

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 5

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&演習 – データ分析 (実践編)

KHCoder よくある質問

- Q1. 表記ゆれを統一したいときは?
- Q2. 単語登録したいときは?
- Q3. 「条件付き確率が同等ないし低下する語も表示」とは?
- Q4. 対応分析の軸の値って何?
- Q5. その他

● Q1. 表記ゆれを統一したいときは?

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

KHCoder よくある質問

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

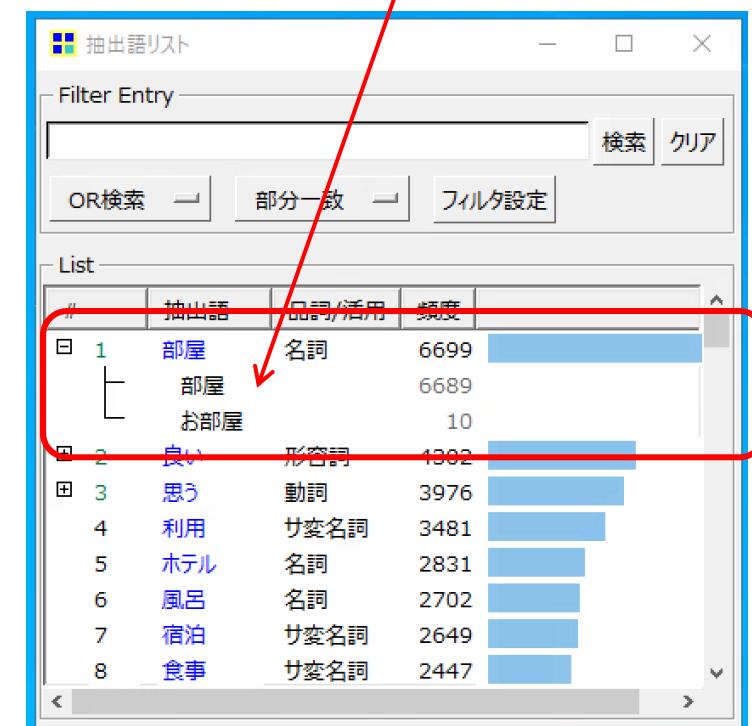
```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '友達' =>
6         [
7             '友人',
8             '旧友',
9             '親友',
10            '盟友',
11            '友',
12        ],
13        '格別' =>
14        [
15            '特別',
16            '格別', # 通常
17        ],
18        '# 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

編集後

```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '部屋' =>
6         [
7             'お部屋',
8         ],
9 };
```

→

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

● Q2. 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

● Q3. 「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 the KHCoder interface with two windows open:

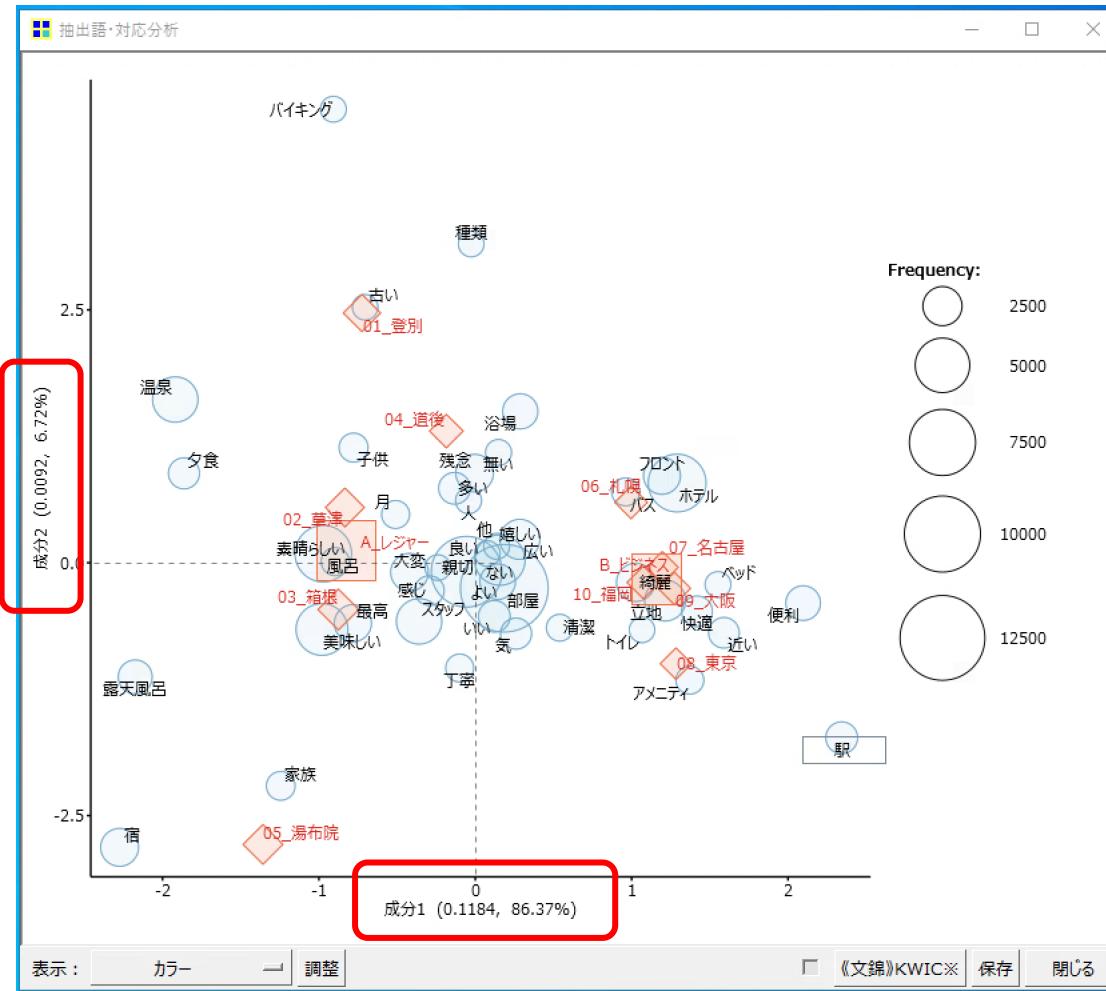
- Main Window (Left):** Shows a search entry of "#直接入力" and a result table. The table has columns: N, 抽出語, 品詞, 全体, 共起, Jaccard. Red boxes highlight the "前提確率" (Premise Probability) column and the "条件付き確率" (Conditional Probability) column. A red arrow points from the "条件付き確率" column to the "Filtra設定" (Filter Settings) button.
