

人文社会ビジネス科学学術院 ビジネス科学研究群 2025年度 春C

テキストマイニングの実践

day 3

スケジュール

day 1

- 講義(後半) – 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
- 講義&演習 – データ理解
- 講義&演習 – テキスト解析 (1)

day 3

- 講義&演習 – テキスト解析 (2)
- 講義&演習 – テキスト分析 (1)

day 4

- 講義&演習 – テキスト分析 (2)

day 5

- テキストマイニングツール紹介 – TMS
- ラップアップ – Q&A

(前回) day 2 – レポート課題

- 以下を PDF ファイルで提出 してください
 - データ集計により作成した「集計表」のキャプチャ (P.XX~XX) ※ページ番号は各スライド右下に記載
 - 作成した「集計結果の整理」の表 (P.XX) ※ページ番号は各スライド右下に記載
- ※ 何らかの事情で上記2つ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

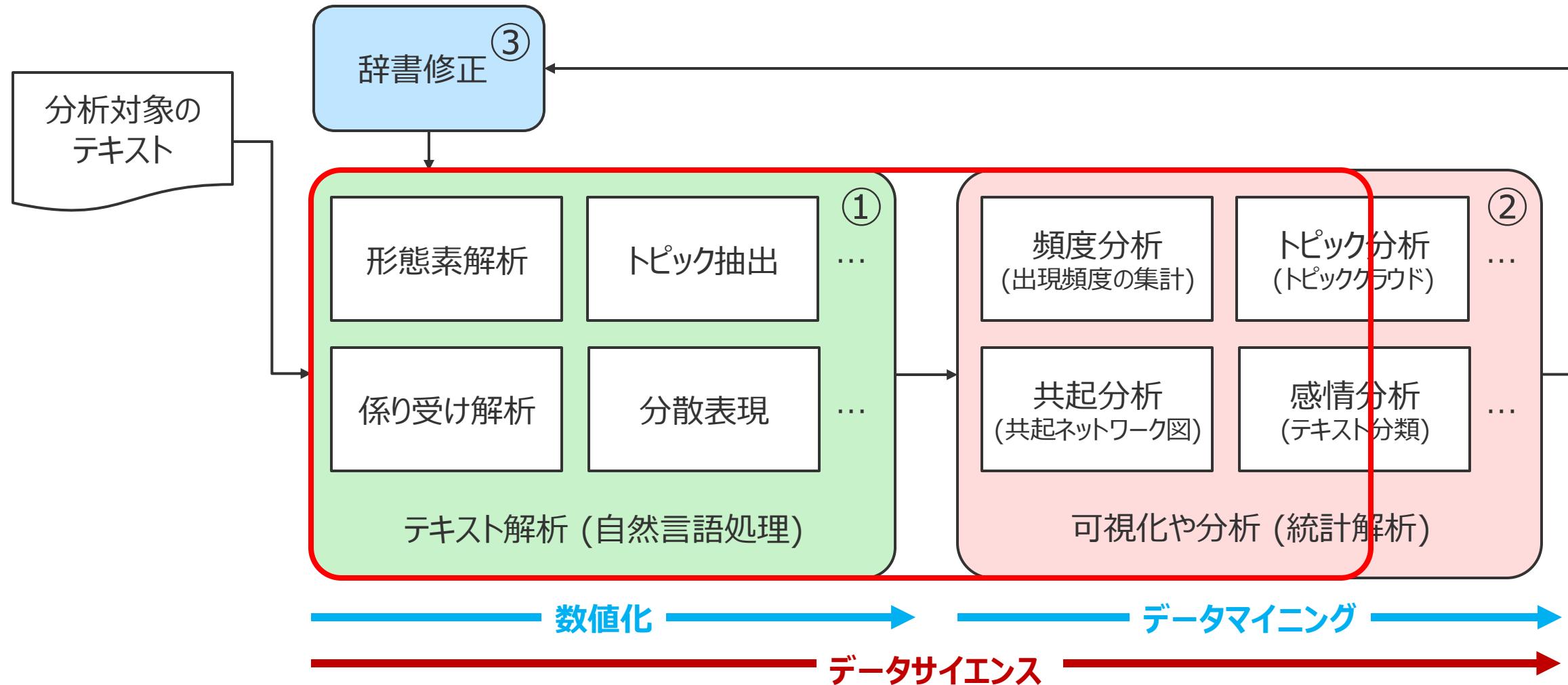
テキスト解析 (2)

(再掲) テキストマイニングの手順

- 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は?
 - 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

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)



- 社会調査データを分析する目的で開発されたフリー(~~商用可能~~)のツール

- 高機能かつ~~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
 - ・ トピックモデル (LDA)

論文検索サービスも提供 → <http://khcoder.net/bib.html>

研究事例リスト

KH Coderを用いたご研究の成果を発表された際には、書誌情報をフォームにご記入いただけますと幸いです。

出版年 :

著者名 :

キーワード :

ヒット件数 : 0200 / 6135

KH Coderを用いた研究事例のリスト 6135件

※2023/6/16 現在

→1646→2042→2695→3741件→4554件→昨年5355件→6135件)

(参考) 2023年12以降のライセンスや料金

2023年12月 無償公開が終了しました

	Starting	Base	Master	Pro
通常ライセンス 販売価格(税込) ※下段は、バージョンアップに追従できるアップデートのサブスクリプション費用(1年間)	無料	59,950 (43,780)	196,900 (43,780)	396,000 (43,780)
アカデミック・ライセンス 販売価格(税込) ※下段は、バージョンアップに追従できるアップデートのサブスクリプション費用(1年間)	無料	24,750 (18,480)	69,850 (18,480)	—
インストール可能台数 ※ライセンス保有者が管理するPCに限る	1台	2台	2台	2台

