

テキストマイニングの実践

—3日目—

2020/7/17

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-202007>



スケジュール

- 1日目: 7/1(水)
 - 説明 — テキストマイニングの手順
 - 実習 — データをよく知る (Excel)
- 2日目: 7/10(金)
 - 説明 — テキストマイニングツールの使い方 (KHCoder)
- **3日目: 7/17(金)**
 - **説明 — データ分析の実践 (KHCoder)**
 - **実習 — データ分析の実践 (KHCoder)**
- **体育の日: 7/24(金)**
- 4日目: 7/31(金)
 - 外部講師 — NTTデータ数理システム
- 5日目: 8/7(金)
 - 発表 — データ分析の実践 (KHCoder)

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

KH Coder で表記ゆれを吸収する (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

注: 最新版ではこのプラグインが
あらかじめインストールされ
ています

(次ページにつづく)

KH Coder で表記ゆれを吸収する (2/2)

- 手順

2. プラグインファイル

z1_edit_words3.pm を編集する

```
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 }
```



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

編集前

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

The screenshot shows the 'Extracted Word List' dialog with the following interface elements:

- Filter Entry:** Displays 'お湯'.
- OR検索** and **部分一致** buttons.
- リスト (List):** A table with columns: #, 抽出語 (Extracted Word), 品詞/活用 (Part of Speech/Usage), and 頻度 (Frequency). It shows three entries:

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

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	4.16	4.20	4.04	3.98	4.22	4.22	4.23
B_ビジネス	3.91	4.24	3.98	3.85	3.69	3.94	4.08

- 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

実践的な分析 — 特徴語の集計

- 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」 → 「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, 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」 → 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

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	.094
風呂	.077
温泉	.062
食事	
美味しい	.062
お部屋	.042
宿	.042
スタッフ	.041
露天風呂	.032
残念	.030
大変	.029

B_ビジネス	数値評価指標
部屋	.102
ホテル	.085
立地	.048
ない	.047
駅	.044
便利	.043
フロント	.038
近い	.038
広い	.030
綺麗	.030

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
部屋	.058	湯畠	.081	良い	.064	温泉	.058	美味しい	.067
良い	.055	温泉	.066	風呂	.056	良い	.053	宿	.064
風呂	.053	良い	.064	美味しい	.053	ホテル	.045	風呂	.063
温泉	.043	風呂	.064	お部屋	.045	ない	.035	スタッフ	.043
美味しい	.039	宿	.040	スタッフ	.041	美味しい	.033	露天風呂	.042
お部屋	.035	美味しい	.040	温泉	.040	立地	.030	温泉	.042
宿	.034	お部屋	.032	宿	.038	フロント	.025	お部屋	.041
スタッフ	.028	立地	.028	露天風呂	.036	よい	.023	家族	.035
最高	.027	スタッフ	.028	残念	.030	浴場	.023	最高	.031
バイキング	.027	残念	.027	大変	.028	便利	.023	夕食	.031

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
部屋	.055	ホテル	.052	部屋	.057	ホテル	.060	部屋	.056
ホテル	.053	部屋	.048	ホテル	.056	部屋	.054	ホテル	.049
立地	.039	駅	.038	駅	.046	便利	.040	立地	.040
ない	.031	便利	.031	便利	.044	立地	.040	フロント	.037
駅	.031	立地	.029	ない	.039	駅	.037	近い	.033
設備・アメニティ		フロント	.028	立地	.035	ない	.034	ない	.032
便利	.031	近い	.027	近い	.034	近い	.031	駅	.032
フロント	.028	綺麗	.026	フロント	.032	フロント	.029	便利	.031
近い	.027	快適	.024	綺麗	.026	綺麗	.028	快適	.025
広い	.026	広い	.022	コンビニ	.022	広い	.025	広い	.023

