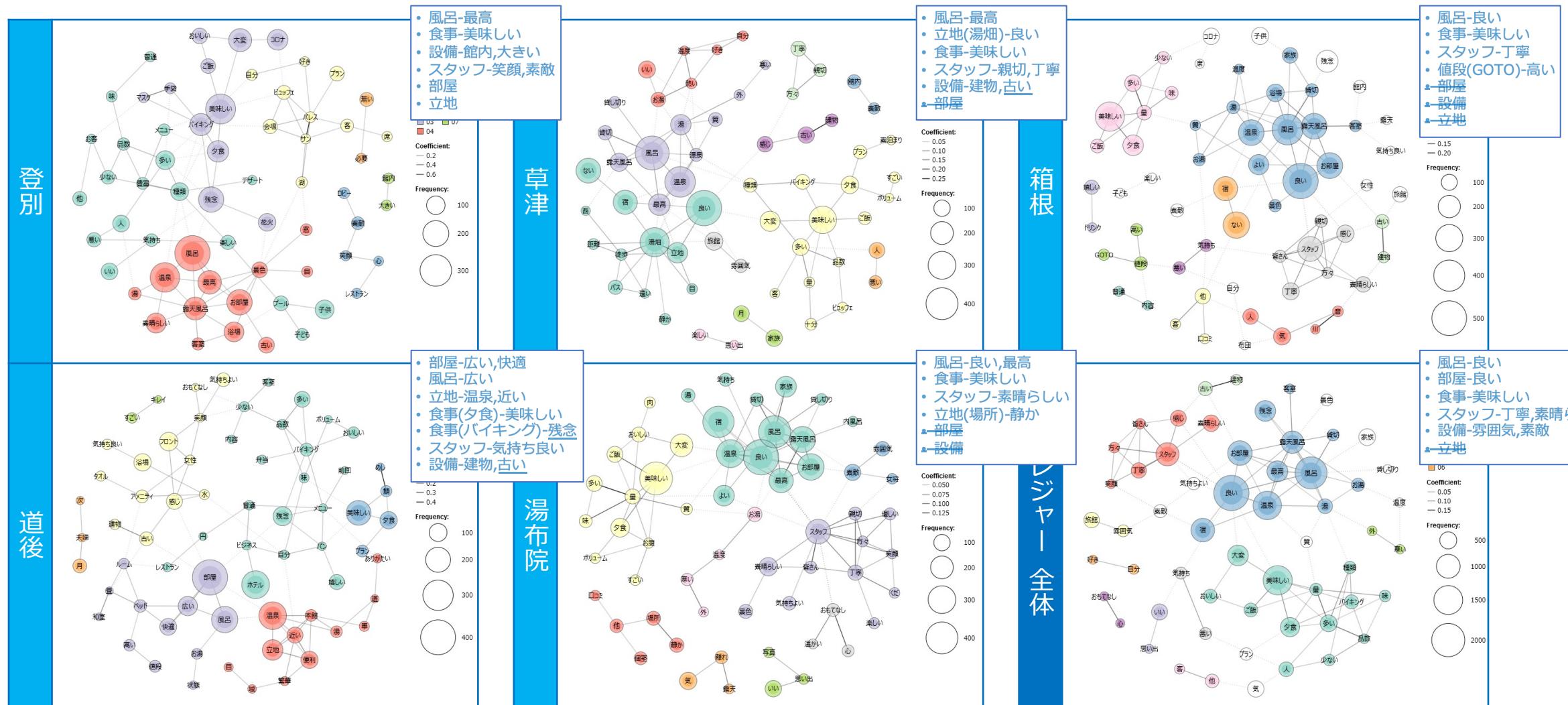
- #1 食事
- #4 風呂
- #2 立地
- #5 部屋(良)
- #3 部屋(悪)
- #6 スタッフ