	Starting	Base	Master	Pro
無料で分析を始められる 一部機能制限ありの公開テスト(beta)版 ・分析対象はデータファイルの最初の100件目まで ・強制抽出語と無視する語の指定はそれぞれ1語のみ ・分析対象の品詞タイプを選択できない	○	—	—	—
機能制限なしの正式版 外部プラグイン「文錦@シリーズ」購入による機能追加も可能	—	○	○	○
アップデートのサブスクリプション(1年間有効) KH Coderのバージョンアップに追従できる便利なパッチを提供	—	○	○	○
インストール／セットアップのサポート KH Coderの導入段階のトラブルシューティング	—	—	○	○
セミナー受講券(1年半以内有効) 「計量テキスト分析公式セミナー」初級編・ステップアップ編に各1回参加可	—	—	1名	1名
使い方のQ&A対応 使い方に関する困り事やご質問にメールでサポート	—	—	—	6回/年

出典: <https://www.screen.co.jp/as/solution/khcoder>

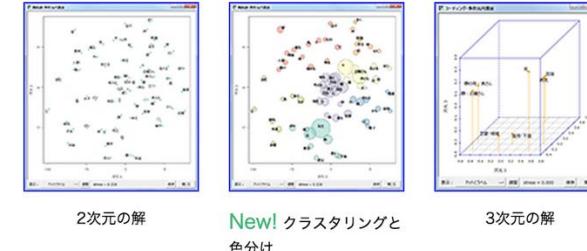
共起ネットワーク

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、すなはち共起関係を線（edge）で表したネットワークを描く機能です。



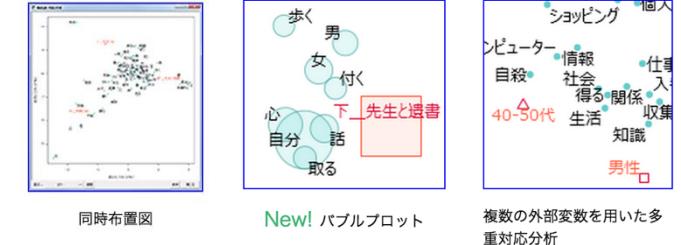
多次元尺度構成法 (MDS)

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。



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。



分析手法

説明

共起ネットワーク

- 同時に出現した単語同士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 係数で評価し、媒介性やグラフクラスタリングを使ってサブグラフも検出できる

多次元尺度構成法 (MDS)

- 出現パターンの似た単語同士を近くに置くよう図示したもの
- 出現パターンとは、ある単語がどの文書に出現したかといった関係を単語ベクトルとして表現したもの
- 似ている（=距離が近い）の計算は Jaccard、ユークリッド、コサイン距離のいずれかで求める

対応分析 (コレ спинデンス分析)

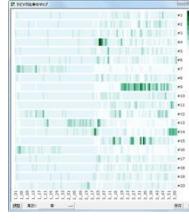
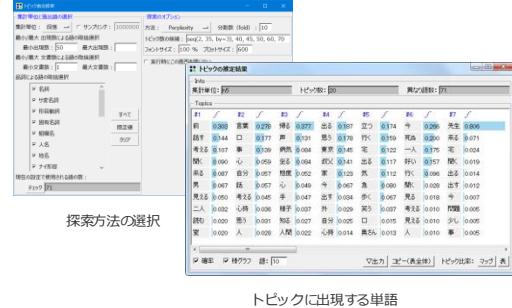
- 出現パターンの似た単語や外部変数を近くに置くよう図示したもの
- 単語と単語または外部変数が同時に出現した頻度をクロス集計し、相関が最大になるような2軸でプロット
- PCA が元の情報をそのまま可視化するのに対して、対応分析は似ているものを近くに表示する
- 外部変数も同時にプロットできる

(参考) KH Coder — 分析手法 (2)

2023年12月 無償公開が終了しました

トピックモデル (LDA)

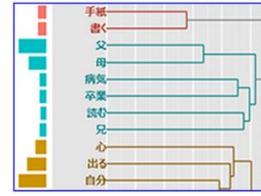
文書ごとにトピックの出現割合を表示したり、各トピックに高い確率で出現する語を表示できます。



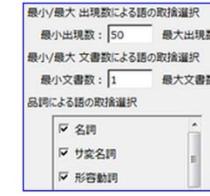
文書ごとのトピック比率

階層的クラスター分析

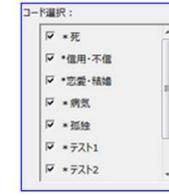
抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけでなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



抽出語は出現数や品詞で選択



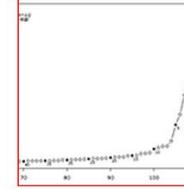
コードはチェックボックスで直接選択

文書のクラスター分析

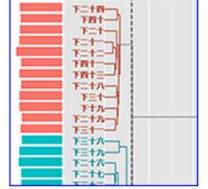
文書の分類を行うクラスター分析です。



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。

分析手法

トピックモデル (LDA)

- 文書が複数のトピックを持つと仮定、文書ごとにトピックの出現割合、各トピックに高確率で出現する語を表示
- R の topicmodels パッケージに含まれる LDA 関数(ギブスサンプリング)を利用 (乱数のシードは固定)
- トピックモデルは教師なし学習**のため、コーディングルールで単語を集約するよりも客観性が高い

階層的クラスター分析

- 出現パターンの似た**単語同士をグルーピング(クラスタリング)**して、樹形図にしたもの
- 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関係を単語ベクトルとして表現したもの
- 似ている(=距離が近い)の計算は Jaccard、ユークリッド、コサイン距離のいずれかで求める

