

テキストマイニングの実習

— 3日目 —

2018/7/18

ビジネス科学研究科
経営システム科学専攻

講義スライド

- <https://github.com/haradatm/lecture/tree/master/gssm-201907>



前回 – KHCoderの主な分析手法

分析手法	解説	データ表	距離尺度
階層的クラスター分析	<ul style="list-style-type: none"> 出現パターンの似た単語をクラスタリングしたもの 出現パターンは,ある単語がどの文書に出現したかといった単語ベクトルで表現 類似度計算には Jaccard, ユークリッド, コサイン距離を使い, いわゆる Ward法, 群平均法, 最遠隣法で樹形図を作成 	「文書-抽出語」表 ↓ 抽出語-文書	<ul style="list-style-type: none"> Jaccard コサイン ユークリッド
多次元尺度構成(MDS)	<ul style="list-style-type: none"> 出現パターンの似た単語を近くに置くよう図示したもの 出現パターンは,ある単語がどの文書に出現したかといった単語ベクトルで表現 類似度計算には Jaccard, ユークリッド, コサイン距離を使い, クラシカル, Kruskal, Sammon 法のいずれかで2次元にプロット 	「文書-抽出語」表 ↓ 抽出語-文書	<ul style="list-style-type: none"> Jaccard コサイン ユークリッド
対応分析	<ul style="list-style-type: none"> 出現パターンの似た単語や外部変数を近くに置くよう図示したもの 単語と単語または外部変数が同時に出現した頻度をクロス集計し, それぞれの相関が最大になるような2変数で数値化し, 2軸上にプロット (PCAが元の情報を<u>そのまま可視化</u>するのに対し, 対応分析は似ているものを近くに表示する) 外部変数も同時にプロット可能 	「文書-抽出語」表 ↓ 外部変数-抽出語 (クロス集計)	• χ^2
共起ネットワーク	<ul style="list-style-type: none"> 同時に出現した単語間をネットワークで結んで図示したもの 同時に出現したかといった共起の有無を集計し, ネットワークを作成 関係の強さ Jaccard 係数で評価, サブグラフは媒介性, クラスタリング精度(エッジ内の密度の高さ)を使って検出 	「文書-抽出語」表 ↓ 抽出語-文書	<ul style="list-style-type: none"> Jaccard コサイン ユークリッド
自己組織化マップ	<ul style="list-style-type: none"> 出現パターンの似た単語を近くに集めて図示したもの ニューラルネットワークを利用して近い単語を集めることで, 距離にはユークリッド距離を使い, クラスタリングは Ward法 	「文書-抽出語」表 ↓ 抽出語-文書	• ユークリッド
文書のクラスター分析	<ul style="list-style-type: none"> 似た文書同士をクラスタリングしたもの 各文書は, 文書中に出現する単語の有無でベクトル化した文書ベクトルで表現 類似度計算には Jaccard, ユークリッド, コサイン距離を使い, いわゆる Ward法, 群平均法, 最遠隣法で階層クラスタを作成 	文書-抽出語	<ul style="list-style-type: none"> Jaccard コサイン ユークリッド

前回 – 「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

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
1	1	0	1	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	

前回 - 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)

「抽出語-文書」頻度表(対応分析以外)

「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

前回 - KH Coder で使われる距離尺度(1)

- KH Coder では Jaccard 距離を多用
 - 語Aと語Bのどちらも出現していない文書(0-0対)が沢山あっても語Aと語Bが類似しているとは見なさない → **スパースなデータ分析向き**

Jaccard 距離	コサイン距離	ユークリッド距離									
<ul style="list-style-type: none">• 1つ文書に含まれる語が少なく、各語が一部の文書中にしか含まれていないスパースデータ向き• 1つの文書の中に語が1回出現した場合も10回出現した場合も単に「出現あり」と見なしてカウントした語と語の共起数を計算	<ul style="list-style-type: none">• 1つひとつの文書が長く、多数の文書に含まれている語が多いデータ向き(各文書中での語の出現回数の大小が重要な場合)• 文書中における語の出現回数(1,000語あたりの出現回数に調整)を計算	<ul style="list-style-type: none">• 増減傾向が似ているかどうかだけを見る場合向き• サイズの差までも見る場合向き									
<table border="1"><tr><td></td><td>1</td><td>0</td></tr><tr><td>1</td><td>n_{11}</td><td>n_{10}</td></tr><tr><td>0</td><td>n_{01}</td><td>n_{00}</td></tr></table> $J\text{ }S = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$		1	0	1	n_{11}	n_{10}	0	n_{01}	n_{00}	$\text{cos } s(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$	$E d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$
	1	0									
1	n_{11}	n_{10}									
0	n_{01}	n_{00}									