実践的な分析 — 関心事の背景を探る

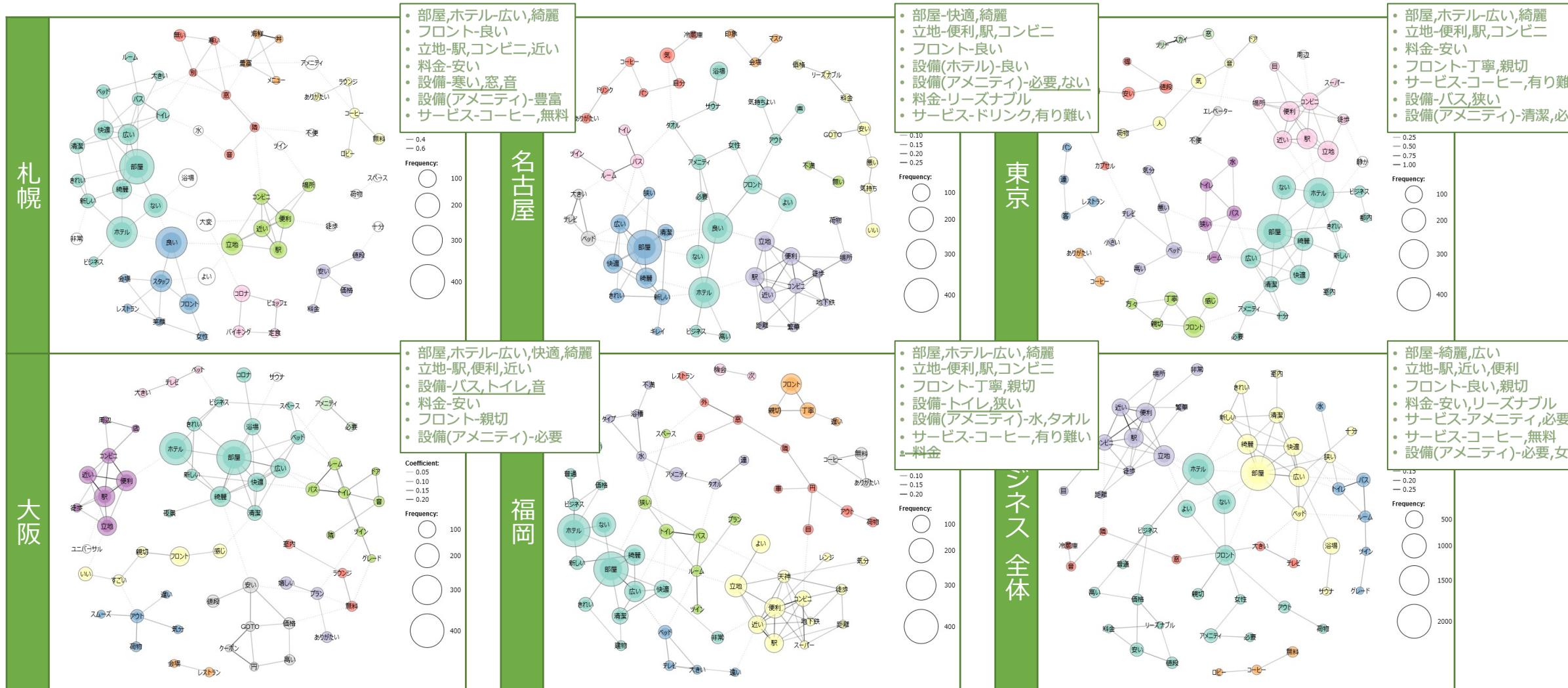
- 宿泊客は、どの項目のどこに注目しているか？
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. カテゴリー「レジャー」(or 「ビジネス」) の 5エリアを比較する
- 手順
 - 特徴語の共起ネットワーク図を作成

「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=エリア」および「値とラベル=01_登別」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→「集計」→「共起ネット」→「調整:上位100,共起関係ほど濃い線に」
 - エリアによって特徴語(とその背景)がどう異なるかを比較
 - 注目する項目の違いを考察する

実践的な分析 — 共起ネットの出力例(1)



実践的な分析 — 共起ネットの出力例(2)



まとめ方の例

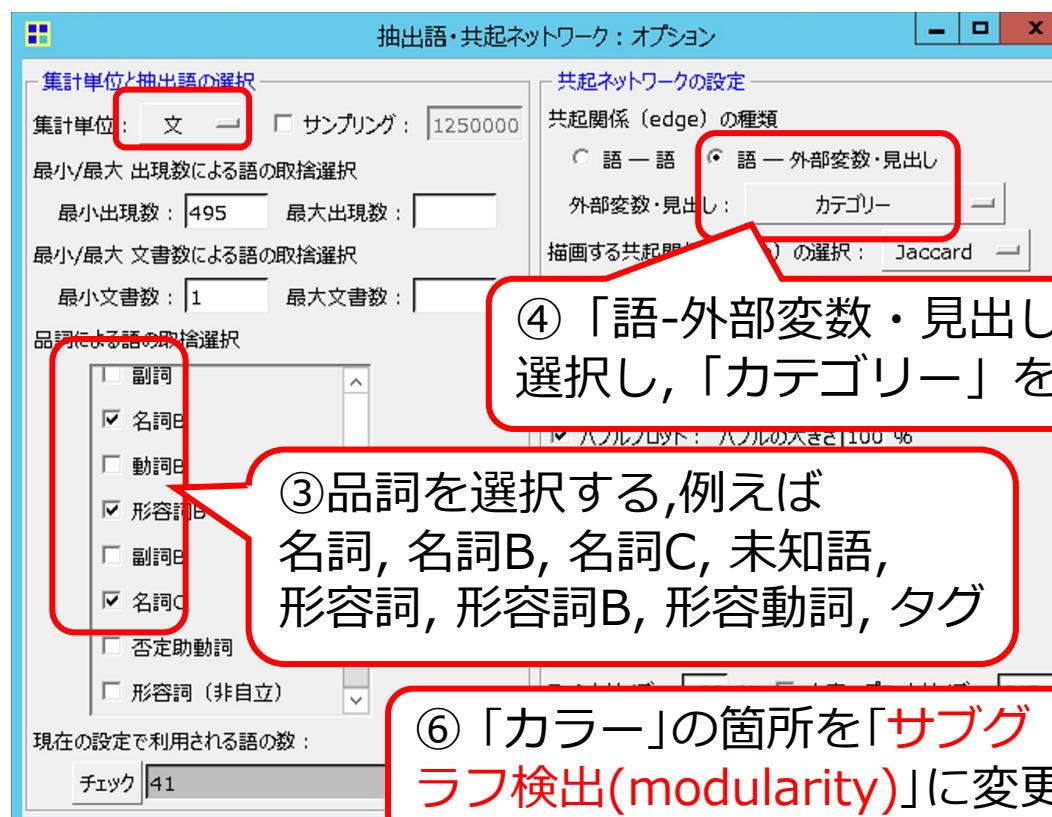
- ・宿泊客が、どの項目のどこに注目しているかを列挙する
 - ・エリアごとに、注目ポイントを列挙
 - ・エリアごとで、注目ポイントを「好評」と「不評」に分類

カテゴリー	エリア	好評	不評
レジャー	XXX	<ul style="list-style-type: none">・風呂が広い・...	<ul style="list-style-type: none">・エアコンが臭い・...