- Filter Dialog (Right):** Shows filter settings for word selection. It includes sections for Part of Speech (名詞, 形容動詞, etc.), document frequency (全体での出現数による語の選択), and top words (上位: 75). A checkbox at the bottom is labeled "条件付き確率が同等ないし低下する語も表示" (Also display words where conditional probability is equal to or lower than premise probability).

N	抽出語	品詞	全体	共起	Jaccard
1	風呂	名詞	2143 (0.214)	323 (0.323)	0.1145
2	温泉	名詞	1298 (0.130)	222 (0.222)	0.1069
3	美味しい	形容詞	1962 (0.196)	255 (0.255)	0.0942
4	良い	形容詞	3182 (0.318)	354 (0.354)	0.0925
5	バイキング	名詞	472 (0.047)	121 (0.121)	0.0896
6	残念	形容動詞	1045 (0.105)	148 (0.148)	0.0780
7	ない	形容詞B	1629 (0.163)	187 (0.187)	0.0766
8	夕食	名詞	678 (0.068)	118 (0.118)	0.0756
9	種類	名詞	454 (0.045)	102 (0.102)	0.0754

- 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ません
- 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**」ではないというのが KHCoder の考え方
- 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ます

KHCoder よくある質問

- Q4. 対応分析の軸の値って何?



- KHCoder の対応分析は R の MASS パッケージにある corresp 関数を使用し、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
 - 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 $86.37 + 6.72 = 93.09\%$ 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ことが分かる

● Q4. その他

Q: 樋口先生のチュートリアル, 夏目漱石の小説を一体どうやって読み込んだの?

A: 青空文庫 (著作権が消滅した作品や著者が許諾した作品のテキストを公開している電子図書館 → <https://www.aozora.gr.jp/>)で公開されています

01_登別	
風呂	.115
温泉	.107
美味しい	.094
良い	.093
バイキング	.090
残念	.078
ない	.077
夕食	.076
種類	.075
露天風呂	.074

Q: 左の最下位の0.074は非常に数値が低いため特徴とは言い難い…という解釈ですか?

A: データは、その収集方法や時期、掲載元の特性などの影響を受けるため、単独で数値の高(頻度や共起率)を議論するのではなく、エリア内/外で比較するなど、常に相対的に見るのがポイントです

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×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			
■ 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×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
B_ビジネス	4.00	4.34	4.10	3.92	3.82	4.06	4.19

