

テキストマイニングの実習

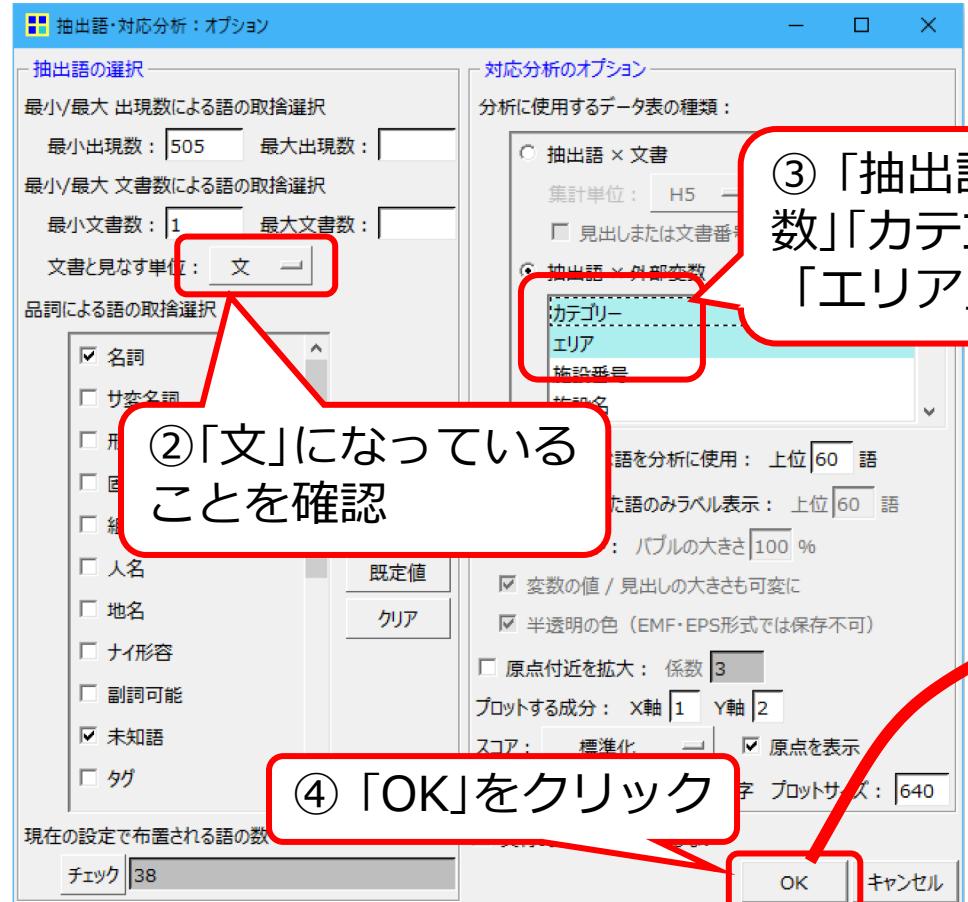
— 3日目 —

2018/7/26

ビジネス科学研究科
経営システム科学専攻

対応分析についての補足【再掲】

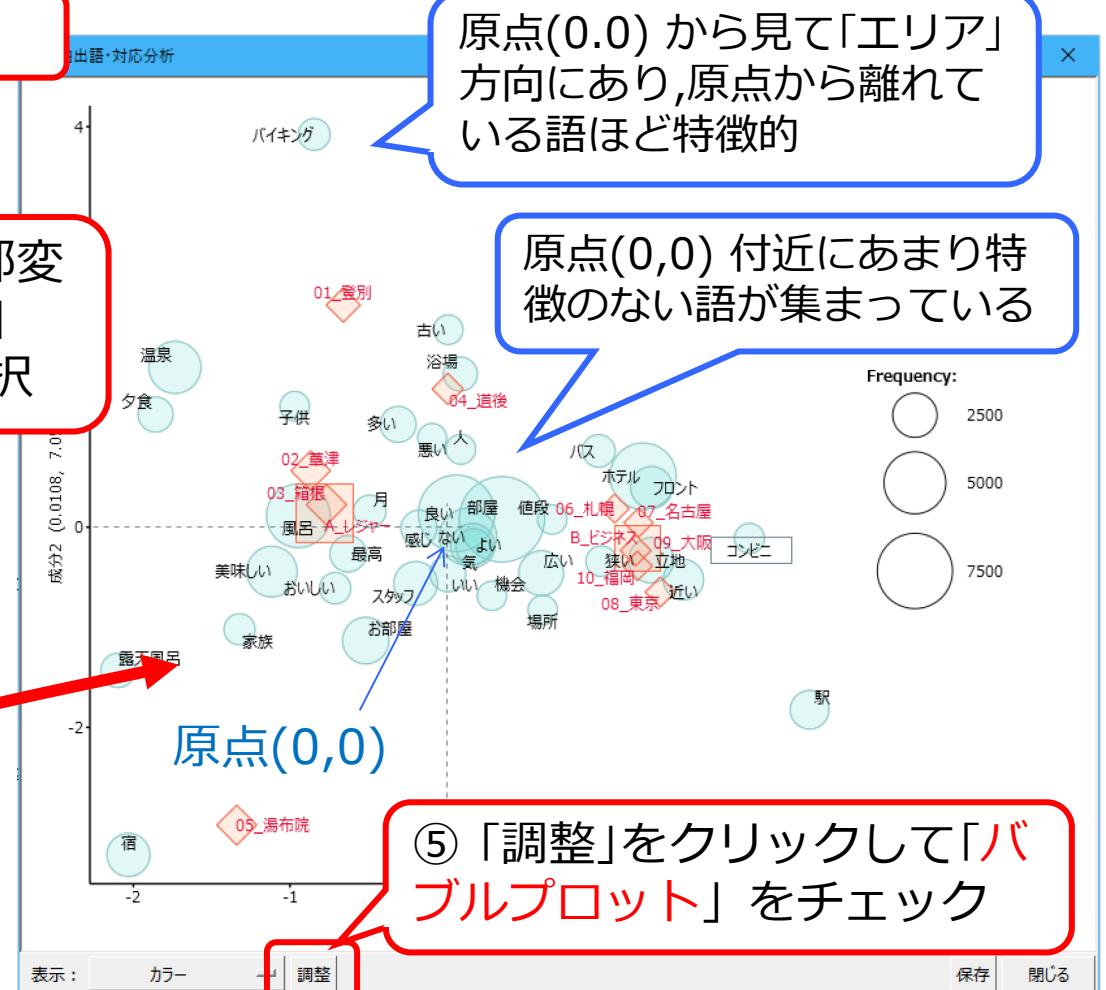
- ①メニューから「ツール」「抽出語」「対応分析」を選ぶ



- ③「抽出語×外部変数」「カテゴリー」「エリア」を選択

②「文」になっていることを確認

- ④「OK」をクリック



- ⑤「調整」をクリックして「バブルプロット」をチェック

対応分析(コレスポンデンス分析)

- クロス集計結果を散布図にして見やすくする手法
 - 基本的な考え方: 行の項目と列項目の関連性が最も強くなるように行や列を並べ替える
 - 関連性が強いもの(パターンが似てるもの)同士が近似するような値を取るように計算する → 関連する=つまり独立でない程度を知りたい
- 対応分析では、変数間や個体間の距離を定義するために χ^2 統計量を用いる

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

観測度数: 変数が互いに独立している場合の期待されるセル内の度数

期待度数: 行の合計と列の合計を掛け合わせ、観測値の合計数で割った値

単語とエリアのクロス集計 (頻度ベース)