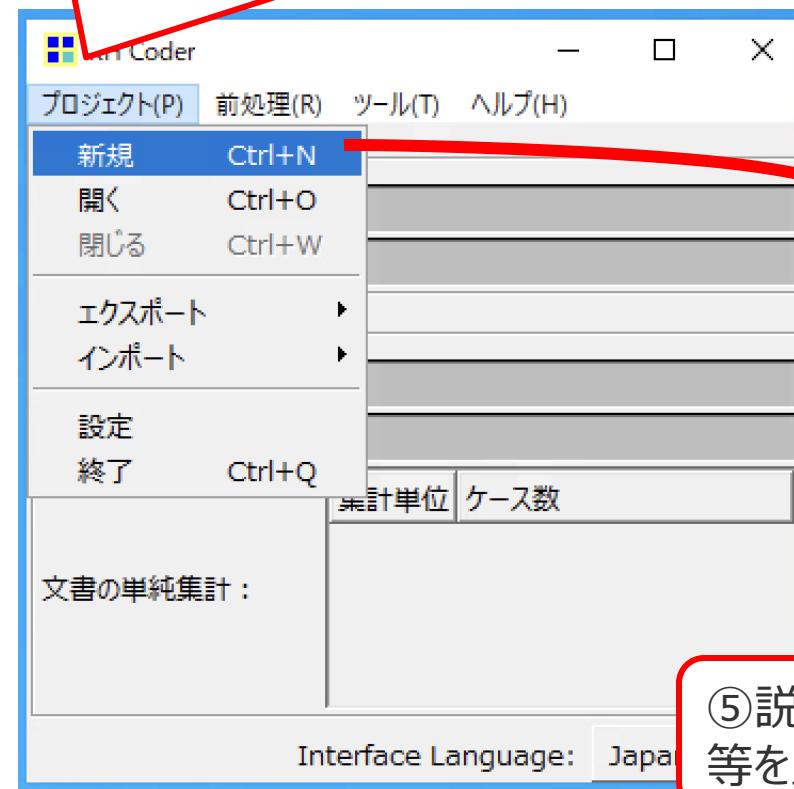
文書のクラスター分析

- 似た**文書同士をグルーピング(クラスタリング)**して、樹形図にしたもの
- 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
- 似ている(=距離が近い)の計算は Jaccard、ユークリッド、コサイン距離のいずれかで求める
- いわゆる Ward法、群平均法、最遠隣法で階層クラスタを作成する

