# 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

# 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 | 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_レジャー | |
|---|---|
| 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 |

数値評価指標

| |
|---|
| 風呂 |
| 部屋 |
| 食事 |
| サービス |
| 設備・アメニティ |
| ~~立地~~ |

| 01_登別 | |
|---|---|
| 風呂 | .056 |
| 温泉 | .043 |
| ない | .035 |
| スタッフ | .030 |
| バイキング | .030 |
| 夕食 | .025 |
| 最高 | .022 |
| 子供 | .022 |
| 露天風呂 | .021 |
| 浴場 | .021 |

| 02_草津 | |
|---|---|
| 湯畑 | .071 |
| 温泉 | .065 |
| 風呂 | .060 |
| 宿 | .044 |
| お部屋 | .033 |
| 湯 | .031 |
| 夕食 | .030 |
| バイキング | .026 |
| よい | .024 |
| 最高 | .024 |

| 03_箱根 | |
|---|---|
| 風呂 | .054 |
| 温泉 | .040 |
| 露天風呂 | .038 |
| お部屋 | .038 |
| スタッフ | .037 |
| 宿 | .036 |
| 夕食 | .026 |
| 感じ | .020 |
| 浴場 | .020 |
| 最高 | .019 |

| 04_道後 | |
|---|---|
| 温泉 | .056 |
| 部屋 | .053 |
| ホテル | .042 |
| 立地 | .034 |
| よい | .025 |
| 浴場 | .022 |
| 本館 | .020 |
| バイキング | .019 |
| 感じ | .017 |
| 夕食 | .017 |

| 05_湯布院 | |
|---|---|
| 宿 | .069 |
| 風呂 | .057 |
| 露天風呂 | .040 |
| 温泉 | .040 |
| お部屋 | .039 |
| スタッフ | .037 |
| 最高 | .031 |
| 家族 | .029 |
| よい | .026 |
| 夕食 | .024 |

| B_ビジネス | |
|---|---|
| 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 |

数値評価指標

| |
|---|
| 風呂 |
| 部屋 |
| ~~食事~~ |
| サービス |
| 設備・アメニティ |
| 立地 |

| 06_札幌 | |
|---|---|
| 部屋 | .059 |
| ホテル | .055 |
| ない | .037 |
| 立地 | .036 |
| フロント | .033 |
| 駅 | .022 |
| 浴場 | .022 |
| ベッド | .017 |
| いい | .017 |
| バス | .016 |

| 07_名古屋 | |
|---|---|
| ホテル | .057 |
| 部屋 | .055 |
| 駅 | .034 |
| フロント | .034 |
| ない | .033 |
| 立地 | .028 |
| よい | .023 |
| アメニティ | .021 |
| コンビニ | .020 |
| ベッド | .017 |

| 08_東京 | |
|---|---|
| ホテル | .054 |
| 部屋 | .052 |
| 駅 | .048 |
| ない | .034 |
| 立地 | .034 |
| フロント | .033 |
| コンビニ | .024 |
| よい | .021 |
| バス | .019 |
| 浴場 | .018 |

| 09_大阪 | |
|---|---|
| 部屋 | .055 |
| ホテル | .053 |
| ない | .039 |
| フロント | .037 |
| 立地 | .036 |
| 駅 | .034 |
| 気 | .019 |
| 浴場 | .018 |
| バス | .018 |
| ベッド | .018 |

| 10_福岡 | |
|---|---|
| ホテル | .064 |
| 部屋 | .057 |
| 立地 | .037 |
| フロント | .030 |
| 駅 | .030 |
| バス | .028 |
| よい | .024 |
| トイレ | .021 |
| コンビニ | .019 |
| 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

