

テキストマイニングの実習

2016/7/7

ビジネス科学研究科
経営システム科学専攻

経歴

- 略歴

- 1992年4月 (株)NTTデータ入社, 以来, 技術開発本部にて, データウェアハウス, CRM および自然言語処理技術に関する技術開発に従事
- 2001年8月 ベンチャー子会社を設立, 役員出向
- 2013年4月～ 現職にて, 自然言語処理と機械学習を応用による技術開発に従事
- 2012年3月 修士(経営学)
- 2015年3月 博士(工学)
- 2015年5月～ 筑波大学 非常勤講師

- 業務経歴

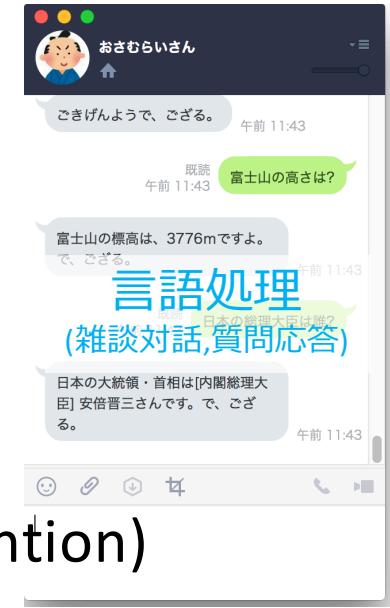
- 2000-2013年 自然言語処理と機械学習技術を応用したEメール自動返信システム
「テクノマークメール」の開発
- 2013-2013年 特許文書向け多言語機械翻訳システムの開発
- 2013-2014年 ソーシャルメディア向け語義曖昧性技術, ノイズ除去技術の開発
- 2014-2015年 リアルタイム英語音声認識を利用した英語会議支援システムの開発
- 2015-2015年 質問応答システムのためのQAコンテンツ自動抽出技術の開発

- 所属学会

- 情報処理学会, 電子情報通信学会, 言語処理学会, 人工知能学会, IEEE各会員

最近は, もっぱら深層学習

- 自然言語処理
 - 単語の表現学習 (word2vec, GloVe)
 - 文書の表現学習 (Recursive Auto Encoder)
 - NN言語モデル (Recurrent NN+LSTM → Attention)
 - 翻訳, 対話生成 (Seq-to-Seq → Attention)
- 画像認識
 - 物体認識・検出 (CNN, Region-CNN)
- 画像言語処理 (画像認識+自然言語処理)
 - 画像キャプション (CNN+RNN)
 - Visual QA (CNN+RNN → Attention)



演習日程

- 7/7
 - 説明 — データ分析の手順
 - 演習 — データの理解 (Excel)
 - 説明 — ツール (KHCoder)
 - 練習 — ツール (KHCoder)
- 7/14, 21
 - 演習 — データの分析 (KHCoder)

テキストマイニング

- ・テキストデータから有益な情報を抽出する技術の総称
大量の文書データに記述されている多種多様な内容を対象として、
その相関関係や出現傾向などから新たな知識を発見する
(那須川,1999)
- ・ビジネスでの活用
 - ・市場調査や販売戦略の立案, 製品やサービス改善, 顧客対応の改善
- ・テキストデータの例
 - ・以前
 - ・営業日報
 - ・自由記述のアンケート
 - ・近年 → CGM(コンシューマー・ジェネレイテッド・メディア)の活用が中心
 - ・レビューサイトの口コミ
 - ・ブログやマイクロブログ (Twitter, Facebook)

CGMが重要な理由 (1/2)

- スマフォやSNSの普及により,SNSの口コミやレビューサイトの重要性が増加
- 商品購入時に参考とする情報・広告:
 - 購入サイト・レビューサイトの口コミが 47.9%
 - SNSでの口コミ: 17.2%
(スマートフォン保有者 n=535)
- 影響を受けやすいSNSがある: 全体の20.8%
 - 特に,若年層を中心にSNSの口コミが浸透
 - 10代女性は49.4%, 10代男性は33.7% で高い傾向
(各 n=83)

(上) 総務省 「ICTの進化がもたらす社会へのインパクトに関する調査研究」(平成26年)

(下) 日本通信販売協会 「ネット通販に関する消費者実態調査2013」

CGMが重要な理由 (2/2)

- ネット上に膨大に蓄積されている製品やサービスに関する消費者の評価情報

消費者: 有用な情報取得・共有ツール

企業: 消費者の評判に関する情報源



- 企業などでは、自社のビジネスに役立ちそうな情報の収集と分析が重要

口コミサイトの例



- ・ホテルの口コミ数: 780万件(昨年)→**836万件** ※年間50万件

The screenshot shows the Rakuten Travel website's customer review section. At the top, it displays a large green banner with the text "お客様の声" (Customer Reviews) and "8,365,916件" (8,365,916 reviews). Below this, there are several user quotes with small profile pictures. A search bar allows users to search for reviews by keyword. The main content area is divided into two sections: "新着! 最新的クチコミ" (Latest Reviews) and "投稿されたクチコミの画像" (Images of posted reviews). The "Latest Reviews" section lists reviews from June 10, 2016, such as "ビジネスホテル秀仙閣のクチコミ" (Business Hotel Showsenkaku Review) with a score of 3.96 and "札幌プリンスホテルのクチコミ" (Sapporo Prince Hotel Review) with a score of 4.15. The "Images of posted reviews" section shows four thumbnail images of hotel rooms or food.

口コミサイトの例



2016/7/7

テキストマイニング

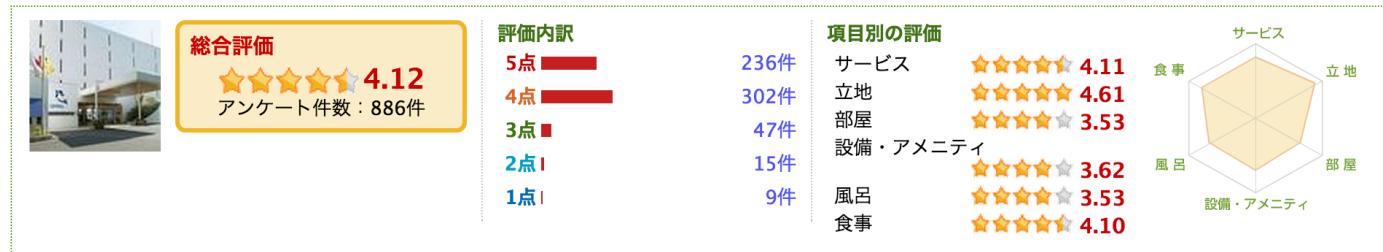
口コミサイトの例



施設紹介 プラン一覧 フォトギャラリー(76) 地図・アクセス お客様の声(886) クーポン一覧 プレゼント

鴨川シーワールドホテルのクチコミ・お客様の声

●ホテル・旅行のクチコミTOPへ



総合 ★★★☆☆ 2
投稿者さんの 鴨川シーワールドホテル のクチコミ (感想)

投稿者さん 2015年06月11日 17:03:57

良かったところ

- ・部屋からの景色（朝日最高でした）
- ・食事（品数多く、朝夕とも良かったです）
- ・フロントの方の対応（お姉さんがとても頑張っていました）以上。

掃除が行き届いているとの口コミを多く見ましたが、そうは思いませんでした。

気にかかることは多々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に救われた思いです。

評価

評価	件数	点数
★★★★☆	2	4.12
★★★★★	236	4.11
★★★★☆	302	4.61
★★★★☆	47	3.53
★★★★☆	15	3.62
★★★★☆	9	3.53
★★★★☆	1	4.10

旅行の目的

- … レジャー

同伴者

- … 家族

宿泊年月

- … 2015年06月

情報)

鴨川シーワールドホテル 2015年06月11日 19:32:50

この度は、ご利用頂きまして誠にありがとうございました。

客室内清掃の件、大変申し訳ございませんでした。重要改善として、早急に対応いたします。

今後は、この様な事の無いように、清掃・点検を強化いたします。

フロントスタッフへのお言葉、誠にありがとうございます。

モチベーションアップに繋がりますので、お客様からの声として、スタッフと共有させて頂きます。

機会がございましたら、またご利用をお待ちしております。

テキストマイニングの手順

1. データの理解

- **データの特徴を把握** → 全体、属性別件数や割合を集計
例)
 - 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?

2. 分析テーマの設定

- 解決すべき課題を設定する → 分析目的の明確化

3. テキストデータの分析

- **データ理解**と課題解決に向けたテキスト分析の実施

ポイント: 新しい発見がポンポン出ることは「まず」ありません
→ 立てた仮説をデータを使って定量的に証明することも重要!

演習 -データの理解

- データの概要
 - 楽天トラベルから収集した「お客様の声」のデータ
 - 宿泊日が2015年の下記10エリアを選択 → 93,325件
 - レジャー: 登別, 草津, 箱根, 道後, 湯布院
 - ビジネス: 札幌, 名古屋, 東京, 大阪, 福岡
- データ項目

施設情報	4項目	分類, エリア, 施設コード, 施設名
口コミ	1項目	テキスト
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別, 投稿回数

演習 -データの理解

- サイズの異なるデータセットを用意しました
 - PC のメモリが少ない等で処理が重い場合は,サンプリングデータ (**small** や **small-A**, **small-B**) もあります

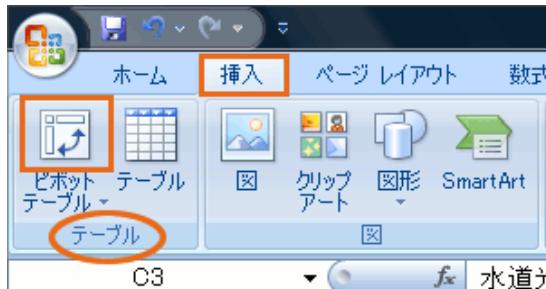
データファイル名	件数	データセット
rakuten-all.xlsx	93,325	全件 (レジャー+ビジネス)
rakuten-small.xlsx	10,000	ランダムサンプリング (レジャー+ビジネス)
rakuten-small-A.xlsx	5,000	ランダムサンプリング, レジャー のみ
rakuten-small-B.xlsx	5,000	ランダムサンプリング, ビジネス のみ

なお,テキスト形式 (拡張子が .txt のファイル) もあります. Rなど EXCEL 以外で分析する場合に使ってください

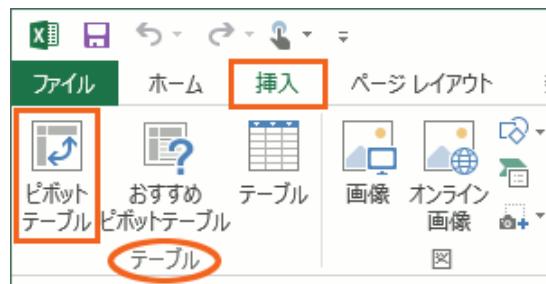
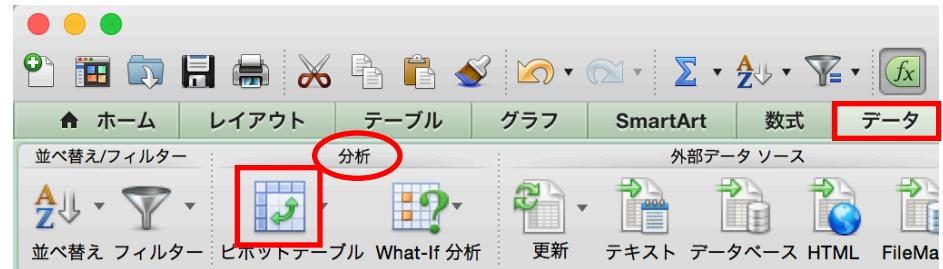
演習 データの理解

- ピボットテーブル(EXCEL)を使ったデータ集計
 - ファイル rakuten-all.xlsx を開く
 - A～S列を選択し、ピボットテーブルを作成する

Windows



Mac



【Windows】 Excel 2007・2010・2013
[挿入] タブ [テーブル] グループの [ピボットテーブル] ボタンをクリックします

【Mac】 Excel 2011
[データ] タブ [分析] グループの [ピボットテーブル] ボタンをクリックします

課題 データの理解

- EXCELを使ってデータ集計を行い,発見した特徴でデータセットを説明(要約)する
 1. 各人でデータ集計を行う
 2. 周囲の4人前後でグループを作る
 3. グループ内でデータ集計で発見した特徴を共有
 4. グループごとにデータセットの説明を発表
- データセットを説明する観点の例
 - 投稿者の属性(年代,性別)は?
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?