<http://mjin.doshisha.ac.jp/R/68/68.html>

前回 - KH Coder で使われる距離尺度(2)

- 対応分析では χ^2 距離を用いる
 - χ^2 距離は、カテゴリー変数間の関連性を測定

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

「観測度数」 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」 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

- 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くなることを示す

前回 - 対応分析

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

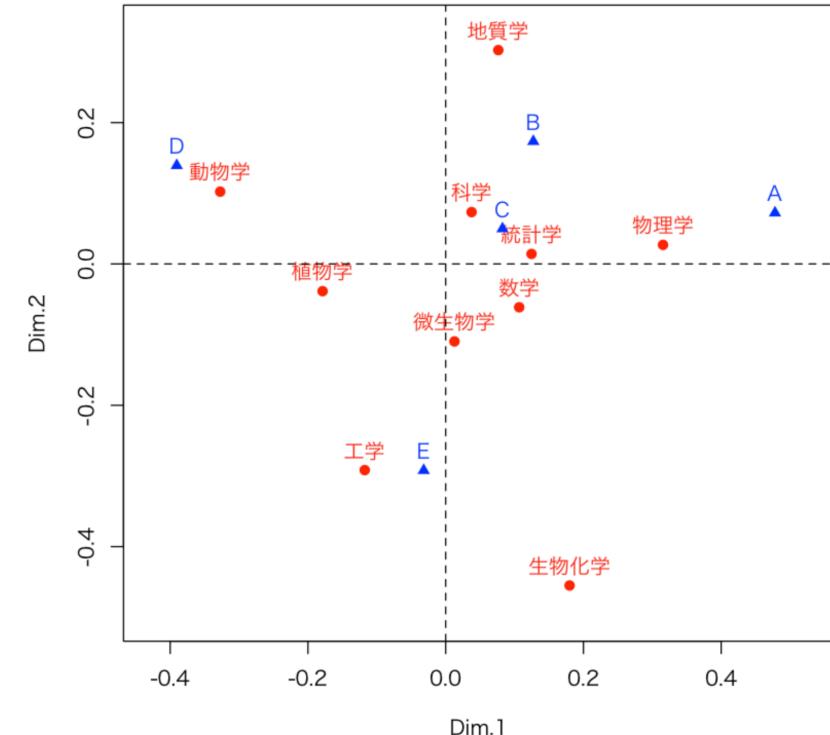
	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410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