外部変数	部屋	ホテル	風呂	温泉	お部屋	スタッフ	フロア	立地	感じ	最高	夕食	浴場	バイ	値段	露天	子供	種類	コン	バス	家族	良い	美味	広い	近い	多い	狭い	無い	悪い	素晴らしい	ない	よい	いい	宿	駅	人	気	月	
A_レジマー	276	103	179	167	88	77	48	55	65	67	81	48	53	34	63	53	46	11	24	51	217	119	66	39	76	25	32	38	48	109	74	54	44	129	14	41	34	39
B_ビジネス	218	153	53	22	42	50	71	59	24	22	3	26	14	30	1	9	14	47	32	4	137	26	57	64	27	36	29	18	8	96	42	28	24	9	68	30	31	18
01_登別	64	21	30	27	14	11	8	4	12	15	17	12	16	10	13	11	12	1	8	6	26	21	12	21	3	7	8	9	20	10	9	7	7	1	6	6	6	
02_草津	52	24	56	48	13	10	10	17	11	15	13	7	13	7	7	16	6	1	7	14	64	23	19	11	23	3	7	14	11	19	22	15	14	38	0	11	6	9
03_箱根	72	26	41	28	26	27	13	7	24	10	21	14	16	9	19	14	19	4	5	10	58	32	16	13	16	10	11	11	16	36	21	13	16	29	8	8	11	11
04_道後	36	28	17	26	11	11	16	18	9	7	15	9	8	6	3	2	7	5	3	4	28	13	10	10	6	6	4	3	4	13	10	10	2	9	2	6	3	2
05_湯布院	52	4	35	38	24	18	1	9	9	20	15	6	0	2	21	10	2	0	1	17	41	30	9	4	10	3	3	2	8	21	11	7	5	46	3	10	8	11
06_札幌	52	31	7	8	7	13	19	13	5	4	0	5	3	4	0	4	3	12	3	0	20	7	11	14	10	5	12	3	1	18	8	7	7	0	9	7	3	3
07_名古屋	46	26	7	6	2	3	23	16	6	1	1	9	5	4	1	0	3	9	3	0	30	5	10	10	3	7	3	3	0	19	13	4	4	0	14	8	7	4
08_東京	32	30	17	2	7	11	8	5	6	6	1	5	3	5	0	4	3	8	9	3	20	5	14	14	5	8	3	1	2	15	4	5	6	3	12	3	6	7
09_大阪	53	29	10	3	10	12	10	11	2	3	1	2	0	9	0	0	1	12	7	1	32	2	8	14	2	10	5	6	3	25	9	7	4	5	13	10	7	1
10_福岡	35	37	12	3	16	11	11	14	5	8	0	5	3	8	0	1	4	6	10	0	35	7	14	12	7	6	6	5	2	19	8	5	3	1	20	2	8	3

並べ替え

外部変数	バイ	露天	子供	夕食	種類	家族	月	感じ	浴場	美味	多い	温泉	最高	素晴らしい	おい	無い	宿	悪い	バス	いい	気	広い	値段	よい	風呂	お部屋	フロア	狭い	人	立地	スタッフ	コン	駅	近い	ない	ホテル	良い	部屋
08_東京	3	0	4	1	3	3	7	6	5	5	5	2	6	2	6	3	3	1	9	5	6	14	5	4	17	7	8	8	3	5	11	8	12	14	15	30	20	32
07_名古屋	5	1	0	1	3	0	4	6	9	5	3	6	1	0	4	3	0	3	3	4	7	10	4	13	7	2	23	7	8	16	3	9	14	10	19	26	30	40
06_札幌	3	0	4	0	3	0	3	5	5	7	10	8	4	1	7	12	0	3	3	7	3	11	4	8	7	7	19	5	7	13	13	12	9	14	18	31	20	52
09_大阪	0	0	0	1	1	1	1	2	2	2	2	3	3	3	4	5	5	6	7	7	8	9	9	10	10	10	10	10	11	12	12	13	14	25	29	32	53	
10_福岡	3	0	1	0	4	0	3	5	5	7	7	3	8	2	3	6	1	5	10	5	8	14	8	8	12	16	11	6	2	14	11	6	20	12	19	37	35	39
04_道後	8	3	2	15	7	4	2	9	9	13	6	26	7	4	2	4	9	3	3	10	3	10	6	10	17	11	16	6	6	18	11	5	2	10	13	28	28	30
01_登別	16	13	11	17	12	6	6	12	12	21	21	27	15	9	7	7	8	8	9	6	12	10	10	30	14	8	3	6	4	11	1	1	20	21	26	64		
05_湯布院	0	21	10	15	2	17	11	9	6	30	10	38	20	8	5	3	46	2	1	7	8	9	2	11	35	24	1	3	10	9	18	0	3	4	21	4	41	52
02_草津	13	7	16	13	6	14	9	11	7	23	23	48	15	11	14	7	38	14	7	15	6	19	7	22	56	13	10	3	11	17	10	1	0	11	19	24	64	52
03_箱根	16	19	14	21	19	10	11	24	14	32	16	28	10	16	11	29	11	5	13	11	16	9	21	41	26	13	10	8	7	27	4	8	13	36	26	58	72	
B_ビジネス	14	1	9	3	14	4	18	24	26	26	27	22	22	8	24	29	9	18	32	28	31	57	30	42	53	42	71	36	30	59	50	47	68	64	96	153	137	218
A_レジマー	53	63	53	81	46	51	39	65	48	119	76	167	67	48	44	32	129	38	24	54	34	66	34	74	179	88	48	25	41	55	77	11	14	39	109	103	217	276

列合計で昇順

行合計で昇順

単語とエリアのクロス集計 (χ^2 値の平方根)