説明

● プロジェクトの作成

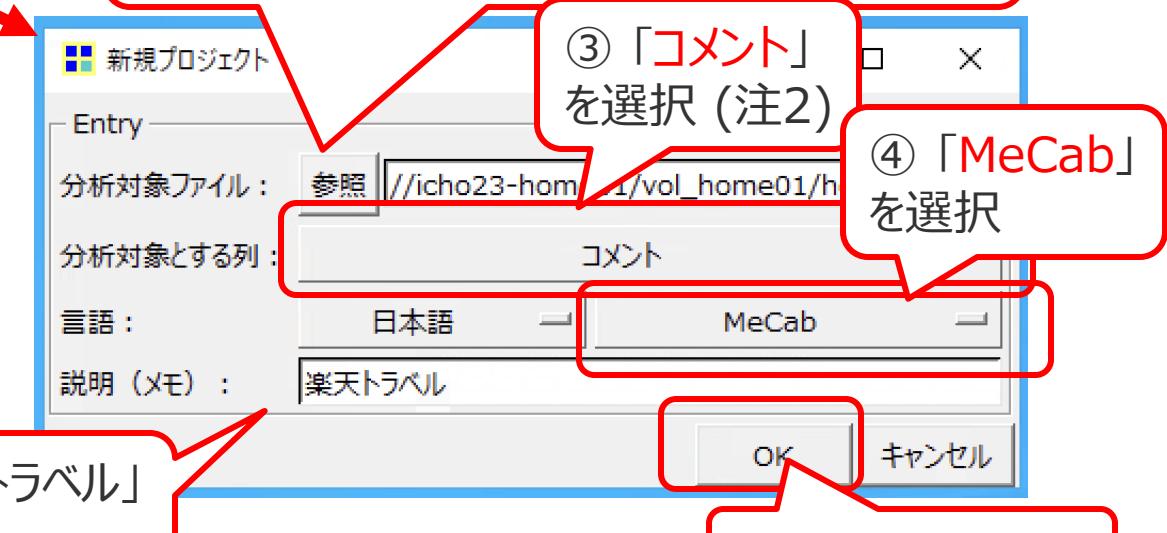
①メニューから「プロジェクト」「新規」を選択（注1）



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

②「参照」をクリックして
「rakuten-1000-2022-2023.xlsx」を開く

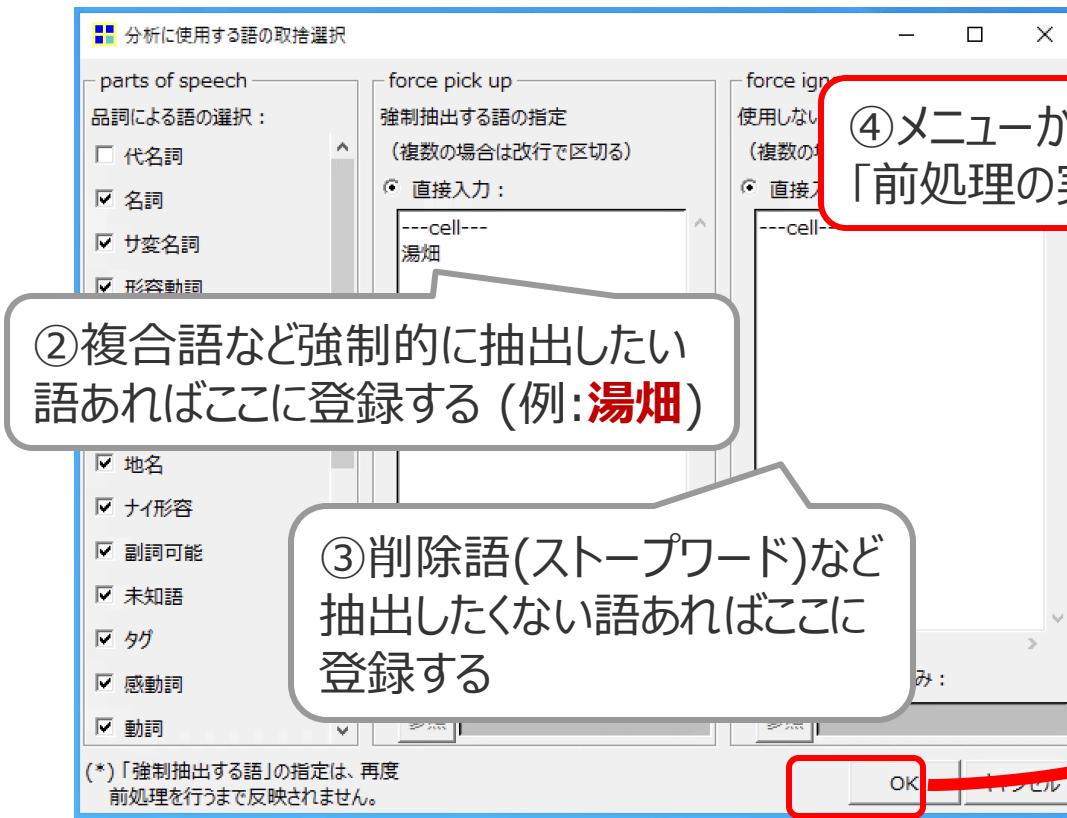


⑤説明「楽天トラベル」
等を入力

⑥「OK」をクリック

● 前処理(形態素解析)の実行

①メニューから「前処理」「語の取捨選択」を選ぶ



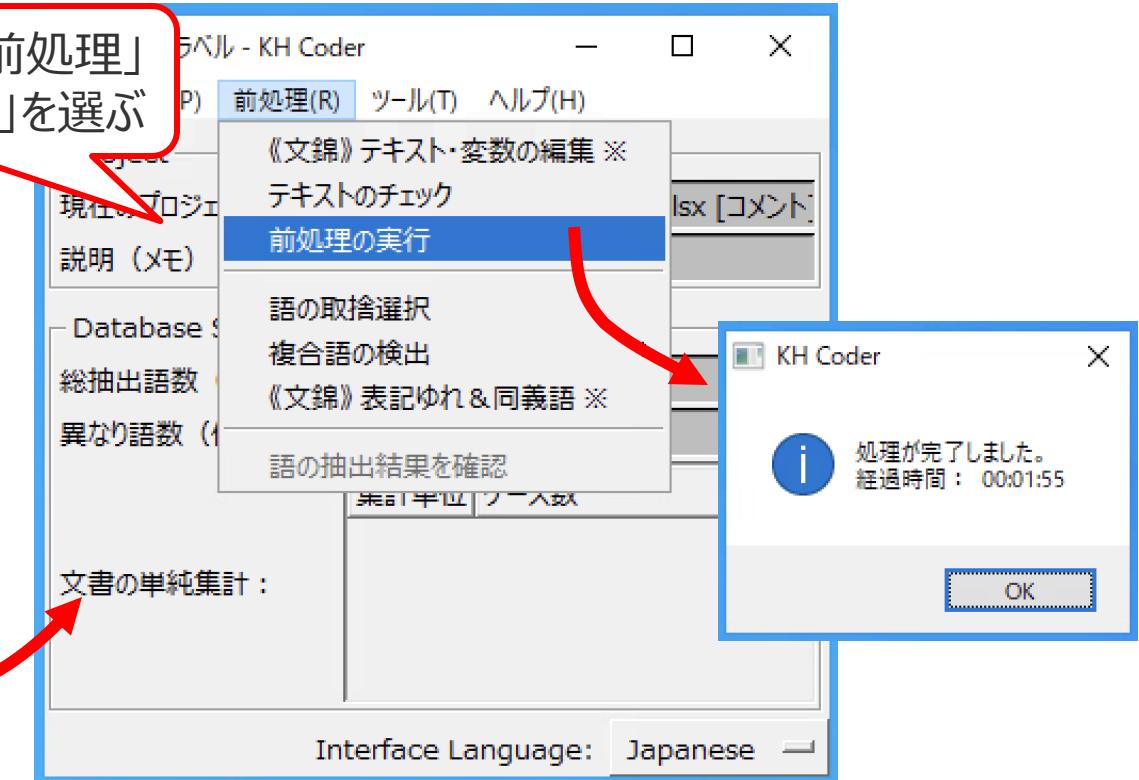
②複合語など強制的に抽出したい語あればここに登録する (例:湯畠)

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

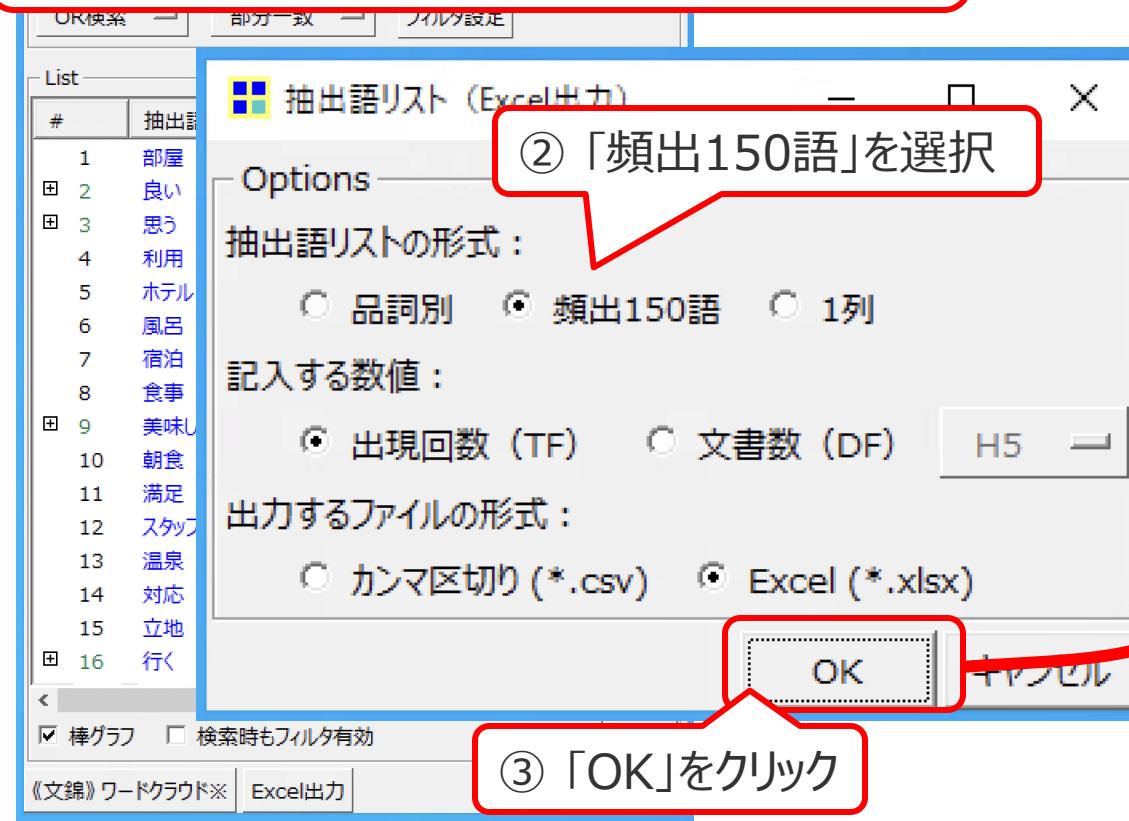
注1: EXCELファイルを読み込んで分析する場合,あらかじめ「---cell---」が入力されています