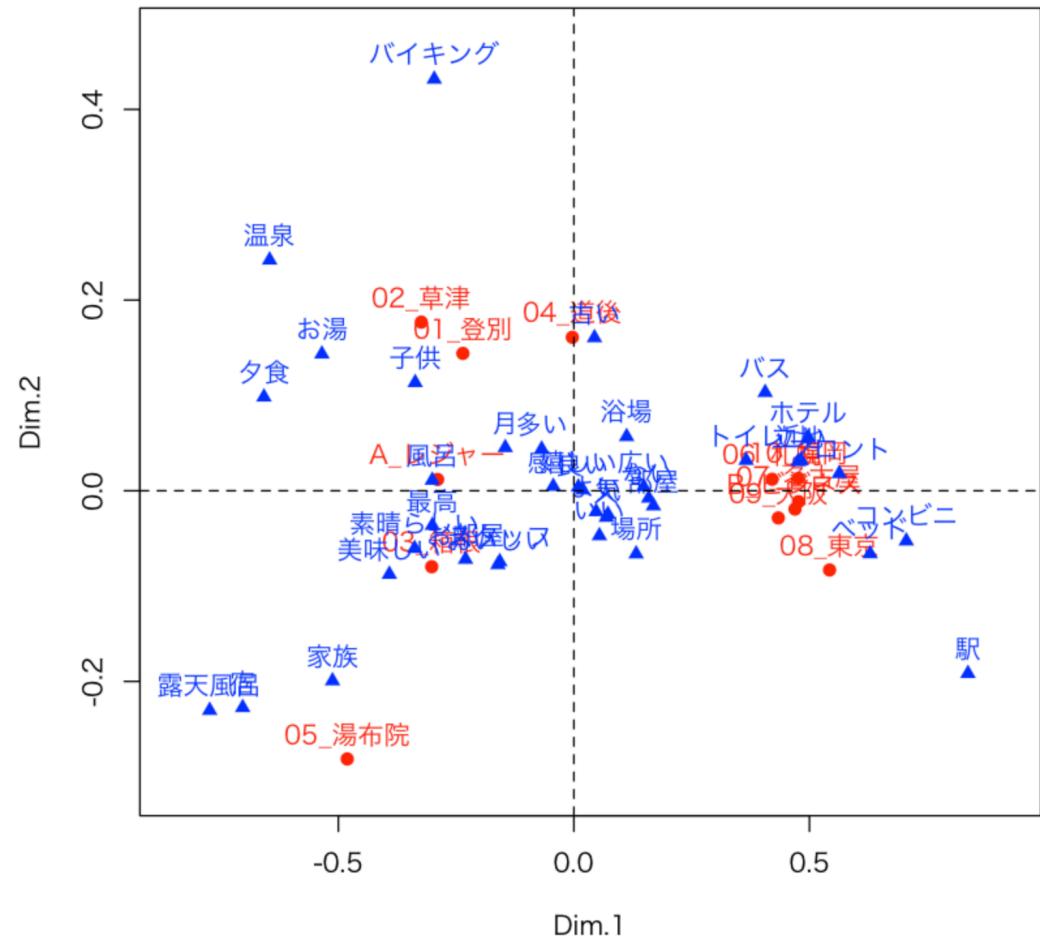


特異値分解してプロット

固有値 = {0.0391, 0.0304, 0.0109, 0.0025, 0}
 寄与率 = {47.2%, 36.66%, 13.11%, 3.03%, 0%}

前回 - 対応分析の読み方

対応分析



- 原点からの距離
遠いほど影響力が大きい
 - 原点からの向き
両端にあるほどカテゴリを対比
 - プロット(点)
近くにある関連性が高い

固有值 = {0.1426, 0.0107, 0.0061, 0.0022, 0.0014, 0.001, 6e-04, 6e-04,
 2e-04, 0, 0, 0}
 寄与率 = {86.14%, 6.46%, 3.71%, 1.36%, 0.84%, 0.63%, 0.39%, 0.34%,
 0.14%, 0%, 0%, 0%}

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/lecture/blob/master/gssm-201808/03-samples/practice-3_sample.ipynb

スケジュール

- 1日目: 7/4
 - 説明 – データ分析の手順
 - 演習 – データの理解 (Excel)
- 2日目: 7/11
 - 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: 7/18
 - 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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05

- ・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, 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」
 - 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

直接入力: [and] の右側に入力する条件

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

集計例 – 特徴語の集計

A_レジヤー	数値評価指標
風呂	.073
温泉	.061
宿	.044
お部屋	.040
スタッフ	.038
露天風呂	.030
よい	.028
夕食	.028
最高	.026
家族	.019

B_ビジネス	数値評価指標
部屋	.106
ホテル	.090
ない	.049
立地	.044
フロント	.043
駅	.040
バス	.023
コンビニ	.021
浴場	.020
ベッド	.019

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
風呂	.056	湯畑	.071	温泉	.054	温泉	.056	宿	.069
温泉	.043	温泉	.065	温泉	.040	部屋	.053	風呂	.057
ない	.035	風呂	.060	露天風呂	.038	ホテル	.042	露天風呂	.040
スタッフ	.030	宿	.044	お部屋	.038	立地	.034	温泉	.040
バイキング	.030	お部屋	.033	スタッフ	.037	よい	.025	お部屋	.039
夕食	.025	湯	.031	宿	.036	浴場	.022	スタッフ	.037
最高	.022	夕食	.030	夕食	.026	本館	.020	最高	.031
子供	.022	バイキング	.026	感じ	.020	バイキング	.019	家族	.029
露天風呂	.021	よい	.024	浴場	.020	感じ	.017	よい	.026
浴場	.021	最高	.024	最高	.019	夕食	.017	夕食	.024

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
部屋	.059	ホテル	.057	ホテル	.054	部屋	.055	ホテル	.064
ホテル	.055	部屋	.055	部屋	.052	ホテル	.053	部屋	.057
ない	.037	駅	.034	駅	.048	ない	.039	立地	.037
立地	.036	フロント	.034	ない	.034	フロント	.037	フロント	.030
フロント	.033	ない	.033	立地	.034	立地	.036	駅	.030
駅	.022	立地	.028	フロント	.033	駅	.034	バス	.028
バス	.022	よい	.023	コンビニ	.024	気	.019	よい	.024
コンビニ	.017	アメニティ	.021	よい	.021	浴場	.018	トイレ	.021
浴場	.017	コンビニ	.020	バス	.019	バス	.018	コンビニ	.019
ベッド	.016	ベッド	.017	浴場	.018	ベッド	.018	ベッド	.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,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

