

人文社会ビジネス科学学術院 ビジネス科学研究群 2024年度 春C

# テキストマイニングの実践

## day 5

## スケジュール

### day 2

- 講義 – 自然言語処理の最新動向
- 講義 – テキストマイニングの手順
- 講義&演習 – データ理解

### day 3

- 講義&演習 – 演習環境の準備
- 講義&演習 – テキスト解析 (1)
- 講義&演習 – テキスト解析 (2)

### day 4

- 講義&演習 – テキスト分析 (1)
- 演習 – テキスト分析 (実践編)

### day 5

- 演習 – テキスト分析 (実践編)

## (前回) day 4 – レポート課題

- 以下を PDF ファイルで提出してください
  - ノートブック **day-4-1.ipynb** の末尾にある「【演習】外部変数を利用したエリアごとの作図」に従って作図した 2.1~2.4 の全てのプロットのキャプチャ

※ 何らかの事情で上記のキャプチャを提出できない場合、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

## テキスト分析（実践編2）

## (再掲) 実践的な分析

- 実践1: カテゴリーやエリアごとの宿泊者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宿泊者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

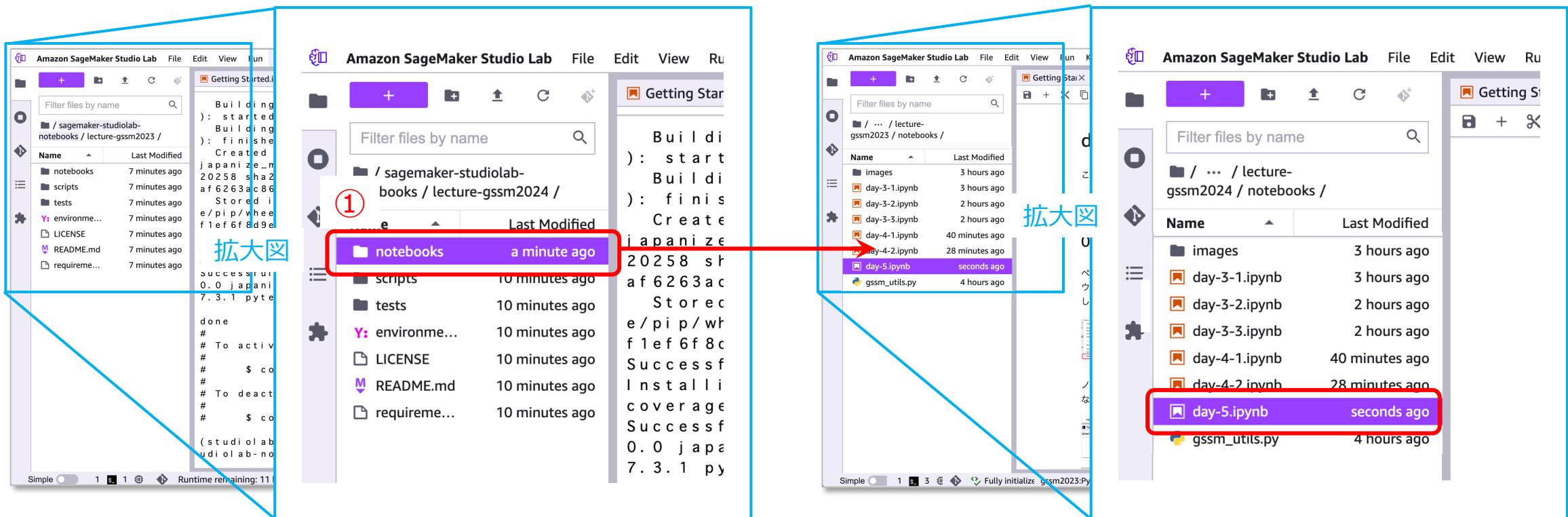
例) 実践3のまとめ方

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,実践3 — テキスト分析

## ● day-5.ipynb を開いてください

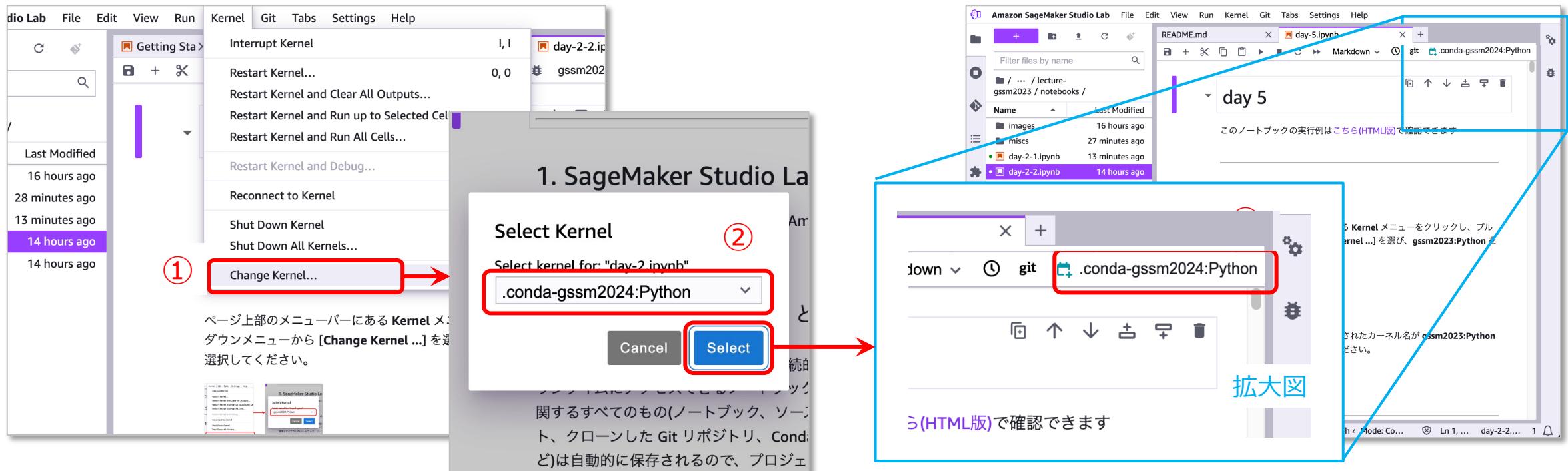
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-5.ipynb** ノートブックを開く