外部変数	部屋	ホテル	風呂	温泉	お部屋	スタッフ	フロント	立地	感じ	最高	夕食	浴場	バイ	値段	露天	子供	種類	コン	バス	家族	良い	美味	広い	近い	多い	狭い	無い	悪い	素晴らしい	ない	よい	いい	おい	宿	駅	人	気	月
A_レジヤー	-0.021	-0.049	0.029	0.047	0.007	-0.003	-0.033	-0.021	0.013	0.016	0.041	0.002	0.018	-0.011	0.038	0.024	0.014	-0.045	-0.020	0.030	-0.004	0.031	-0.014	-0.034	0.015	-0.023	-0.011	0.005	0.023	-0.019	0.001	0.004	0.002	0.048	-0.056	-0.006	-0.011	0.006
B_ビジネス	0.027	0.063	-0.038	-0.061	-0.009	0.005	0.043	0.027	-0.017	-0.020	-0.054	-0.003	-0.023	0.014	-0.050	-0.031	-0.019	0.058	0.026	-0.039	0.005	-0.040	0.018	0.044	-0.019	0.030	0.014	-0.006	-0.030	0.024	-0.002	-0.005	-0.003	-0.063	0.072	0.008	0.015	-0.007
01_登別	0.013	-0.015	0.009	0.014	-0.001	-0.009	-0.015	-0.026	0.007	0.017	0.027	0.014	0.033	0.012	0.023	0.017	0.022	-0.023	0.008	0.000	-0.023	0.013	-0.005	-0.033	0.030	-0.015	0.001	0.008	0.012	-0.006	-0.009	0.000	-0.002	-0.023	-0.029	-0.007	-0.005	-0.001
02_草津	-0.026	-0.024	0.039	0.040	-0.015	-0.022	-0.019	0.000	-0.006	0.005	0.002	-0.013	0.010	-0.009	-0.009	0.024	-0.010	-0.028	-0.005	0.022	0.017	0.003	0.002	-0.012	0.021	0.021	-0.007	0.021	0.010	-0.022	0.012	0.009	0.013	0.041	-0.037	0.002	-0.012	0.002
03_箱根	-0.012	-0.027	0.004	-0.007	0.010	0.013	-0.016	-0.029	0.025	-0.013	0.020	0.005	0.015	-0.006	0.027	0.012	0.030	-0.019	-0.015	0.003	-0.002	0.017	-0.011	-0.003	-0.001	0.003	0.006	0.023	0.003	0.004	-0.002	0.015	0.013	-0.016	-0.012	0.000	0.005	
04_道後	-0.011	0.013	-0.007	0.026	-0.001	0.000	0.019	0.028	0.005	-0.003	0.031	0.011	0.010	0.002	-0.011	-0.015	0.009	0.000	-0.009	-0.004	-0.005	0.001	-0.002	0.004	-0.010	0.003	-0.006	-0.009	-0.004	-0.012	0.000	0.012	-0.017	-0.009	-0.020	-0.001	-0.012	-0.014
05_湯布院	-0.008	-0.050	0.016	0.036	0.024	0.009	-0.037	-0.012	-0.005	0.032	0.018	-0.009	-0.030	-0.021	0.053	0.011	-0.020	-0.028	-0.023	0.044	0.000	0.034	-0.015	-0.025	-0.006	-0.016	-0.019	0.006	-0.006	-0.007	-0.009	-0.011	0.079	-0.023	0.006	0.002	0.018	
06_札幌	0.025	0.028	-0.027	-0.018	-0.010	0.011	0.035	0.015	-0.007	-0.011	-0.027	-0.003	-0.010	-0.004	-0.008	0.038	-0.007	-0.022	-0.014	-0.013	0.006	0.023	0.008	0.002	0.036	-0.007	-0.017	0.006	-0.003	0.003	0.008	-0.034	0.012	0.007	-0.009	-0.007		
07_名古屋	0.019	0.019	-0.025	-0.022	-0.025	-0.021	0.053	0.029	-0.001	-0.023	-0.022	0.017	0.001	-0.003	-0.018	-0.022	-0.007	0.026	-0.005	-0.021	0.010	-0.018	0.004	0.010	-0.017	0.014	-0.007	-0.005	-0.021	0.012	0.018	-0.008	-0.004	-0.033	0.036	0.014	0.012	0.000
08_東京	-0.002	0.033	0.004	-0.032	-0.006	0.009	0.000	-0.010	0.000	0.000	-0.021	0.000	-0.008	0.004	-0.022	-0.001	-0.005	0.022	0.029	-0.004	-0.008	-0.016	0.021	0.029	-0.008	0.020	-0.006	-0.015	0.010	0.003	-0.014	-0.002	0.007	-0.022	0.029	-0.009	0.008	0.017
09_大阪	0.026	0.023	-0.020	-0.032	0.000	0.008	0.003	0.008	-0.020	-0.016	-0.023	-0.016	-0.024	0.020	-0.024	-0.023	-0.018	0.038	0.014	-0.017	0.010	-0.029	-0.005	0.023	-0.005	0.025	0.002	0.009	-0.018	0.029	0.021	0.010	-0.017					
10_福岡	-0.007	0.039	-0.016	-0.033	0.019	0.003	0.005	0.017	-0.008	0.004	-0.027	-0.004	-0.011	0.014	-0.024	-0.019	-0.004	0.007	0.028	-0.022	0.014	-0.014	0.014	-0.004	0.006	0.006	0.003	-0.012	0.007	-0.006	-0.011	-0.032	0.056	-0.016	0.013	-0.008		