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,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

直接入力: [and] の右側に入力する条件

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

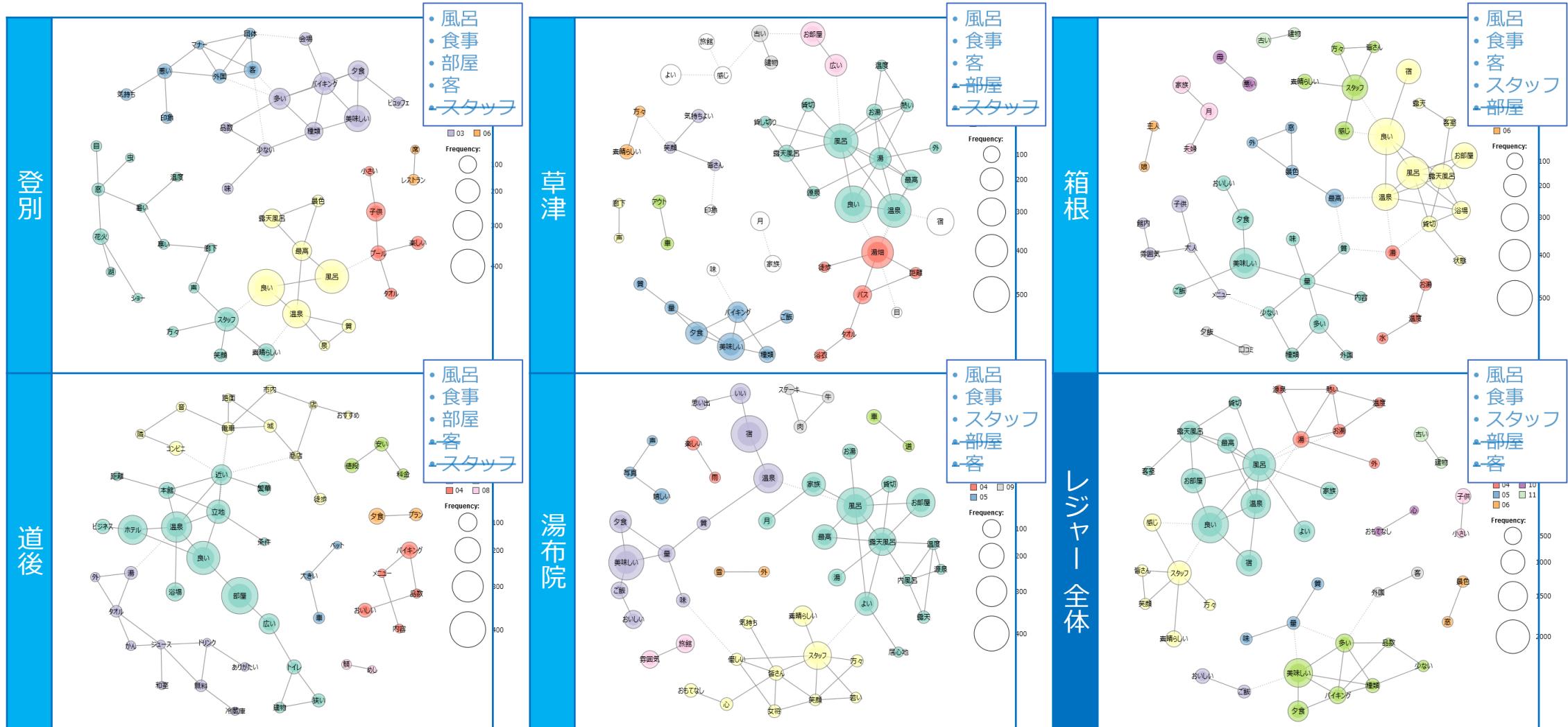
<>エリア-->07_名古屋

<>エリア-->08_東京

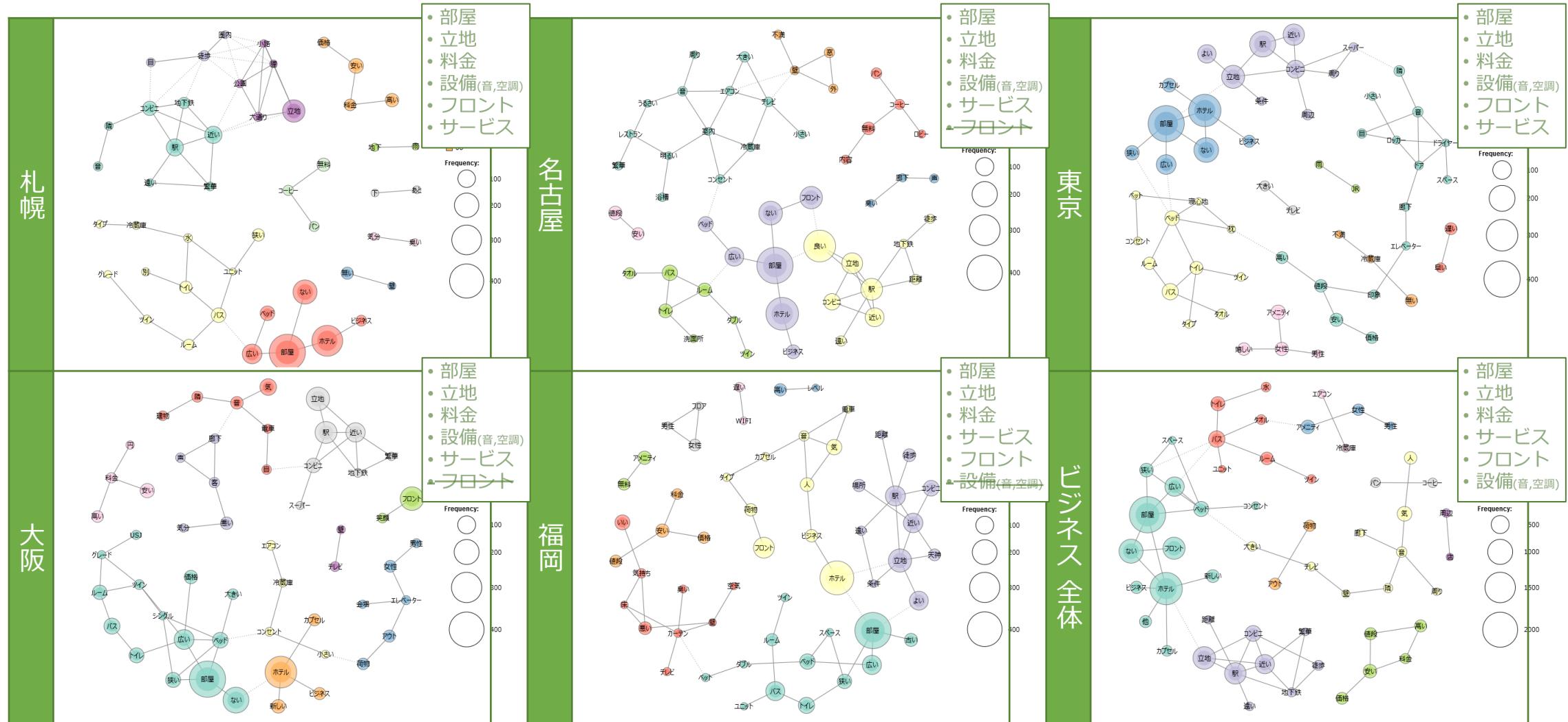
<>エリア-->09_大阪

<>エリア-->10_福岡

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)



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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05

- ・ 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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
01_登別	3.87	4.13	3.82	3.78	4.22	3.94	4.00
02_草津	4.18	4.27	4.04	3.91	4.30	4.16	4.25
03_箱根	4.18	4.10	4.05	3.97	4.16	4.27	4.18
04_道後	4.03	4.28	4.00	3.89	3.97	4.12	4.17
05_湯布院	4.50	4.27	4.38	4.28	4.46	4.60	4.51
B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05
06_札幌	3.91	4.19	4.00	3.83	3.73	3.92	4.10
07_名古屋	3.85	4.11	3.95	3.81	3.71	3.84	4.03
08_東京	3.85	4.28	3.94	3.76	3.64	3.89	4.01
09_大阪	3.88	4.33	3.96	3.83	3.72	3.96	4.10
10_福岡	3.88	4.19	3.89	3.80	3.70	3.89	4.00