実践的な分析 — 共起ネットの出力例(3)

①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

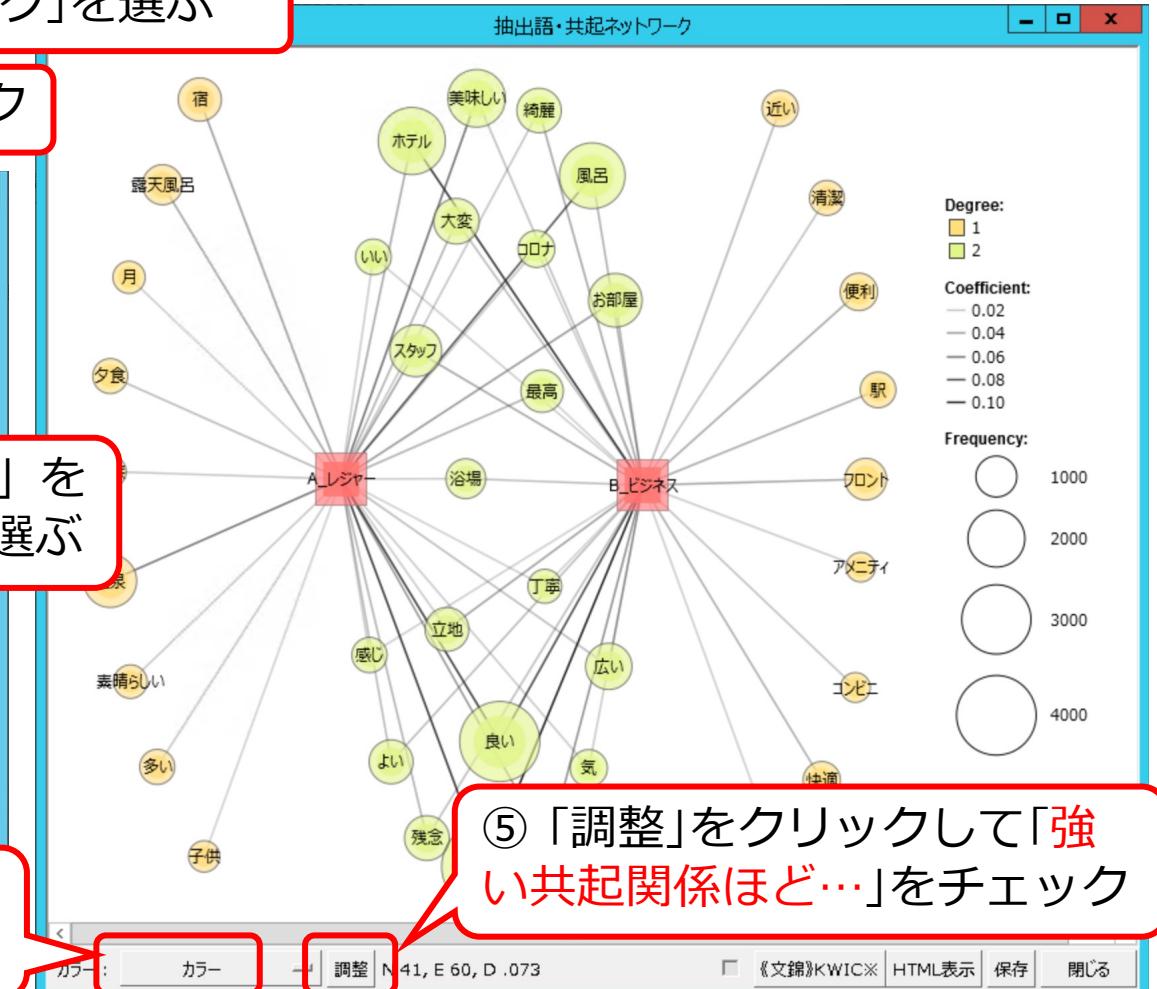
②「集計単位」として「文」を選んで「OK」をクリック



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

⑥「カラー」の箇所を「サブグラフ検出(modularity)」に変更



実践的な分析 — 改善案を提案する(1/2)

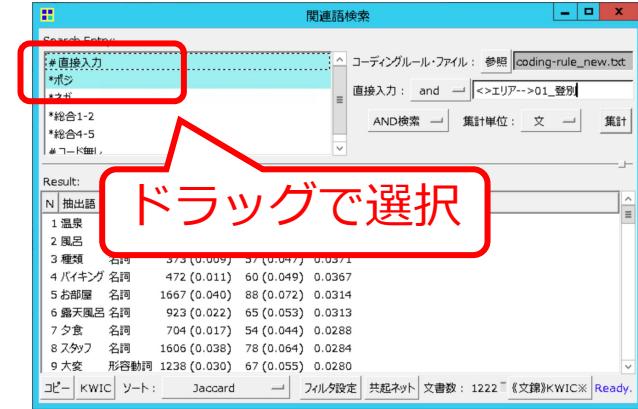
- ・ユーザーは何をどう高評価しているか?