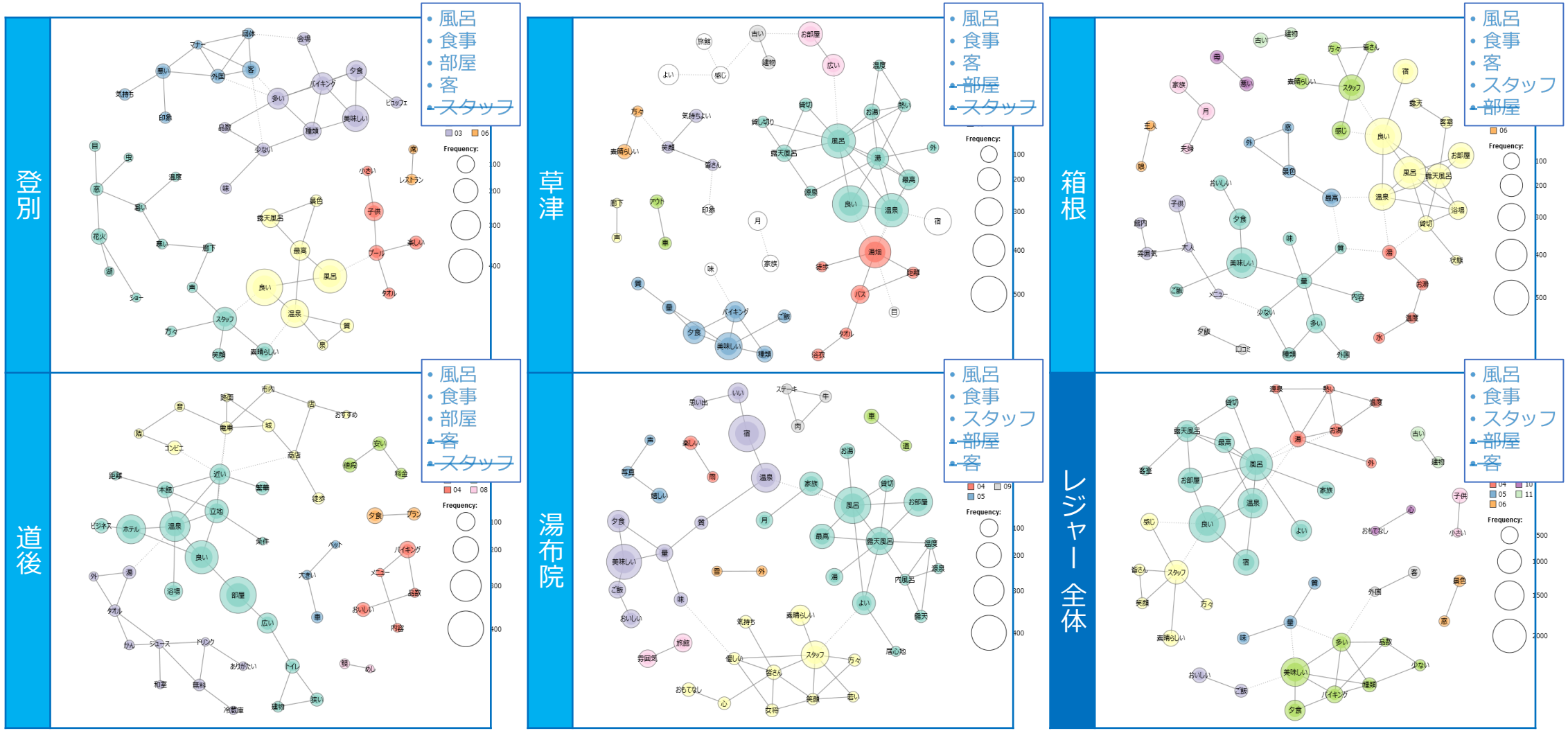
<>エリア-->01_登別
<>エリア-->02_草津
<>エリア-->03_箱根
<>エリア-->04_道後
<>エリア-->05_湯布院