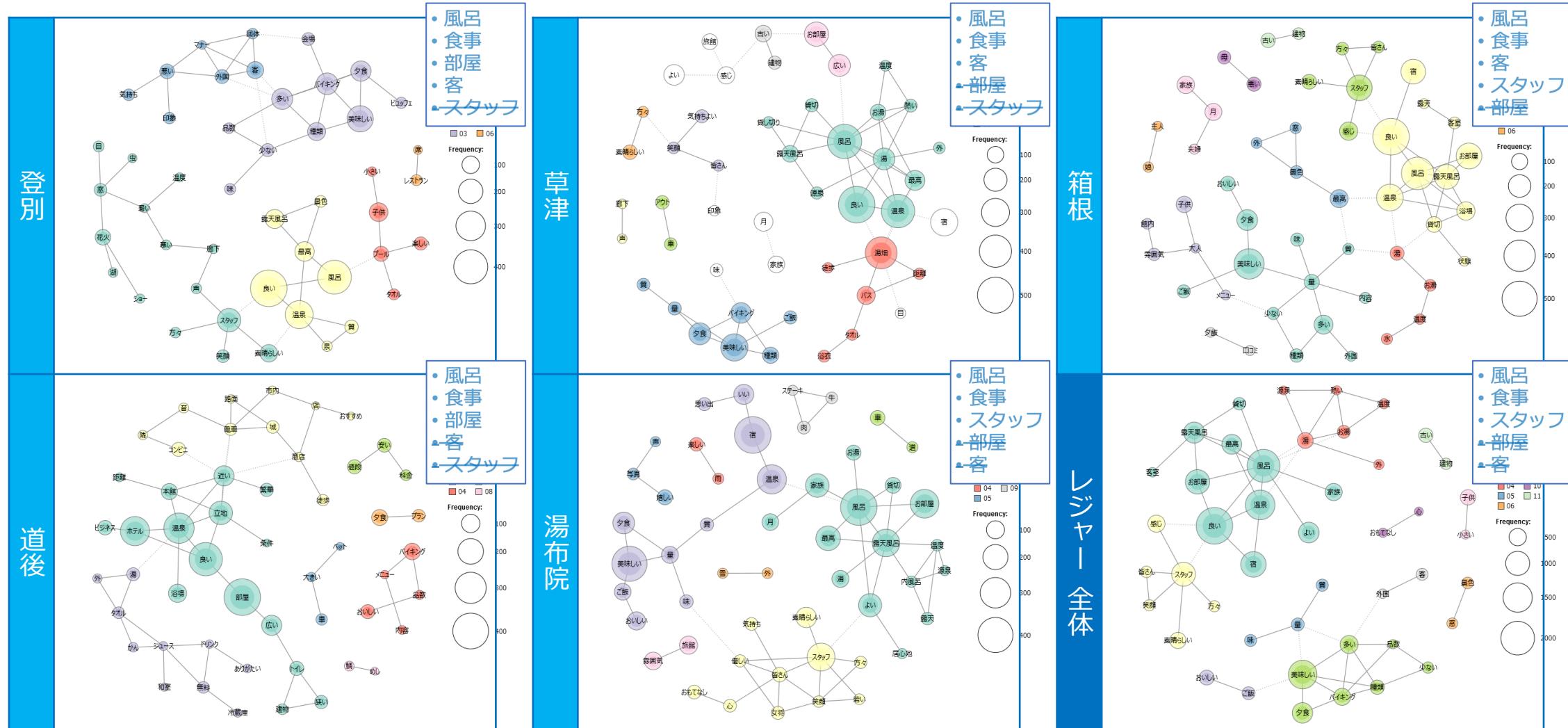
Tips: 「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=カテゴリー」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→「フィルタ設定」→「品詞=名詞,形容動詞,未知語,タグ,形容詞,名詞B,形容詞B,名詞C」を選択→「▽特徴語」→「一覧(EXCEL形式)」で連続実行