1. カテゴリー「レジャー」と「ビジネス」を比較
2. 対照的な2エリアを比較

- ・手順

- ・特徴語とポジティブ意見の共起ネットワーク図を作成

「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=エリア」および「値とラベル=01_登別」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→ドラッグしながら「#直接入力(and)」と「Search Entry:*ポジ」の両方を選択→「AND検索」「集計単位:文」「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」→「共起ネット」→「調整:上位=100,共起関係ほど濃い線に」



- ・エリアによってポジティブ意見(とその背景)どう異なるかを比較
- ・何がどう評価されているかを考察する

実践的な分析 — 改善案を提案する(2/2)

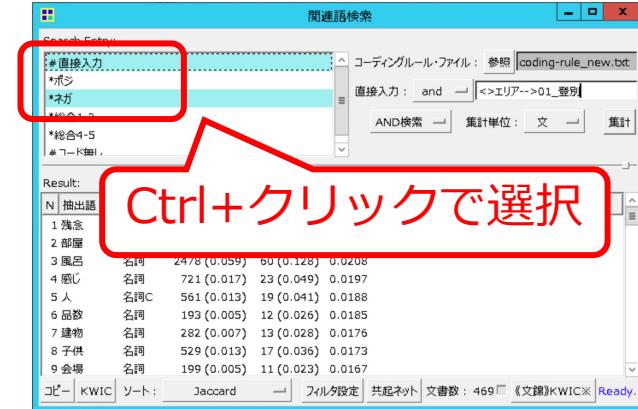
- ユーザーは何をどう**低評価**しているか?

- カテゴリー「レジャー」と「ビジネス」を比較
- 対照的な2エリアを比較

- 手順

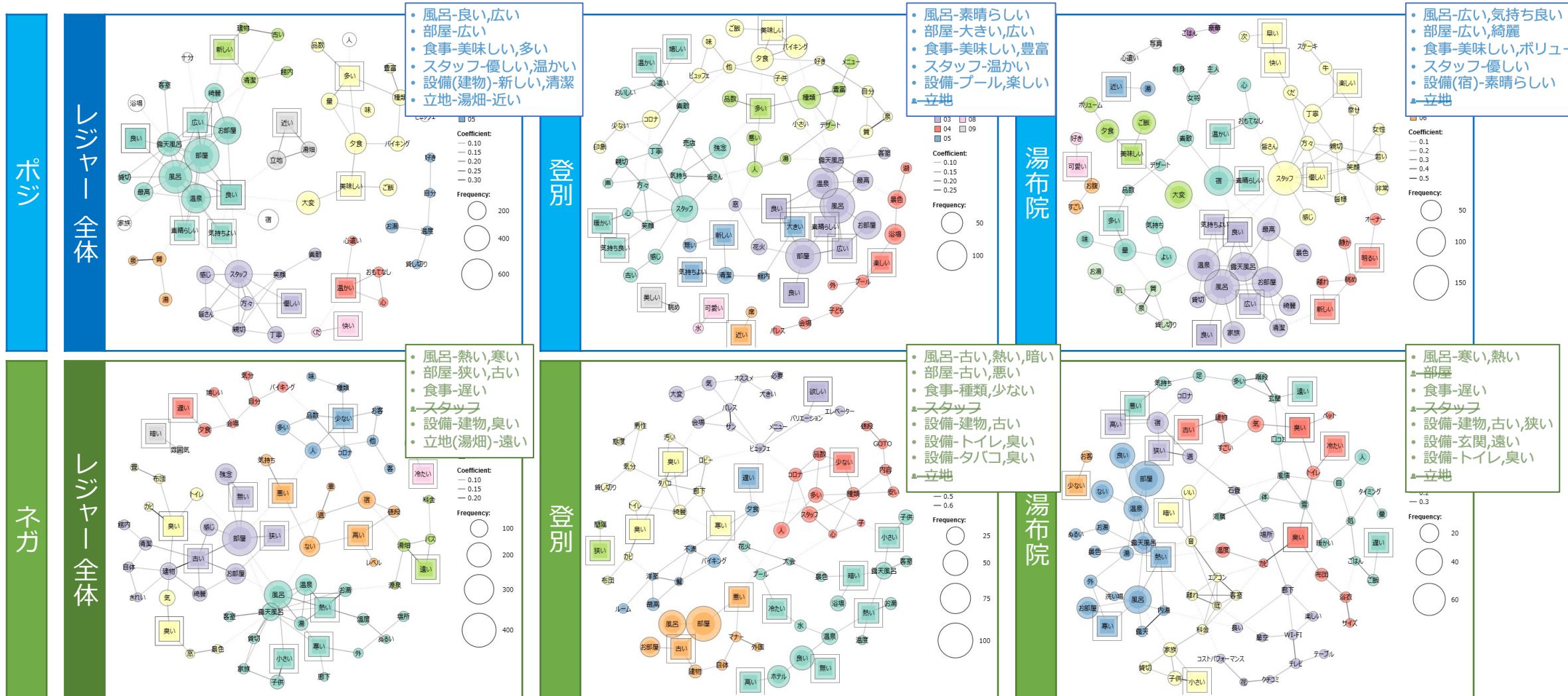
- 特徴語と**ネガティブ意見**の共起ネットワーク図を作成

「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=エリア」および「値とラベル=01_登別」を選択→
「▽特徴語」→「選択した値」→「関連語検索画面」→Ctrlキーを押しながら「#直接入力(and)」と「Search Entry:***ネガ**」の両方を選択→「AND検索」「集計単位:文」「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C」を選択→「集計」→「共起ネット」→「調整:上位=100,共起関係ほど濃い線に」