**ビジネス:**

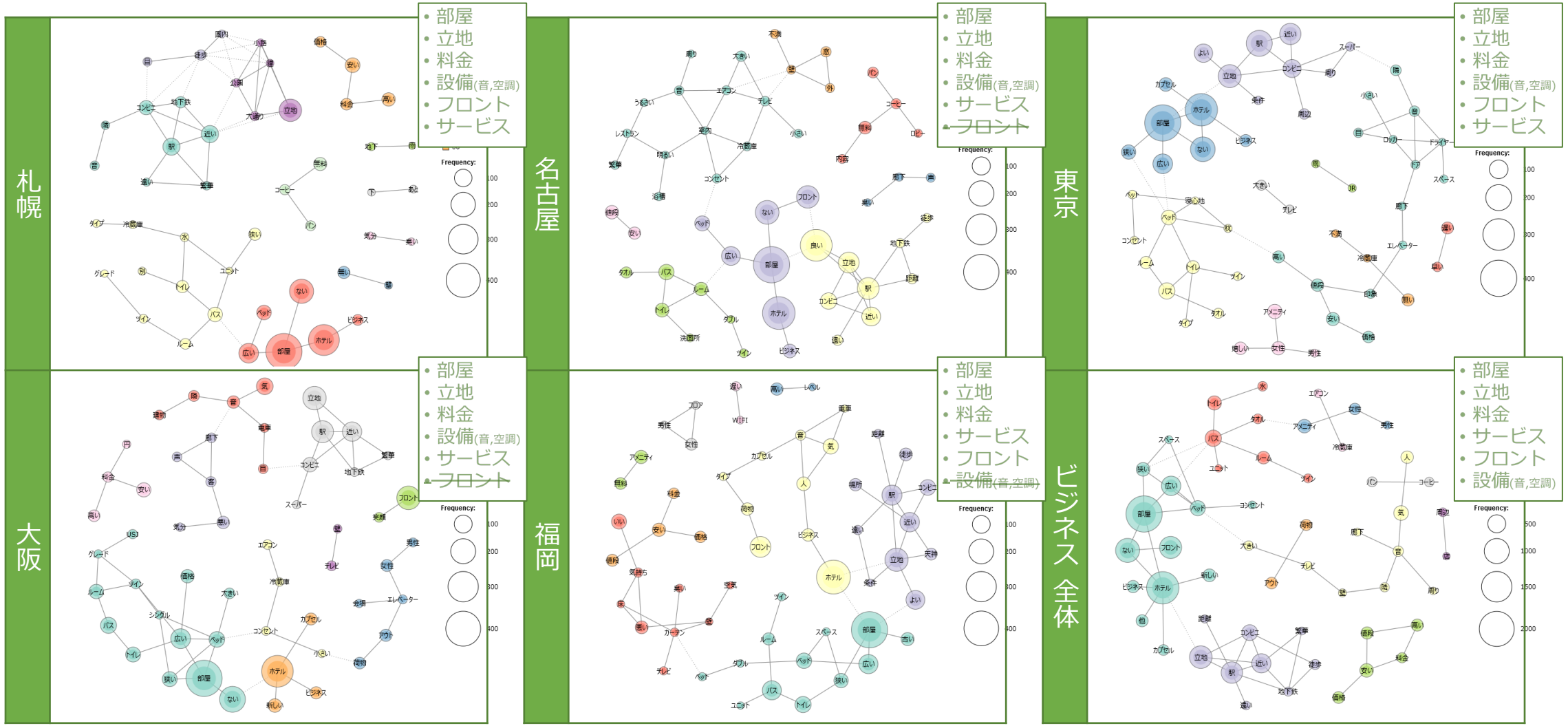
<>カテゴリー-->B_ビジネス

<>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

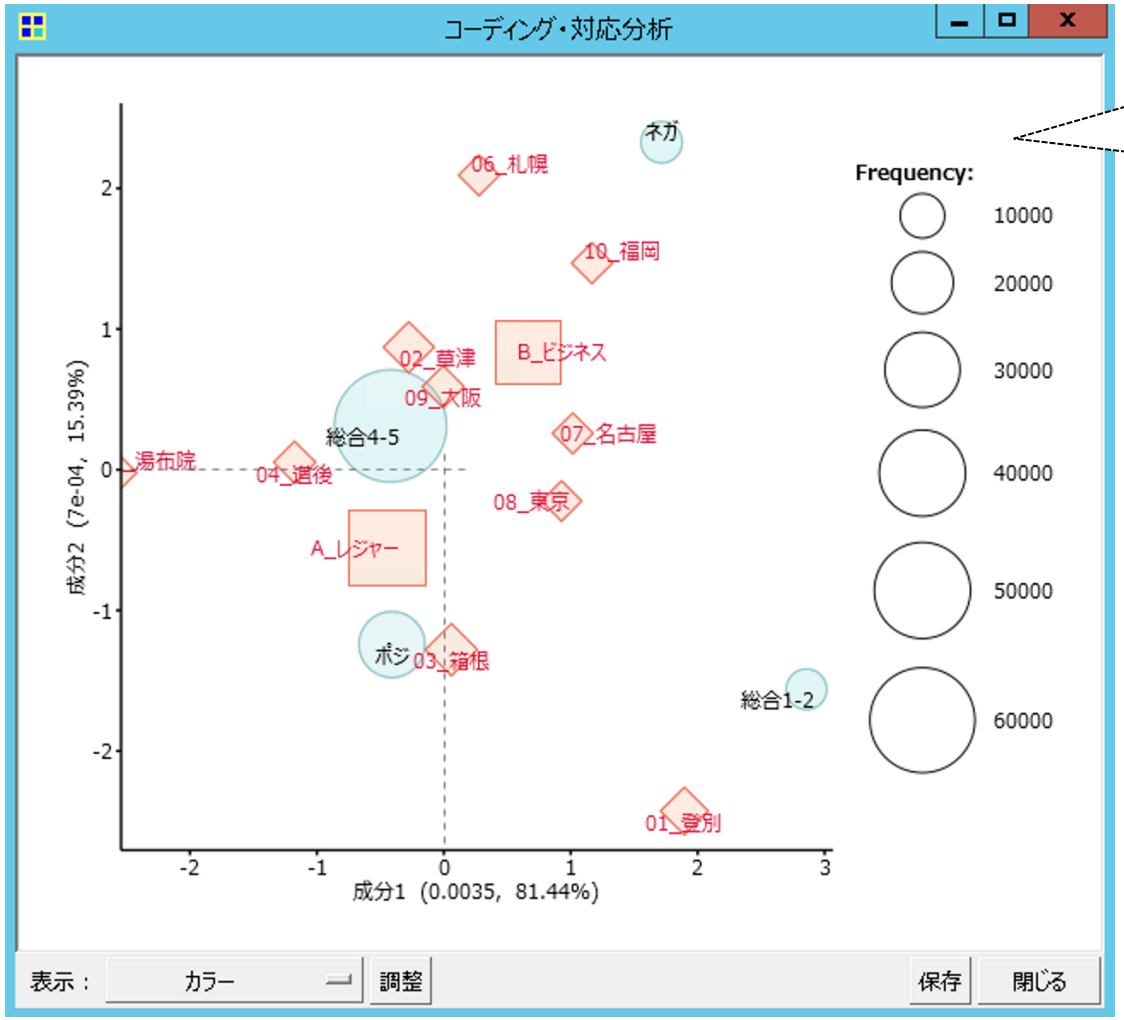
- ユーザーがどの項目に注目しているかを議論する