(再掲) 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2 – ユーザー注目ポイントの評価を見る

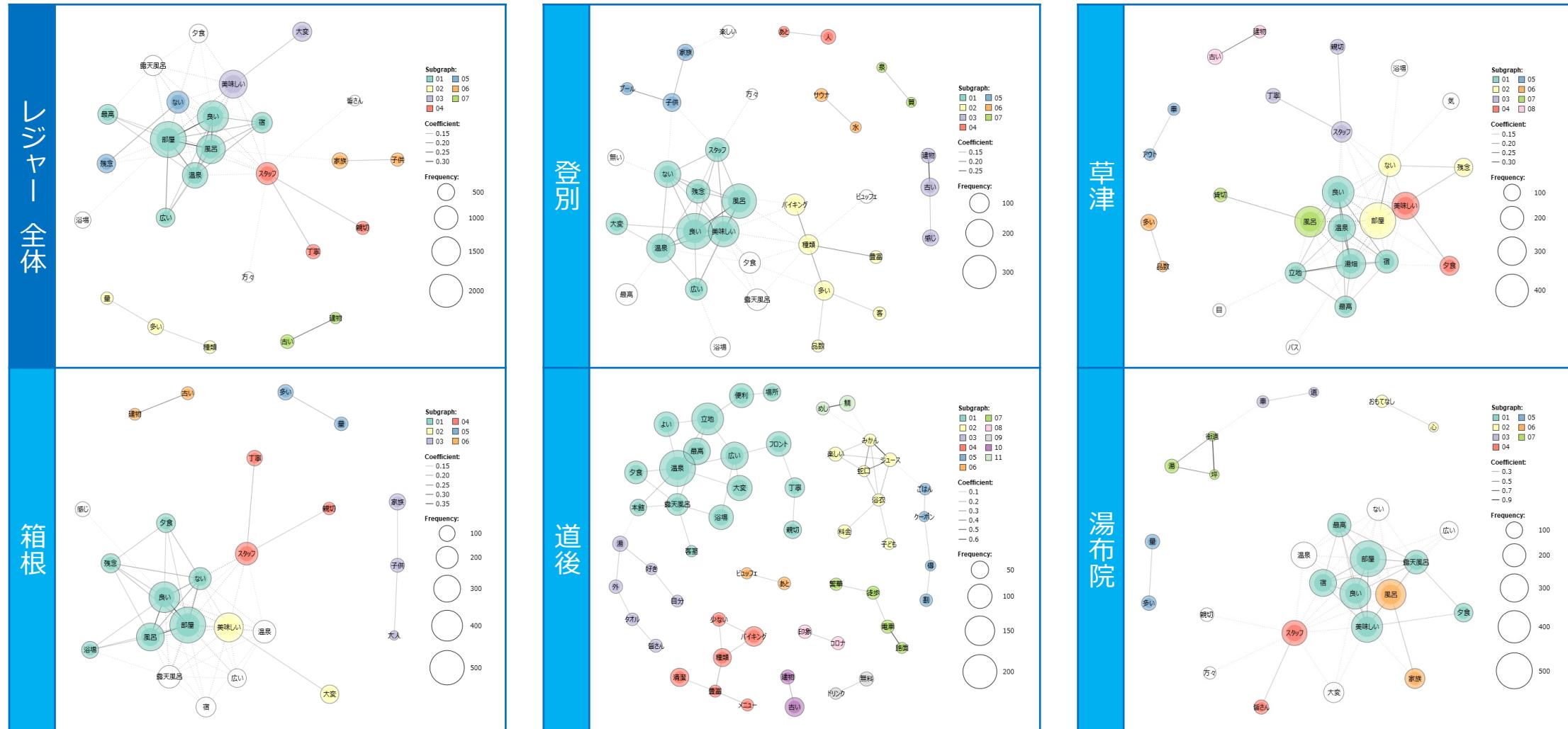
- カテゴリーやエリアごとでの注目する観点の評価の違いを確認する
 - カテゴリー「レジャー」と「ビジネス」を比較する
 - カテゴリー「レジャー」(or「ビジネス」) の 5エリアを比較する
- 手順の一例:
 - カテゴリーやエリアごとの共起NWから、どの観点をどう評価しているか調べる

「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]”<>カテゴリ-->A_レジャー”」「集計単位:H5」→「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」

「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,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」

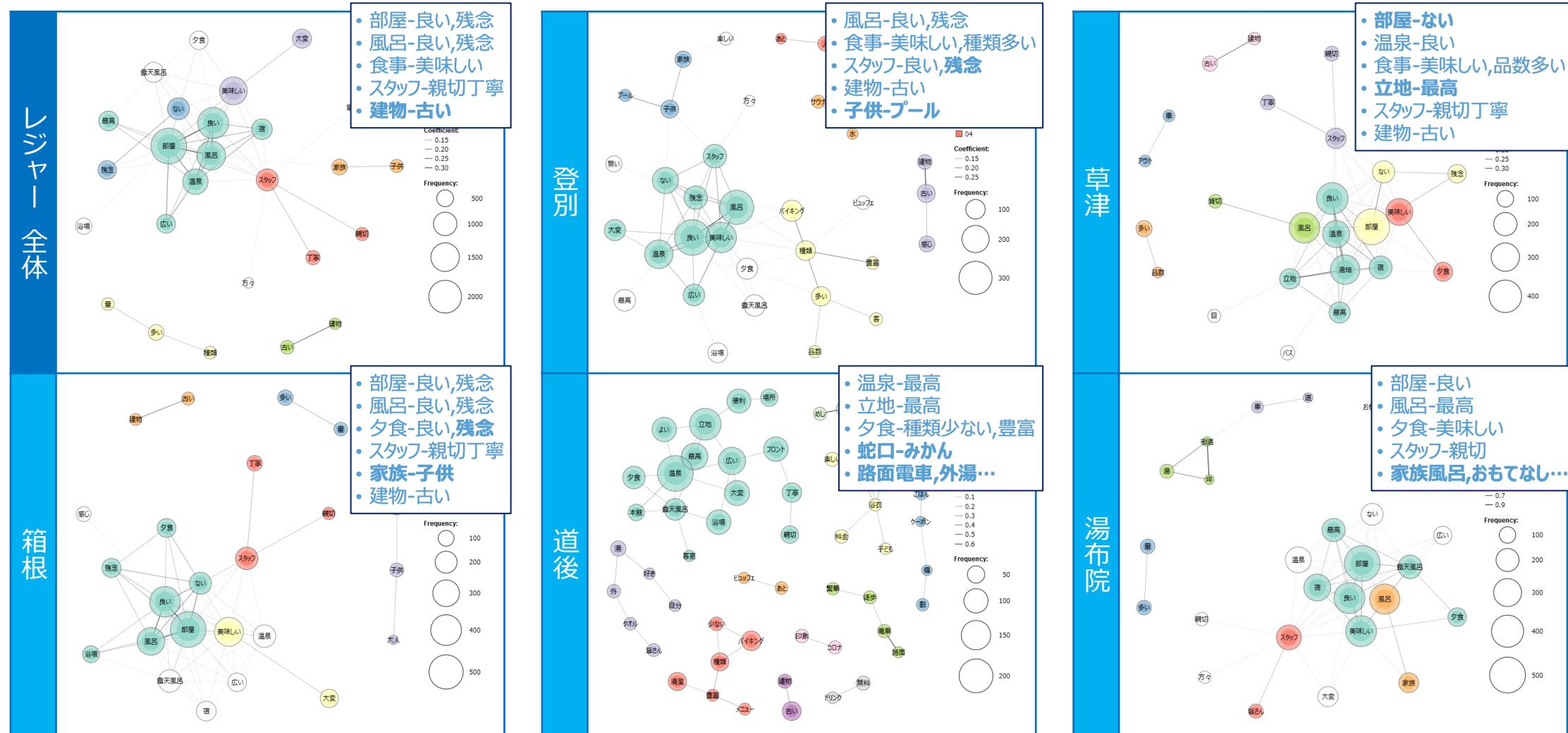
実践2 – ユーザー注目ポイントの評価を見る

● レジヤーとエリアごとの共起ネットワーク



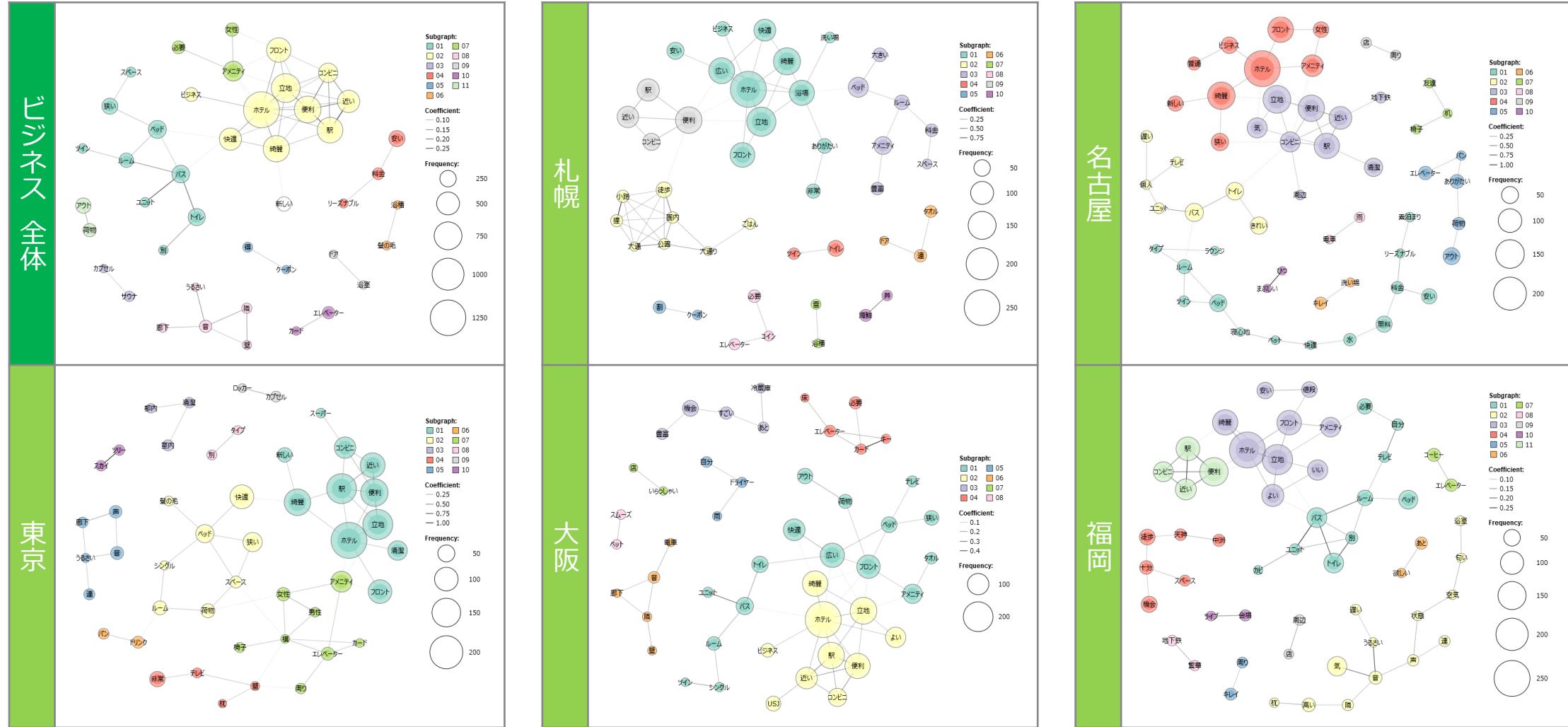
実践2 – ユーザー注目ポイントの評価を見る

● レジャーとエリアごとの共起ネットワーク



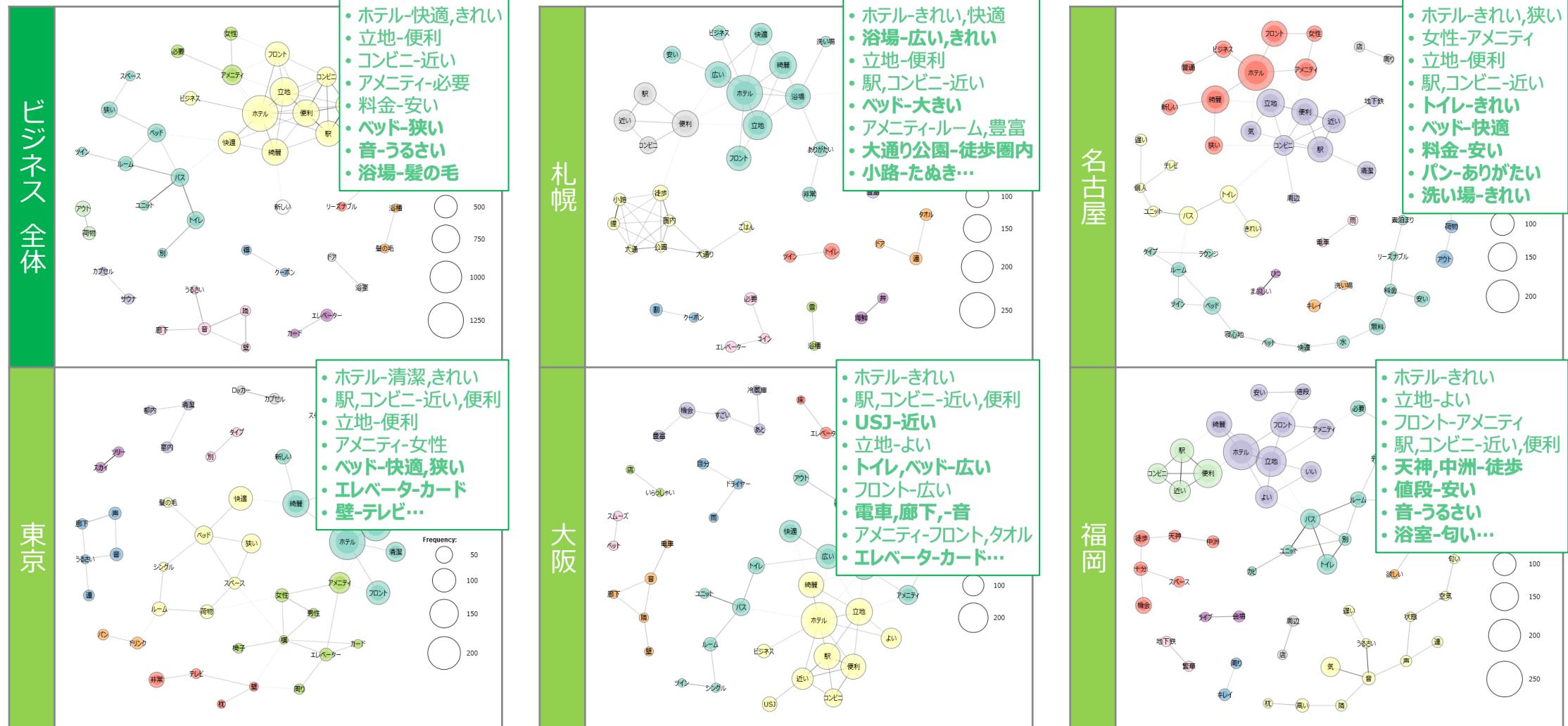
実践2 – ユーザー注目ポイントの評価を見る

● ビジネスとエリアごとの共起ネットワーク



実践2 – ユーザー注目ポイントの評価を見る

● ビジネスとエリアごとの共起ネットワーク



実践3 – 改善プランを提案する

- カテゴリーやエリアごとのポジ/ネガ意見の違いを確認する
 - 対照的な2エリアを選択する
 - 選択したエリアと、そのカテゴリー（「レジャー」or「ビジネス」）を比較する
- 手順の一例：