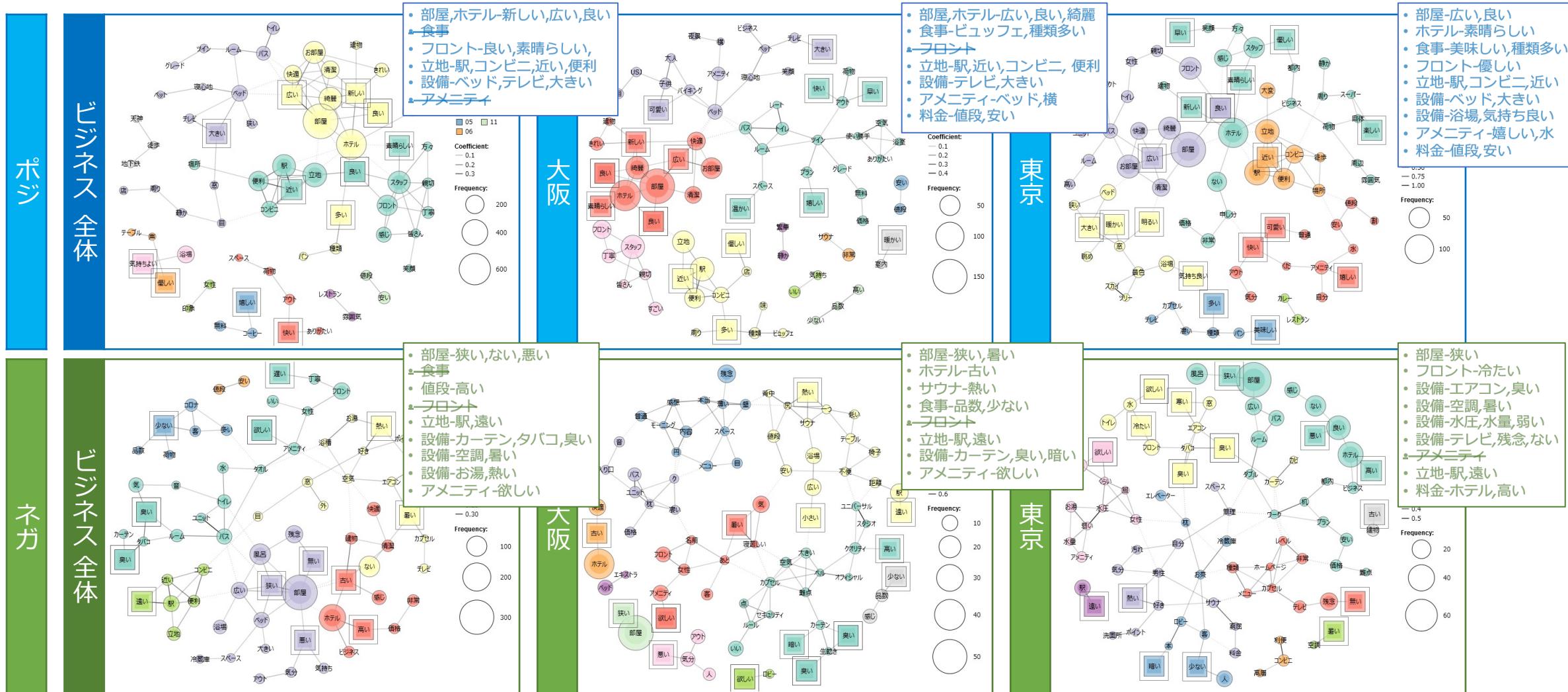


- エリアによって**ネガティブ意見**(とその背景)どう異なるかを比較
- エリアの課題を考察する

実践的な分析 — 登別と湯布院のポジネガ比較



実践的な分析 — 大阪と東京のポジネガ比較



まとめ方の例

- ・主張を支持する図とユーザーの生の声(原文)を使って議論する
 - ・エリア X が評価されている点は何か?
 - ・エリア Y の課題は何か?
 - ・エリア Y の改善に向けた提案?

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: ... ・...	・... ・...

自己紹介 10分 (2021/7/8 更新)

グループ1	202040081
	202040166
	202130165
	202140051
	202140065
グループ2	202140063
	202140064
	202140074
	202140084
	202140085
グループ3	202040176
	202140073
	202140078
	202140083
	202140404

グループ4	202040173
	202140053
	202140059
	202140067
	202140075
グループ5	201921622
	201945008
	202140068
	202140072
	202140077
グループ6	202140056
	202140062
	202140069
	202140080
	202140082

グループ7	202040153
	202140057
	202140066
	202140071
	202140076
グループ8	201845006
	202040174
	202140054
	202140079
グループ9	202140058
	202140060
	202140061
	202140070
	202140407