参考 -集計例1

データ件数 (エリア別)

行ラベル	データの個数
□ A_レジャー	10931
01_登別	1425
02_草津	1632
03_箱根	3119
04_道後	3082
05_湯布院	1673
□ B_ビジネス	82394
06_札幌	7718
07_名古屋	9856
08_東京	37475
09_大阪	17678
10_福岡	9667
総計	93325

データ件数 (年代別, 性別)

行ラベル	データの個数 / 施設コード	列ラベル	男性	女性	na	総計
10才未満			0.00%	0.00%	0.00%	0.00%
10代			0.02%	0.03%	0.00%	0.05%
20代			0.73%	0.99%	0.00%	1.72%
30代			3.95%	2.62%	0.00%	6.57%
40代			11.67%	4.07%	0.00%	15.73%
50代			10.67%	2.86%	0.00%	13.53%
60代			3.09%	0.63%	0.00%	3.71%
70代			0.45%	0.07%	0.00%	0.52%
80代			0.04%	0.01%	0.00%	0.05%
90代			0.00%	0.00%	0.00%	0.00%
100代			0.00%	0.00%	0.00%	0.00%
110代			0.03%	0.01%	0.00%	0.04%
na			0.00%	0.00%	58.07%	58.07%
総計			30.65%	11.29%	58.07%	100.00%

投稿者の傾向 (エリア別, 年代)

行ラベル	データの個数 / 列ラベル											B_ビジネス 集 総計
	□ A_レジャー	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	
10才未満	0.00%	0.00%	0.03%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
10代	0.07%	0.12%	0.10%	0.00%	0.12%	0.07%	0.06%	0.03%	0.05%	0.07%	0.03%	0.05%
20代	1.82%	3.06%	3.14%	1.20%	3.17%	2.42%	1.49%	1.46%	1.61%	1.88%	1.49%	1.63%
30代	7.86%	6.62%	6.64%	6.72%	10.40%	7.39%	6.43%	5.99%	6.30%	6.61%	7.29%	6.46%
40代	13.96%	14.89%	12.76%	14.34%	12.55%	13.65%	18.27%	17.25%	15.00%	15.77%	17.31%	16.01%
50代	14.18%	12.01%	12.98%	14.76%	11.72%	13.30%	14.01%	13.06%	13.83%	13.29%	13.19%	13.56%
60代	5.89%	5.82%	6.00%	6.55%	3.53%	5.74%	4.03%	3.37%	3.58%	2.96%	3.41%	3.45%
70代	1.40%	0.80%	1.51%	0.81%	0.78%	1.08%	0.65%	0.25%	0.47%	0.46%	0.32%	0.44%
80代	0.21%	0.06%	0.03%	0.10%	0.12%	0.09%	0.03%	0.02%	0.05%	0.02%	0.07%	0.04%
90代	0.07%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%
100代	0.00%	0.00%	0.00%	0.00%	0.12%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
110代	0.00%	0.00%	0.06%	0.03%	0.06%	0.04%	0.06%	0.05%	0.04%	0.02%	0.08%	0.04%
na	54.53%	56.62%	56.75%	55.48%	57.44%	56.19%	54.98%	58.52%	59.07%	58.90%	56.79%	58.32%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

参考 -集計例2

投稿者の傾向 (エリア別, 性別)

行ラベル	データの個数 / 1列ラベル										B_ビジネス 集総計
	A_レジャー					B_ビジネス					
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	
男性	30.46%	26.96%	25.55%	32.38%	23.85%	28.07%	35.51%	32.31%	29.58%	29.85%	33.60% 30.99% 30.65%
女性	15.02%	16.42%	17.70%	12.13%	18.71%	15.74%	9.51%	9.17%	11.35%	11.25%	9.61% 10.69% 11.29%
na	54.53%	56.62%	56.75%	55.48%	57.44%	56.19%	54.98%	58.52%	59.07%	58.90%	56.79% 58.32% 58.07%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00% 100.00% 100.00%

旅行目的の傾向 (エリア別)

行ラベル	データの個数 / 1列ラベル										B_ビジネス 集総計
	A_レジャー					B_ビジネス					
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岀	
ビジネス	9.12%	0.92%	0.96%	34.72%	1.02%	11.55%	36.76%	47.05%	44.19%	38.27%	42.81% 42.41% 38.79%
レジャー	87.58%	97.73%	97.21%	59.51%	97.07%	85.38%	56.28%	47.88%	48.76%	56.36%	51.82% 51.35% 55.34%
その他	3.30%	1.35%	1.83%	5.78%	1.91%	3.07%	6.96%	5.07%	7.04%	5.36%	5.38% 6.24% 5.87%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00% 100.00% 100.00%

同伴者の傾向 (エリア別)

行ラベル	データの個数 / 1列ラベル										B_ビジネス 集総計
	A_レジャー					B_ビジネス					
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	
一人	25.68%	10.48%	12.73%	52.86%	11.00%	25.13%	66.43%	71.25%	71.22%	65.41%	66.27% 68.94% 63.81%
家族	58.46%	68.38%	65.44%	33.03%	65.93%	55.91%	21.94%	16.55%	18.01%	21.08%	20.33% 19.13% 23.44%
友達	5.33%	7.48%	8.40%	4.61%	8.31%	6.78%	4.77%	4.08%	3.87%	5.44%	4.27% 4.36% 4.65%
恋人	6.53%	11.89%	11.03%	3.73%	13.21%	8.85%	3.34%	4.21%	3.81%	4.64%	4.50% 4.07% 4.63%
仕事仲間	2.95%	1.29%	1.41%	4.93%	0.96%	2.52%	2.86%	3.28%	2.42%	2.71%	3.91% 2.80% 2.77%
その他	1.05%	0.49%	0.99%	0.84%	0.60%	0.82%	0.66%	0.64%	0.68%	0.72%	0.72% 0.68% 0.70%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00% 100.00% 100.00%

参考 -集計例3

数値評価の構成 (エリア別)

行ラベル	データの個数 / 列ラベル										B_ビジネス 集計	
	A_レジャー					B_ビジネス						
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
1	2.25%	1.29%	0.80%	0.88%	0.42%	1.02%	1.10%	1.27%	1.04%	1.40%	0.89%	1.13%
2	4.07%	2.21%	2.73%	1.85%	0.84%	2.29%	2.80%	3.15%	3.00%	3.74%	2.80%	3.13%
3	12.84%	9.74%	8.59%	10.61%	3.47%	9.10%	12.05%	15.32%	14.16%	16.62%	13.30%	14.53%
4	40.56%	39.95%	40.01%	48.80%	23.55%	40.03%	46.18%	49.06%	48.96%	48.11%	47.57%	48.36%
5	40.28%	46.81%	47.87%	37.87%	71.73%	47.55%	37.87%	31.21%	32.85%	30.12%	35.43%	32.84%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

数値評価の平均 (エリア別)

分類	エリア	平均 / サービ		平均 / 立地		平均 / 部屋		平均 / 設備・		平均 / 風呂		平均 / 食事		平均 / 総合	
		平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
A_レジャー	01_登別	4.0	4.2	3.9	3.8	4.2	4.2	4.1	4.1	4.2	4.1	4.1	4.1	4.1	4.1
	02_草津	4.2	4.3	4.0	3.9	4.3	4.3	4.2	4.2	4.3	4.3	4.2	4.3	4.2	4.3
	03_箱根	4.3	4.1	4.2	4.0	4.0	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3
	04_道後	4.1	4.3	4.0	3.9	4.0	4.0	4.1	4.1	4.1	4.2	4.1	4.2	4.1	4.2
	05_湯布院	4.6	4.3	4.5	4.3	4.6	4.6	4.7	4.7	4.7	4.7	4.6	4.7	4.7	4.7
B_ビジネス	06_札幌	4.0	4.2	4.1	3.9	3.8	3.8	4.0	4.0	4.0	4.2	4.0	4.2	4.0	4.2
	07_名古屋	3.9	4.2	3.9	3.8	3.6	3.6	3.9	3.9	3.9	4.1	3.9	4.1	3.9	4.1
	08_東京	3.9	4.3	4.0	3.8	3.7	3.7	3.8	3.8	3.8	4.1	3.9	4.1	3.9	4.1
	09_大阪	3.9	4.2	3.9	3.7	3.6	3.6	3.9	3.9	3.9	4.0	3.9	4.0	3.9	4.0
	10_福岡	4.0	4.3	4.0	3.8	3.7	3.7	4.0	4.0	4.0	4.1	4.0	4.1	4.0	4.1

数値評価の平均 (レジャー, ビジネス別)

分類	平均 / サービ		平均 / 立地		平均 / 部屋		平均 / 設備・		平均 / 風呂		平均 / 食事		平均 / 総合		
	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	
A_レジャー	4.2	4.2	4.1	4.0	4.2	4.2	4.3	4.3	4.2	4.2	4.3	4.3	4.3	4.3	4.3
B_ビジネス	3.9	4.3	3.9	3.8	3.7	3.7	3.9	3.9	3.7	3.7	4.1	4.1	4.1	4.1	4.1

論文紹介－辻井・津田,2012

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)

分類	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・施設	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.2	4.2	4.1	4.0	4.2	4.3	4.3
B_ビジネス	3.9	4.3	3.9	3.8	3.7	3.9	4.1

- ・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

数値評価のみから違いを見つけるのは難しい

→ レジャーとビジネスでは、評価すべき項目も異なる

→ 同じ点数でも、テキストと対応付ければ差異がある

KH Coder –立命館の樋口先生が開発

社会調査データを分析するために開発された
フリーのテキストマイニングツール

- 高機能でも商用可能でフリー
- Rを用いた多変量解析と可視化
- 実装されている分析手法
 - 階層的クラスター分析
 - 多次元尺度構成法(MDS)
 - 対応分析
 - 共起ネットワーク
 - 自己組織化マップ
 - 文書のクラスター分析

論文検索サービスも提供 →

[http://khc.sourceforge.net/bib.html?
year=2015&auth=all&key=](http://khc.sourceforge.net/bib.html?year=2015&auth=all&key=)

[KH Coderを用いた研究事例]

[KH Coderに戻る]

KH Coderを用いた研究事例のリストです。※KH Coderを用いたご研究の成果を発表された際には、書誌情報をお送りいただけますと幸いです。

出版年: すべて -2005 06 07 08 09 10 11 12 13 14 2015-

著者名: すべて あ か さ た な は ま や ら わ A-Z

キーワード: クリア

ヒット件数: 058 / 961

荒瀬雅子 2015 「災害時の『やさしい日本語』を教室教材として使用する方法を探る
—ラジオ放送用災害時音声素材を中心に—」 『龍谷大学国際センター研究年報』
24: 21-34

[KH Coderを用いた研究事例のリスト](#) 1206件

※ 2016/6/18 現在 (昨年同時期961件)

KH Coder の情報

・ ホームページ

<http://khc.sourceforge.net/>

Index

概要

KH Coderとは、テキスト型（文章型）データを統計的に分析するためのフリーソフトウェアです。アンケートの自由記述・インタビュー記録・新聞記事など、さまざまな社会調査データを分析するため制作しました。「計量テキスト分析」または「テキストマイニング」と呼ばれる方法に対応。

- 主な機能と分析の手順
- KH Coderを用いた研究事例のリスト [961件](#)

機能紹介

スクリーンショット :

- データ中の言葉を探索する
- 文書の検索とコーディング
- Rを用いた多変量解析と可視化 [New!](#)
- 複数の言語・環境に対応

スライド (slideshare) :

- 漱石「こころ」チュートリアル
- アンケート自由回答の分析例ほか
- 英語テキストのKWIC検索と分析
- カスタマイズと自動化

KH Coderの入手

- KH Coderのダウンロード ([Ver. 2.Beta.32g](#), 2015 07/06, RSS, Twitter)
- 必要なソフトウェア / ハードウェア
- バージョンアップ履歴
- 使用許諾

サポート

- 掲示板 (ユーザーフォーラム)
- チュートリアル & ヒント [New!](#)
- よくある質問 (FAQ)
- 関連リンク

科 研 費
R RITSUMEIKAN

・ スライドも充実

<http://www.slideshare.net/khcoder/presentations>

フリーソフトウェア「KH Coder」を使った計量テキスト分析

KH Coder チュートリアル

—手軽なマウス操作による分析からプラグイン作成へ

漱石「こころ」を題材に【スライド版】

・ 参考書



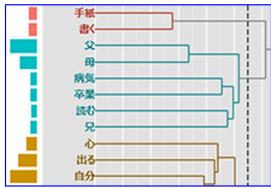
KH Coder の主な分析手法

分析手法	解説
階層クラスタリング	<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, コサイン距離を用い,いわゆる Ward法, 群平均法, 最遠隣法で樹形図を作成
多次元尺度構成法	<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, コサイン距離を用い, クラシカル, Kruskal, Sammon 法のいずれかで2次元にプロット
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軸上にプロット外部変数も同時にプロット可能
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
自己組織化マップ(SOM)	<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方法で, 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
文書のクラスター	<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

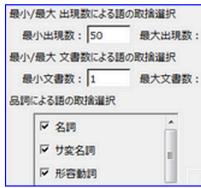
KH Coder –スナップショット

階層的クラスター分析

抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけではなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



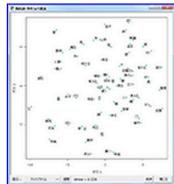
抽出語は出現数や品詞で選択



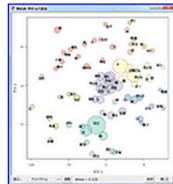
コードはチェックボックスで直接選択

多次元尺度構成法（MDS）

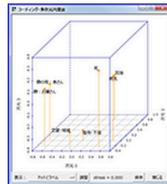
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



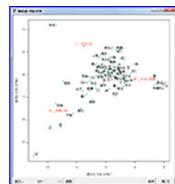
New! クラスタリングと色分け



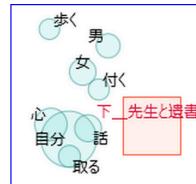
3次元の解

対応分析

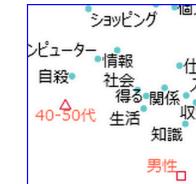
同じく抽出語またはコードを用いての、対応分析です。



同時布置図



New! バブルプロット



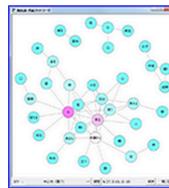
複数の外部変数を用いた多重対応分析

2016/7/7

テキストマイニング

共起ネットワーク

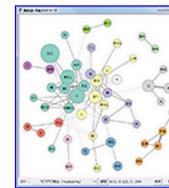
抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



共起の程度が非常に強いものだけを線で結んだ図



やや弱い共起関係も画面に含め、自動的にグループ分け（色分け）



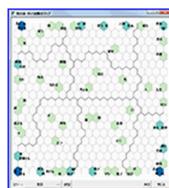
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

自己組織化マップ

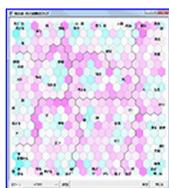
抽出語またはコードを用いての、自己組織化マップです。



クラスター色分け



頻度のプロット



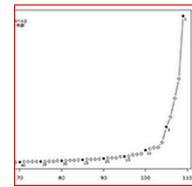
U-Matrix

文書のクラスター分析

文書の分類を行うクラスター分析です。

文書クラスター分析	
各文書ごとに割り当てる文書	
クラスター番号	文書番号
1	1-106 -10 9.007
2	2 -107 3 9.311
3	3 -10 9.311
4	4 -75 -76 9.375
5	5 -50 -51 9.456
6	6 -93 -92 10.547
7	7 -99 -92 10.703
8	8 -96 -97 10.703
9	9 -100 2 10.992
10	10 -106 2 10.992
11	11 -107 3 10.992