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



● 頻出語を確認する

- ①メニューから「ツール」「抽出語」「抽出語リスト」
→右下「EXCEL出力」ボタンを選択

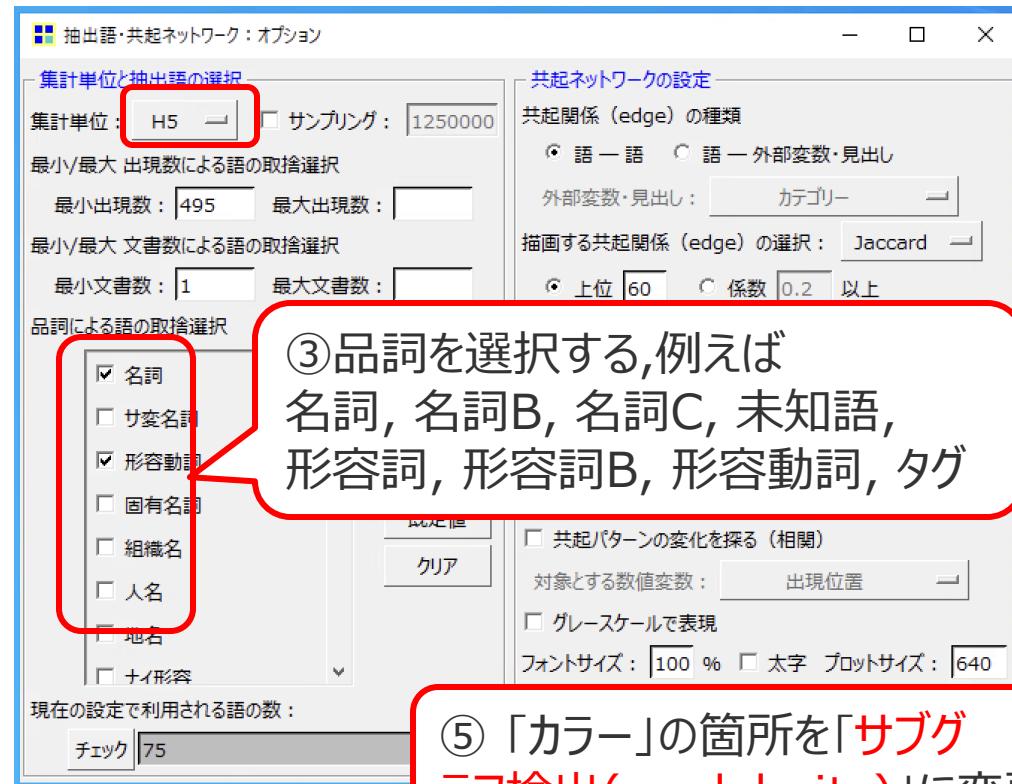


A	B	C	D	E	F	G	H
抽出語	出現回数	抽出語	出現回数	抽出語	出現回数	抽出語	出現回数
1 部屋	6689	2 子供	661	3 プラン	389		
2 良い	4302	4 過ごす	657	5 見える	388		
3 思う	3976	6 家族	648	7 機会	387		
4 利用	3481	8 予約	636	9 設備	387		
5 ホテル	2831	10 過ごせる	626	11 旅館	386		
6 風呂	2702	12 駐車	613	13 置く	384		
7 宿泊	2649	14 素晴らしい	612	15 きれい	377		
8 食事	2447	16 月	611	17 歩く	368		
9 美味しい	2249	18 バス	610	19 湯	359		
10 朝食	2172	20 丁寧	610	21 施設	345		
11 満足	1785	22 アメニティ	609	23 無料	345		
12 スタッフ	1712	24 清潔	556	25 新しい	340		
13 温泉	1705	26 入れる	544	27 楽しい	335		
14 対応	1603	28 使う	536	29 掃除	335		
15 立地	1374	30 初めて	523	31 気持ち	328		
16 行く	1334	32 行く	521	33 雰囲気	328		
17 広い	1314	34 無い	521	35 女性	323		
18 綺麗	1193	36 人	520	37 シャワー	321		
19 宿	1171	38 パイキング	515	39 建物	316		
20 大変	1157	40 嬉しい	515	41 高い	316		
21 少し	1156	42 ベッド	514	43 問題	316		
22 残念	1155	44 他	504	45 全体	314		
23 最高	1118	46 親切	503	47 大きい	313		
24		48 種類	502				

● 共起ネットワークの作成(1)