並べ替え

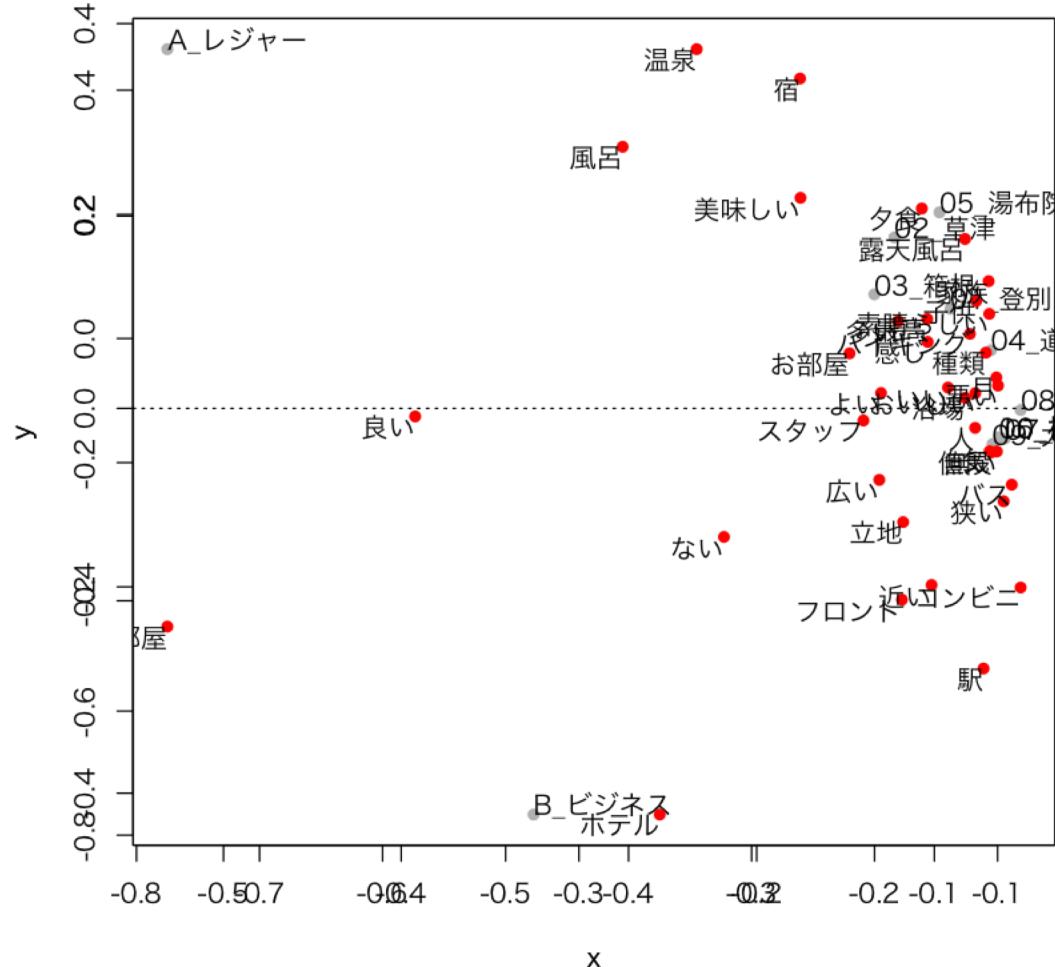
外部変数	宿	温泉	露天	夕食	風呂	美味	家族	子供	素晴らしい	バ	多い	種類	感じ	最高	悪い	おい	お部屋	月	よい	いい	良い	浴場	スタ	人	無い	気	値段	広い	ない	バス	部屋	狭い	立地	近い	フロ	コン	ホテル	駅
B_ビジネス	-0.063	-0.061	-0.050	-0.054	-0.038	-0.040	-0.039	-0.031	-0.030	-0.023	-0.019	-0.019	-0.017	-0.020	-0.006	-0.003	-0.009	-0.007	-0.002	-0.005	0.005	0.008	0.014	0.015	0.014	0.018	0.024	0.026	0.027	0.027	0.044	0.043	0.058	0.063	0.072			
09_大阪	-0.018	-0.032	-0.024	-0.023	-0.020	-0.029	-0.017	-0.023	-0.007	-0.024	-0.022	-0.018	-0.020	-0.016	0.009	-0.006	0.000	-0.017	0.000	0.003	0.010	-0.016	0.008	0.021	0.002	0.010	0.027	0.026	0.026	0.026	0.038	0.023	0.021	0.021				
07_名古屋	-0.033	-0.022	-0.018	-0.022	-0.025	-0.018	-0.021	-0.022	-0.021	0.001	-0.017	-0.007	-0.001	-0.023	-0.005	-0.004	-0.025	0.000	0.018	-0.008	0.010	-0.017	-0.014	-0.007	0.012	-0.003	0.004	0.012	-0.005	0.019	0.014	0.029	0.010	0.053	0.026	0.019	0.03	
06_札幌	-0.034	-0.018	-0.023	-0.027	-0.027	-0.013	-0.022	-0.004	-0.017	-0.010	0.008	-0.008	-0.007	-0.011	-0.007	0.008	-0.010	-0.007	-0.003	0.011	-0.014	-0.003	0.007	0.036	-0.009	-0.004	0.006	-0.007	0.025	0.002	0.015	0.023	0.035	0.038	0.028	0.01		
10_福岡	-0.032	-0.033	-0.024	-0.027	-0.016	-0.014	-0.022	-0.019	-0.012	-0.011	-0.004	-0.004	-0.008	0.004	0.003	-0.011	-0.019	-0.008	-0.004	-0.006	0.014	-0.004	0.003	-0.016	0.006	0.013	0.014	0.007	0.006	0.006	0.006	0.007	0.039	0.05				
05_湯布院	0.079	0.036	0.053	0.018	0.016	0.034	0.044	0.011	0.006	-0.030	-0.006	-0.020	-0.005	-0.032	-0.019	-0.011	0.024	0.018	-0.007	-0.009	0.000	-0.009	0.009	0.006	-0.016	0.002	-0.021	-0.005	-0.006	-0.023	-0.008	-0.016	-0.012	-0.025	-0.037	-0.028	-0.050	-0.02
04_道後	-0.009	0.026	-0.011	0.031	-0.007	0.001	-0.015	-0.004	0.010	0.010	0.009	0.005	-0.003	-0.009	-0.017	-0.001	-0.014	0.000	0.012	-0.005	0.011	0.000	-0.001	-0.001	-0.002	-0.002	-0.012	-0.009	-0.011	0.003	0.028	0.004	0.019	0.000	0.013	-0.02		
02_草津	0.041	0.040	-0.009	0.002	0.039	0.003	0.022	0.017	0.010	0.010	0.021	-0.010	-0.006	0.005	0.021	0.013	-0.015	0.002	0.012	0.009	0.017	-0.017	-0.022	0.002	-0.007	-0.012	-0.009	0.002	-0.022	-0.021	0.000	-0.012	-0.019	-0.028	-0.024	-0.03		
08_東京	-0.022	-0.032	-0.022	-0.021	0.004	-0.016	-0.004	-0.001	-0.010	-0.008	-0.005	0.000	0.000	-0.015	0.007	-0.006	0.017	0.000	-0.002	-0.008	0.000	0.009	-0.006	0.008	0.004	0.021	0.003	0.029	-0.002	0.020	-0.010	0.029	0.000	0.022	0.033	0.02		
A_レジヤー	0.048	0.047	0.037	0.041	0.029	0.031	0.030	0.024	0.023	0.018	0.015	0.014	0.013	0.016	0.005	-0.002	0.007	0.006	0.001	0.004	-0.004	0.002	-0.003	-0.006	-0.011	-0.011	-0.014	-0.019	-0.020	-0.021	-0.023	-0.021	-0.033	-0.045	-0.03			
01_登別	-0.023	0.014	0.023	0.027	0.009	0.013	0.000	0.017	0.012	0.033	0.020	-0.022	-0.007	0.017	0.008	-0.002	-0.001	-0.009	0.000	-0.023	0.014	-0.009	-0.007	0.001	-0.005	0.012	-0.005	-0.006	0.008	0.013	-0.015	-0.026	-0.033	-0.015	-0.023	-0.015	-0.029	
03_箱根	0.013	-0.007	0.027	0.020	0.004	0.017	0.012	0.017	0.023	0.015	-0.003	0.030	0.025	-0.013	0.006	0.015	0.010	0.005	0.004	-0.002	-0.002	0.005	0.013	-0.012	0.003	0.000	-0.006	-0.011	-0.011	-0.016	-0.019	-0.027	-0.016	-0.016	-0.019			