# 実践2,実践3 — テキスト分析

## ● カーネル **.conda-gssm2024:Python** を選択してください **!重要!**

- ① ページ上部の **Kernel** メニューから「**Change Kernel ...**」を選ぶ
- ② ポップアップ画面から「**.conda-gssm2024:Python**」を選択し、「**Select**」を押す
- ③ 右上隅にカーネル名「**.conda-gssm2024:Python**」が表示されていることを確認する



# 実践2,実践3 — テキスト分析

## ● テキスト解析と可視化

The screenshot shows the Amazon SageMaker Studio Lab interface. On the left, there is a file browser with a list of notebooks and Python files. In the center, there is a code editor window titled "0. はじめに" containing some text and code snippets. A context menu is open over the code editor, with the "Run" tab selected. The "Run" menu contains several options: "Run Selected Cells", "Run Selected Cells and Insert Below", "Run Selected Cells and Do not Advance", "Run Selected Text or Current Line in Console", "Run All Above Selected Cell", "Run Selected Cell and All Below", "Render All Markdown Cells", "Run All Cells", and "Restart Kernel and Run All Cells...". At the bottom of the interface, there is a toolbar with various icons and status information like "Fully initialize gssm2023:Python..." and "Runtime remaining: 3 h 1 Mode: Com...".

演習:

① ページ上部の Run メニューから  
「Run All Cells」を選択

この後、Step-by-step で解説します

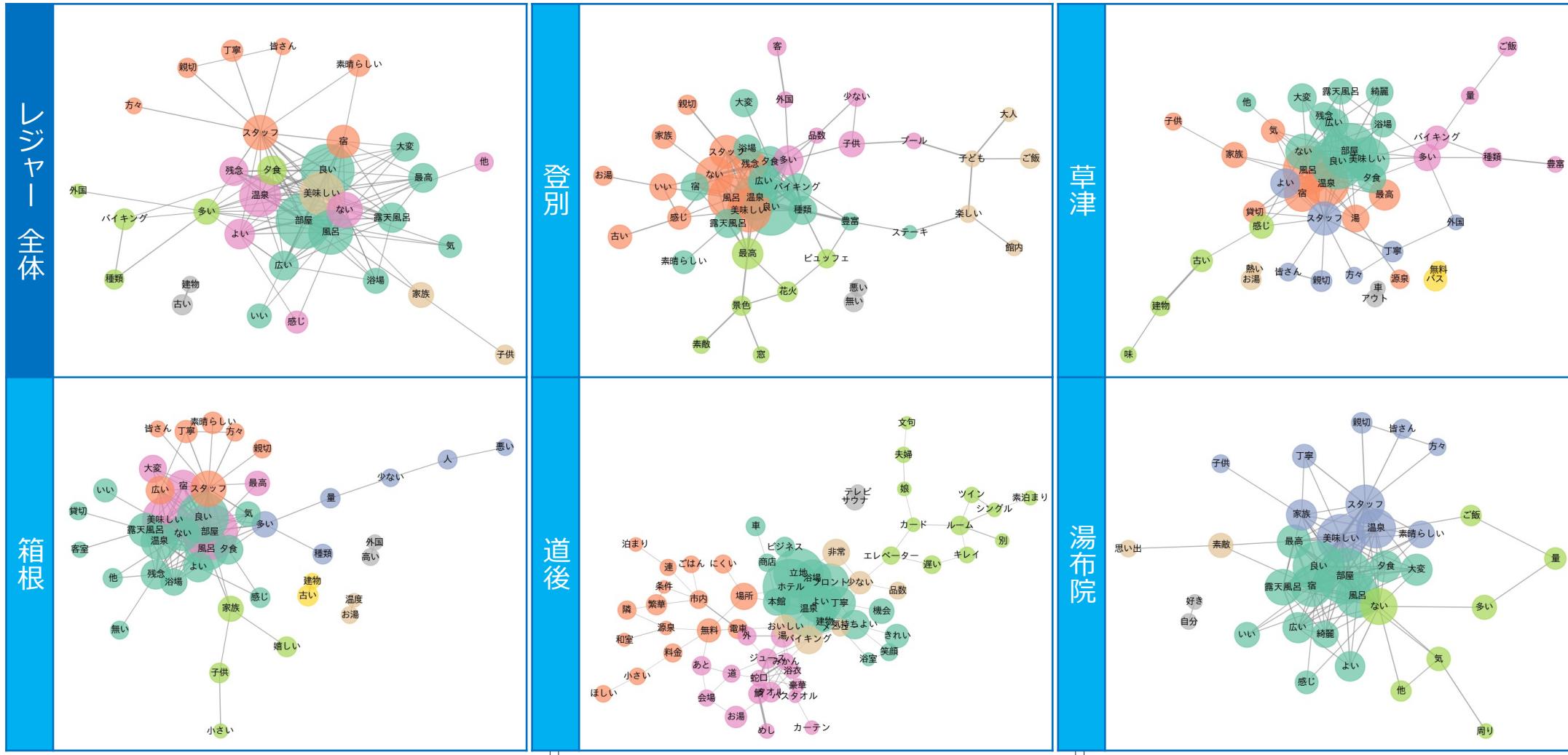
## 実践2 – ユーザー注目ポイントの評価を見る

---

- 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

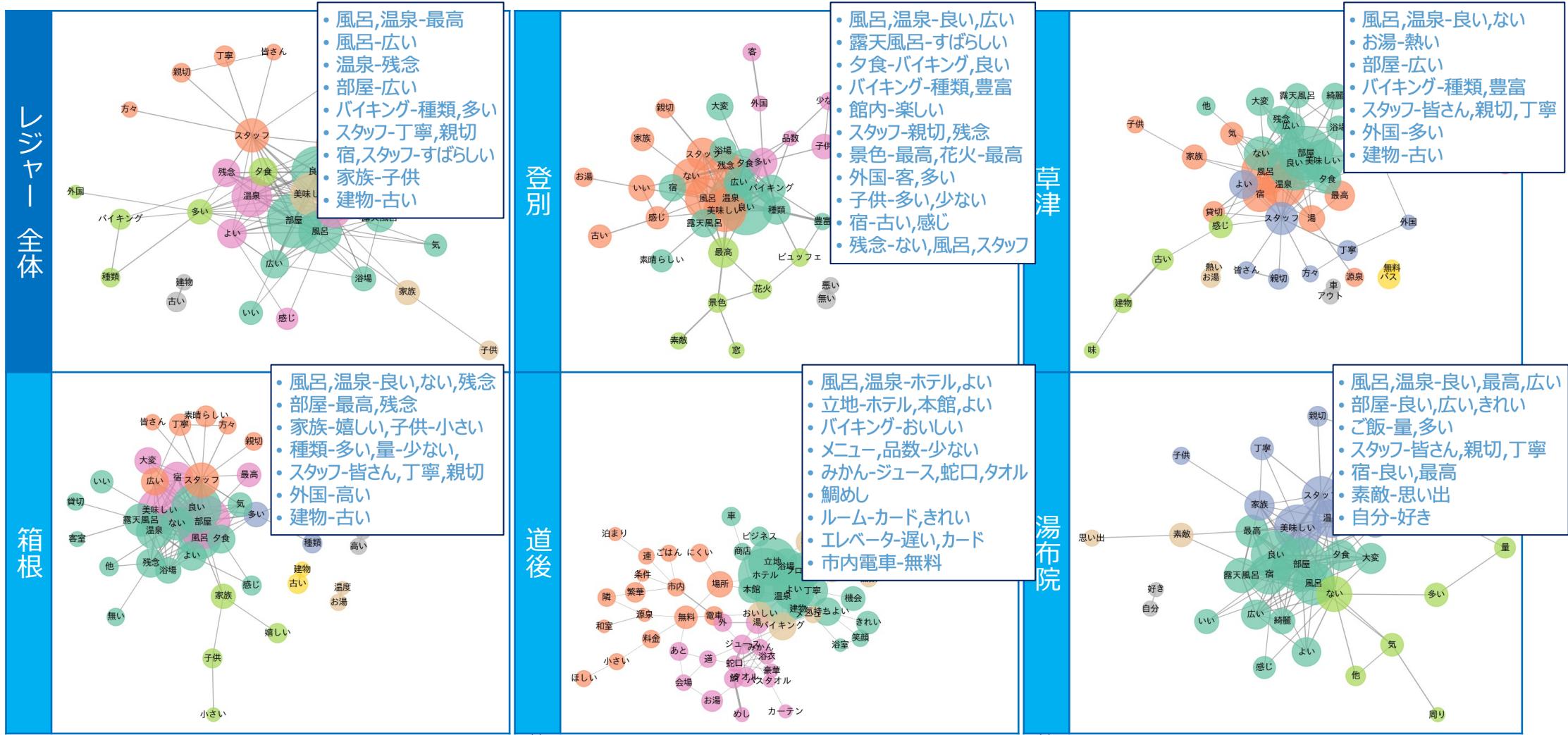
## 実践2 – ユーザー注目ポイントの評価を見る

#### ● レジャーとエリアごとの共起ネットワーク



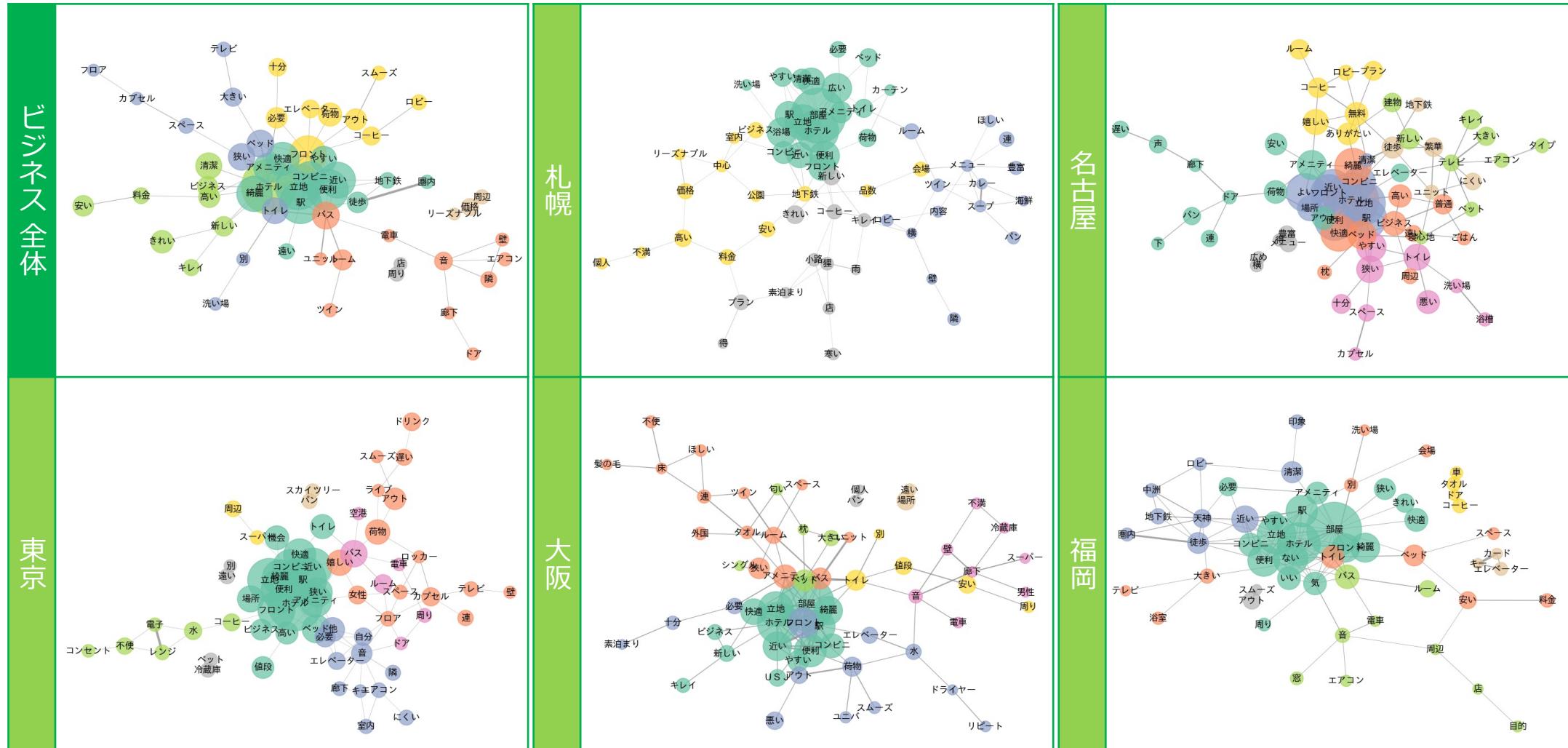
# 実践2 – ユーザー注目ポイントの評価を見る

## ● レジャーとエリアごとの共起ネットワーク



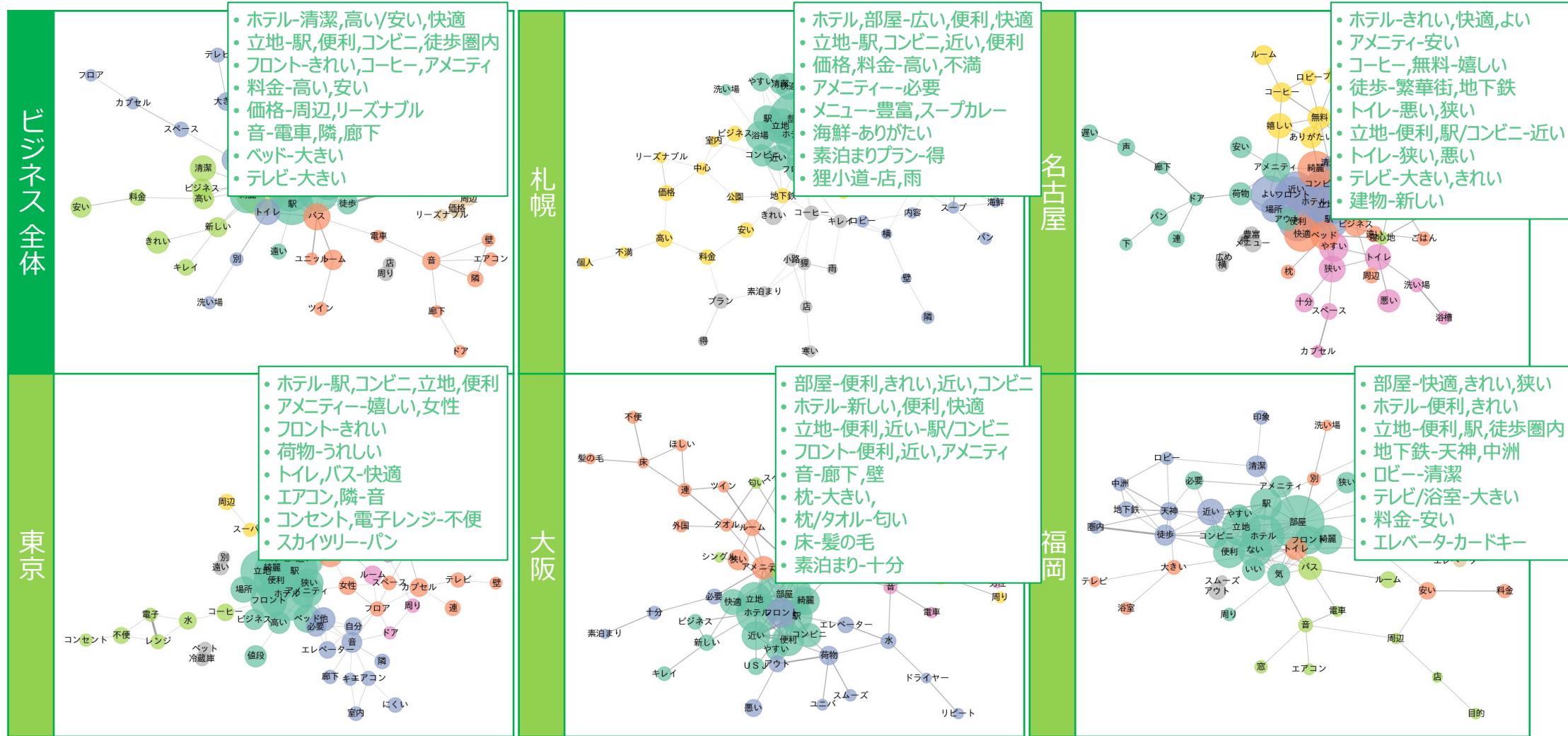
# 実践2 – ユーザー注目ポイントの評価を見る

## ● 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

# 実践3 – 改善プランを提案する

## ●コーディングルール(KH Coder用語)を作成する

### 単語のまとめ上げ

```
# ポジティブワード
```

```
coding_pos = ["良い","美味しい","広い","多い","素晴らしい","嬉しい","気持ちよい","楽しい","近い","大きい","気持ち良い","温かい","早い","優しい","新しい","暖かい","快い","明るい","美しい","可愛い","満足"]
```

```
# ネガティブワード
```

```
coding_neg = ["古い","無い","高い","悪い","小さい","狭い","少ない","寒い","遅い","熱い","欲しい","暑い","冷たい","遠い","臭い","暗い","うるさい","ない","無い","残念","改善","不満"]
```