  - カテゴリー「レジャー」と「ビジネス」の対比
  - 「レジャー」5エリアの対比
  - 「ビジネス」5エリアの対比

# 実践 – エリアの改善案を提案する

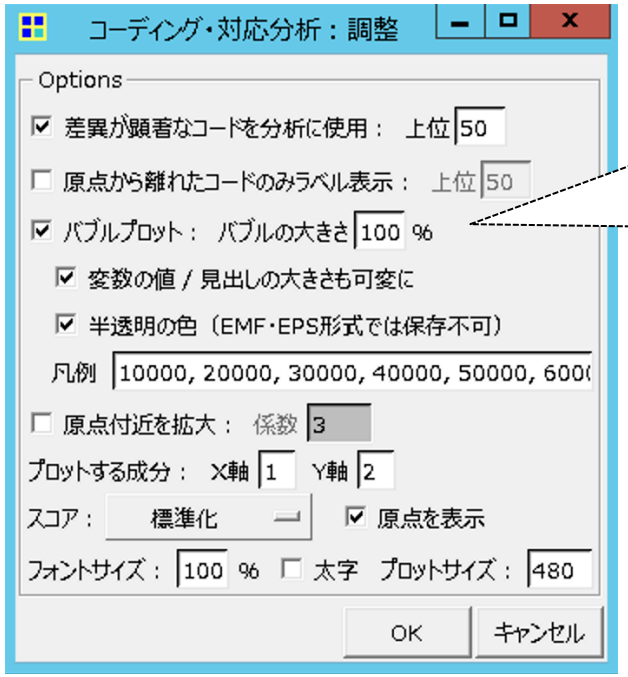
- 対称的な2エリアを選択し,てポジティブ/ネガティブの両方の意見から,比較先エリアと比較し,改善案を議論
- 主張を支持する図とユーザーの生の声(原文)を使って説明する