列合計で昇順

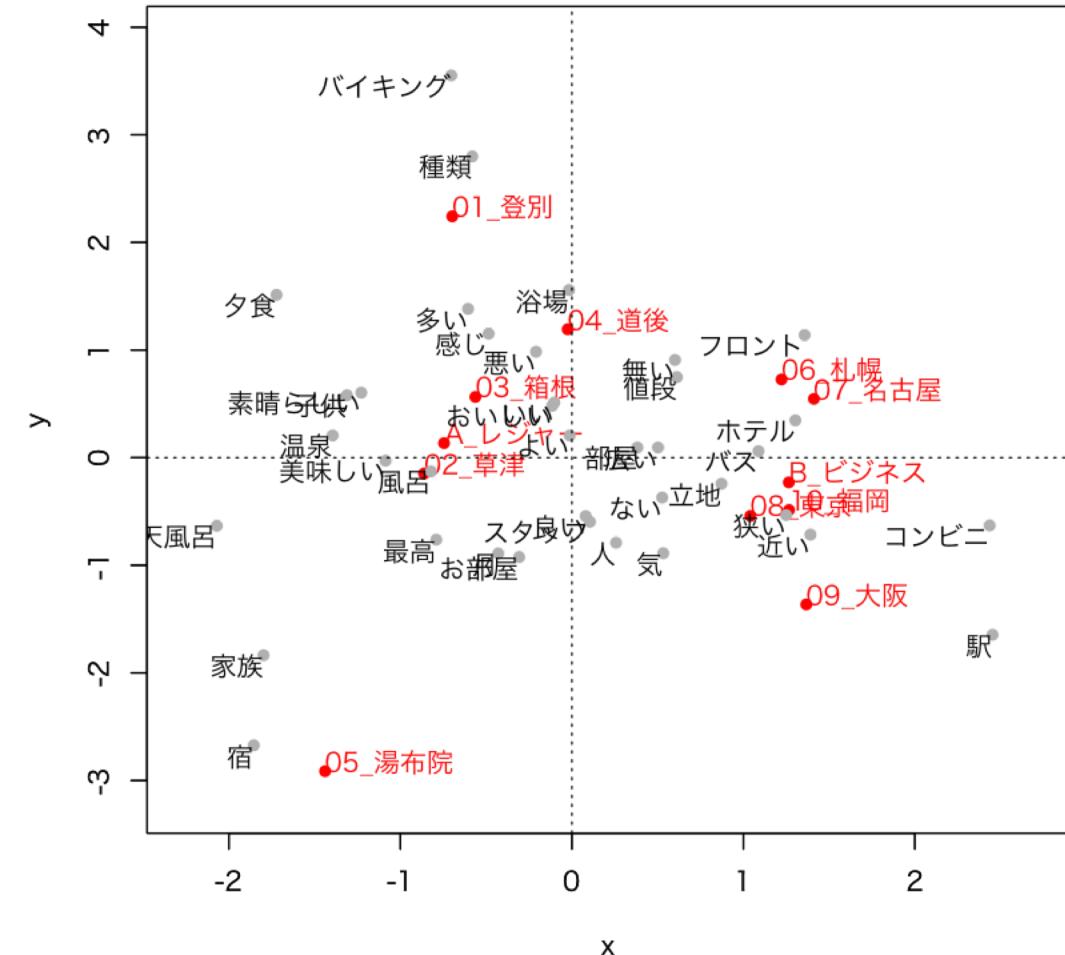
行合計で昇順

特異値分解によるプロット

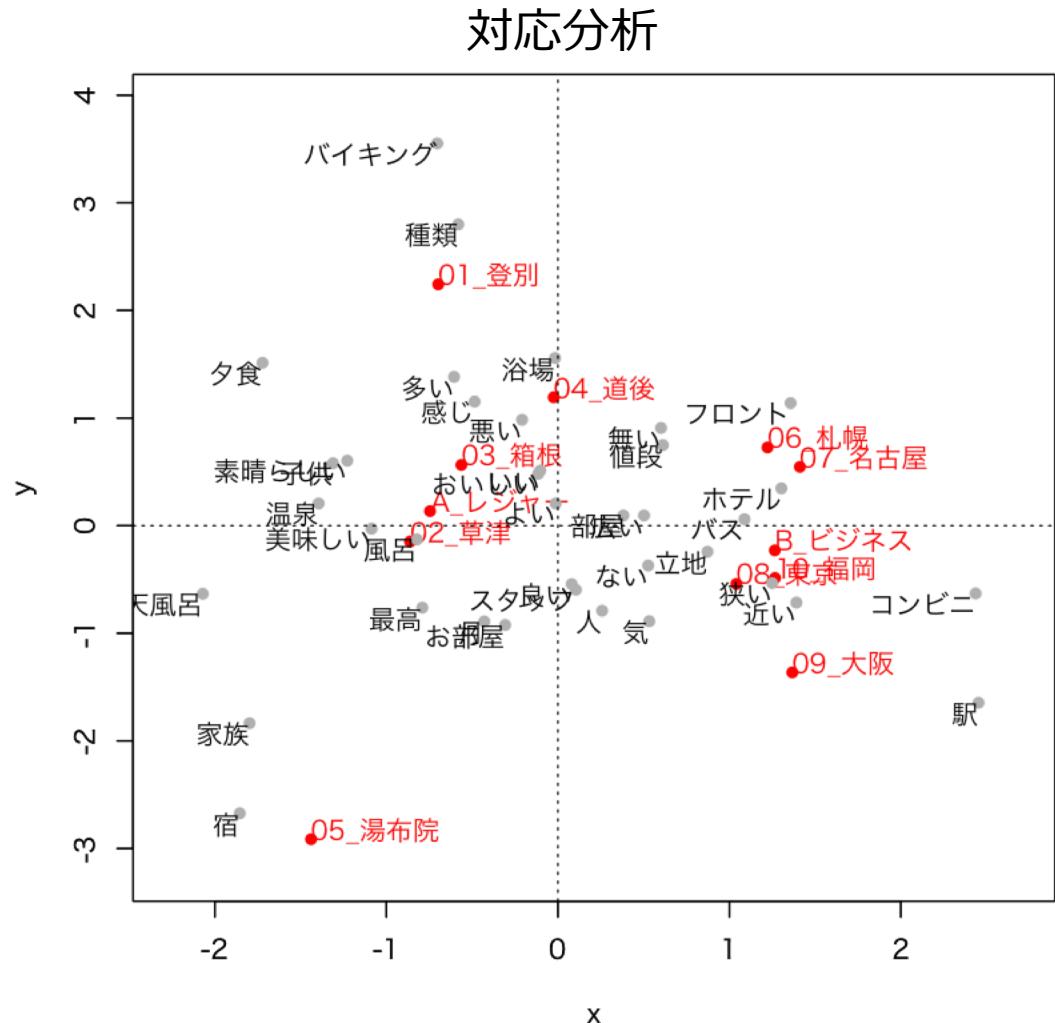
頻度ベース



χ^2 の平方根 (対応行列)



対応分析の読み方



- 原点からの距離
遠いほど違いがある(特徴の強さ)
- 原点からの向き
近いほど項目間の関連性が強い
- プロット(点)
近くにプロットされるほど類似度が高い

参考文献

(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/gssm-201807/blob/master/03-samples/practice-3_sample.ipynb

スケジュール

- 1日目: 7/5
 - 説明 – データ分析の手順
 - 演習 – データの理解 (Excel)
- 2日目: 7/12
 - 説明 – テキストマイニングツールの使い方 (KHCoder)
 - 練習 – テキストマイニングツールの使い方 (KHCoder)
- 3日目: 7/26
 - 演習 – データ分析の実践 (KHCoder)

関連研究(再掲)

- 辻井康一 and 津田和彦「テキストマイニングを用いた宿泊レビューからの注目情報抽出方法」, デジタルプラクティス 3.4 (2012): 289-296.

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・アメニティ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.08	4.16	3.97	3.89	4.16	4.16	4.16
B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06

- 数値評価のみから違いを見つけるのは難しい!!
 - ユーザーの8割が4~5の評価, 1~2をつけない
 - ユーザーは注目の有無に関係なくすべての項目に回答
- レジャーとビジネスでは, 評価すべき項目も異なることを確認した
- テキストと対応付ければ, 同じ点数でも差異があることを確認した

演習 – 特徴語の集計

- ・ユーザーは,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」→「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:文」→「フィルタ設定」→「品詞=名詞,未知語,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」
 - ・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

集計例 – 特徴語の集計

A_レジヤー	数値評価指標
良い	.085
風呂	.074
温泉	.062
美味しい	.052
宿	.041
お部屋	.035
スタッフ	.030
夕食	.029
露天風呂	.027
最高	.023

B_ビジネス	数値評価指標
部屋	.104
ホテル	.088
ない	.045
立地	.045
駅	.042
フロント	.038
近い	.037
広い	.035
よい	.028
コンビニ	.023

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
部屋	.057	温泉	.067	良い	.059	温泉	.059	宿	.071
風呂	.055	風呂	.063	風呂	.054	部屋	.053	風呂	.059
温泉	.042	良い	.060	美味しい	.049	良い	.050	美味しい	.052
美味しい	.036	宿	.043	ない	.038	ホテル	.047	温泉	.043
宿	.035	美味しい	.040	温泉	.037	立地	.033	お部屋	.041
お部屋	.033	湯	.026	お部屋	.033	フロント	.024	露天風呂	.038
スタッフ	.028	夕食	.026	露天風呂	.031	近い	.023	スタッフ	.031
夕食	.026	お部屋	.026	夕食	.029	広い	.023	最高	.026
最高	.023	最高	.025	スタッフ	.028	よい	.023	家族	.025
浴場	.020	よい	.025	宿	.027	スタッフ	.022	夕食	.025

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
部屋	.053	部屋	.056	ホテル	.056	部屋	.059	ホテル	.058
ホテル	.053	ホテル	.051	部屋	.054	ホテル	.058	部屋	.050
ない	.050	フロント	.033	駅	.044	立地	.035	駅	.040
立地	.039	駅	.032	近い	.036	駅	.035	立地	.039
駅	.035	立地	.031	立地	.030	ない	.033	近い	.032
フロント	.032	近い	.026	フロント	.027	フロント	.030	フロント	.029
近い	.032	広い	.026	広い	.026	近い	.028	広い	.027
広い	.030	よい	.024	コンビニ	.022	広い	.025	よい	.025
よい	.024	コンビニ	.022	よい	.022	安い	.019	コンビニ	.020
コンビニ	.022	狭い	.018	アメニティ	.021	値段	.019	バス	.019