 - カテゴリーやエリアごとのコーディングや数値評価から、対照的な2エリアを選ぶ

対応分析を用いる例：

「ツール」→「コーディング」→「**対応分析**」→「コーディングルール・ファイル:**coding-rule_new2.txt**」「コーディング単位:**H5**」「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: **エリア**」※ **coding-rule_new2.txt** については次頁で後述

クロス集計を用いる例：

「ツール」→「コーディング」→「**クロス集計**」→「コーディングルール・ファイル:**coding-rule_new2.txt**」「コーディング単位:**H5**」「コード選択: ***ポジ, *ネガ, *総合1-2, *総合4-5**」→「集計」ボタンをクリック→「ヒート」ボタンをクリック※ **coding-rule_new2.txt** については次頁で後述

実践3 – 改善プランを提案する

- コーディングルールをテキストエディタで編集する（例：メモ帳、Notepad++）

coding-rule.txt の中身

*ポジ

良い or 美味しい or 広い or 多い or 素晴らしい or 嬉しい
or 気持ちよい or 楽しい or 近い or 大きい or 気持ち良い
or 温かい or 早い or 優しい or 新しい or 暖かい or 快い
or 明るい or 美しい or 可愛い

*ネガ

古い or 無い or 高い or 悪い or 小さい or 狹い or 少ない
or 寒い or 遅い or 熱い or 欲しい or 暑い or 冷たい or
遠い or 臭い or 暗い

*風呂1-2

<>風呂-->1 | <>風呂-->2

*風呂4-5

<>風呂-->4 | <>風呂-->5

coding-rule_new2.txt の中身

*ポジ

良い or 美味しい or 広い or 多い or 素晴らしい or 嬉しい
or 気持ちよい or 楽しい or 近い or 大きい or 気持ち良い