クラスター分析の結果画面



併合水準のプロット。クラスター数5付近から併合水準が急上昇。10でも少し上がっているので、この場合クラスター数は11が良いか。



文書のデンドログラム。左の棒グラフは各文書の長さをあらわす。なお、文書数が500を超える場合、デンドログラムは表示不可。

演習 - KH Coderを使う

- ・ダウンロードとインストール
 - ・<http://khc.sourceforge.net/dl.html>



- ① ここをクリックすると遷移先のページからダウンロードが始まります
- ② 指示に従いインストール

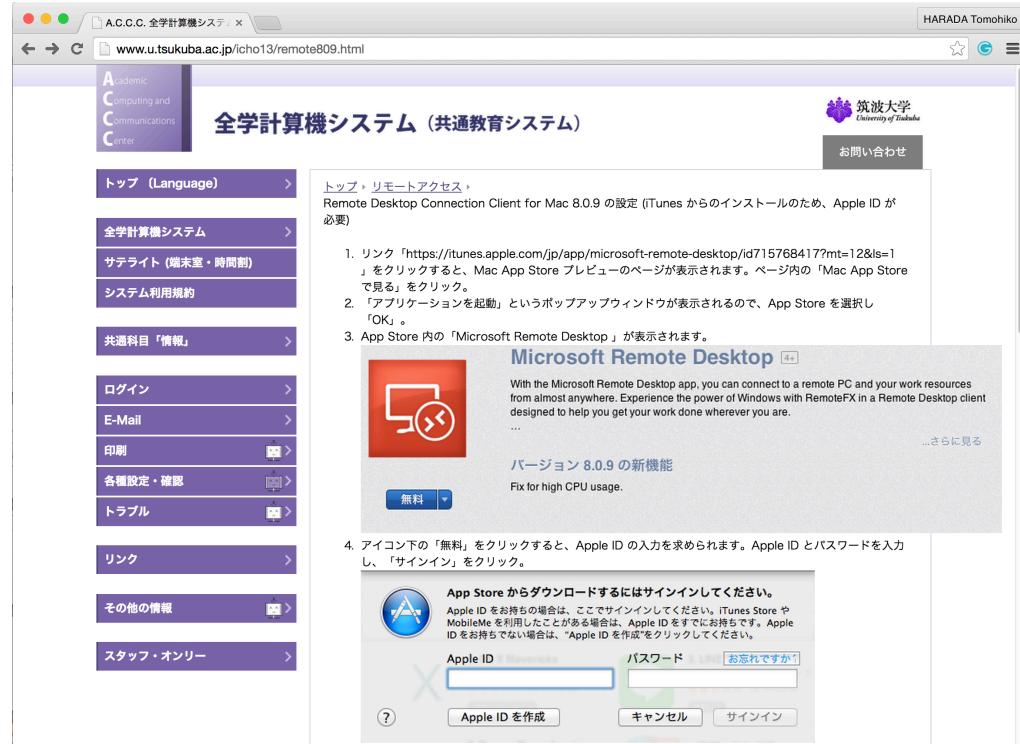
自己解凍ファイルです。このファイルを実行（ダブルクリック）し、開いたWindowの「Unzip」ボタンをクリックすると、（特に変更しなければ）「C:\khcoder」というフォルダにすべてのファイルが解凍されます。解凍されたkh_coder.exeを実行すると、KH Coderが起動します。

注意:

全学のRDPの場合は、ログイン後のデスクトップ上に「khcoder」というフォルダを作成して、その中に解凍してください

演習 – Windows環境のないMac

- 全学計算機システム(RDP)を使います
 - <http://www.u.tsukuba.ac.jp/icho13/remote809.html>



左記のページにある説明に従って、

- ツール (MS Remote Desktop) のインストール
- 全学計算機システム (Windows)へのログイン

を行ってください

演習データの公開場所

- <https://github.com/haradatm/gssm-201607>

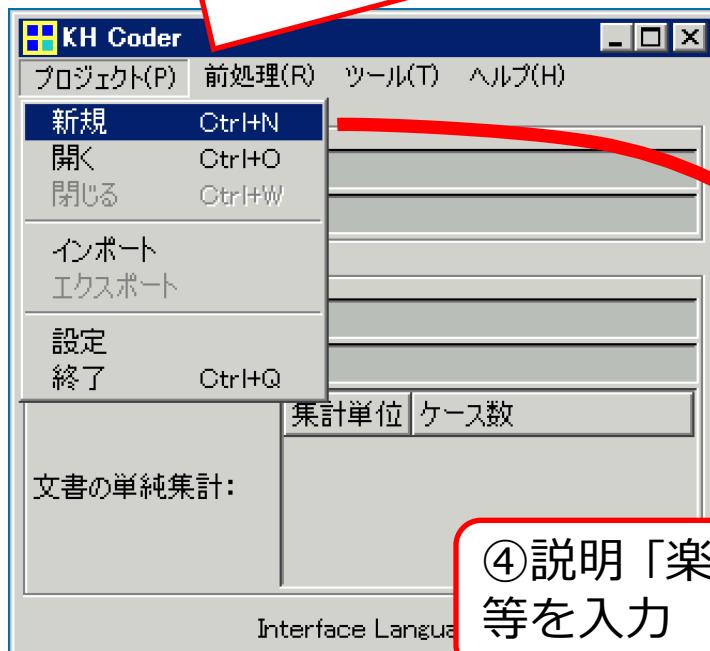
The image shows a GitHub repository interface with three main sections:

- Left Panel (Master Branch):** Shows the root directory structure:
 - first commit
 - Tomohiko HARADA authored 9 minutes ago
 - 00-slides
 - 01-data** (highlighted with a red box)
 - 02-tools
 - 03-samples
 - 04-docs** (highlighted with a red box)
 - .gitignore
 - README.md
- Middle Panel (01-data Branch):** Shows the contents of the 01-data folder:
 - first commit
 - Tomohiko HARADA authored 6 minutes ago
 - ..
 - coding-rule.txt.zip** (highlighted with a red box) KH Coder演習で使用(要解凍)
 - rakuten-eval-sjis.txt
 - rakuten-eval-utf8.txt
 - rakuten-eval.xlsx** (highlighted with a red box) 演習用データ(変更なし)
- Bottom Panel (04-docs Branch):** Shows the contents of the 04-docs folder:
 - first commit
 - Tomohiko HARADA authored 12 minutes ago
 - ..
 - khcoder-slides.zip** (highlighted with a dashed blue box) KH Coder のチュートリアル
(from SlideShare)

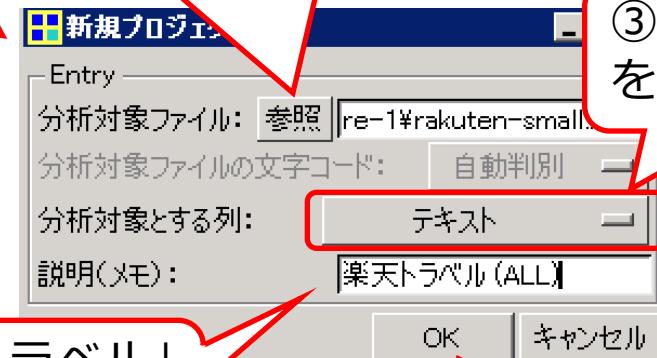
演習 —プロジェクトの作成

- ・ファイル **rakuten-small.xlsx** を開く

①メニューから「プロジェクト」「新規」を選択 (注1)



②「参照」をクリックして
「rakuten-small.xlsx」を開く



③「テキスト」
を選択 (注2)

④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

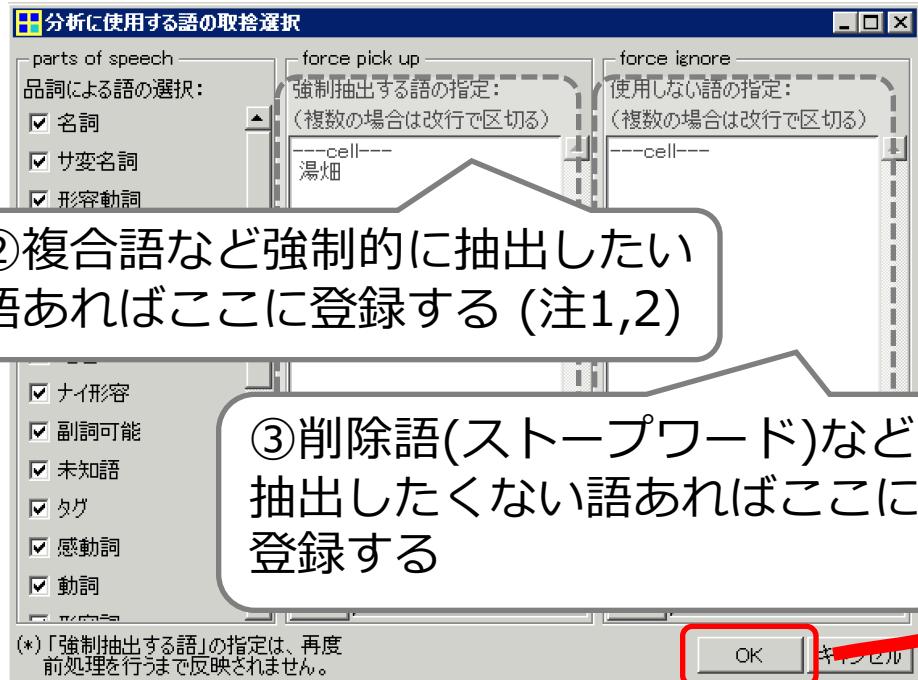
注1: 次回 KH Coderを起動した時は「新規」ではなく「開く」を選択します

注2: ②のファイル選択後,ここに「テキスト」等の選択項目が表示されるまで数分がかかります

演習 - 前処理

・形態素解析を行う

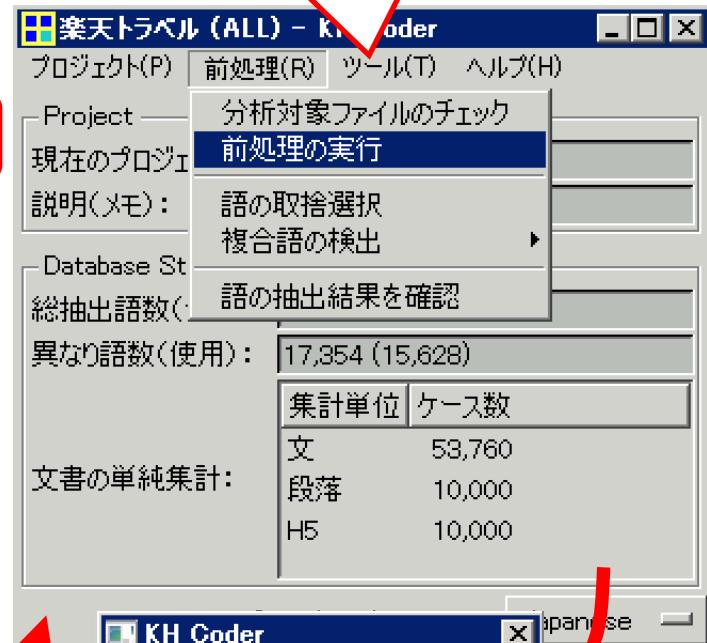
①メニューから「前処理」「語の取捨選択」を選ぶ



②複合語など強制的に抽出したい語あればここに登録する (注1,2)

③削除語(ストップワード)など抽出したくない語あればここに登録する

④メニューから「前処理」「前処理の実行」を選ぶ



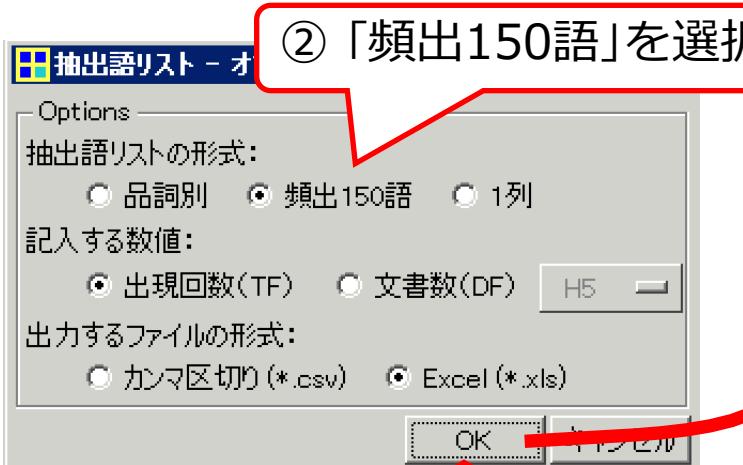
注1: EXCELファイルを読み込んで分析する場合, あらかじめ「---cell---」が入力されています

注2: メニューから「前処理」「複合語の検出」を選ぶと, 複合語候補の一覧を出力できます

演習 一頻出語を確認する

- ・ 頻出語リストを出力する

①メニューから「ツール」「抽出語」「抽出語リスト」を選択



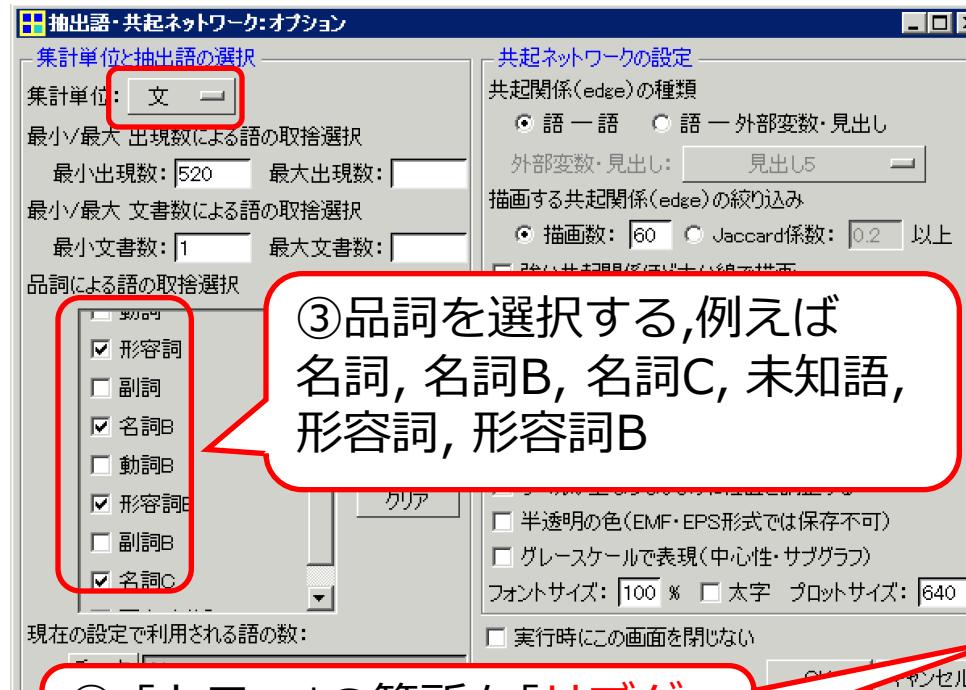
③ 「OK」をクリック

A	B	C	D	E	F	G	H
1 抽出語	出現回数	抽出語	出現回数	抽出語	出現回数		
2 部屋	4681	チェックイン	682	ベッド	419		
3 思う	4485	バイキング	659	気持ちよい	419		
4 利用	4035	初めて	648	十分	418		
5 良い	3984	旅行	648	チェック	416		
6 ホテル	3110	清潔	644	女性	413		
7 風呂	2921	値段	640	湯畑	413		
8 宿泊	2811	古い	639	トイレ	410		
9 食事	2425	バス	629	お湯	399		
10 朝食	2308	子供	585	悪い	398		
11 満足	2267	夜	575	アメニティ	395		
12 温泉	2031	過ごす	572	来る	394		
13 美味しい	1725	入れる	569	接客	384		
14 行く	1501	場所	564	価格	382		
15 お部屋	1467	機会	563	料金	371		
16 対応	1448	本当に	546	期待	369		
17 立地	1427	狭い	545	従業	367		
18 広い	1410	素晴らしい	543	見える	366		
19 宿	1304	丁寧	541	置く	365		
20 大変	1300	プラン	534	雰囲気	361		
21 スタッフ	1193	駐車	531	もう少し	360		
22 サービス	1179	人	529	草津	357		
23 フロント	1145	月	524	施設	351		
24 料理	1109	安い	523	荷物	350		
25 近い	1099	コンビニ	520	ビジネス	345		
26 少し	1082	過ごせる	517	出張	341		
27 便利	1082	些	515	II. 一	327		