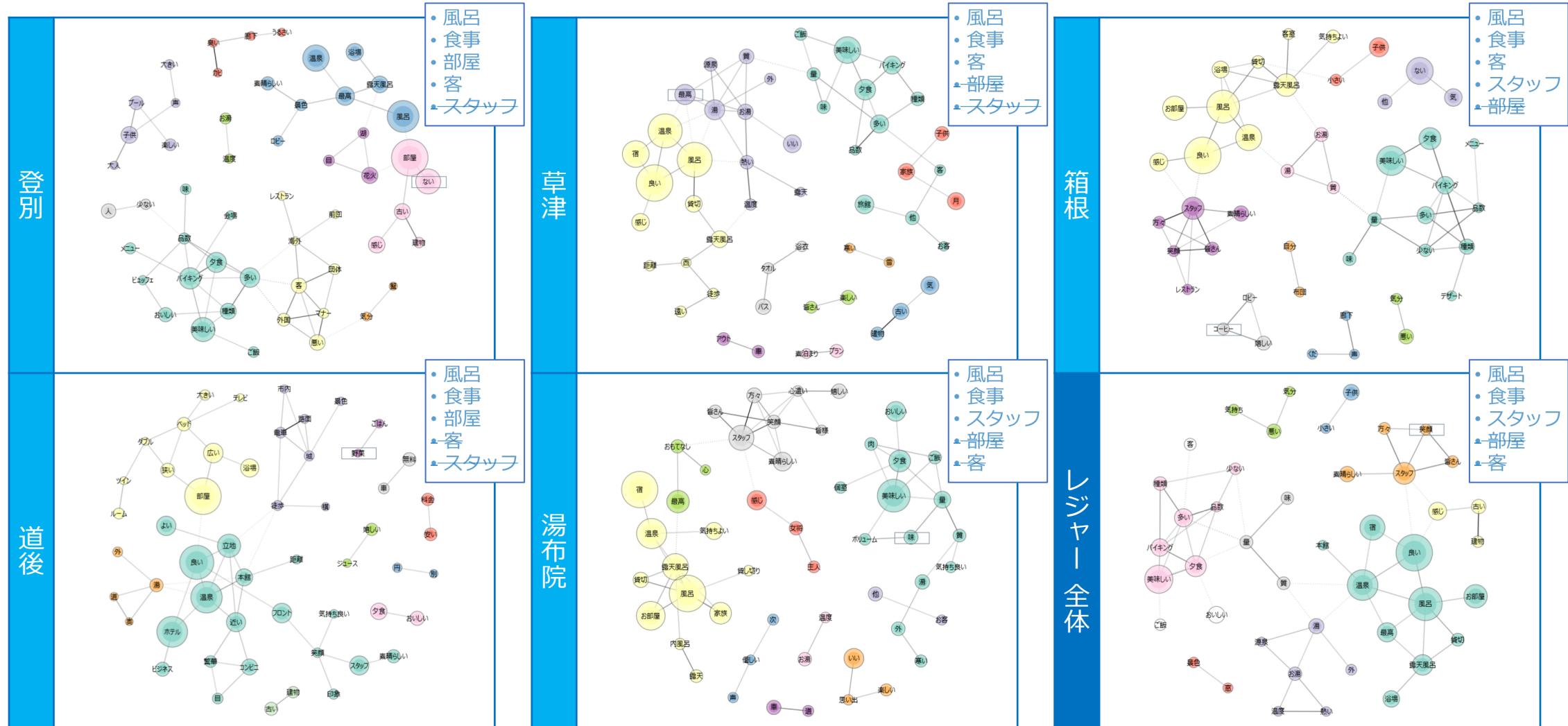
Tips: 「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=カテゴリー」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→「フィルタ設定」→「品詞=名詞, 未知語, 形容詞, 名詞B, 形容詞B, 名詞C」を選択→「▽特徴語」→「一覧(EXCEL形式)」で連続実行

演習 – 特徴語の共起ネットワーク

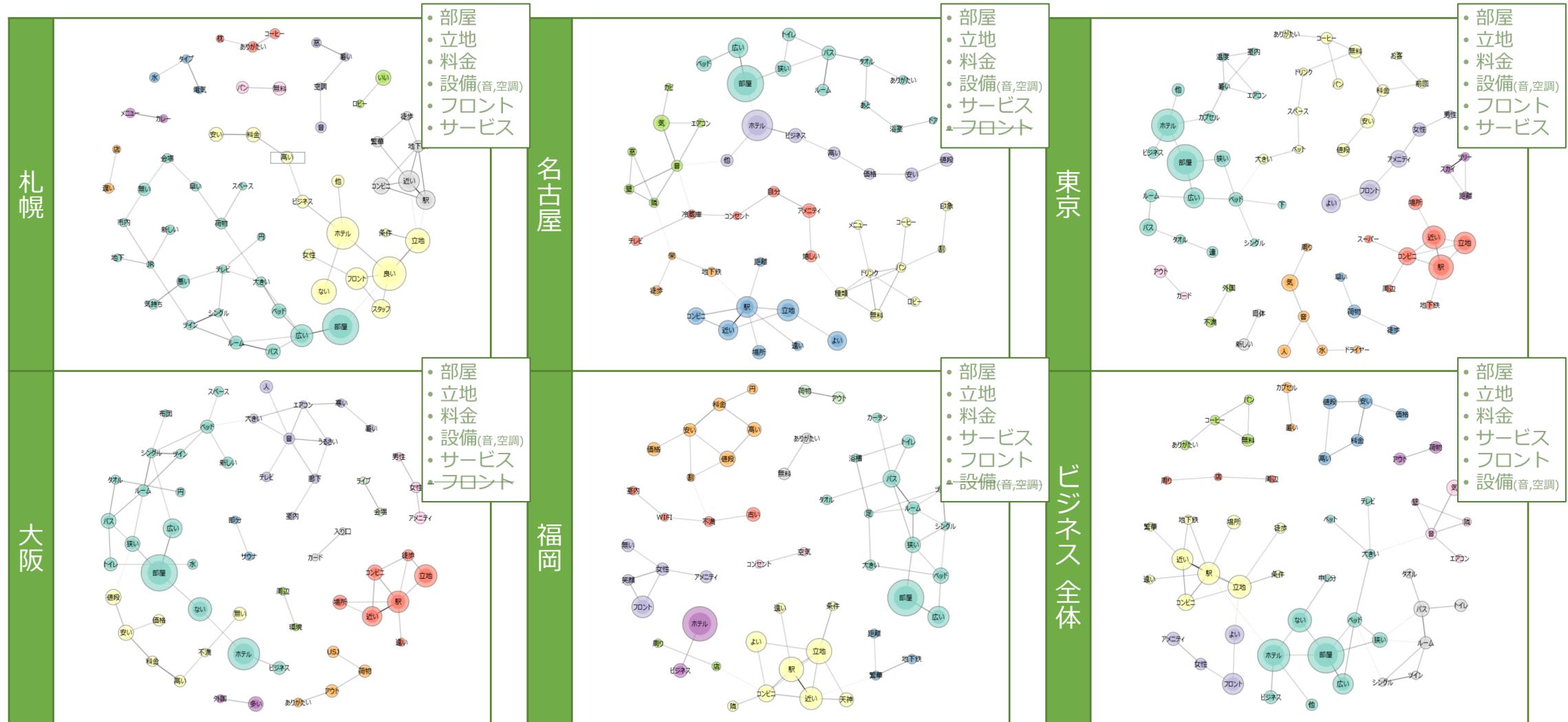
- ユーザーは,どの項目に注目しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. カテゴリー「レジャー」(or 「ビジネス」)の5エリアを比較する
- 手順
 - 特徴語の共起ネットワーク図を作成

「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:文」→「フィルタ設定」→「品詞=名詞,未知語,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60,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

出力例 – 特徴語の共起ネットワーク(1)



出力例 – 特徴語の共起ネットワーク(2)



議論1

- ・ユーザーがどの項目に注目しているかを議論する
 - ・カテゴリー「レジャー」と「ビジネス」の対比
 - ・「レジャー」5エリアの対比
 - ・「ビジネス」5エリアの対比

参考 – 数値評価の平均

- ・ カテゴリー「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	4.08	4.16	3.97	3.89	4.16	4.16	4.16
B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06