- 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

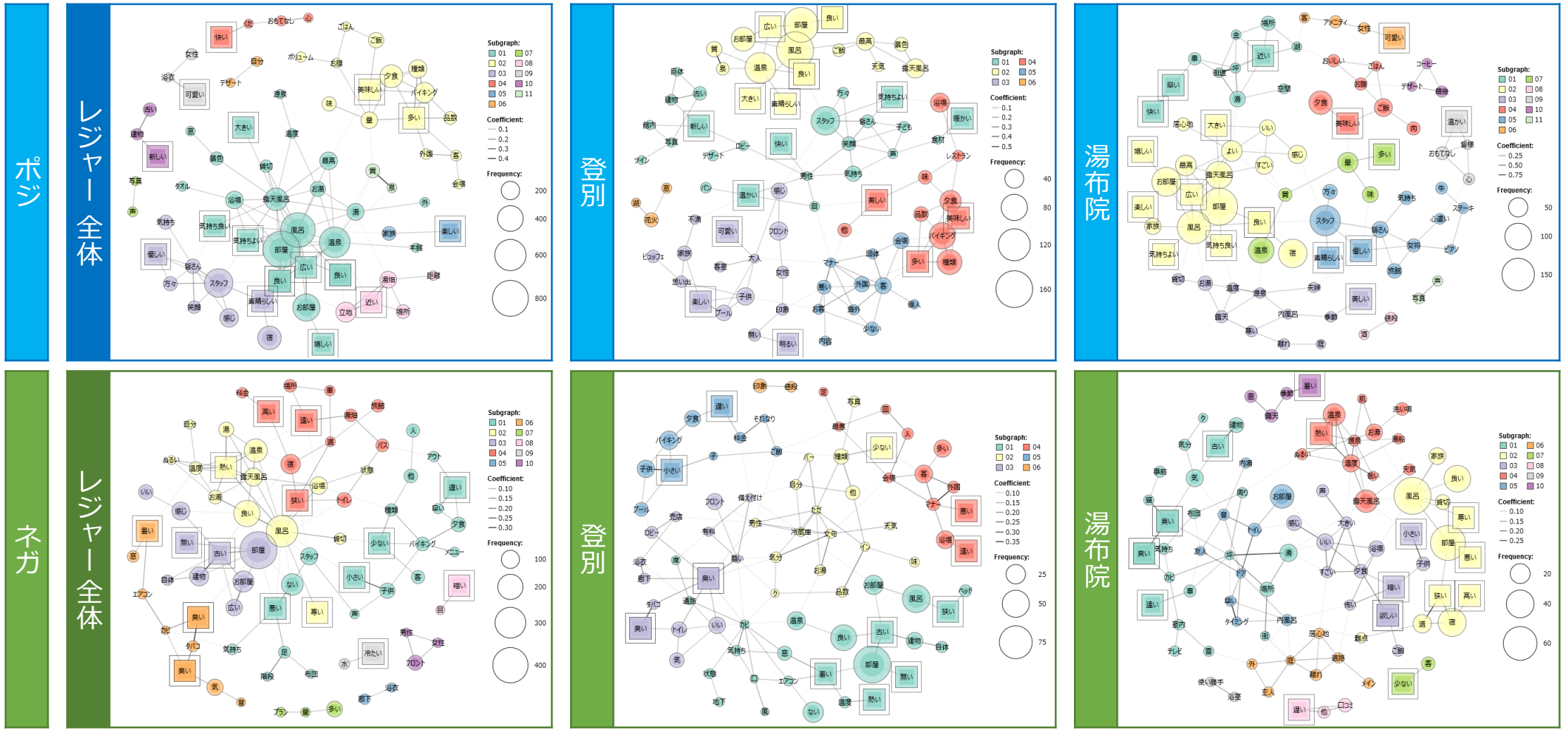
- ユーザーは何をどう評価しているか?
  1. カテゴリー「レジャー」と「ビジネス」を比較する
  2. 対照的な2エリアを比較する

- 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:*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