演習－共起ネットワークの作成

①メニューから「ツール」「抽出後」「共起ネットワーク」を選ぶ

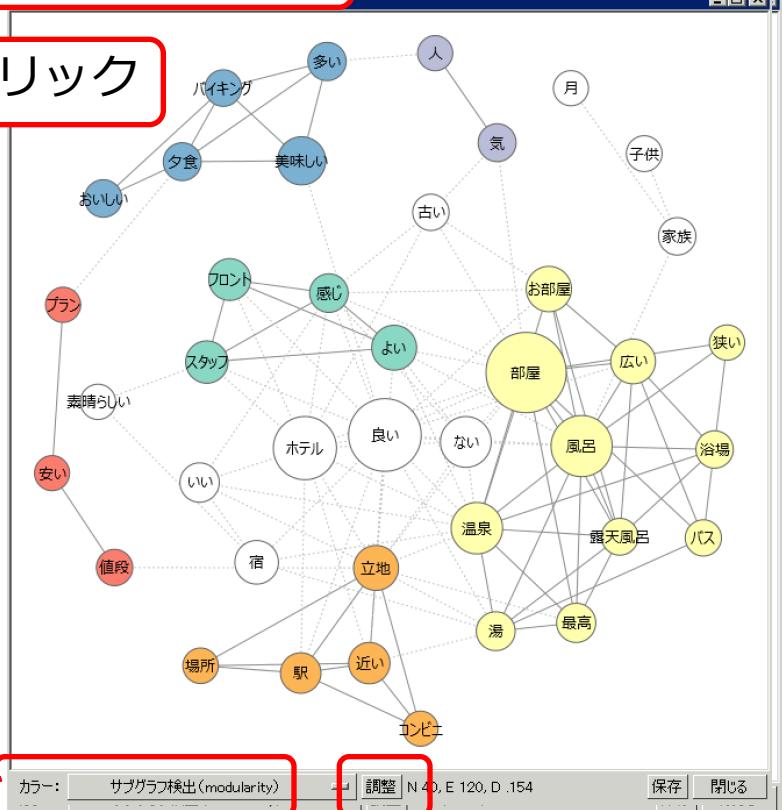
②「集計単位」として「文」を選んで「OK」をクリック



③品詞を選択する, 例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B

⑤「カラー」の箇所を「サブグ
ラフ検出(modularity)」に変更

④「調整」をクリックして「描画数」に120 を
入力し、「出現数の多い語ほど…」をチェック



KH Coder の品詞体系

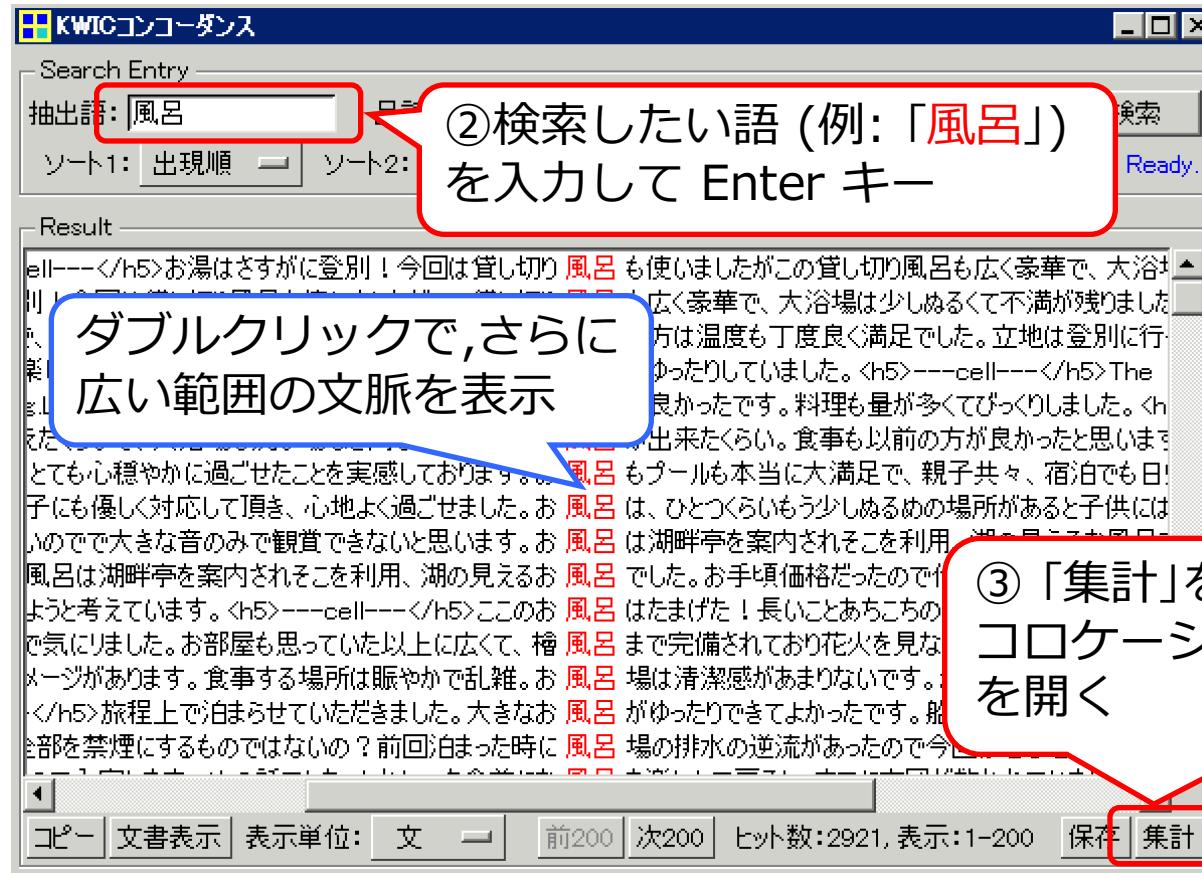
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般（漢字を含む 2 文字以上の語）
名詞 B	名詞一般（平仮名のみの語）
名詞 C	名詞一般（漢字 1 文字の語）
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立（漢字を含む語）
動詞 B	動詞-自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平仮名のみの語）
副詞	副詞（漢字を含む語）
副詞 B	副詞（平从名のみの語）
否定助動詞	助動詞「ない」「まい」「ぬ」「ん」
形容詞（非自立）	形容詞-非自立（「がたい」「つらい」「にくい」等）
その他	上記以外のもの

「KH Coder 2.x リファレンス・マニュアル」P.11 より

演習 -KWICコンコーダンス1

- テキスト中でその語がどう使われているか

①メニューから「ツール」「抽出後」「KWICコンコーダンス」を選ぶ



練習 – KWICコンコンコードанс2

①前のページの手順でコロケーション統計を開く

The screenshot shows the 'Kotoko' software interface for frequency analysis. The main window displays a table of words and their co-occurrence statistics. A callout box highlights the word '広い' (Wide) appearing 82 times two positions after '風呂'.

Result Table:

N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	スコ
1	良い	形容詞	227	80	147	42	12	13	13	0	2	64	39	26	16	78.9
2	広い	形容詞	183	48	135	10	11	10	16	1	1	82	27	14	11	73.1
3	よい	形容詞B	94	38	56	23	7	4	4	0	0	20	15	8	12	30.0
4	狭い	形容詞	57	6	51	7	7	7	7	7	7	15	5	4	22.4	
5	ない	形容詞B	63	26	37	11	11	11	11	11	11	3	10	12	21.9	
6	気持ちよい	形容詞	33	9	24	6	6	6	6	6	6	9	7	10	10.7	
7	熱い	形容詞	34	9	25	8	8	8	8	8	8	6	5	12	12.4	
8	大きい	形容詞	31	11	20	3	3	1	4	0	0	10	6	1	2	11.5
9	気持ち良い	形容詞	25	7	18	2	1	0	0	0	0	8	7	2	1	10.1
10	いい	形容詞B	35	18	17	8	3	3	4	0	0	6	4	2	5	11.1
11	古い	形容詞	23	5	15	2	1	3	1	1	0	7	4	3	1	8.9
12	素晴らしい	形容詞	18	4	15	3	0	1	0	0	0	8	1	2	4	6.5
13	小さい	形容詞	19	5	14	0	3	0	0	0	1	10	2	1	0	8.6

Filter Settings (Top Right):

- 品詞による語の選択:
 - 形容詞
 - 副詞
 - 名詞B
 - 動詞B
 - 形容詞B
 - 副詞B
 - 名詞O
- 「合計」列による語の選択:
 - 最小: 1
- 表示する語の数:
 - 上位: 200

Buttons at the Bottom Left:

- コピー
- フィルタ設定 (highlighted)
- ソート: 右合計 (highlighted)

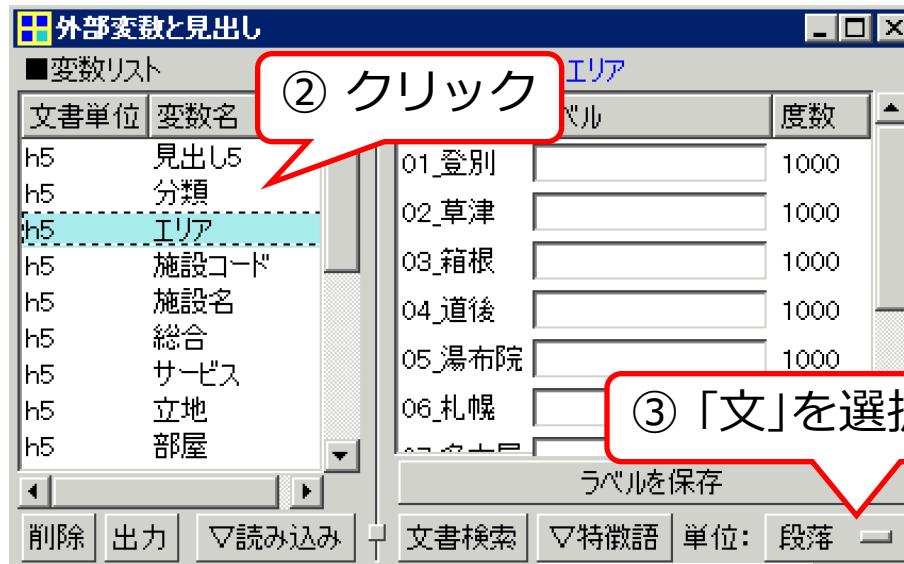
Annotations:

- 「右1」は右側の1つ目(=直後)に出現していた回数
- 「広い」は「風呂」の2つ後に 82 回出現
- ② 「右合計」でソート
- ③ 表示する語を品詞(例: 形容詞, 形容詞B)とともに選択

練習 - KWICコンコーダンス2

- 外部変数(エリア)ごとの特徴語を確認する

①メニューから「ツール」「外部変数と見出し」「リスト」を開く



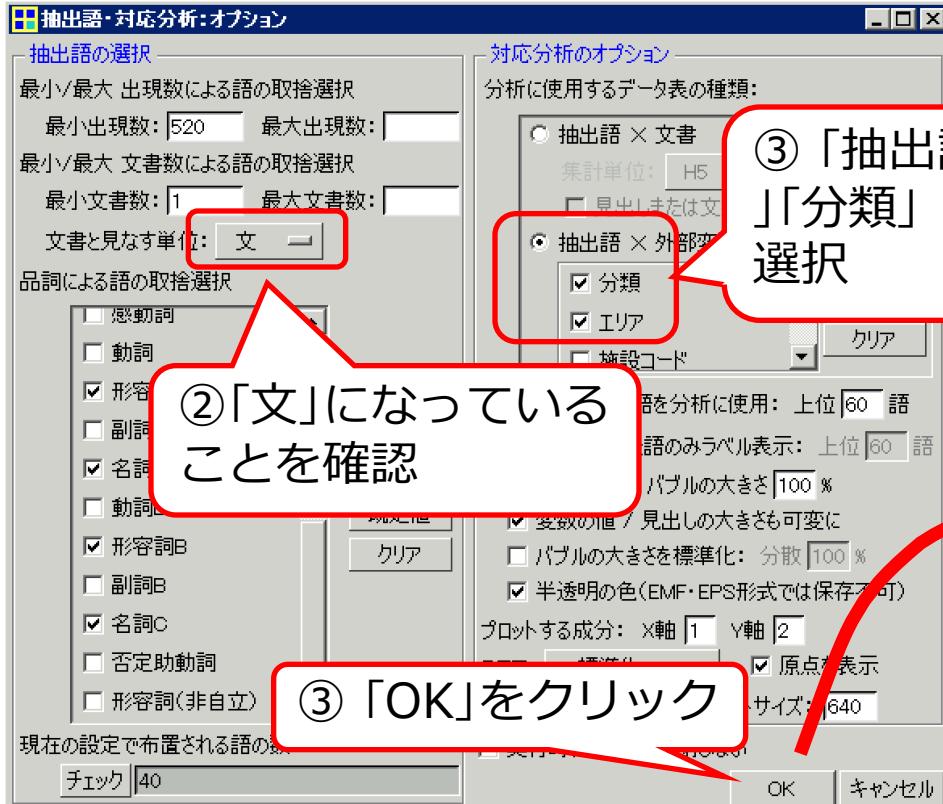
A	B	C	D	E	F	G	H	I	J	K
1	01_登別	02_草津	03_箱根	04_道後						
2										
3	食事	.132	畠	.343	食事	.157	道後	.191		
4	風呂	.105	湯	.326	箱根	.146	温泉	.122		
5	温泉	.101	草津	.278	風呂	.136	松山	.117		
6	バイキング	.098	温泉	.166	温泉	.126	朝食	.086		
7	部屋	.092	食事	.154	夕食	.121	本館	.085		
8	宿泊	.089	風呂	.149	料理	.118	立地	.080		
9	思う	.084	宿	.124	美味しい	.116	ホテル	.077		
10	満足	.082	良い	.111	満足	.115	美味しい	.064		
11	花火	.081	料理	.108	宿	.110	対応	.064		
12	夕食	.080	満足	.107	良い	.110	便利	.061		
13	05_湯布院		06_札幌		07_名古屋		08_東京			
14	宿	.178	札幌	.171	名古屋	.202	駅	.116		
15	料理	.164	朝食	.101	朝食	.096	利用	.100		
16	食事	.160	ホテル	.092	利用	.091	便利	.100		
17	湯布院	.159	利用	.087	ホテル	.089	ホテル	.088		
18	美味しい	.157	立地	.075	駅	.086	部屋	.088		
19	露天風呂	.127	便利	.073	立地	.078	近い	.085		
20	風呂	.126	すすきの	.071	近い	.082	東京	.082		
21	温									
22	あ									
23	フ									
24	ナ									
25	メ									