- ・ エリア別

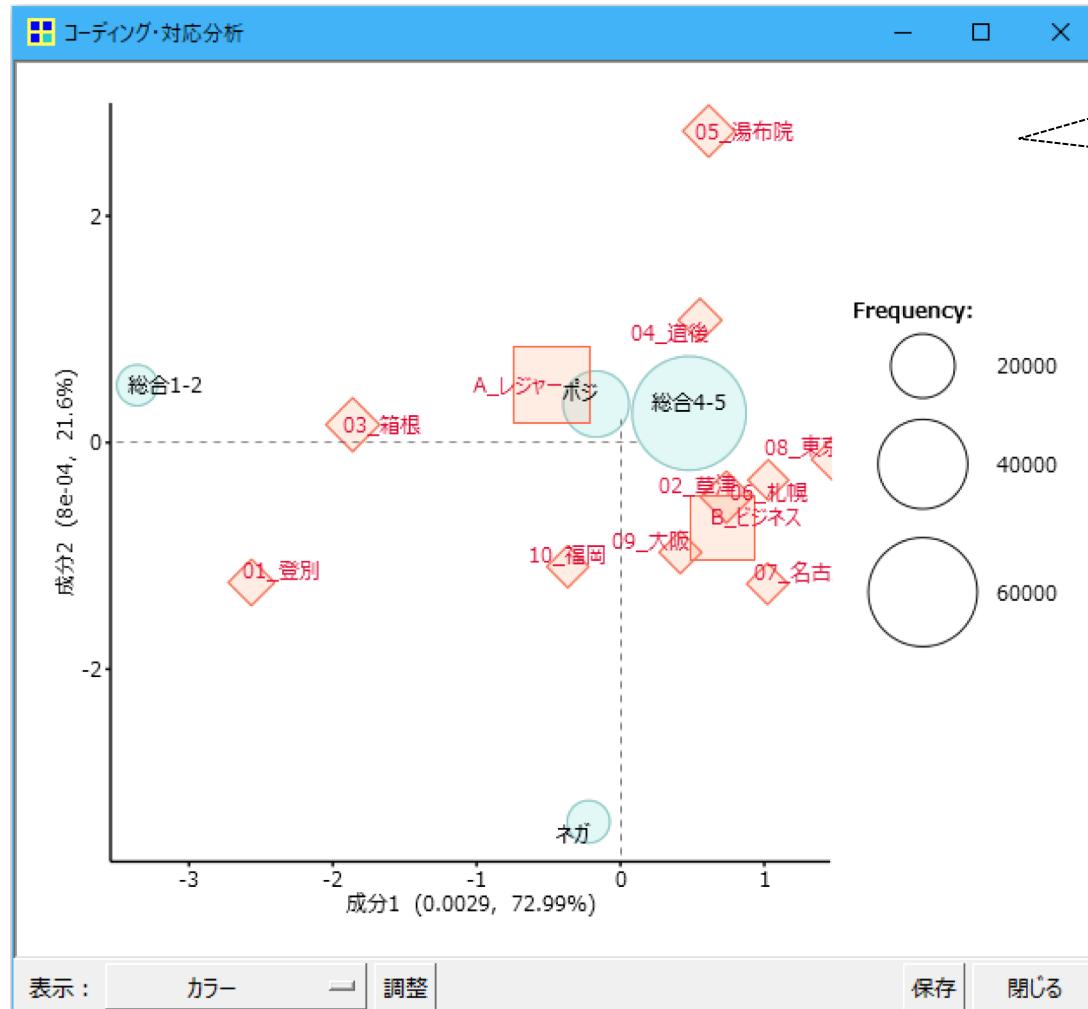
行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.08	4.16	3.97	3.89	4.16	4.16	4.16
01_登別	3.83	4.12	3.71	3.69	4.16	3.97	4.00
02_草津	4.11	4.23	3.90	3.79	4.22	4.12	4.17
03_箱根	4.09	4.01	4.00	3.88	4.10	4.16	4.07
04_道後	3.93	4.21	3.89	3.82	3.92	3.98	4.08
05_湯布院	4.44	4.21	4.37	4.26	4.40	4.51	4.47
B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06
06_札幌	3.98	4.22	3.96	3.86	3.78	3.91	4.10
07_名古屋	3.88	4.17	3.88	3.75	3.59	3.85	4.00
08_東京	3.94	4.38	3.99	3.87	3.68	3.91	4.12
09_大阪	3.89	4.25	3.94	3.77	3.71	3.93	4.07
10_福岡	3.86	4.21	3.86	3.71	3.57	3.83	3.99

実践 – エリアの改善案を提案する

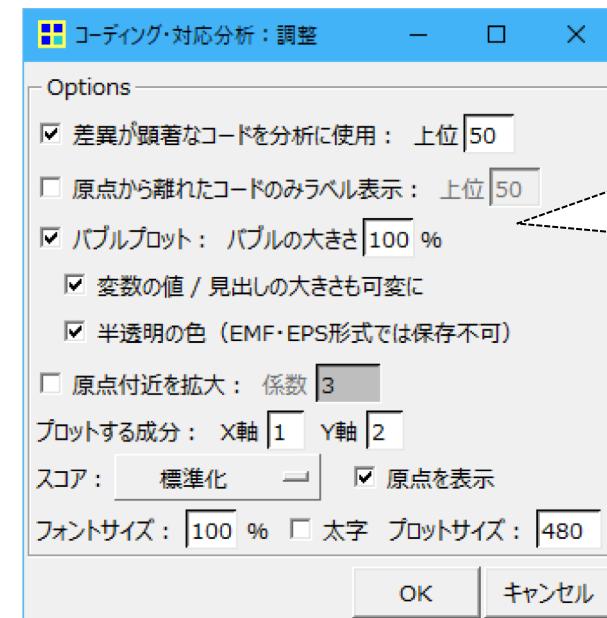
- ・対称的な2エリアを選択してポジティブ/ネガティブの両方の意見から,比較先エリアと比較し,改善案を議論
- ・主張を支持する図とユーザーの生の声(原文)を使って説明する
- ・手順
 - ・「数値評価の総合点」および「ポジティブ/ネガティブの両方の意見」から対照的な2エリアを選択(対応分析)
 - ・対象エリアについて,ポジティブ/ネガティブの両方の意見から,比較先エリアと比較し,改善すべき点を考察する(共起ネットワーク)

「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<>エリア-->01_登別”」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,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=120」

出力例 – 対称的なエリアを見つける



① 「ツール」 → 「コーディング」 → 「対応分析」 → 「コーディング単位:文」「コード選択: *ポジ,*ネガ,*総合1-2,*総合4-5」「コードx外部変数: カテゴリー,エリア」



② 「調整」をクリックして「バブルプロット」をチェック

演習 – ポジティブ意見の共起NW

- ・ユーザーは何をどう評価しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する

- ・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」→「抽出語」→「関連語検索」→「#直接入力(and)“<>エリア-->01_登別”」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,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=120,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