実践的な分析 — 特徴語の共起ネット

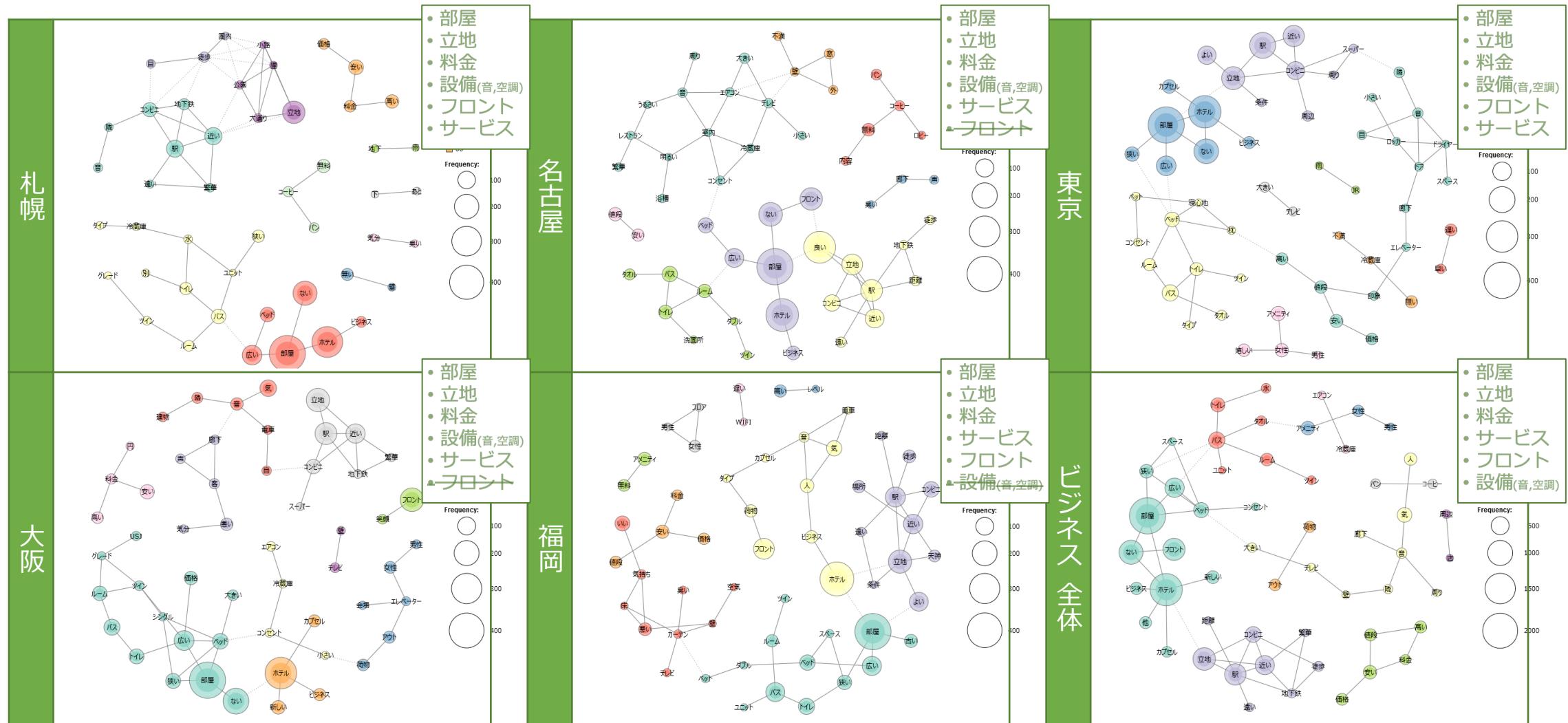
- 宿泊客は、どの項目のどこに注目しているか？
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. カテゴリー「レジャー」(or 「ビジネス」) の 5エリアを比較する
- 手順
 - 特徴語の共起ネットワーク図を作成

「ツール」→「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

実践的な分析 – 特徴語の共起ネット(1)