実習 — グループワーク

- テキストマイニングの手順に倣い、テキストデータの分析を進めてください
例: データによく知る → テーマを設定する → テキスト分析に取り組む
- 本演習はグループワークです。グループごとに分析や議論を進めてください
例: 個人で分析する(本日) → グループでもちより選択&議論する(7/16)
- **7/30** は発表会です。グループごとに発表準備をお願いします
 - グループごと、説明 10分、質疑5分

(再掲) 実習用のデータ

データファイル名	件数	データセット	備考
rakuten-1000-2020-2021.xlsx	10,000	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアエリアごと 1,000件 (ランダムサンプリング)期間: 2020/1/1~2021/5/12	<ul style="list-style-type: none">本講義の全体を通して利用する
rakuten-1000-2018-2019.xlsx	10,000	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアエリアごと 1,000件 (ランダムサンプリング)期間: 2018/1/1~2019/12/31	<ul style="list-style-type: none">実習用 (3~4日目)
covid19-10000.xlsx	10,000	<ul style="list-style-type: none">ハッシュタグ「#新型コロナ」がついたツイートSearch API (1%サンプリング) で取得した 32万 →10,000件 (ランダムサンプリング)期間: 2020/4/24~2021/5/31	<ul style="list-style-type: none">実習用 (3~4日目)

(再掲) データの取得方法

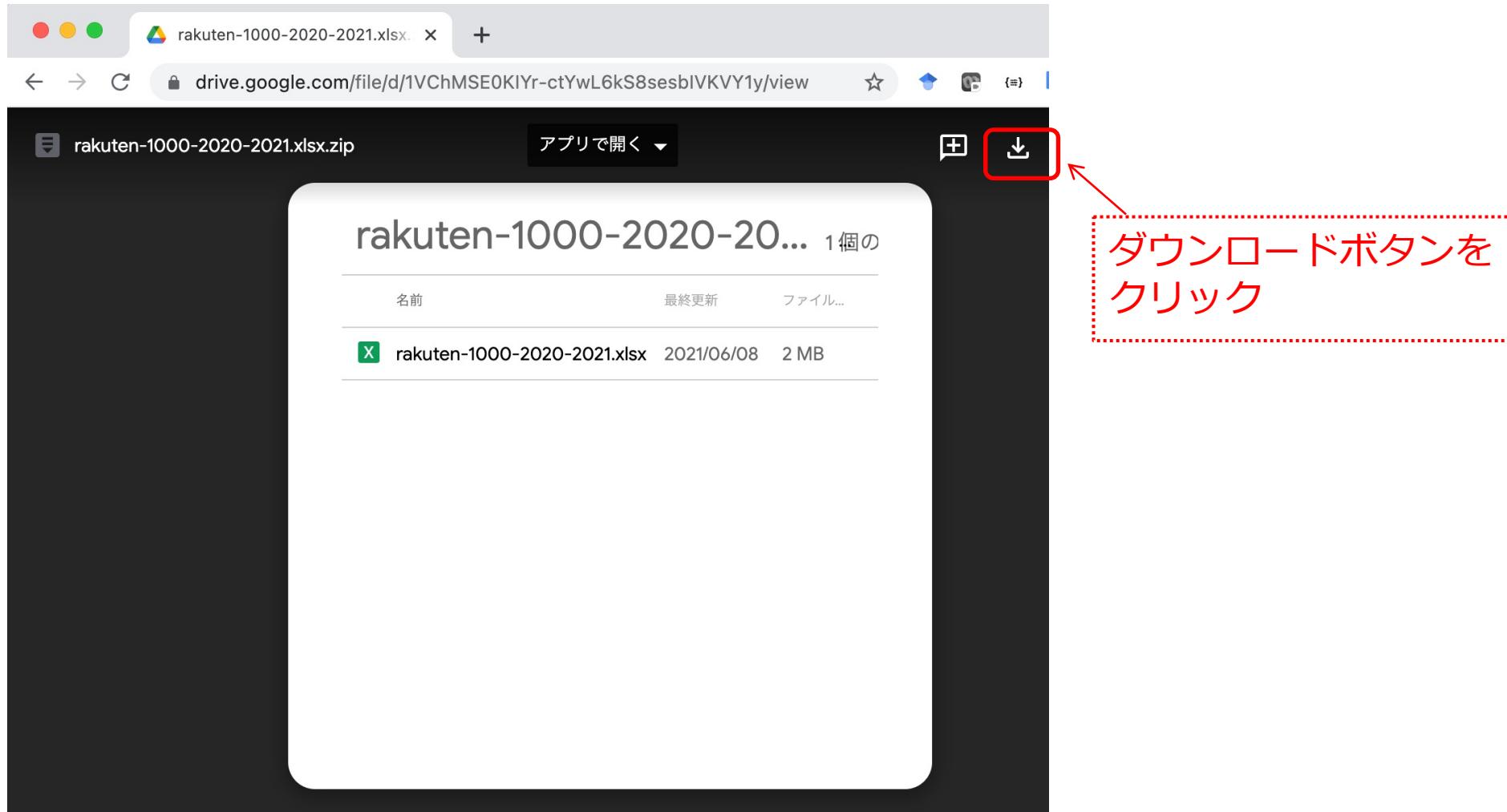
- <https://github.com/haradatm/lecture/tree/master/gssm-202107>