演習 – ネガティブ意見の共起NW

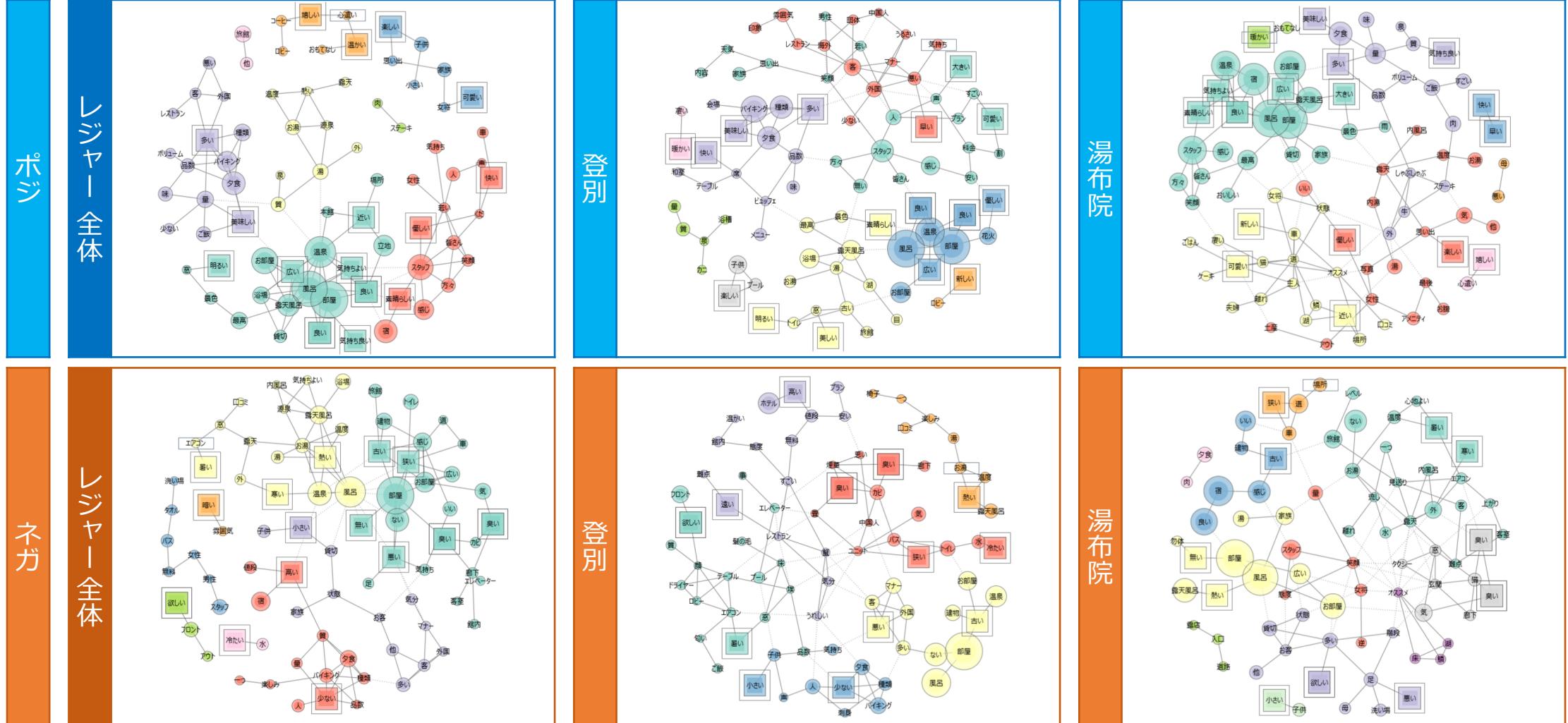
- ・ユーザーは何をどう評価しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する

- ・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」→「抽出語」→「関連語検索」→「#直接入力(and)“<>エリア-->01_登別”」「Search Entry:*ボジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,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=120,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

出力例 – 登別と湯布院のポジネガ比較



議論2

- ・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

補足1 – KH Coder で単語登録する

- 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」「畳」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

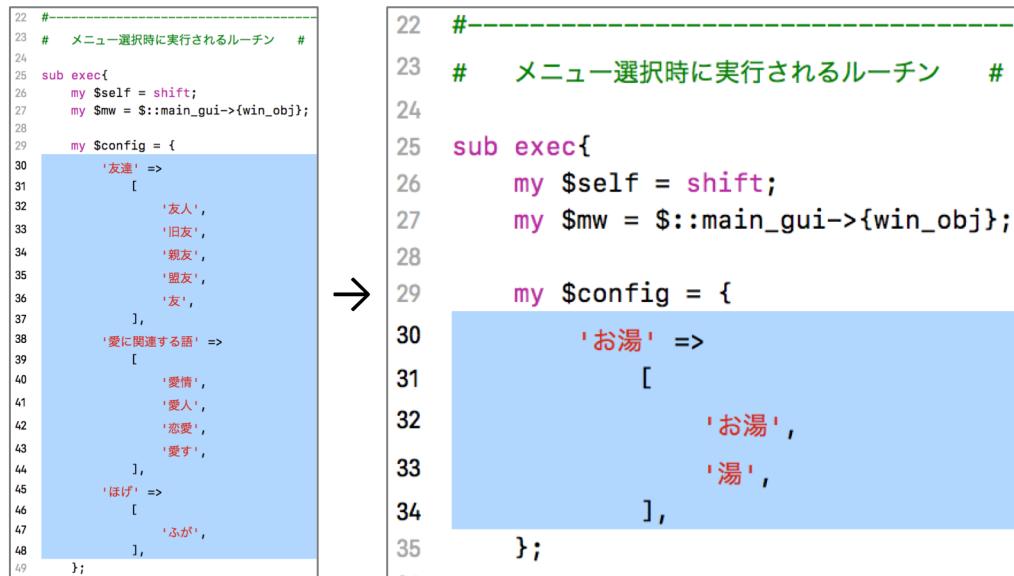
補足2 – KH Coder で同義語登録する (1/2)

- 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] http://koichi.nihon.to/psnl/tmp/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)

補足2 – KH Coder で同義語登録する (2/2)

- 手順

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



The image shows two code snippets side-by-side, separated by a right-pointing arrow. The left snippet is labeled "編集前" (before edit) and the right one is labeled "編集後" (after edit). Both snippets are in Perl syntax. The code defines a subroutine `exec` which contains several hash references (`$config`) mapping strings to arrays of words. In the "before edit" version, the arrays for each key contain multiple words. In the "after edit" version, the array for the key 'お湯' contains only the word 'お湯', while the array for '湯' contains only the word '湯'. This indicates that the script has been modified to merge the two words into a single entry.

```
22 #-----  
23 # メニュー選択時に実行されるルーチン #  
24  
25 sub exec{  
26     my $self = shift;  
27     my $mw = $::main_gui->{win_obj};  
28  
29     my $config = {  
30         '友達' =>  
31         [  
32             '友人',  
33             '旧友',  
34             '親友',  
35             '盟友',  
36             '友',  
37         ],  
38         '愛に関連する語' =>  
39         [  
40             '愛情',  
41             '愛人',  
42             '恋愛',  
43             '愛す',  
44             '1',  
45             'ほげ' =>  
46             [  
47                 'ふが',  
48             ],  
49     };  
50 };
```

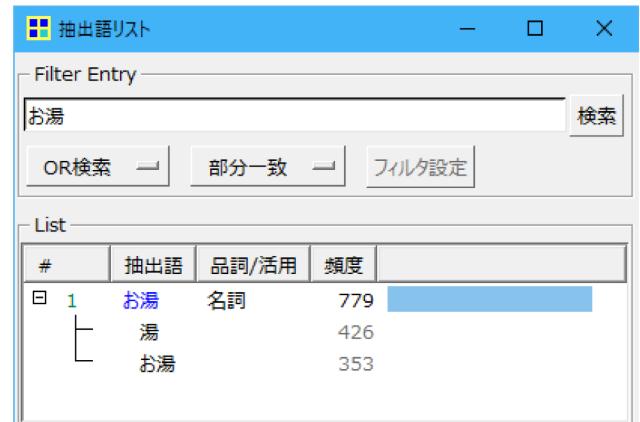
```
22 #-----  
23 # メニュー選択時に実行されるルーチン #  
24  
25 sub exec{  
26     my $self = shift;  
27     my $mw = $::main_gui->{win_obj};  
28  
29     my $config = {  
30         'お湯' =>  
31         [  
32             'お湯',  
33             '湯',  
34         ],  
35     };  
36 }
```

↓

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



参考書

(Rを使った参考書)

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

- [6] 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.
- [7] 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)

- [8] 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.

参考書

(KH Coder)

- [1] 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－. ナカニシヤ出版, 京都, 2014.
- [2] 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.

(Windows環境によるCGM収集の参考に)

- [3] 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/%E7%AC%AC10%E5%88%86%E7%A7%91%E4%BC%9A_1.pdf