実践的な分析 — 特徴語の共起ネット(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

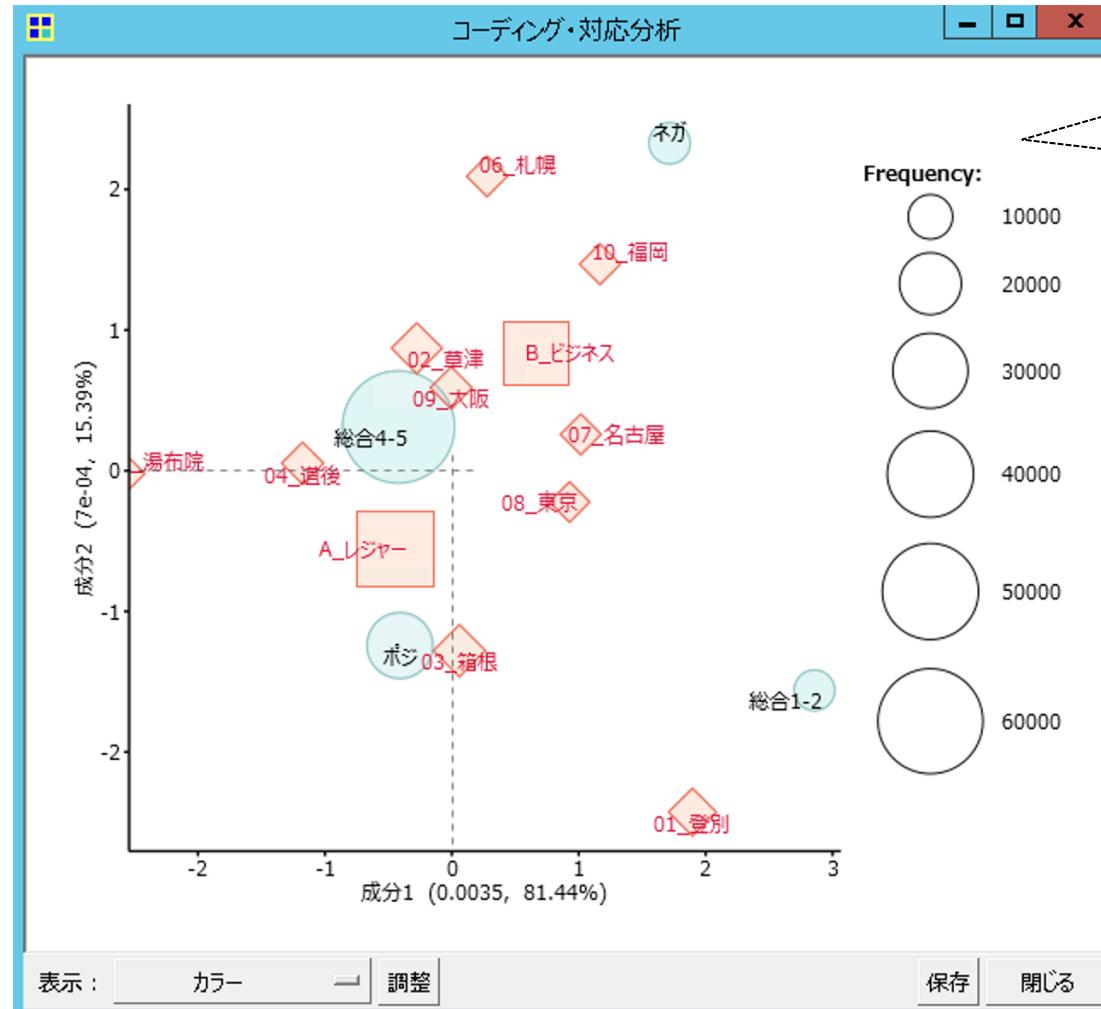
- ・ エリア別

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

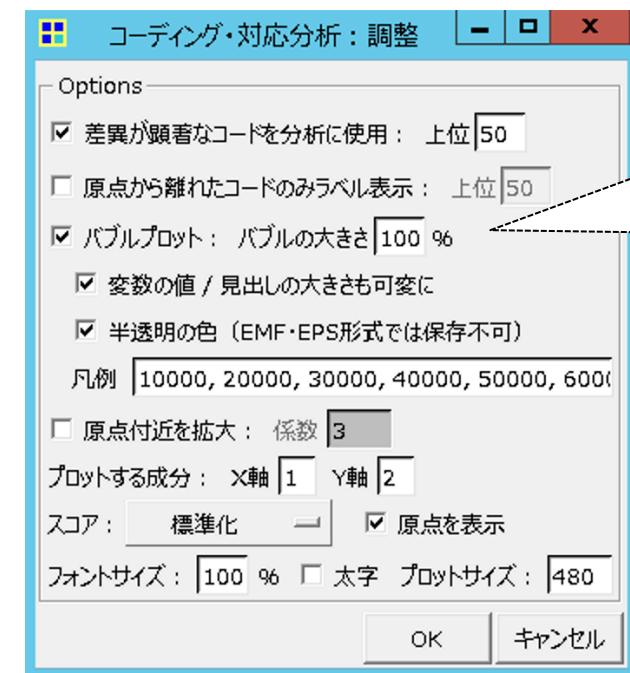
実践的な分析 — 改善案を提案する

- ・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)