or 温かい or 早い or 優しい or 新しい or 暖かい or 快い
or 明るい or 美しい or 可愛い **or 満足**

*ネガ

古い or 無い or 高い or 悪い or 小さい or 狹い or 少ない
or 寒い or 遅い or 熱い or 欲しい or 暑い or 冷たい or
遠い or 臭い or 暗い **or 不満 or 残念**

*総合1-2

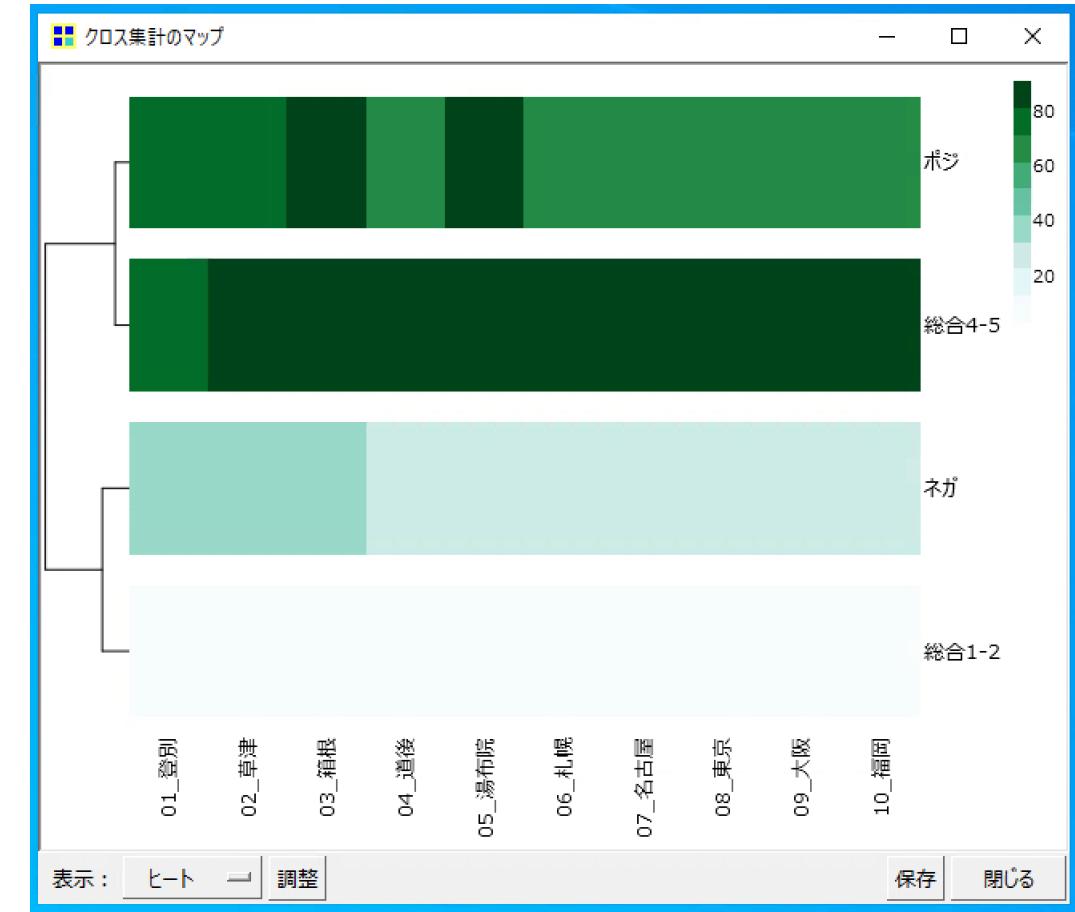
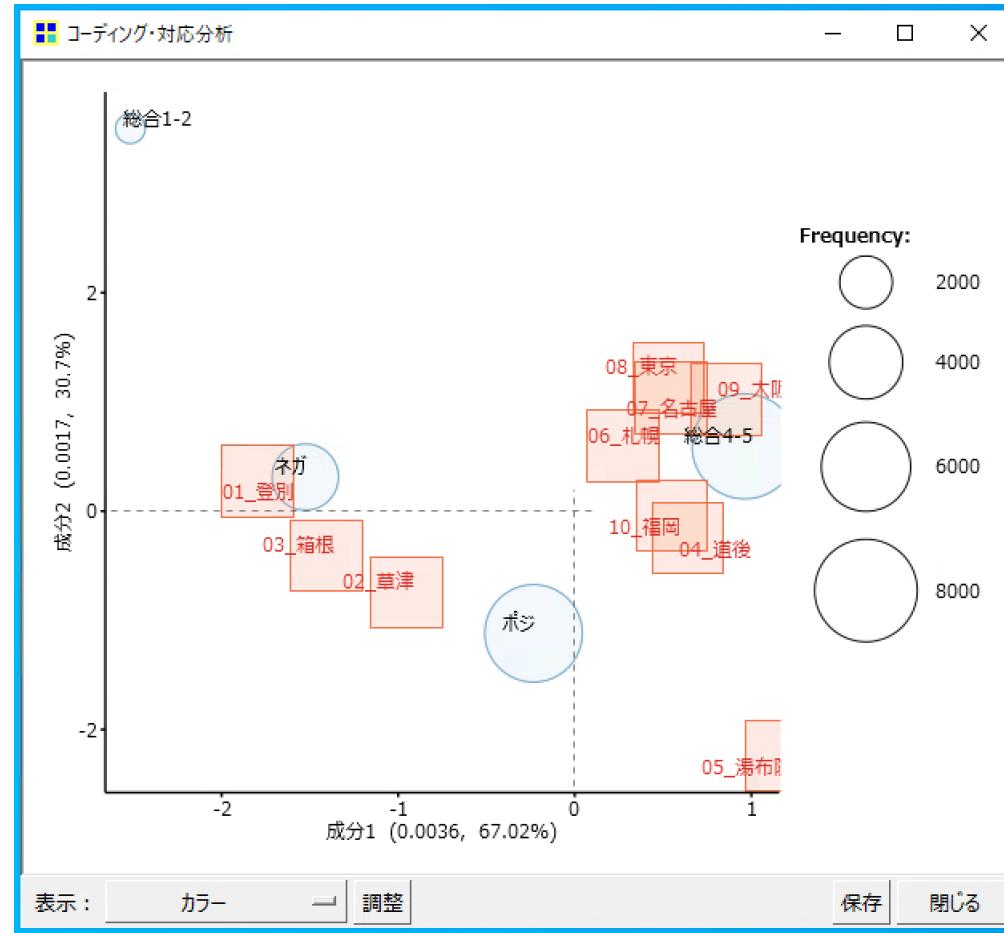
<>総合-->1 | <>総合-->2

*総合4-5

<>総合-->4 | <>総合-->5

実践3 – 改善プランを提案する

● 対照的な2エリアを選択する（出力例）



実践3 – 改善プランを提案する

- カテゴリーやエリアごとのポジティブ意見の違いを確認する
 - 対照的な2エリアを選択する
 - 選択したエリアと、そのカテゴリー（「レジャー」or「ビジネス」）を比較する
- 手順の一例：
 - カテゴリーやエリアごとの共起NWから、何を高(好)評価しているかを調べる

「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”」
「Search Entry: *ポジ」「AND検索」「集計単位:H5」→「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」

「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:H5」→「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」

実践3 – 改善プランを提案する

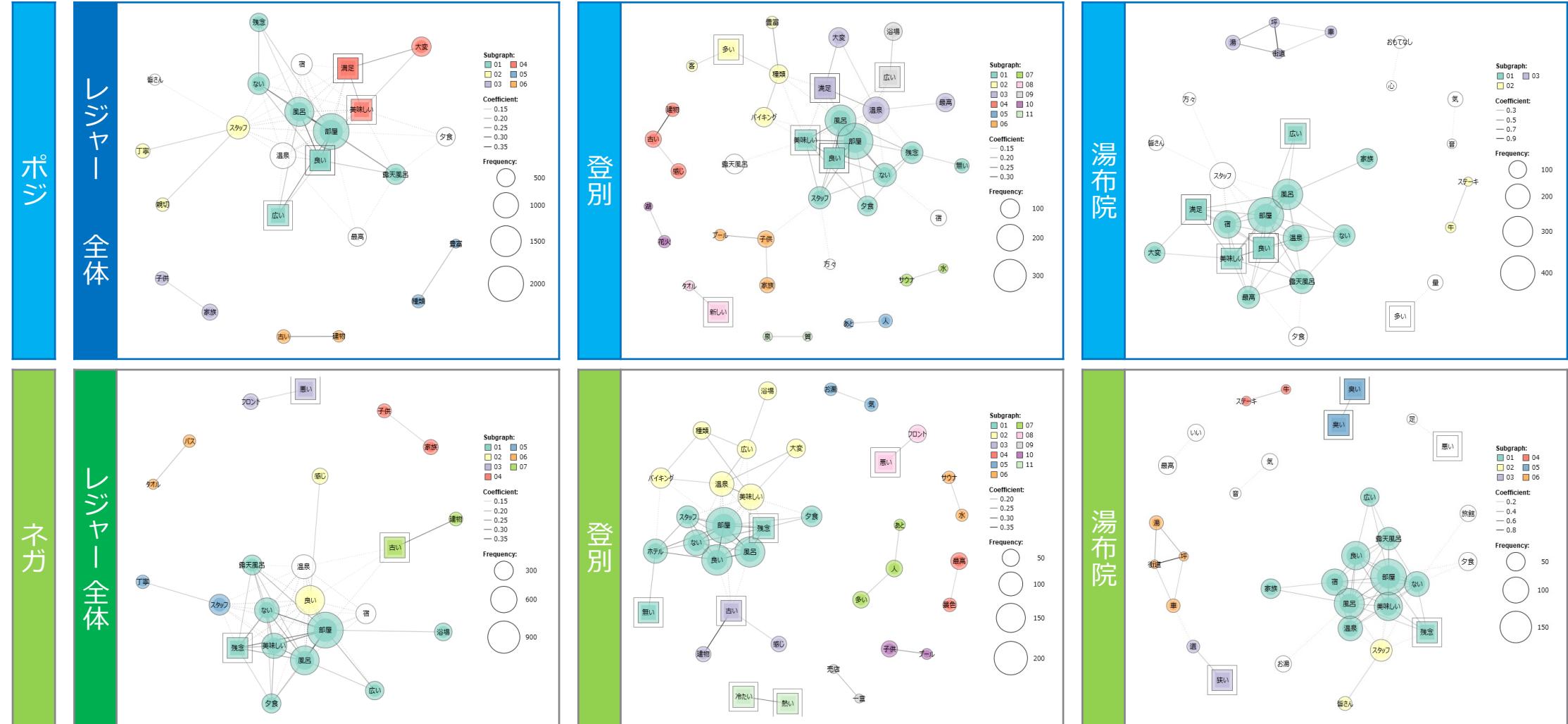
- カテゴリーやエリアごとのネガティブ意見の違いを確認する
 - 対照的な2エリアを選択する
 - 選択したエリアと、そのカテゴリー（「レジャー」or「ビジネス」）を比較する