### 外部変数のまとめ上げ

- 数値評価「総合」の値が 1 or 2 → 「総合1~2」
- 数値評価「総合」の値が 4 or 5 → 「総合4~5」

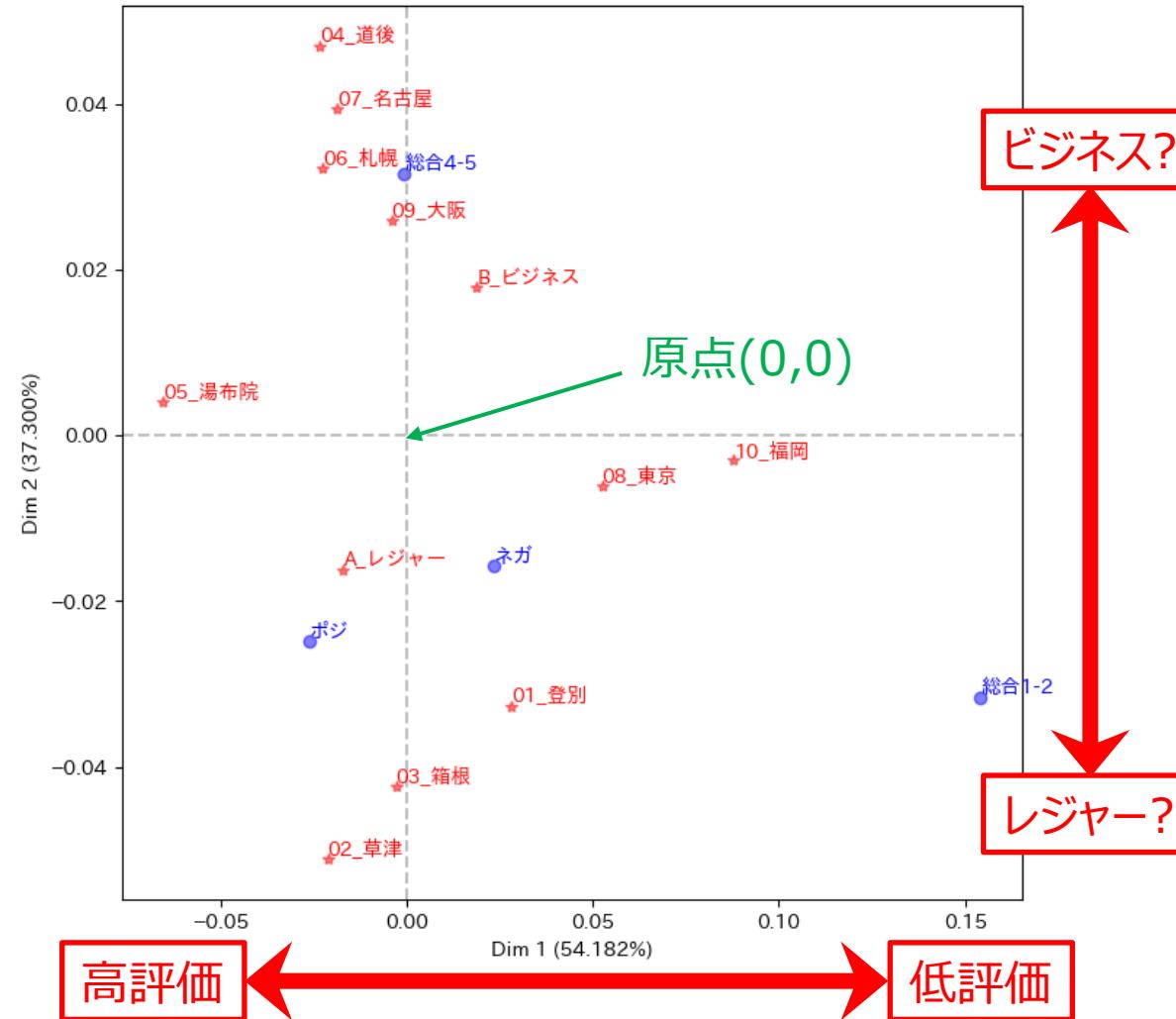


カテゴリ	エリア	文書ID	表層	ポジ	ネガ	総合1-2	総合4-5
			1	2	3	4	5
A_レジャー	01_登別	1	1	0	0	1	
		2	1	1	0	0	1
		3	0	0	0	0	1
		4	1	0	1	0	0
		5	1	1	0	0	1
...		...	...	...	...	...	...
B_ビジネス	10_福岡	9996	1	0	0	0	1
		9997	0	1	0	0	1
		9998	0	0	0	0	1
		9999	1	0	0	0	1
		10000	1	0	0	0	0

9914 rows × 4 columns

# 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）



### 見方:

- 原点付近は特徴がない
- 原点から遠い方が特徴的（以下は仮説であって、原文で確認すること）
  - 湯布院は圧倒的に高評価？
  - 福岡は圧倒的に低評価か？
  - 道後はレジヤーよりもビジネスに近い（出張で行く場所か）
  - 草津は圧倒的にレジヤー
- 第2固有値までの累積寄与率は  $54.18 + 37.30 = 91.5\%$  で非常に高く、これら2軸でデータをよく表現できている