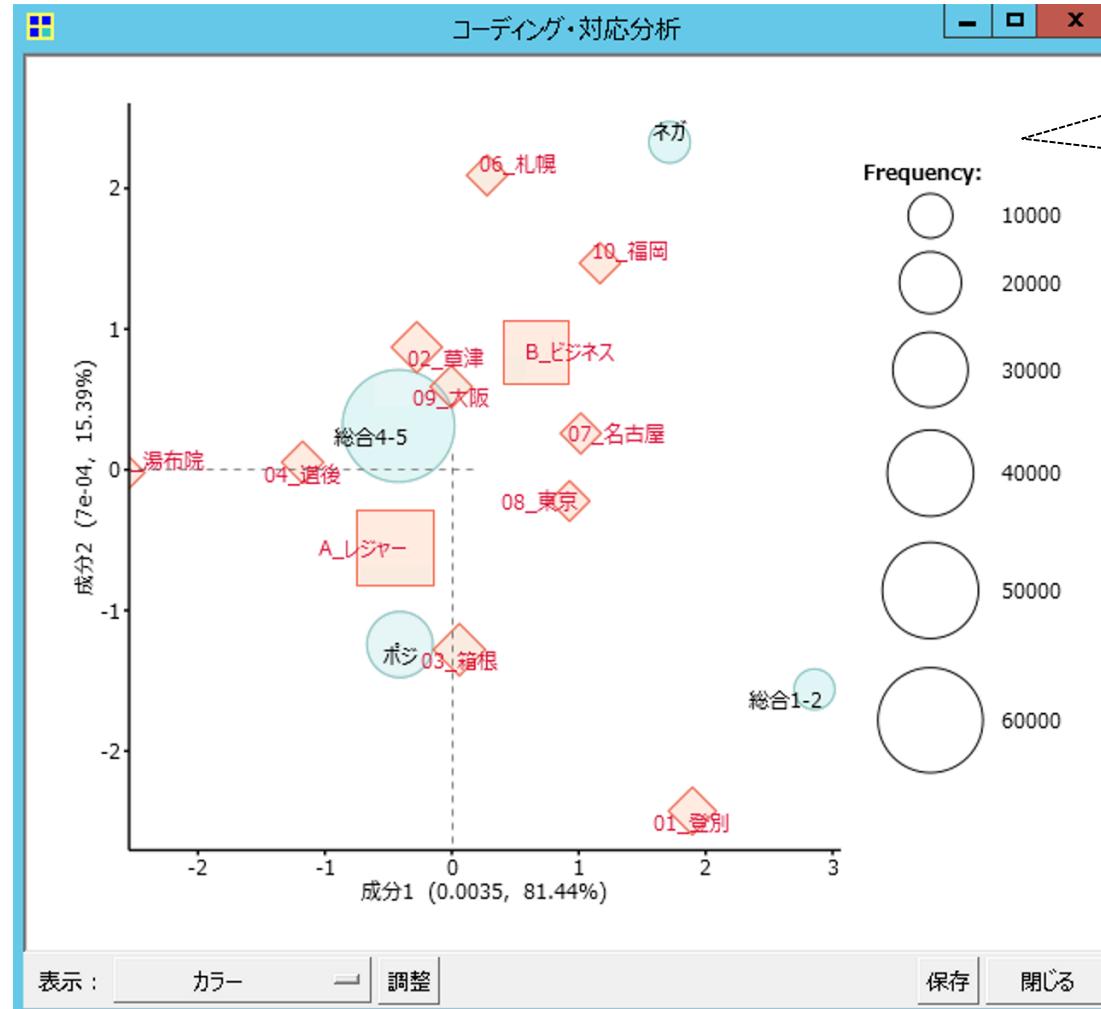
討論1

- ・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

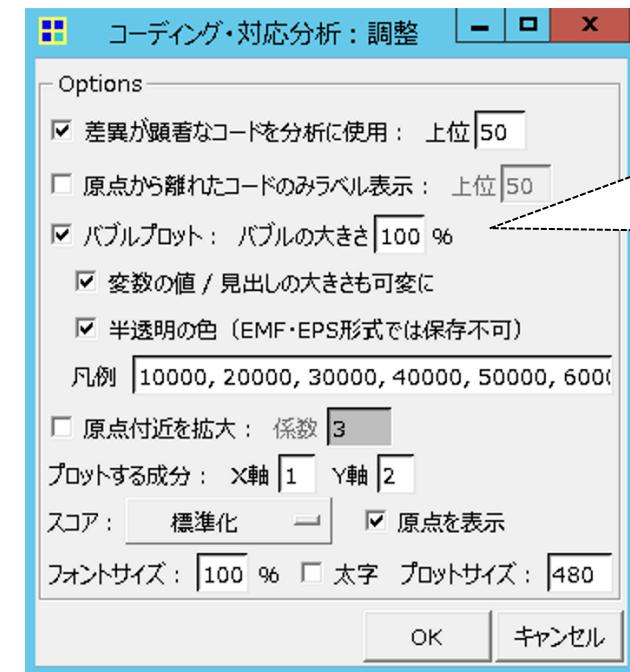
実践 – エリアの改善案を提案する

- ・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1
 - ・「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)
- ・手順2
 - ・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)

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