手順1 — 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

手順2-A — 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

手順2-B — ネガティブ意見の共起NW

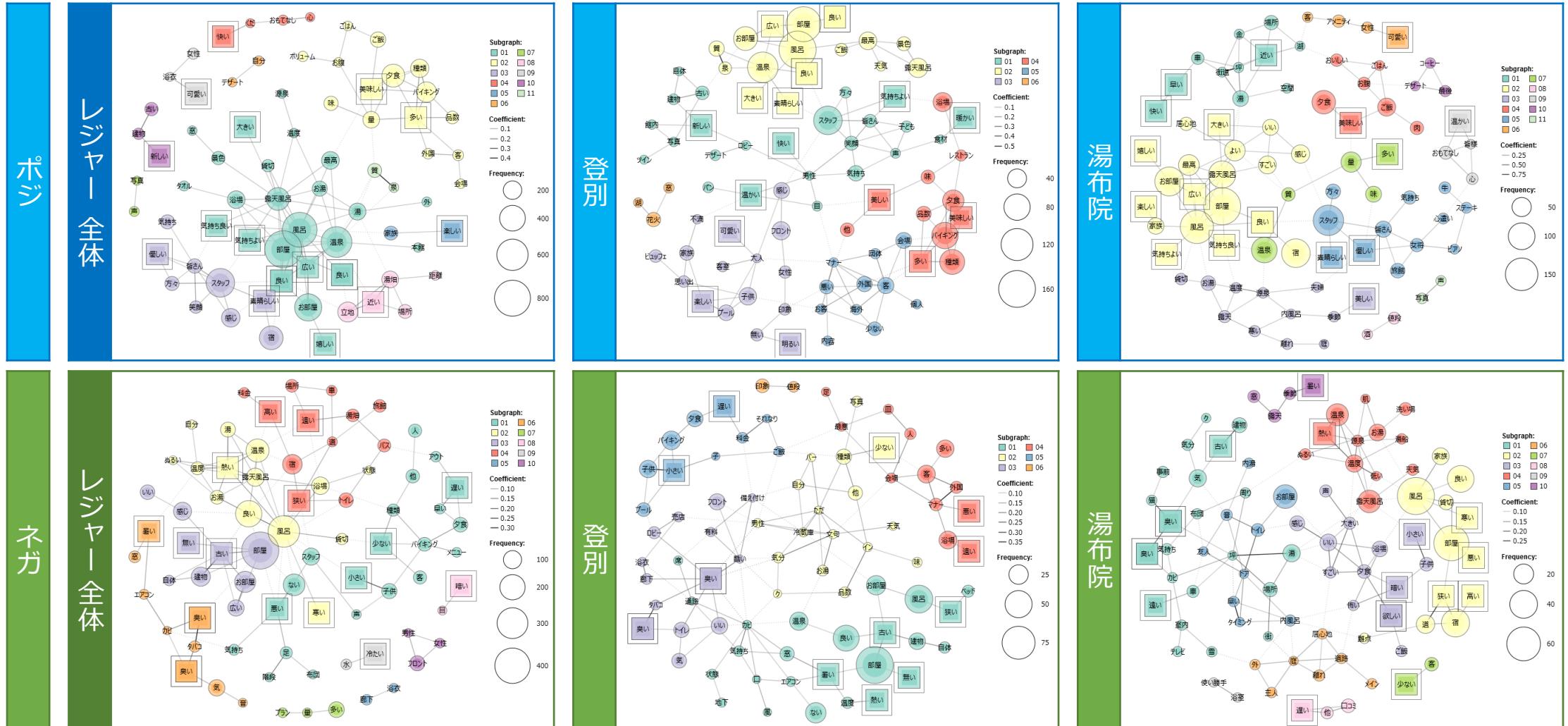
- ・ユーザーは何をどう評価しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する