The image shows two screenshots of a GitHub repository. The left screenshot shows the repository structure for the 'master' branch. A red arrow points from the '01-data' folder in the left sidebar to the right screenshot. The right screenshot shows the contents of the '01-data' folder, including a README.md file and three zip files: 'rakuten-1000-2020-2021.xlsx.zip', 'rakuten-1000-2018-2019.xlsx.zip', and 'covid19-10000.xlsx.zip'. A red box highlights the first zip file. A red dotted box surrounds the table and the highlighted file name. A red arrow points from the highlighted file name in the table to the same file name in the list below. A red box also surrounds the text '本講義で主として使用'.

file name	# records	size (zipped)	period
rakuten-1000-2020-2021.xlsx.zip	10,000	2.4 MB	2020/1/1~2021/5/12
rakuten-1000-2018-2019.xlsx.zip	10,000	2.4 MB	2018/1/1~2019/12/31
covid19-10000.xlsx.zip	10		

本講義で主として使用

(再掲) ダウンロード方法



(再掲) A. クチコミデータ

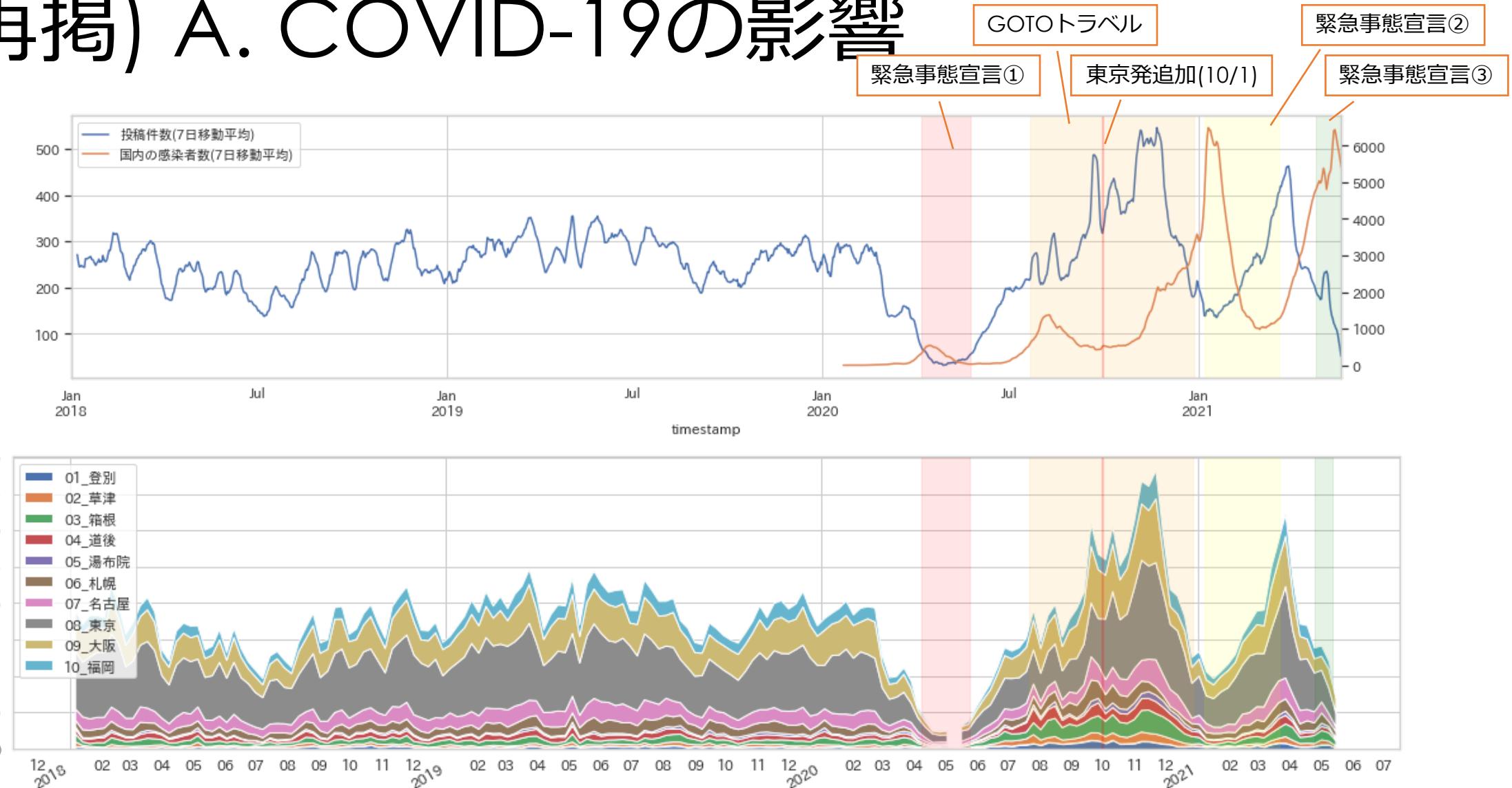
- ・楽天トラベルから収集した「お客様の声」のデータ
 - ・宿泊日が **2018-2019年** および **2020-2021年** (~5/12), 下記10エリア

レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件 × 10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- ・データ項目