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

注: Jaccard係数は共起尺度のひとつで、共通要素の数を少なくとも一方にある数で割ったもの

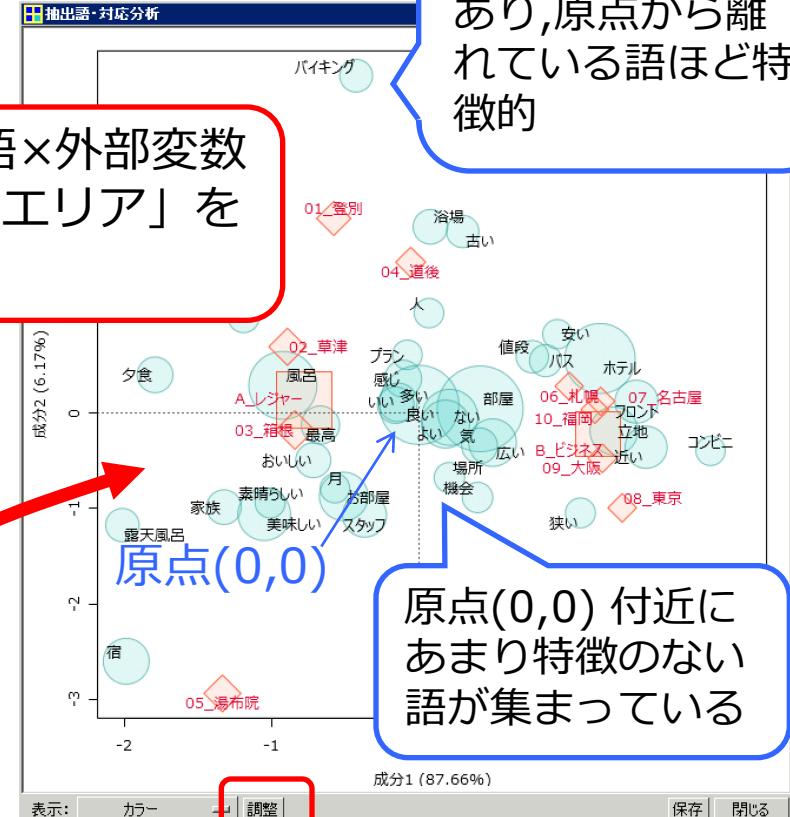
練習 一対応分析による探索1

- ①メニューから「ツール」「抽出語」「対応分析」を選ぶ



原点(0.0)から見て「エリア」方向にあり、原点から離れている語ほど特徴的

- ③「抽出語×外部変数」「分類」「エリア」を選択



原点(0,0)付近にあまり特徴のない語が集まっている

- ④「調整」をクリックして「バブルプロット」をチェック

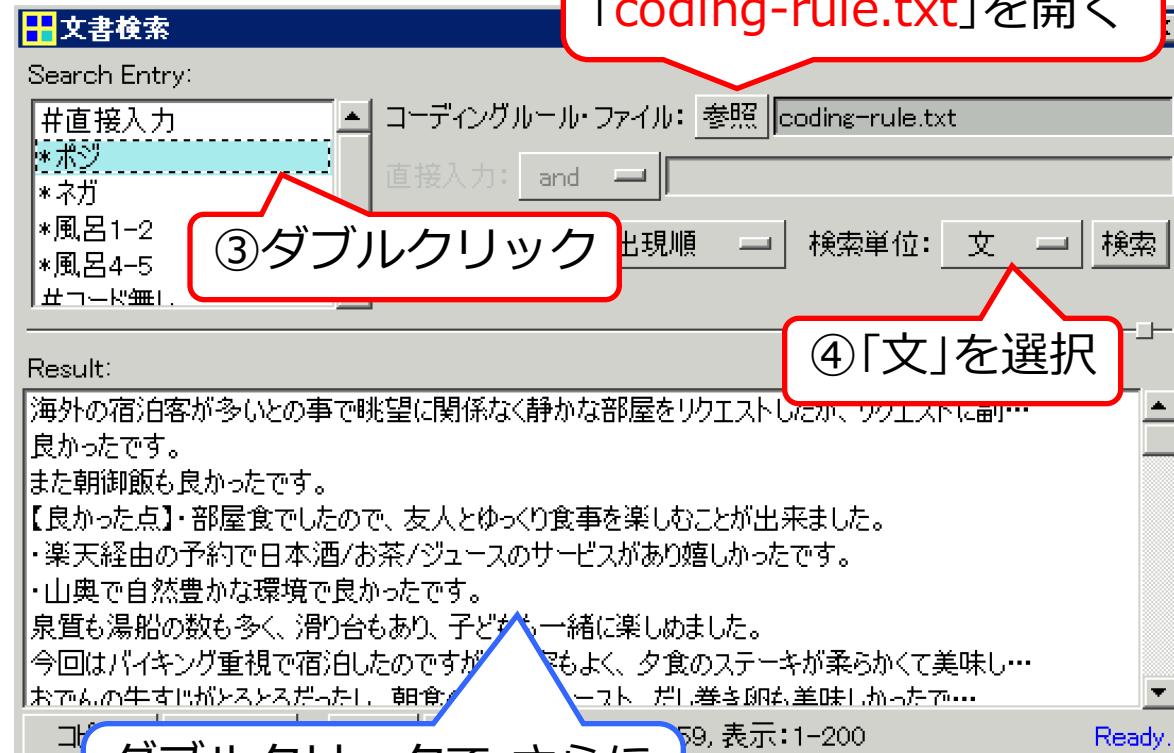
練習－コーディングルール1

①メニューから「ツール」「文書」「文書検索」を選ぶ

②「参照」をクリックして
「coding-rule.txt」を開く

③ダブルクリック

④「文」を選択



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

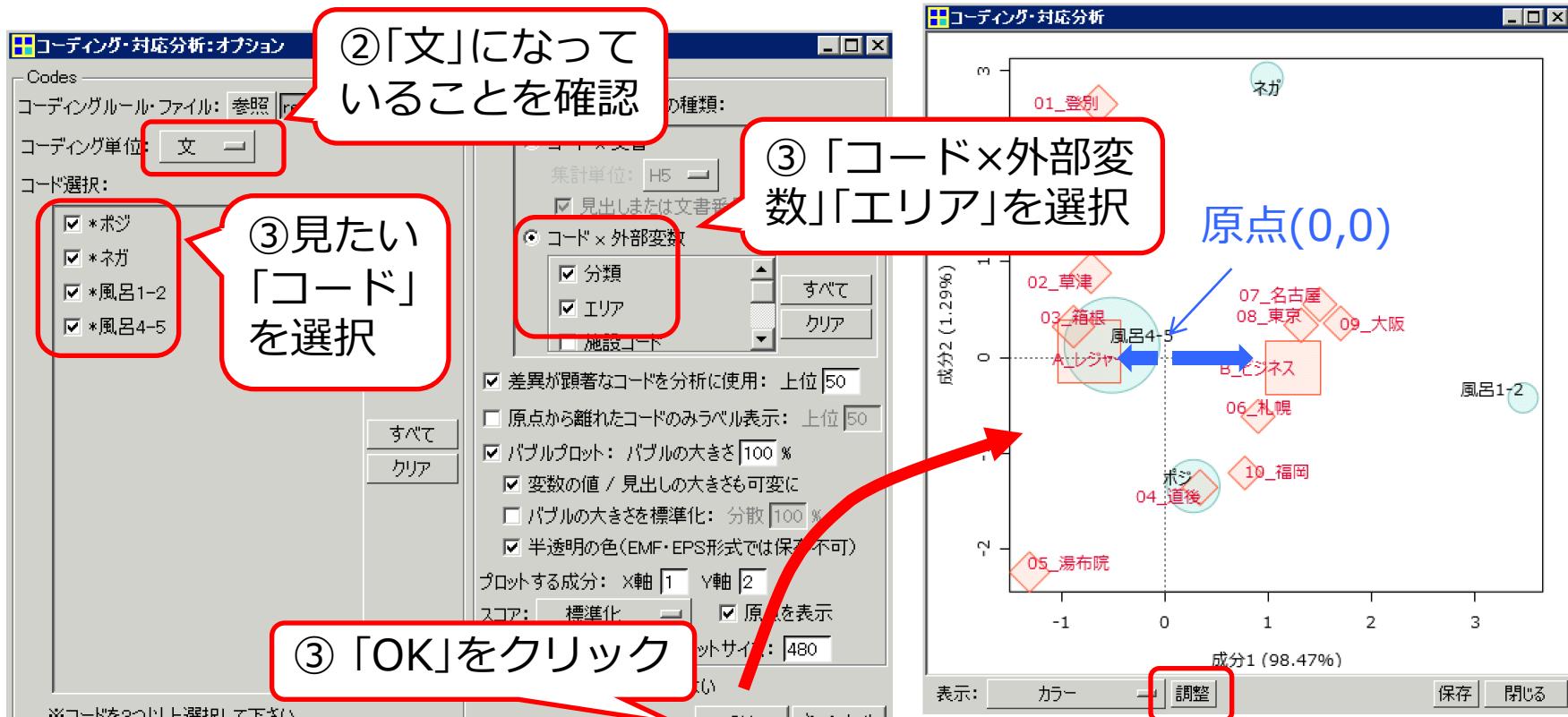
外部変数

*風呂4-5

<>風呂-->4 | <>風呂-->5

練習 一対応分析による探索2

①メニューから「ツール」「コーディング」「対応分析」を選ぶ



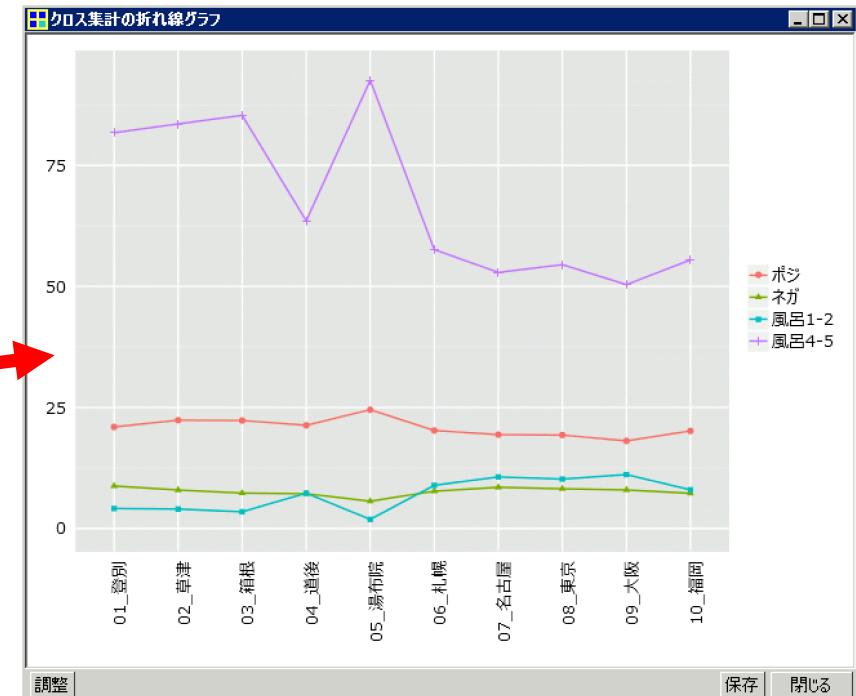
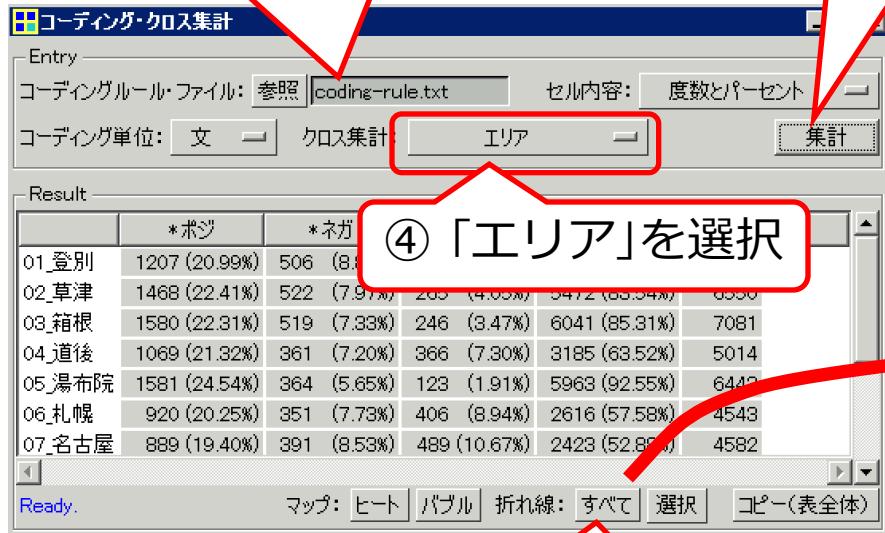
④ 「調整」をクリックして「バブルプロット」をチェック

練習－クロス集計1

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

②「参照」をクリックして
「coding-rule.txt」を開く

⑤「集計」を
クリック



練習 -コーディングルール2

- 数値評価と口コミの傾向比較
 - 前ページで紹介したクロス集計を用いて集計し,エリアごとのポジ・ネガ意見の傾向と,数値評価の総合点を比較し,違いについて考察する