## 実践3 – 改善プランを提案する

---

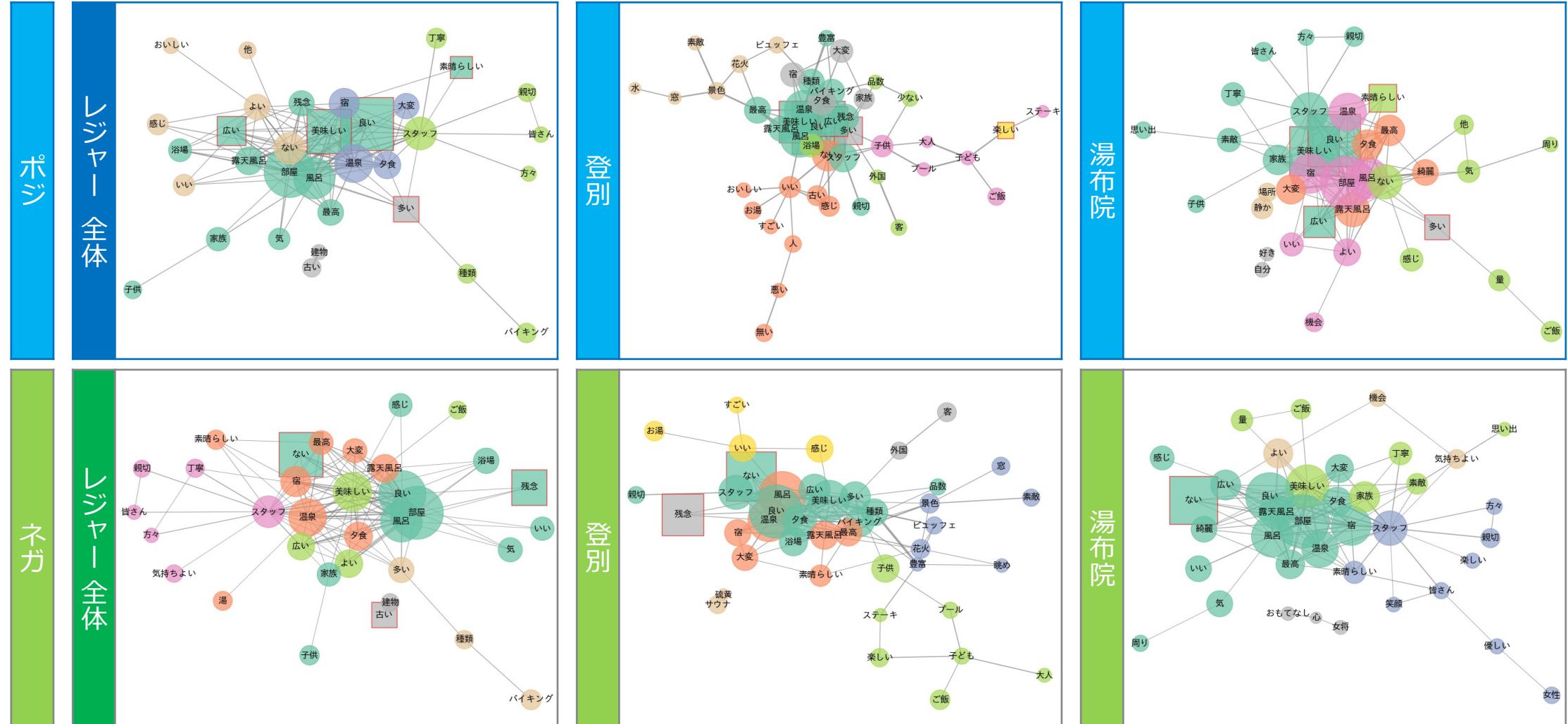
- カテゴリーやエリアごとのポジティブ意見の違いを確認する
  - 対照的な2エリアを選択する
  - 選択したエリアと、そのカテゴリー（「レジャー」or「ビジネス」）を比較する
- 手順の一例：
  - カテゴリーやエリアごとの共起NWから、何を高(好)評価しているかを調べる