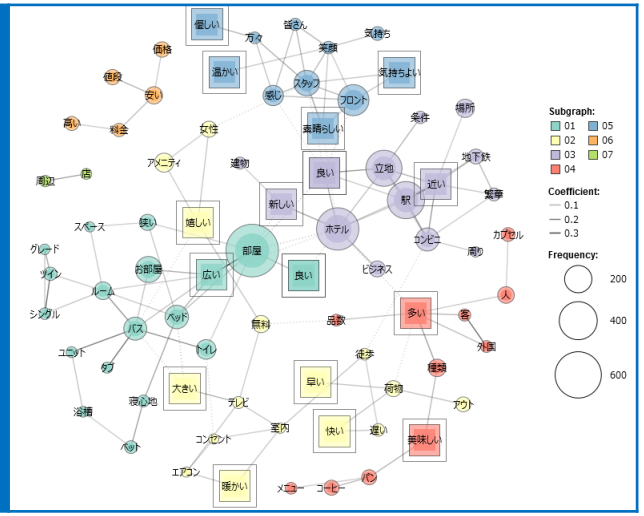
# 出力例 – 登別と湯布院のポジネガ比較



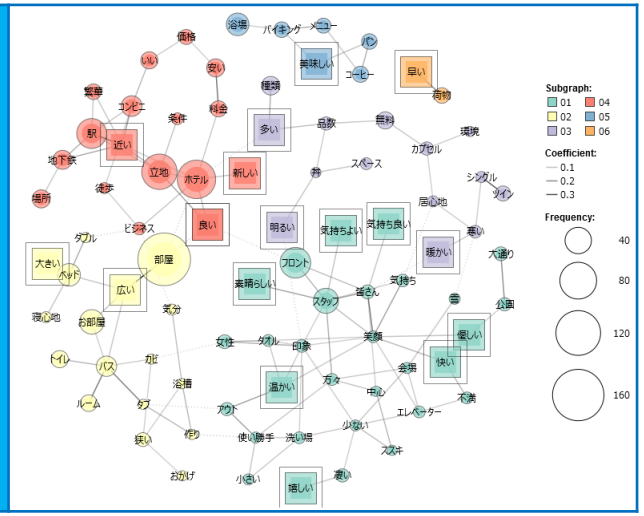
# 出力例 – 東京と札幌のポジネガ比較

ポジ

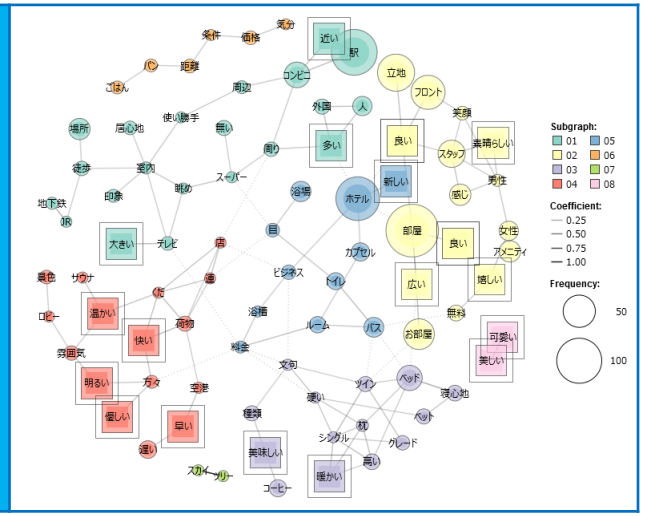
ビジネス 全体



札幌

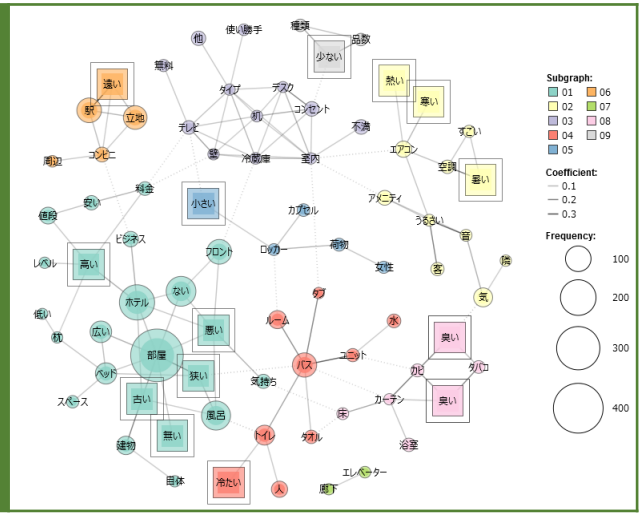


東京

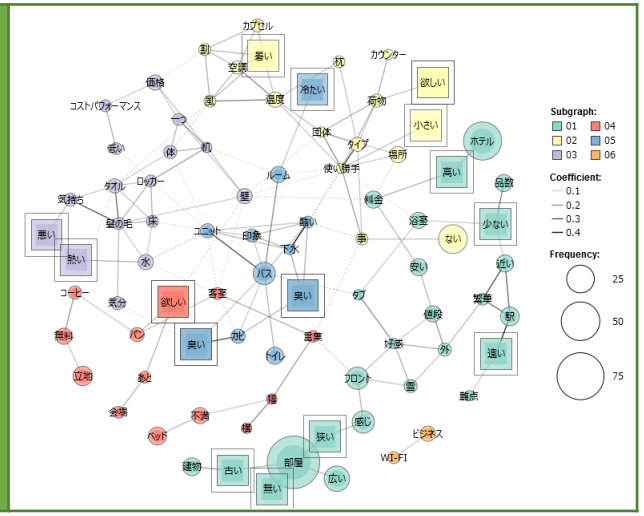


ネガ

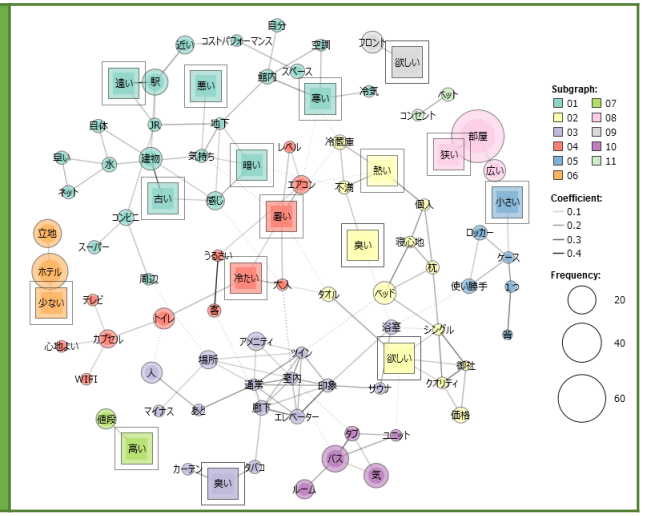
ビジネス 全体



札幌



東京



# 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」「畑」の 2単語 →「湯畑」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

```
#-------------------------------------
#   メニュー選択時に実行されるルーチン   #

sub exec{
    my $self = shift;
    my $mw = $::main_gui->{win_obj};

    my $config = {
        '友達' =>
            [
                '友人',
                '旧友',