- ・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

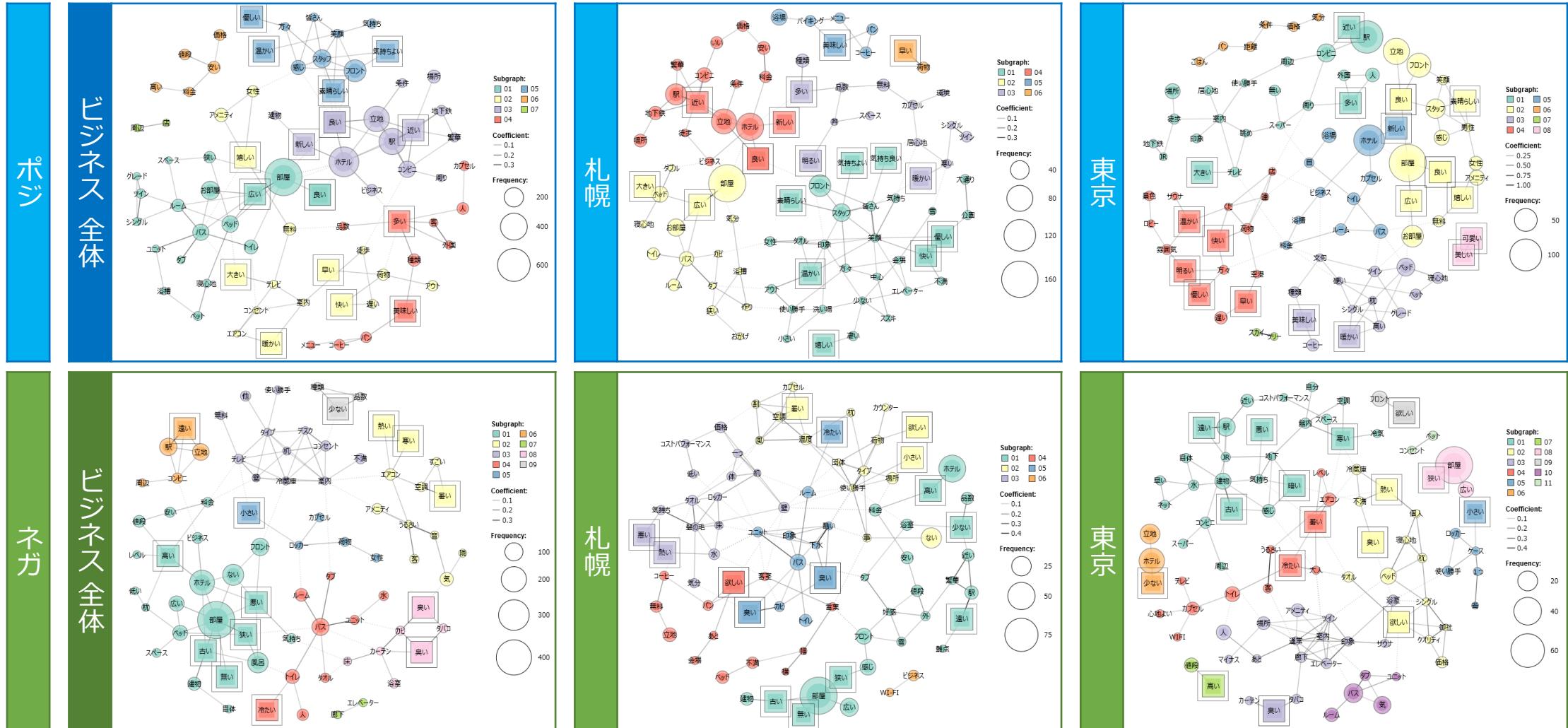
「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)"<>エリア-->01_登別"」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:上位=120,共起関係ほど濃い線に」

- ・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 — 東京と札幌のポジネガ比較



演習 — グループワーク

- テキストマイニングの手順 (1日目) に倣って、テキストデータの分析を進めてください
 - 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。6グループ (各5人) に分かれて、グループ単位で分析や議論を進めてください
- 時間配分:
 - 時間: 次回発表まで
 - 発表: 次回 (グループごとに発表)

演習で使用できるデータ

データファイル名	件数	データセット	備考
rakuten_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)EXCEL 形式 (シート名「2019」)	<ul style="list-style-type: none">本講義の全体を通して利用する
rakuten_2020.xlsx	8,518	<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 (登別,草津,由布院は 1,000件以下のため全数, それ以外はランダムサンプリング)EXCEL 形式 (シート名「2020年」)	<ul style="list-style-type: none">演習用 (3~4日目)
corona_2020.xlsx	10,000	<ul style="list-style-type: none">2020/4/27~5/30 のハッシュタグ「#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EXCEL 形式 (シート名「corona」)	<ul style="list-style-type: none">演習用 (3~4日目)

※ 自身の業務課題を題材にしても構いませんが、個人情報や機密情報など**守秘義務に特に留意**してください。

まとめ方の一例

- ・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・...

まとめ方の一例

- ・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: ... ・...	・... ・...

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.