## 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

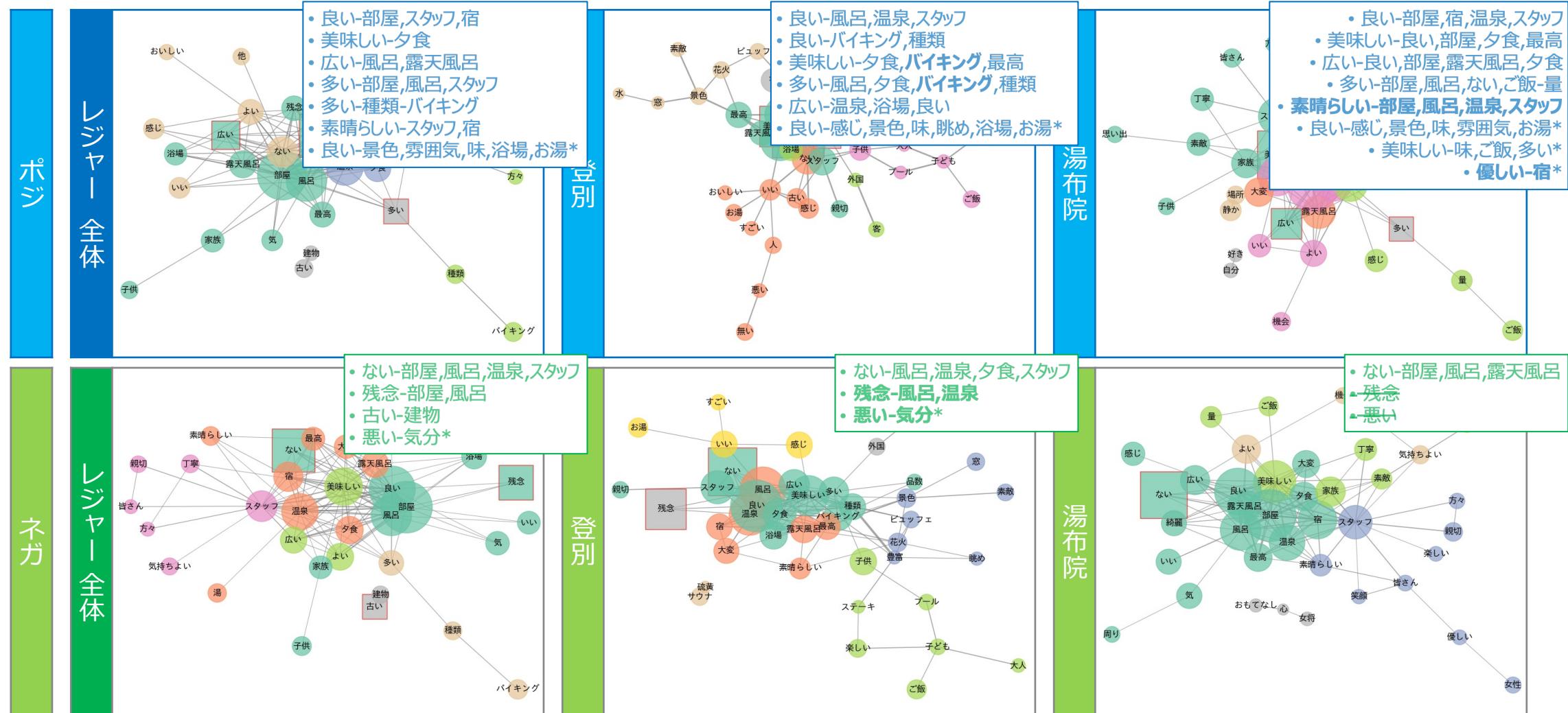
# 実践3 – 改善プランを提案する

## ● 例: 登別と湯布院のポジネガ比較



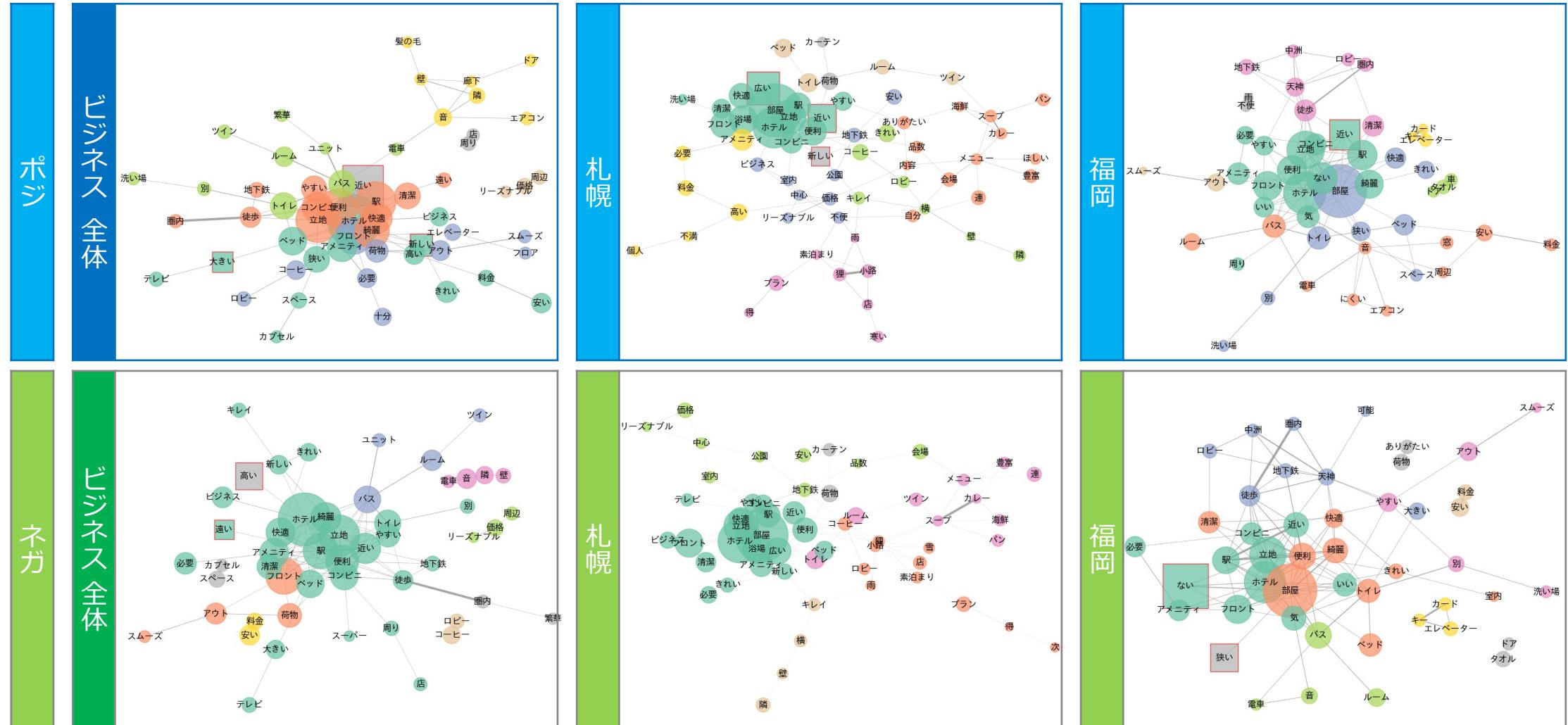
# 実践3 – 改善プランを提案する

## ● 例：登別と湯布院のポジネガ比較



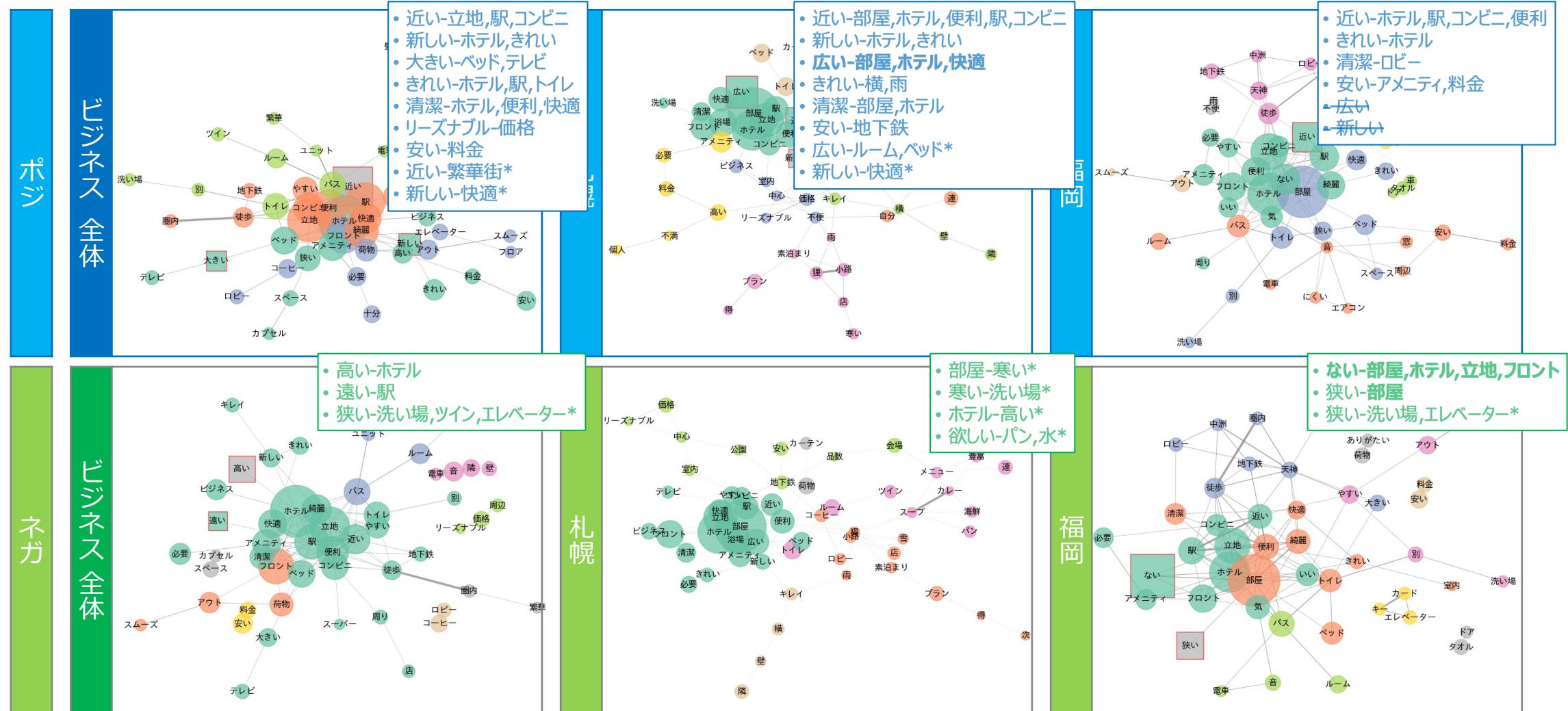
# 実践3 – 改善プランを提案する

## ● 例: 札幌と福岡のポジネガ比較



# 実践3 – 改善プランを提案する

## ● 例：札幌と福岡のポジネガ比較



# 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.XXX)
- 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.XXX)
- 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.172)  