ヒント: 「風呂1-2」「風呂4-5」を参考に「総合1-2」「総合4-5」を追加し,クロス集計する

練習－クロス集計2

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

②「参照」をクリックして
「coding-rule_new.txt」を開く

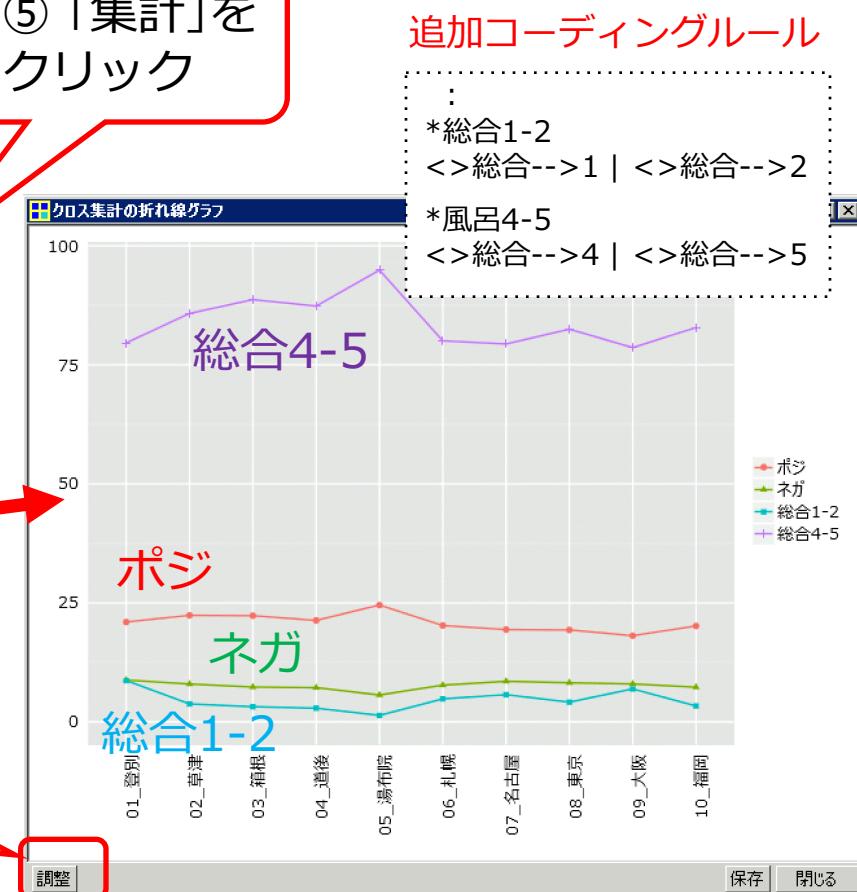
⑤「集計」を
クリック

④「エリア」を選択

⑥「すべて」をクリック

	*ポジ	*ネガ	
01_登別	1207 (20.99%)	506 (8.1%)	
02_草津	1468 (22.41%)	522 (7.97%)	248 (3.79%) 3185 (48.63%) 398 (6.06%)
03_箱根	1580 (22.31%)	519 (7.83%)	226 (3.19%) 3487 (49.24%) 388 (5.48%)
04_道後	1060 (17.90%)	361 (6.20%)	145 (2.29%) 2000 (34.46%) 144 (2.87%)
05_湯布院			19% 133 (2.03%) 47 (1.9%)
06_札幌			82% 238 (5.24%) 16 (0.7%)
07_名古屋	889 (19.40%)	391 (8.53%)	262 (4.38%) 1072 (36.45%) 280 (6.11%) 15 (0.5%)

⑦「ポジ」「ネガ」「総合1-2」「総合4-5」の4つを選択



課題 – テキスト分析(概要)

- ユーザーの声の可視化と生の声(原文)をもとに, 数値評価では見えないニーズや課題を発見する
 1. 各人でデータ分析を行う
 2. 周囲の4人前後でグループを作る
 3. グループ内で考察や支持する図,原文を共有し議論
 4. グループごと図と原文を使って考察内容を発表
- 課題は,下記の4つ
 1. ユーザーがどの項目に注目しているか (特徴語)
 2. ユーザーがどの項目に注目しているか (共起NW)
 3. ユーザーが何をどう評価しているか (共起NW)
 4. 分析結果にもとづいて改善案を提案してみる

※ 課題2-4は,作図はすべて,考察はレジヤーかビジネスのいずれか

課題1

- ・ユーザーがどの項目に注目しているかを確認する(1)
 - ・分類=レジャーと分類=ビジネスについて,テキスト中の**特徴語**を集計し,エリアによって**特徴語**がどう異なるかを比較, 2つの**注目する項目の違い**を考察する
 - ・分類=レジャーの5エリアについて,テキスト中の**特徴語**を集計し,エリアによって**特徴語**がどう異なるかを比較, 5エリアで**注目する項目の違い**を考察する
 - ・分類=ビジネスの5エリアについて,テキスト中の**特徴語**を集計し,エリアによって**特徴語**がどう異なるかを比較, 5エリアで**注目する項目の違い**を考察する

参考 – 出力例1

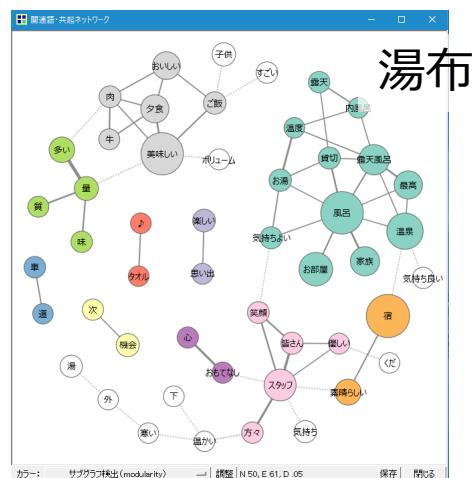
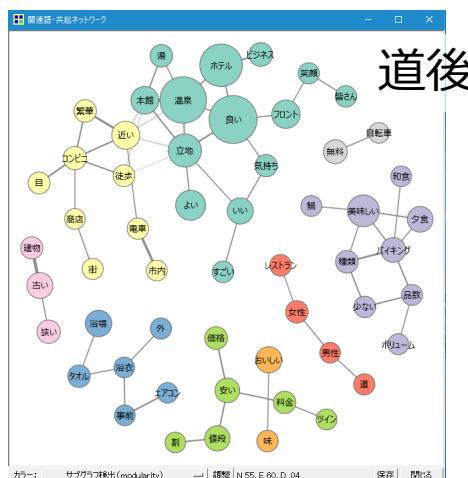
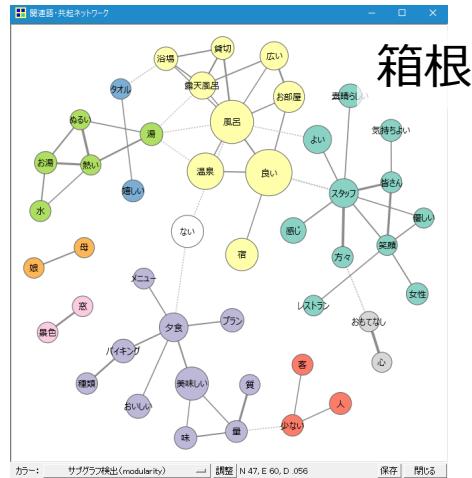
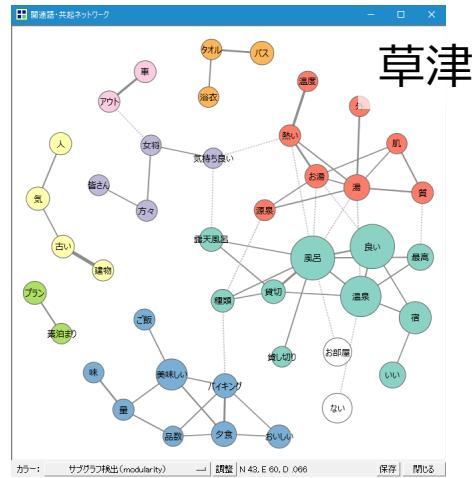
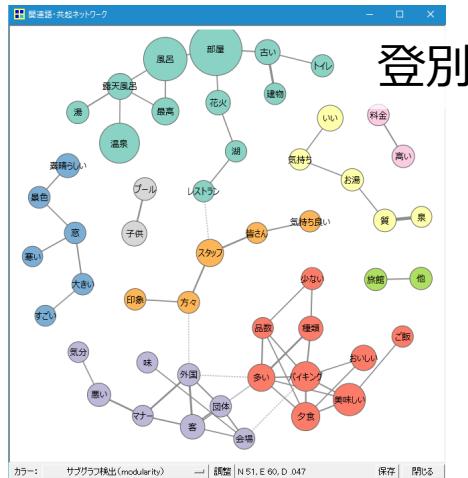
A_レジャー		01_登別		02_草津		03_箱根		04_道後		05_湯布院	
風呂	.298	風呂		温泉	.105	温泉	.166	温泉	.136	温泉	.122
良い	.273	部屋		温泉	.101	風呂	.149	温泉	.126	本館	.085
温泉	.269	食事		バイキング	.098	宿	.124	夕食	.121	立地	.080
部屋	.255	サービス		部屋	.092	良い	.111	美味しい	.116	ホテル	.077
美味しい	.216	設備・アメニティ		花火	.081	美味しい	.100	宿	.110	美味しい	.064
宿	.163	立地		夕食	.080	湯	.098	良い	.110	フロント	.058
ない	.155			美味しい	.080	最高	.087	お部屋	.101	近い	.057
お部屋	.145			ホテル	.074	貸切	.084	露天風呂	.100	浴場	.052
夕食	.129			多い	.072	家族	.083	部屋	.097	古い	.052
スタッフ	.127			ない	.070	ない	.080	スタッフ	.095	おいしい	.049
B_ビジネス		06_札幌		07_名古屋		08_東京		09_大阪		10_福岡	
ホテル	.224	立地		ホテル	.092	ホテル	.089	駅	.116	駅	.113
立地	.158	サービス		立地	.075	駅	.086	ホテル	.088	ホテル	.093
駅	.150	風呂		広い	.070	立地	.078	部屋	.088	部屋	.085
近い	.130	部屋		フロント	.064	近い	.073	近い	.085	立地	.078
フロント	.123	設備・アメニティ		近い	.059	フロント	.066	フロント	.076	近い	.078
コンビニ	.073	食事		駅	.057	狭い	.049	立地	.075	駅	.061
値段	.061	値段		コンビニ	.045	コンビニ	.048	コンビニ	.064	コンビニ	.050
狭い	.057			安い	.044	安い	.046	狭い	.052	天神	.049
バス	.056			浴場	.042	ベッド	.044	バス	.049	バス	.044
安い	.053			ベッド	.041	値段	.044	ベッド	.042	ビジネス	.040
								ない	.067	徒步	.038

操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<>分類-->A_レジャー”」 「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」→結果を選択し「コピー」

課題2

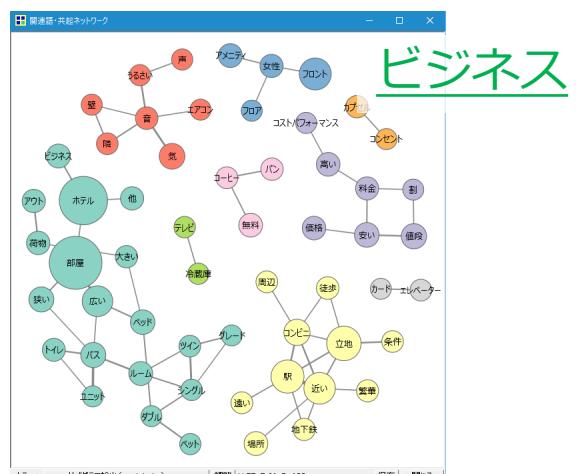
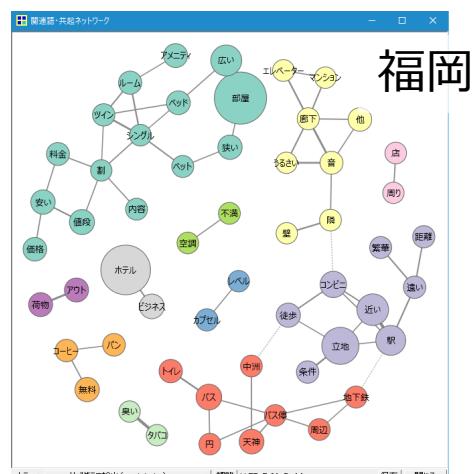
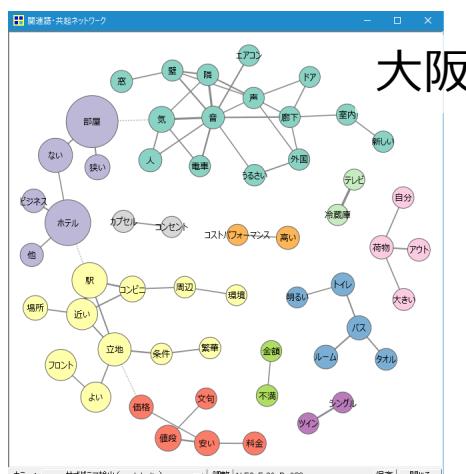
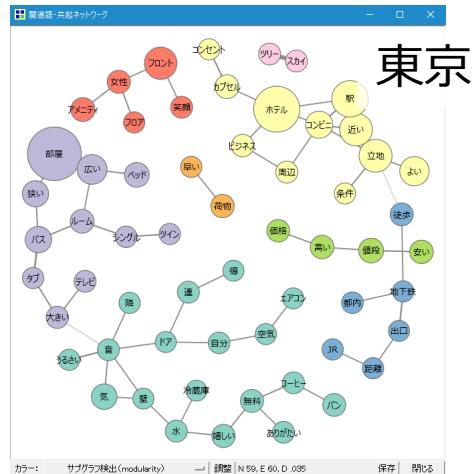
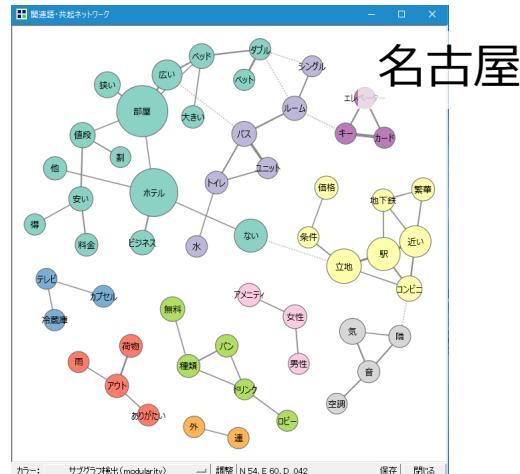
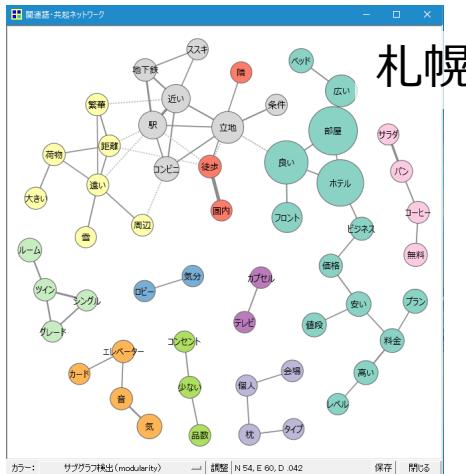
- ・ユーザーがどの項目に注目しているかを確認する(2)
 - ・分類=レジャーの5エリアについて,特徴語の共起ネットワーク図を作成し,エリアによって特徴語がどう異なるかを比較し,5エリアで注目する項目の違いを考察する
 - ・分類=ビジネスの5エリアについて,特徴語の共起ネットワーク図を作成し,エリアによって特徴語がどう異なるかを比較し,5エリアで注目する項目の違いを考察する
 - ・分類=レジャーと分類=ビジネスについて,テキスト中の特徴語を集計し,エリアによって特徴語がどう異なるかを比較し,5エリアで注目する項目の違いを考察する