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1
 - ・「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)
- ・手順2
 - ・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)

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

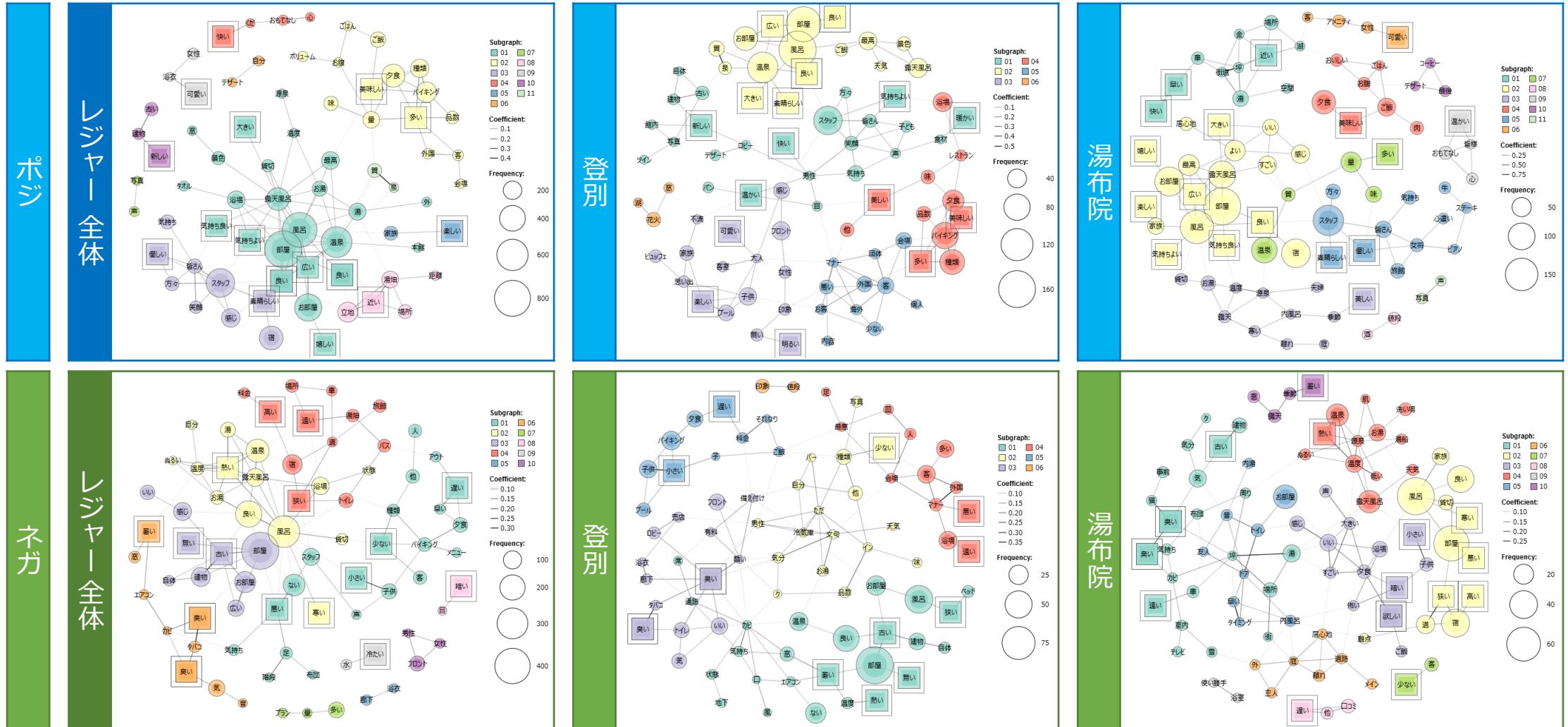
- ・ユーザーは何をどう評価しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する