→ 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

# レポート課題

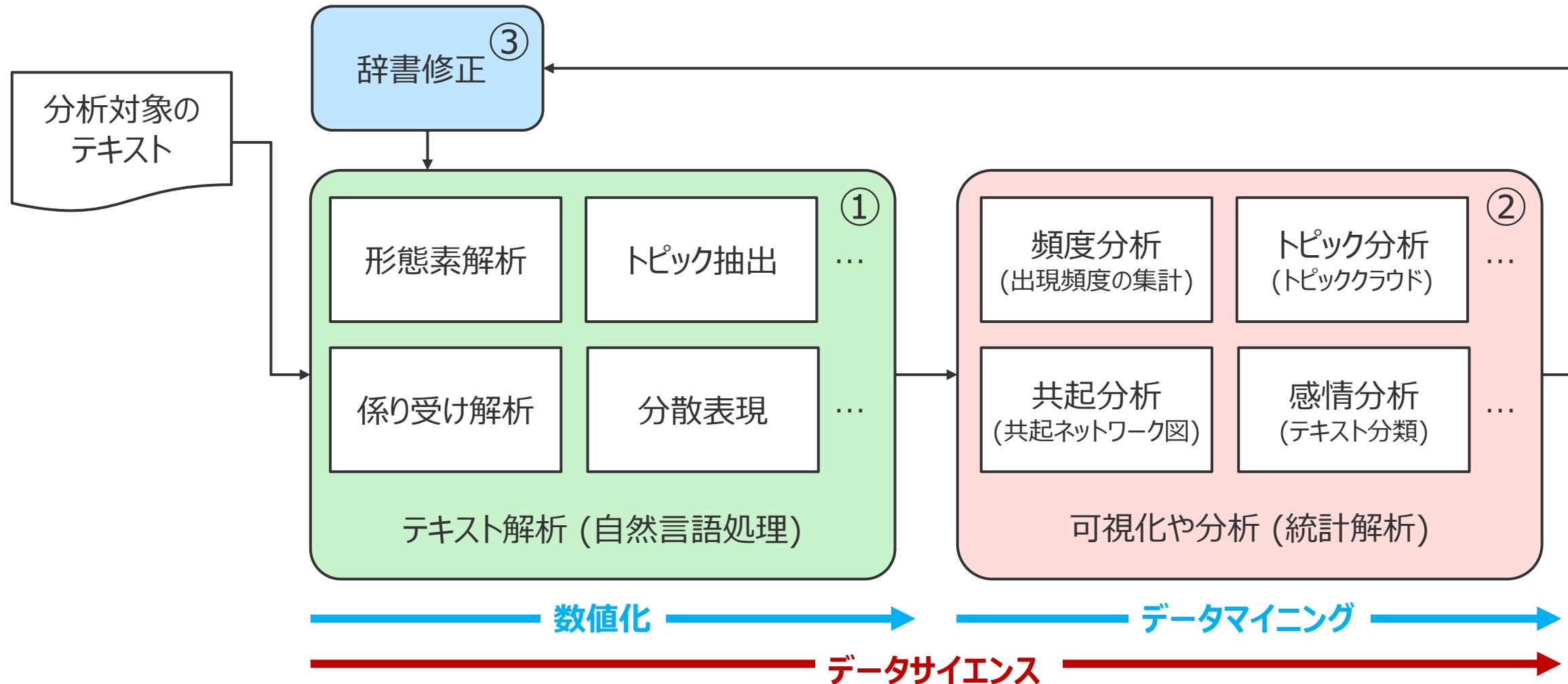
- 以下を PDF ファイルで提出してください
    - 演習で作成した共起ネットワーク図(P.26,28)と結果の整理(P.30)を用いて、選択したエリアの改善プランを提案してください
- ※ 「プロット」のキャプチャは Jupyter の出力でも EXCEL でも構いません
- ※ 何らかの事情で上記の提出ができない場合は、本日の講義の感想を文章で記述してください

レポート形式	提出先	期限
PDF	manaba	8/4 ~18:20

- 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是?
    - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
  - これら課題を解決するために、テキスト分析を実施

# (再掲) テキスト分析の手順

①自然言語処理によりテキストを数値化する → ②統計解析や可視化を行う → ③結果を読み解きながら解析のための辞書を編纂する → 分析のサイクルを回していく(①へ)



お疲れ様でした!