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

(再掲) A. COVID-19の影響

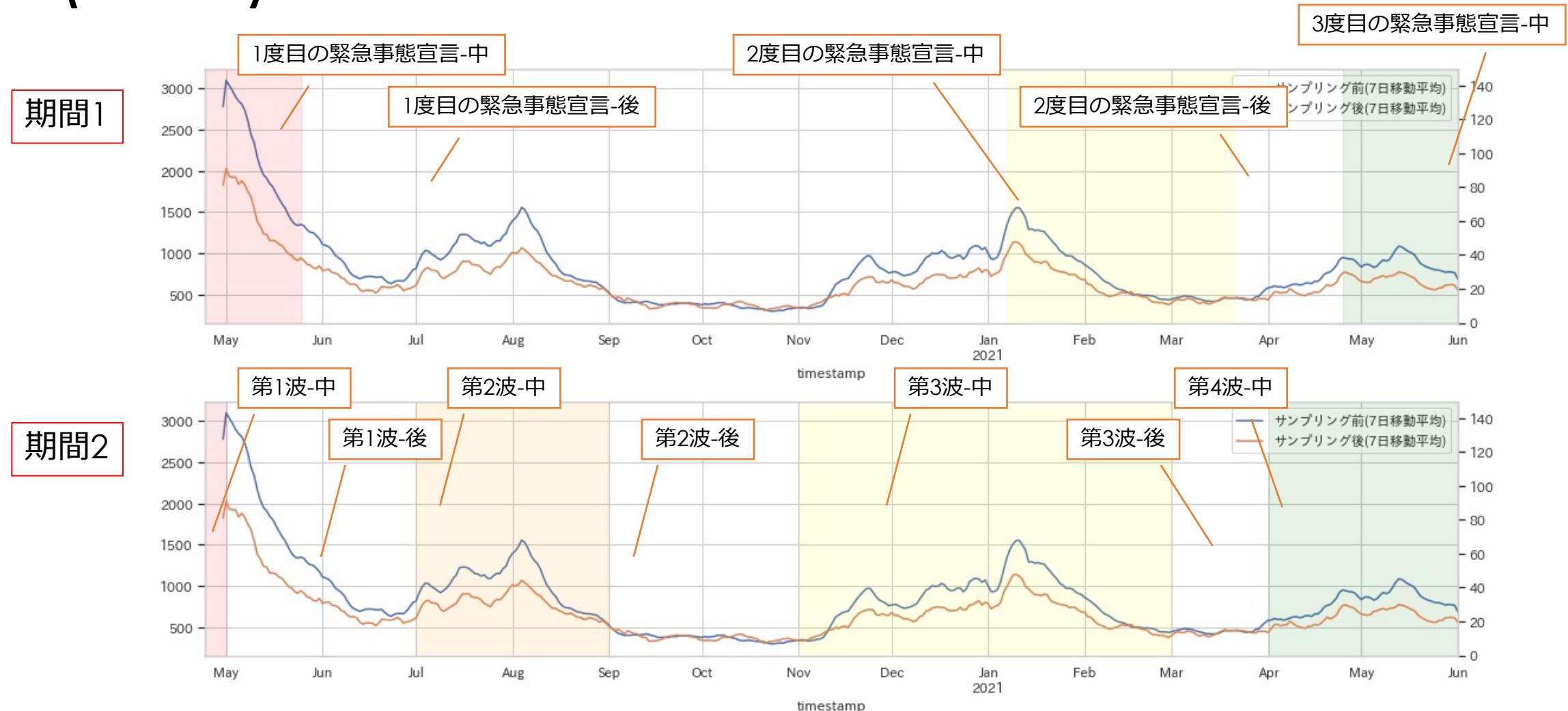


(再掲) B. Twitter データ

- ・ハッシュタグ「#新型コロナ」で投稿されたツイート
 - ・収集期間は **2020-04-24 ~ 2021-05-05** の 32万→**1万件をサンプリング**
 - ・データ項目

投稿日時	ツイート情報 5項目	ユーザID	ユーザ情報 4項目
ツイートの内容		ユーザのフォロワー数	
お気に入り数		ユーザのフォロー数	
リツート数		ユーザのツイート数	
言語			
期間1	追加の属性情報 2項目	{1,2,3,4}度目の緊急事態宣言-{中,後}	
期間2		第{1,2,3}波-{中,後}	

(再掲) B. 追加属性 — 期間1, 期間2



発表内容

1. テーマ設定

例) 「A. クチコミデータ」であれば、好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

例) 「B. Twitterデータ」であれば、シーズン(期間1,期間2)ごとの国民の心情の変化を分析し、新たな施策を提案する

2. 分析結果 (プロットおよび考察)

- ・テーマや仮説にもとづくストーリーで、分析を進めるのがベター
- ・支持する**プロット**とユーザーの**生の声(原文)**を使って主張する

課題 (3日目)

- ・「KH Coder で表記ゆれを吸収する」を参考に,任意の表記ゆれをまとめ,確認した「抽出語リスト」画面を提出してください
 - ・形式: PPT(PDF), 提出先: manaba, 期限: 次回 7/16(金) 21:00

予告:

- ・グループワークについて, 所属するグループ名とで取り上げる分析テーマ(候補)を記載して提出してください
 - ・形式: PPT(PDF), 提出先: manaba, 期限: 7/23(祝) 21:00

参考書

(KH Coder)

- [1] 横口耕一. 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.
- [2] 横口耕一. 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.
- [3] 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

(Windows環境によるデータ収集方法の参考に)

- [4] 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. http://www.shijo-sozo.org/news/第10分科会_1.pdf

参考書

(Rを使った参考書)

- [5] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [6] 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)

- [7] 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.
- [8] 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)

- [9] 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.