参考 - 出力例2 (特徴語の共起NW:レジャー)



操作例: 「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,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:太い線,大きい円」

参考 - 出力例3 (特徴語の共起NW:ビジネス)

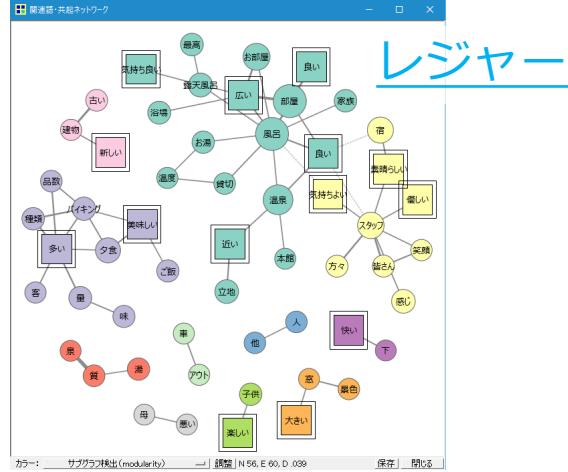
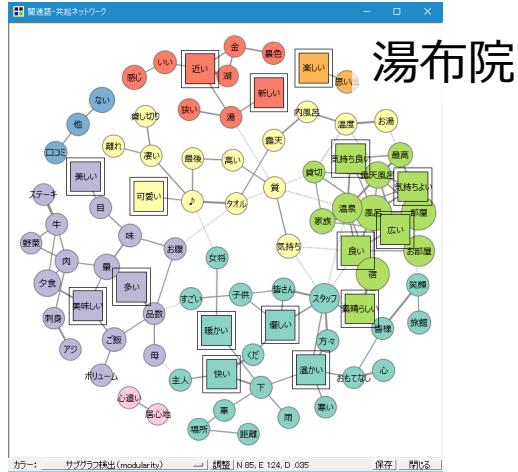
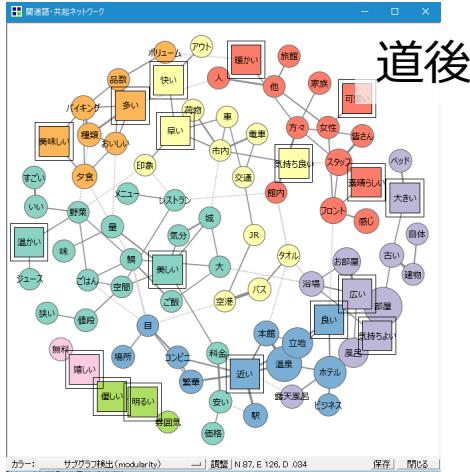
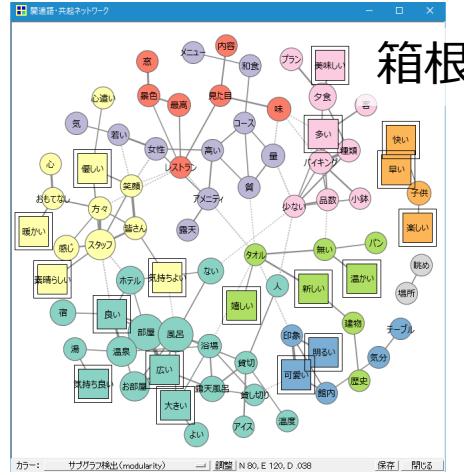
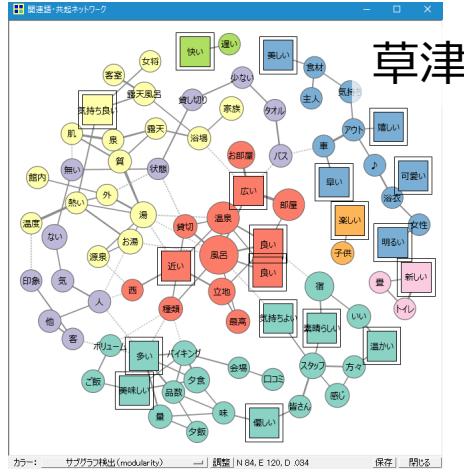
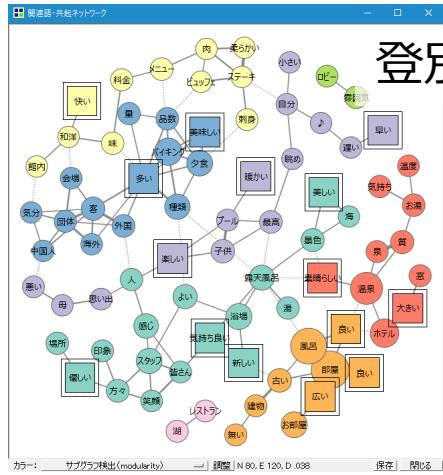


操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<エリア-->01_登別”」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:太い線,大きい円」

課題3

- ・ユーザーが何をどう評価しているかを確認する
 - ・分類=レジャーの5エリアについて,特徴語とポジティブ意見の共起NW図を作成して比較し,5エリアの何がどう評価されている点を考察する
 - ・分類=レジャーの5エリアについて,特徴語とネガティブ意見の共起NW図を作成して比較し,5エリアの課題を考察する
 - ・分類=ビジネスの5エリアについて,特徴語とポジティブ意見の共起NW図を作成して比較し,5エリアの何がどう評価されている点を考察する
 - ・分類=ビジネスの5エリアについて,特徴語とネガティブ意見の共起NW図を作成して比較し,5エリアの課題を考察する

参考 - 出力例4 (ポジの共起NW:レジヤー)



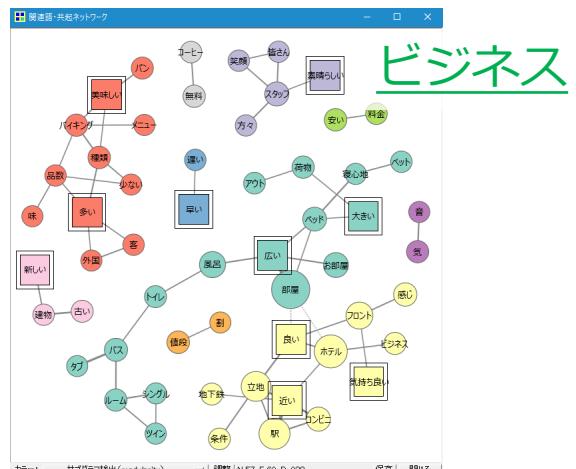
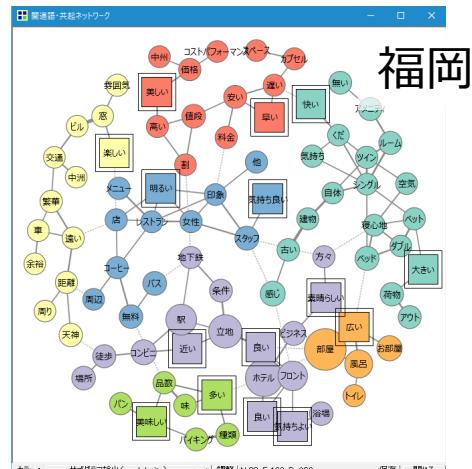
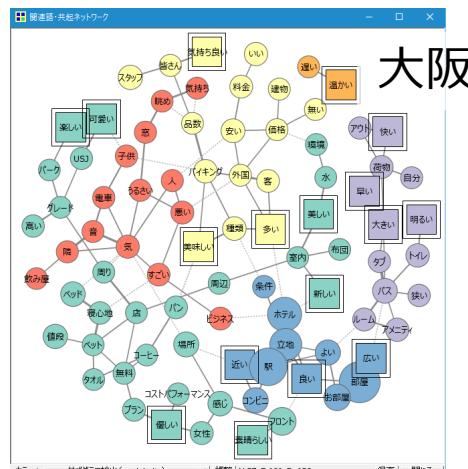
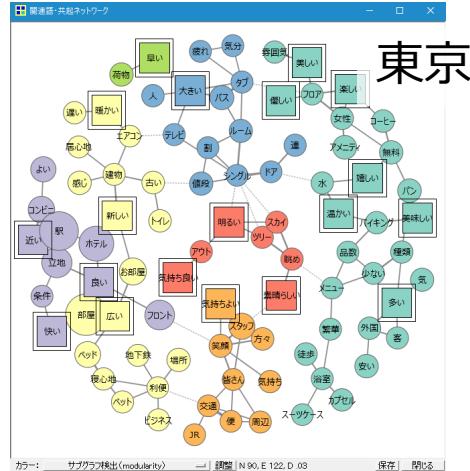
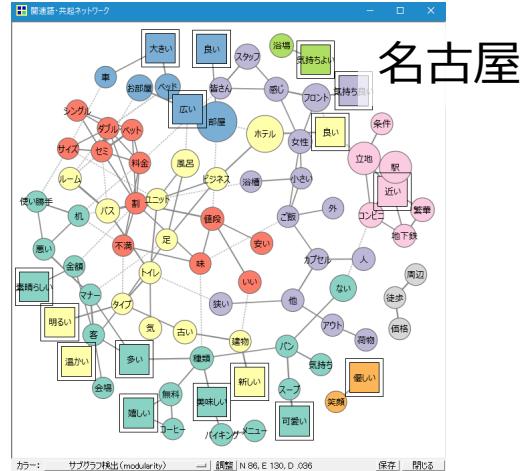
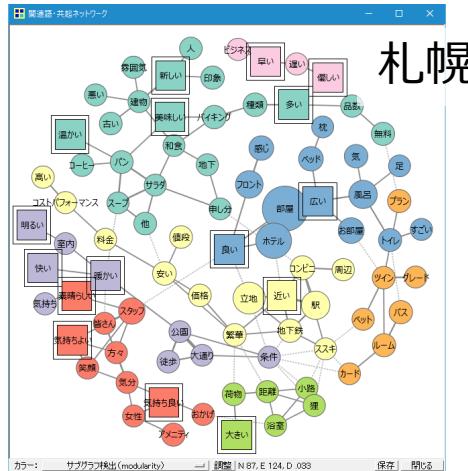
操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)"<エリア-->01_登別"」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:太い線,大きい円,描画数=120」

参考 - 出力例5 (ネガの共起NW:レジヤー)



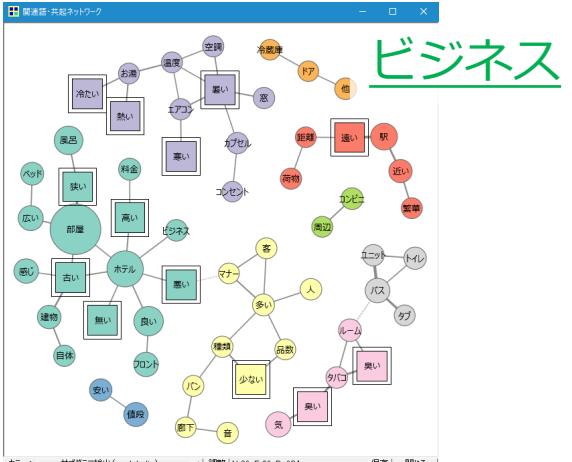
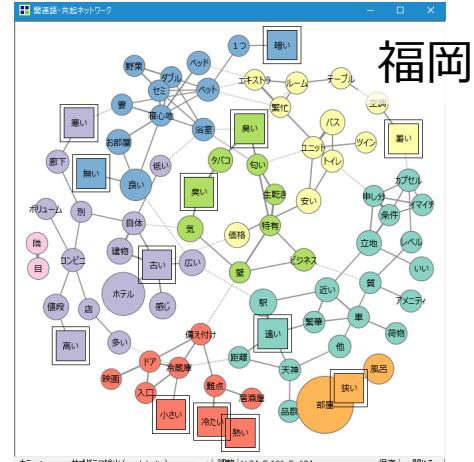
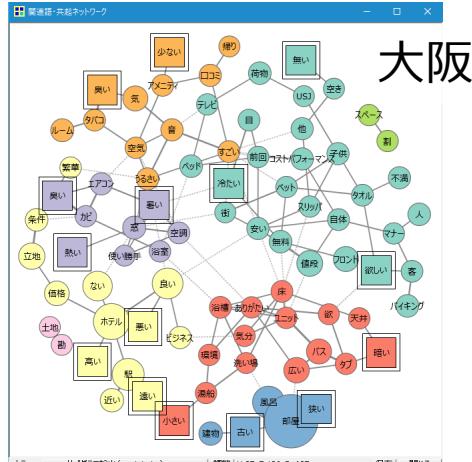
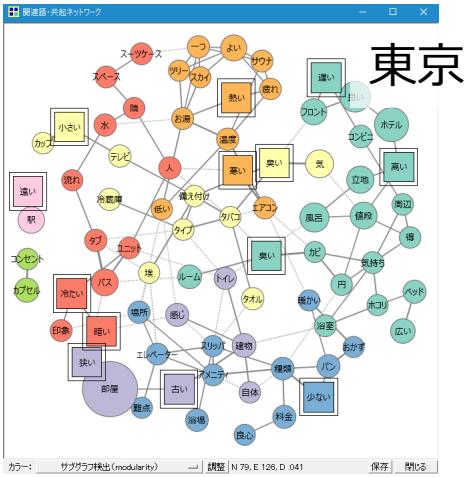
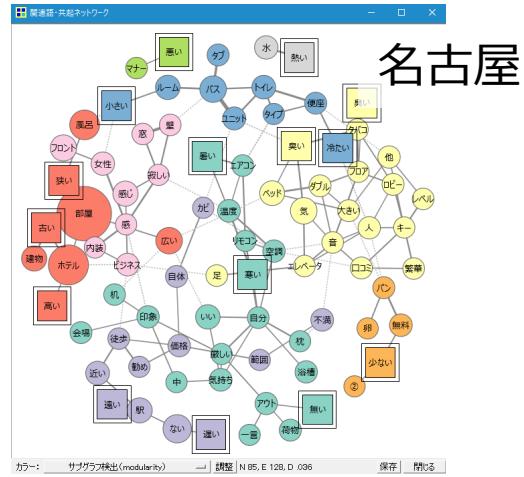
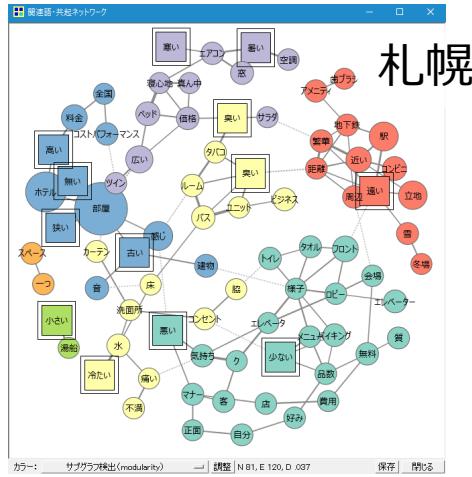
操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<エリア-->01_登別”」「Search Entry:*ネガ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:太い線,大きい円,描画数=120」