①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「H5」を選んで「OK」をクリック



⑤「カラー」の箇所を「サブグラフ検出(modularity)」に変更

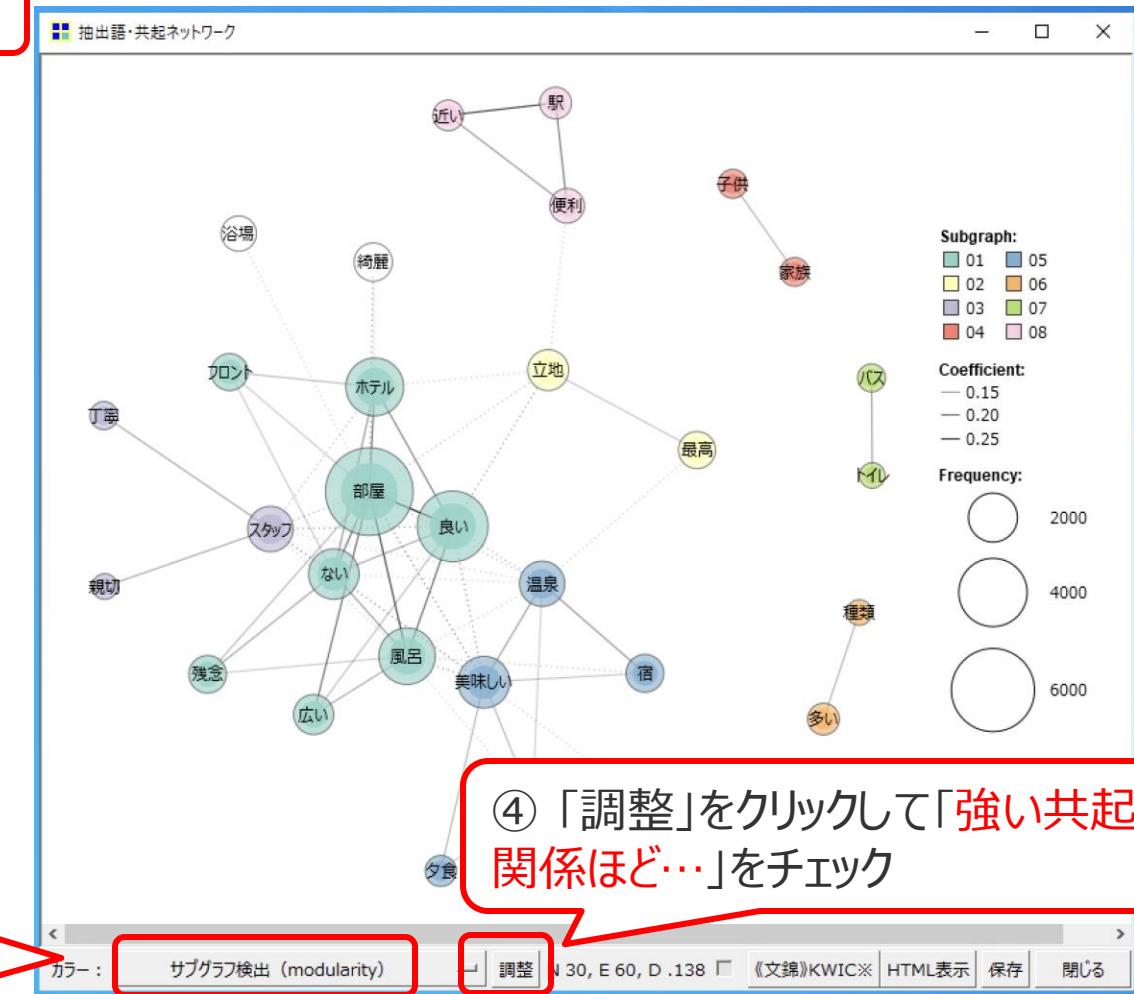


表 A.1 KH Coder の品詞体系

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 (平仮名のみの語)
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