                '親友',
                '盟友',
                '友',
            ],
        '愛に関連する語' =>
            [
                '愛情',
                '愛人',
                '恋愛',
                '愛す',
            ],
        'ほげ' =>
            [
                'ふが',
            ],
    };
```

編集前

→

```
#-------------------------------------
#   メニュー選択時に実行されるルーチン   #

sub exec{
    my $self = shift;
    my $mw = $::main_gui->{win_obj};

    my $config = {
        'お湯' =>
            [
                '湯',
            ],
    };
```

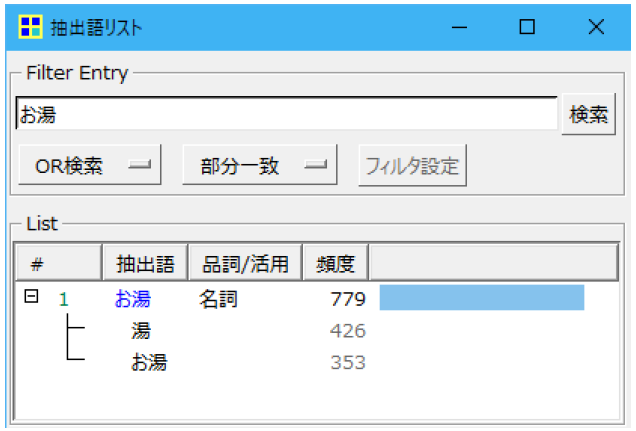
編集後

↓

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



# 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.