- 手順の一例：
 - カテゴリーやエリアごとの共起NWから、何を低(悪)評価しているかを調べる

「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”」
「Search Entry: *ネガ」「AND検索」「集計単位:H5」→「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」

「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]”<>エリア-->01_登別”」
「Search Entry: *ネガ」「AND検索」「集計単位:H5」→「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」

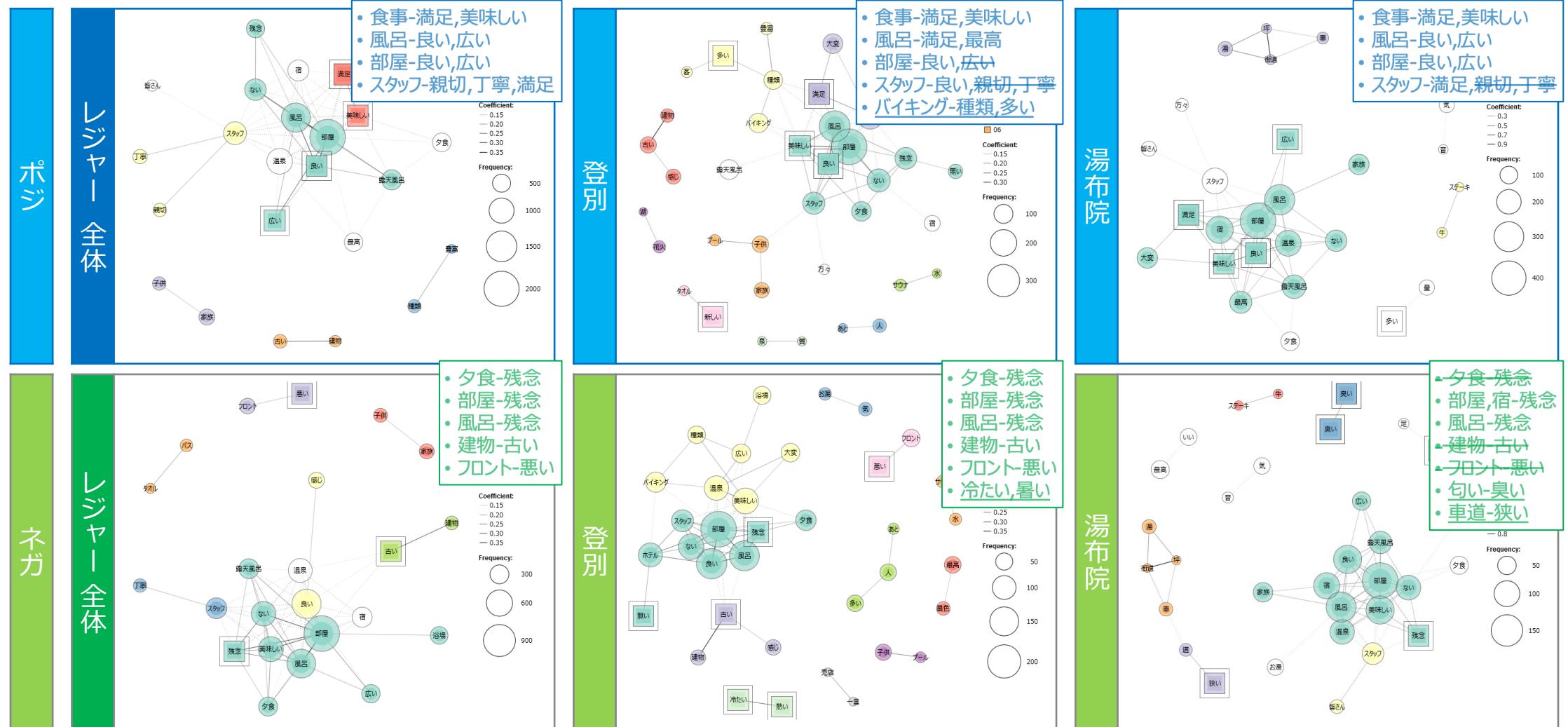
実践3 – 改善プランを提案する

● 例: 登別と湯布院のポジネガ比較



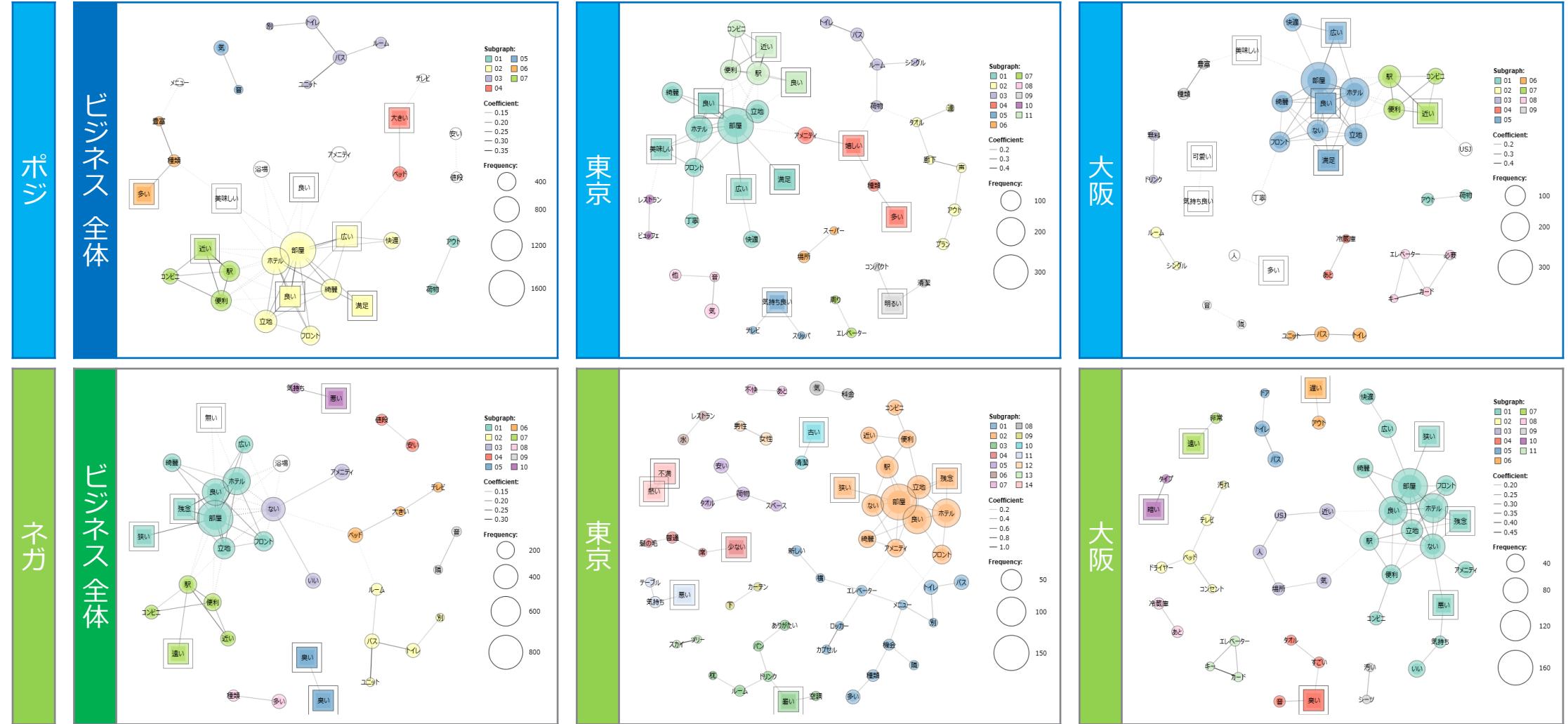
実践3 – 改善プランを提案する

● 例: 登別と湯布院のポジネガ比較



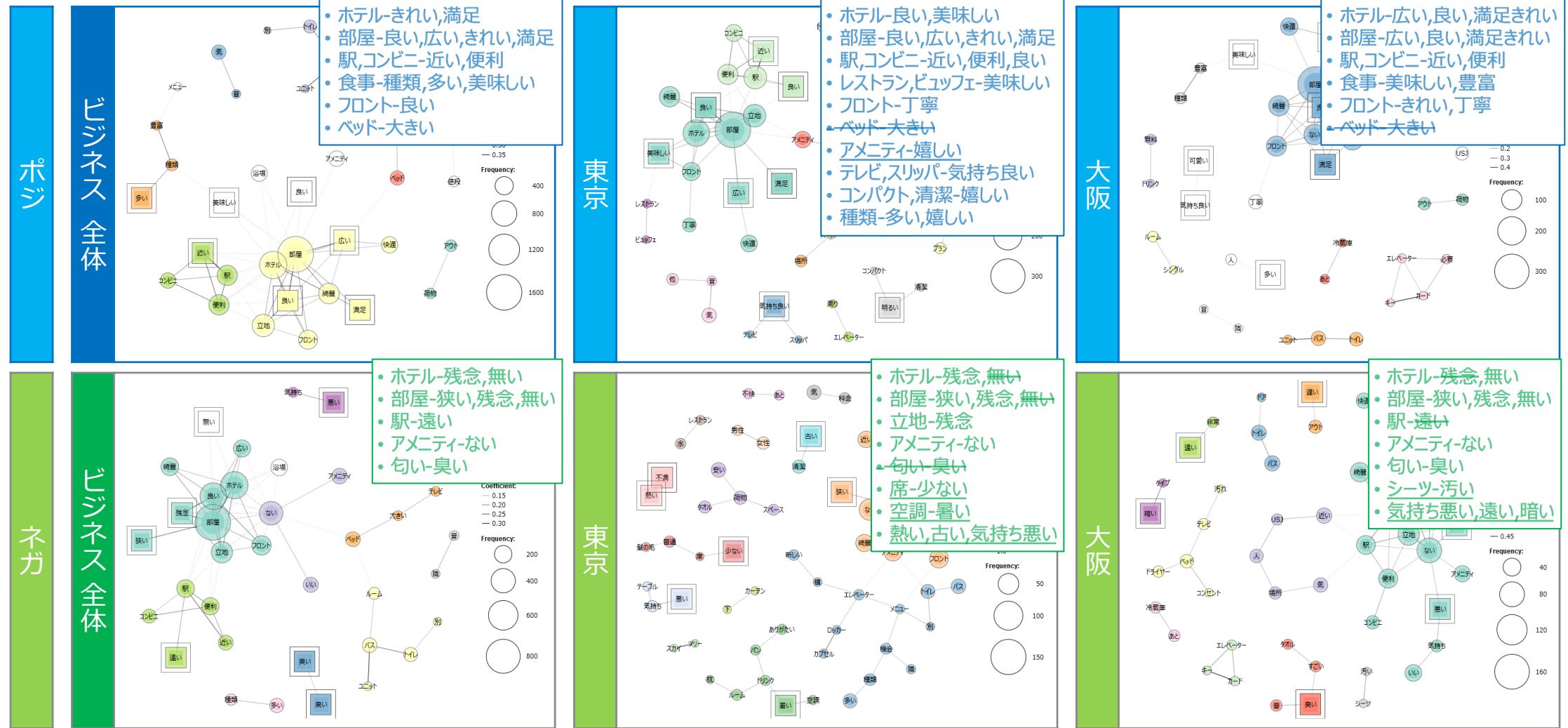
実践3 – 改善プランを提案する

● 例: 東京と大阪のポジネガ比較



実践3 – 改善プランを提案する

● 例: 東京と大阪のポジネガ比較



実践3 – 改善プランを提案する

● 結果の整理

- 主張を支持するプロットと、ユーザーの生の声(原文)を使って解釈する
 - 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?

例)

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: ... • ...	• ... • ...

演習 — 改善プランを提案する

- 特徴語とポジティブ意見の共起ネットワーク図を作成し、エリアによってポジティブ意見(とその背景)がどう異なるかを比較することで、何がどう評価されているかを確認する(→P.26)
- 特徴語とネガティブ意見の共起ネットワーク図を作成し、エリアによってネガティブ意見(とその背景)がどう異なるかを比較することで、何がどう評価されているかを確認する(→P.28)
- 高評価エリアに倣って、低評価エリアを改善プランを提案する(→P.30)
→ 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

演習 — 改善プランを提案する

● 個人ワーク (~20:20)

- ・ 共起ネットワーク図を作成し、何がどう評価されているかを確認する
- ・ 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

● グループ討論 (~20:50)

- ・ 個人ワークで発見した、2エリアの特徴や改善プランについてグループ内で討論する
- ・ グループ討論の内容を反映し、結果の整理 をブラッシュアップする

day 5 – レポート課題

- 以下を PDF ファイルで提出してください
 - 演習で作成した共起ネットワーク図(P.26,28)と結果の整理(P.30)を用いて、選択したエリアの改善プランを提案してください
- ※ 「プロット」のキャプチャは Jupyter の出力でも EXCEL でも構いません
- ※ 何らかの事情で上記の提出ができない場合は、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	8/4 ~18:20

Q&A

参考資料

● KH Coder

- ・ 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.
- ・ 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一. 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
- ・ 横口耕一. 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.