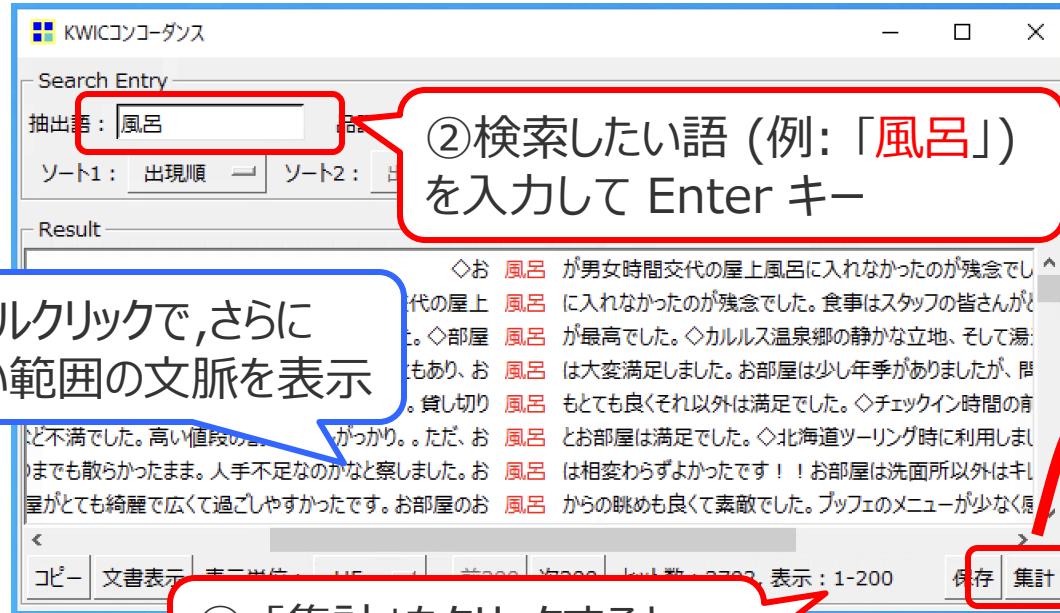
出典: KH Coder 3 リファレンス・マニュアル

注: どの品詞を選択すべきかは、分析対象のデータや分析目的により異なります。

分析結果を確認しながら、適宜、適切な品詞選択を検討することが重要です。

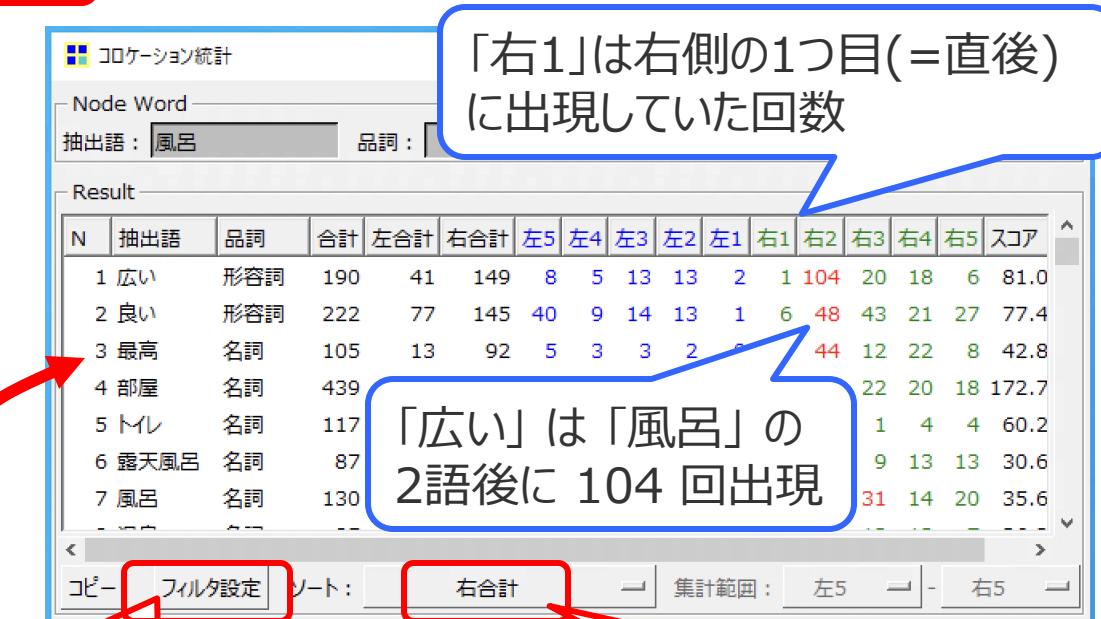
● 前後文脈を確認する

- ①メニューから「ツール」「抽出語」「KWICコンコーダンス」を選ぶ



- ③「集計」をクリックするとコロケーション統計(右)を開く

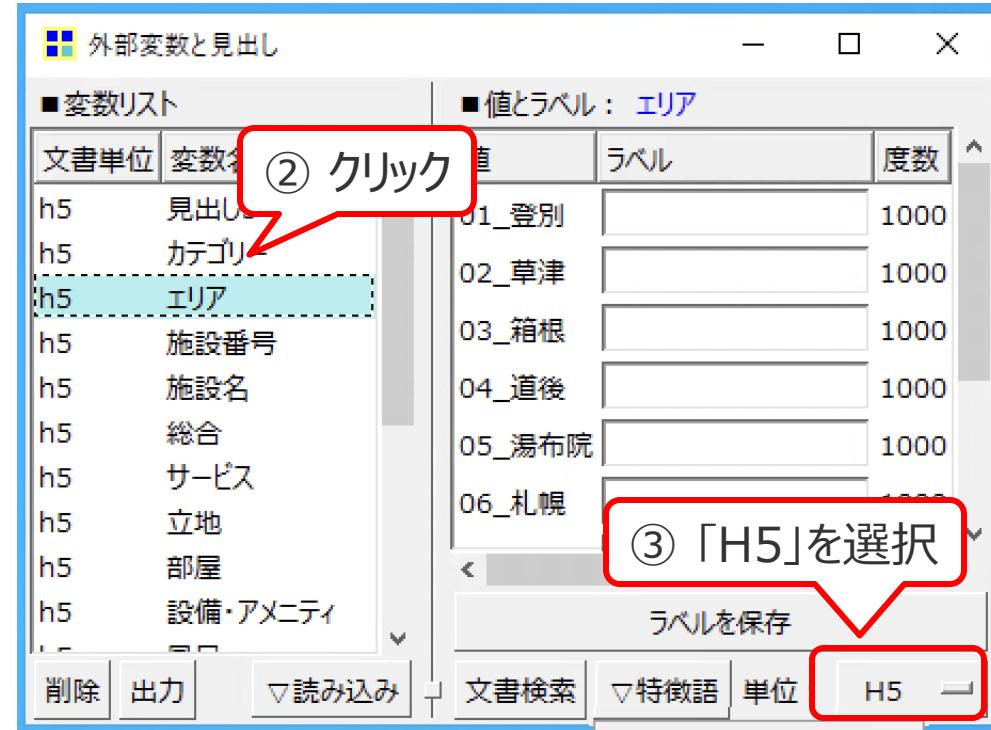
注: 共起ネットワーク上で「風呂」をクリックすると①②と同じ操作となります(V3以降)



- ④表示する語の品詞を選択
(例: 形容詞, 形容詞B, 形容動詞)