参考 - 出力例6 (ポジの共起NW:ビジネス)



操作例: 「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:太い線,大きい円,描画数=120」

参考 - 出力例7 (ネガの共起NW:ビジネス)



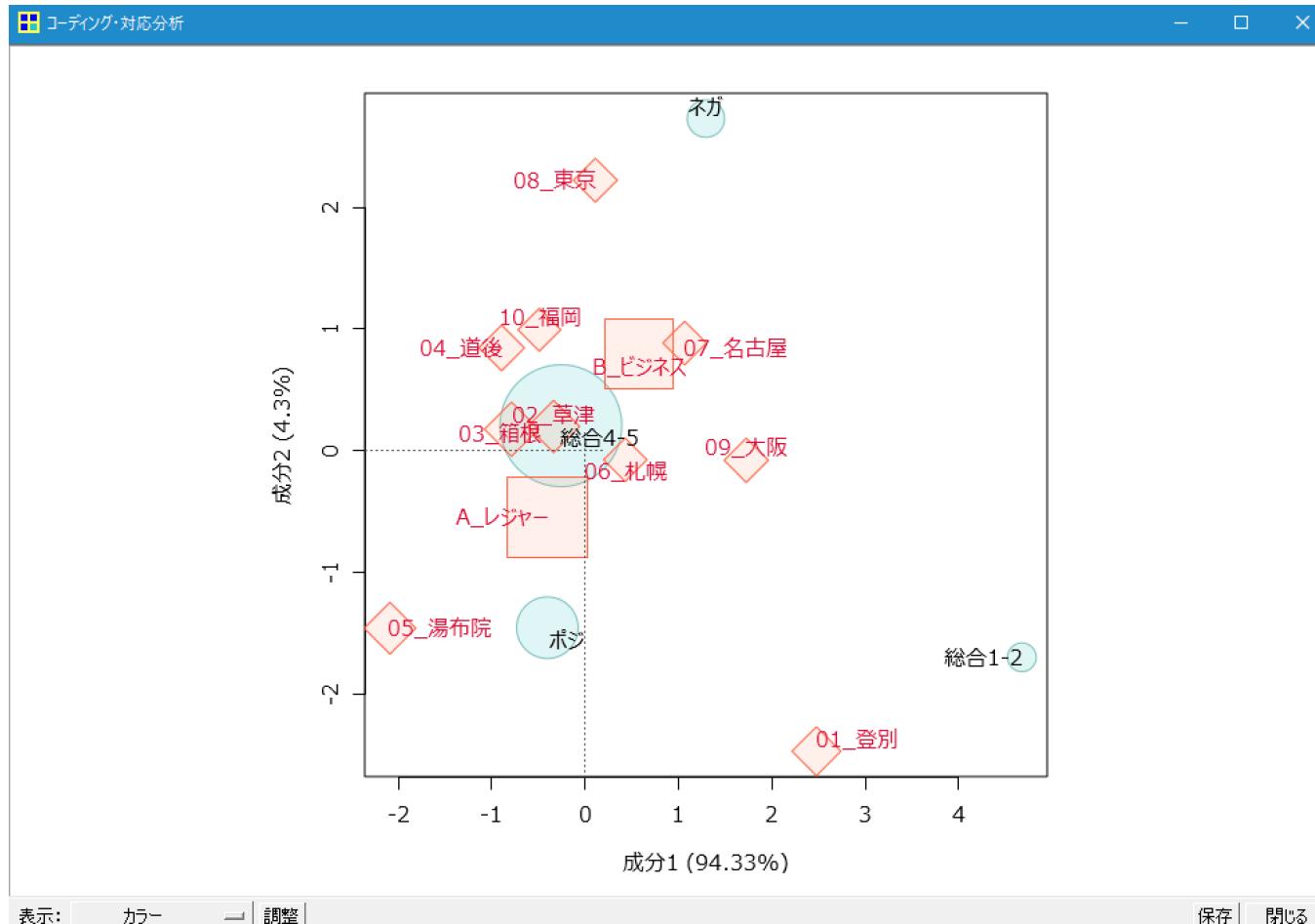
操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)"<エリア-->01_登別"」「Search Entry:*ネガ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:太い線,大きい円,描画数=120」

課題4

- あなたは観光地域振興担当者です。地域の改善案を提案してください
 - レジャーエリアのうち, **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**, **担当エリアと比較先エリア**とする
 - ビジネスエリアのうち, **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**, **担当エリアと比較先エリア**とする
 - **担当エリア**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

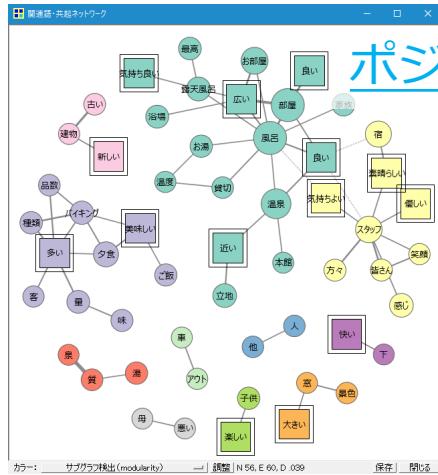
注意: 主張を支持する**図**とユーザーの**生の声**(原文)の両方を使って説明してください

参考 – 出力例8 (総合評価 × ポジとネガ)

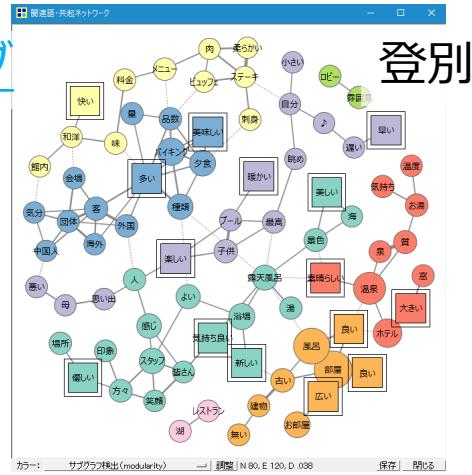


操作例: 「ツール」→「コーディング」→「対応分析」→「コーディング単位:文」「コード選択: *ポジ,*ネガ,*総合1-2,*総合4-5」「コードx外部変数: 分類,エリア」

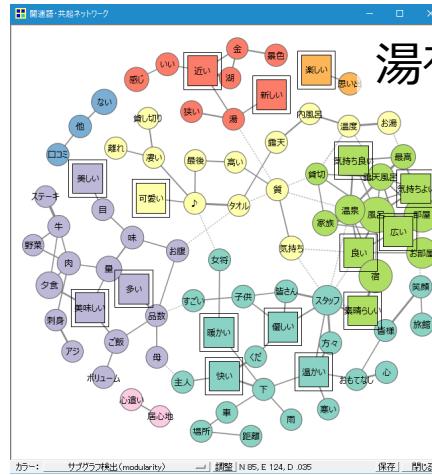
参考 - 出力例9 (ポジとネガ: 登別-湯布院の比較)



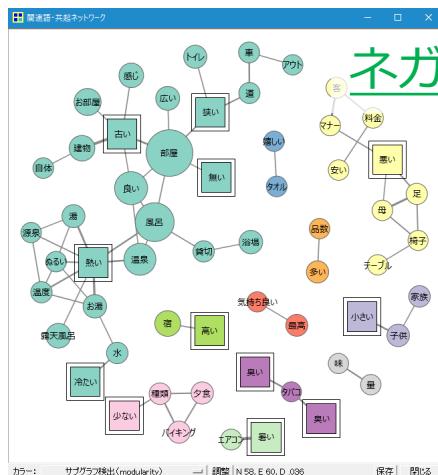
ポジティブ



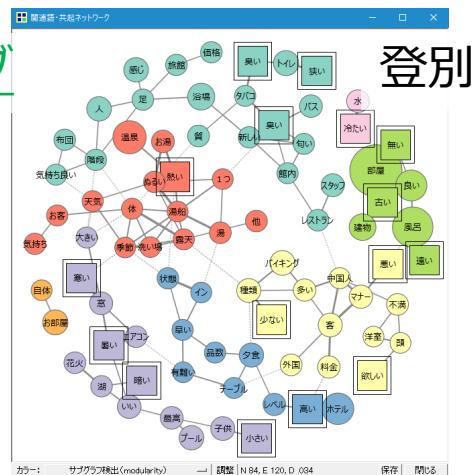
登別



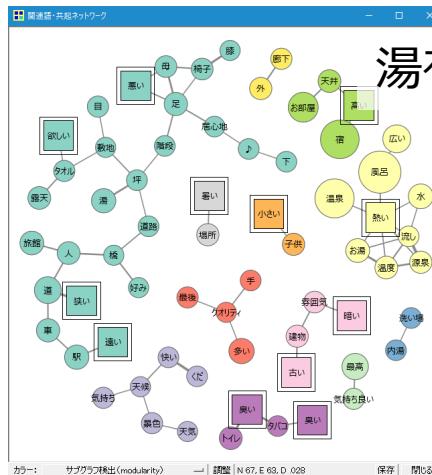
湯布院



ネガティブ



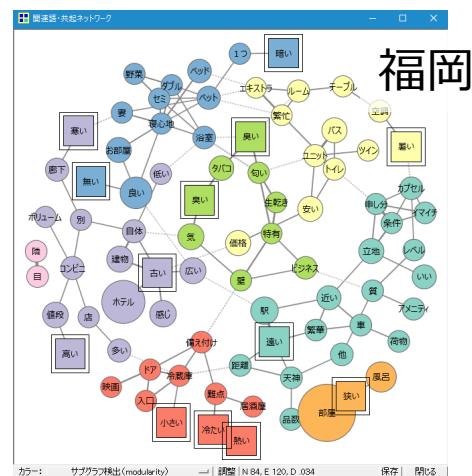
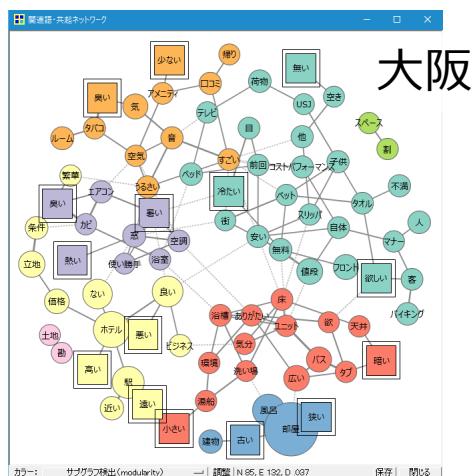
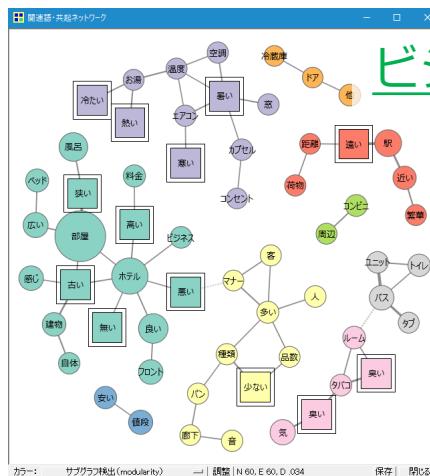
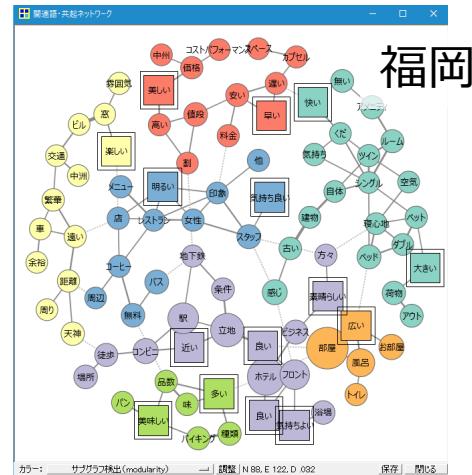
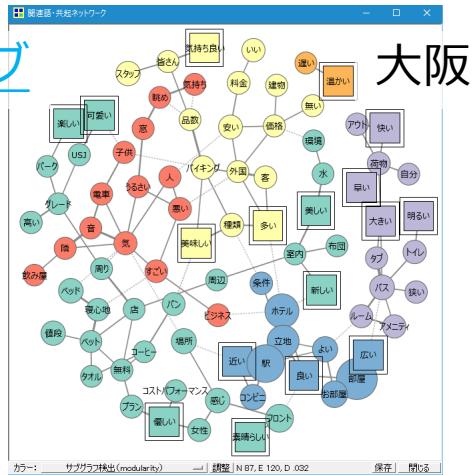
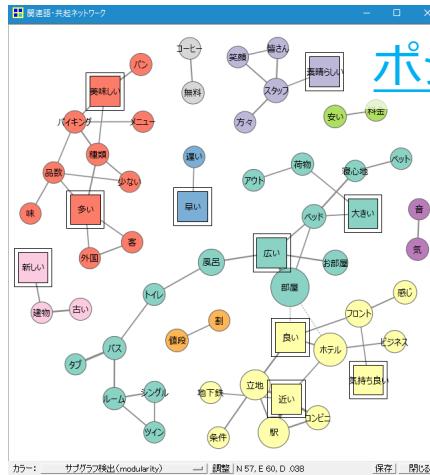
登別



湯布院

操作例: 「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:太い線,大きい円,描画数=120」

参考 - 出力例10 (ポジとネガ: 大阪-福岡の比較)



操作例: 「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)"<エリア-->06_大阪"」「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:太い線,大きい円,描画数=120」

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, 京都, 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.

(Windows環境によるCGM収集の参考に)

- [3] 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)

- [4] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [5] 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)

- [6] 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.
- [7] 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)

- [8] 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.