- ・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

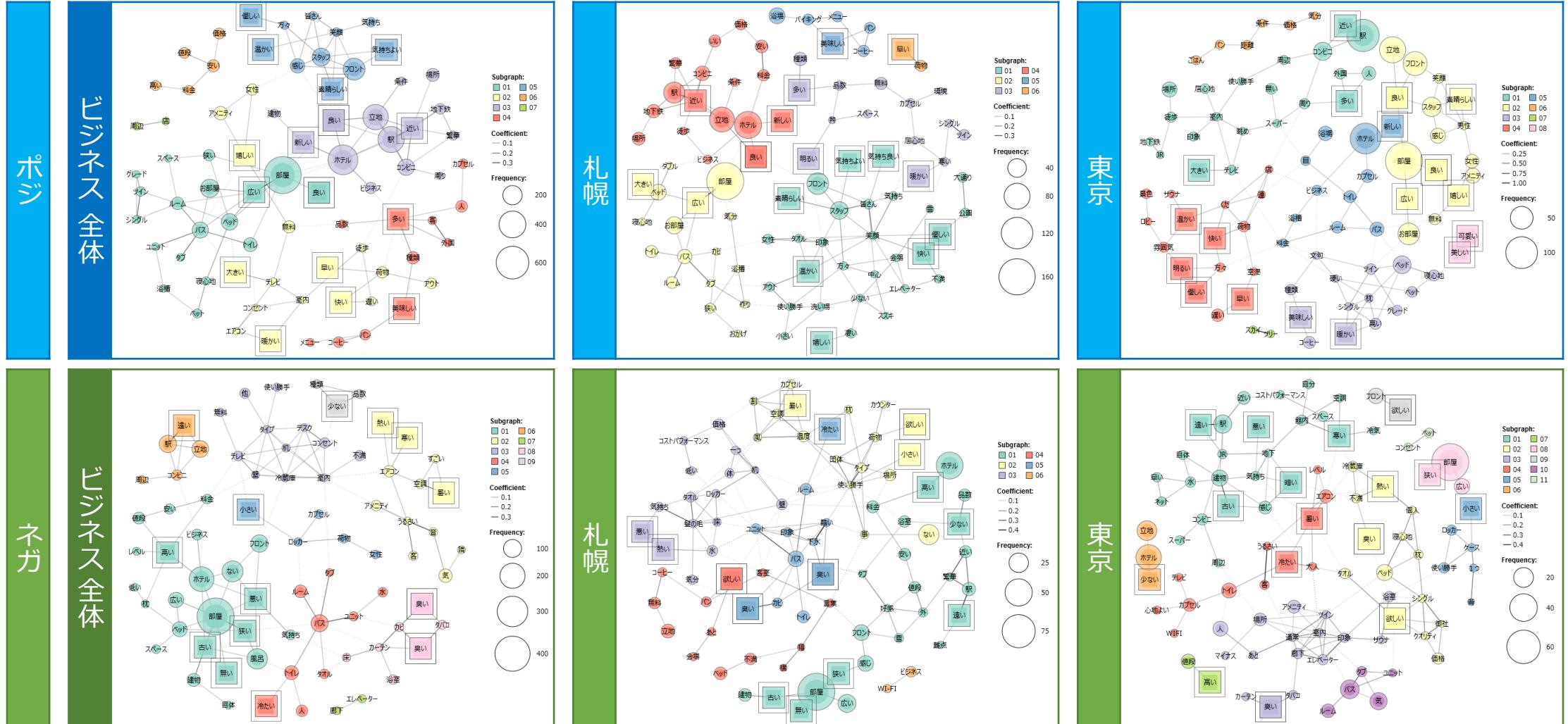
「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<>エリア-->01_登別”」「Search Entry:*ボジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:上位=120,共起関係ほど濃い線に」

- ・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



出力例 – 東京と札幌のポジネガ比較



討論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

Tips 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」「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

Tips 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)

Tips 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

```
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         ],  
45         'ほげ' =>  
46         [  
47             'ふが',  
48         ],  
49     };  
};
```



編集前

```
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             '湯',  
33         ],  
34     };  
};
```

編集後

- 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

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, 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.
- New [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CGM収集の参考に)

- [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/%E7%AC%AC10%E5%88%86%E7%A7%91%E4%BC%9A_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.