● 外部変数を利用する

① メニューから「ツール」「外部変数と見出し」を開く



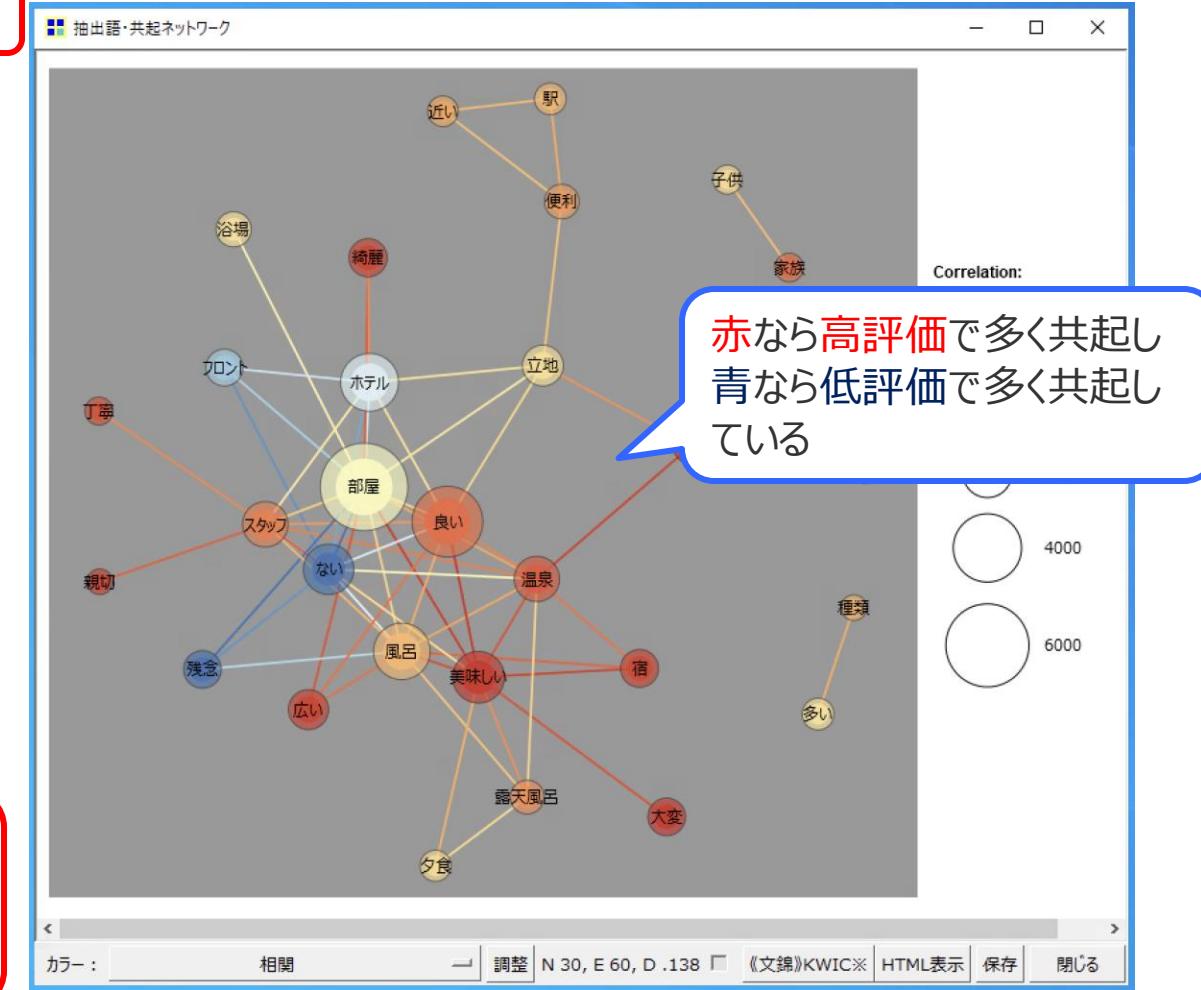
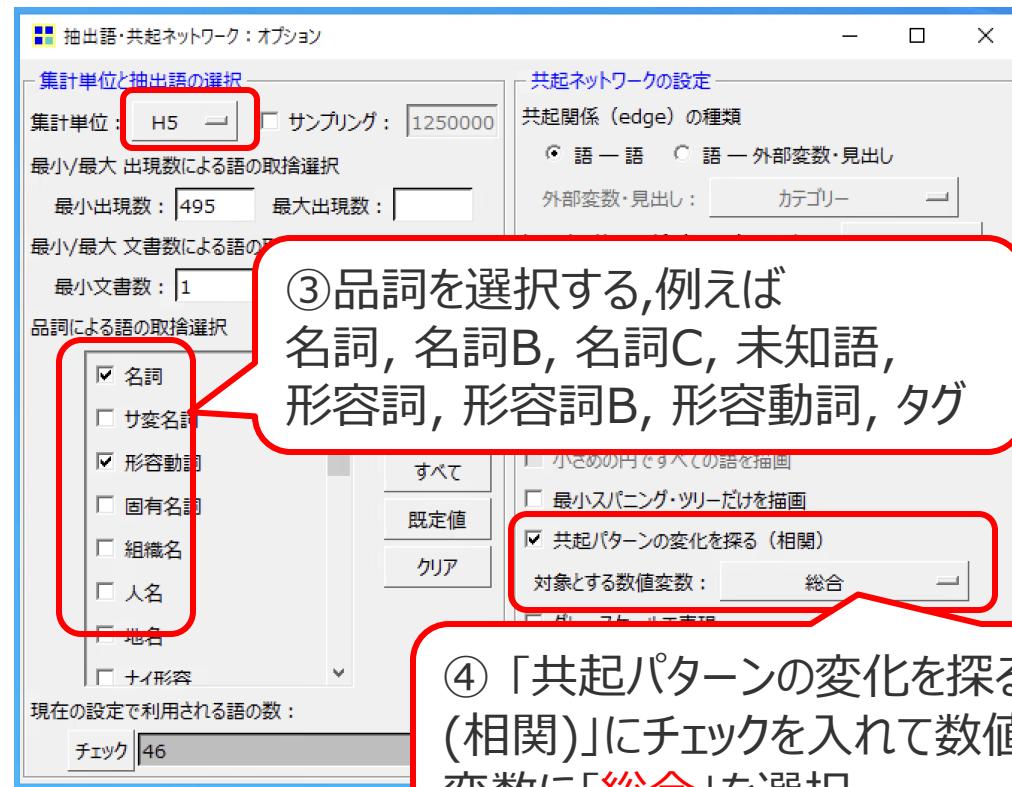
④ 「特徴語」「一覧(Excel形式)」を選択

A	B	C	D	E	F	G	H	I	J	K
1										
2	01_登別		02_草津		03_箱根		04_道後			
3	風呂	.115	湯畑	.327	美味しい	.136	温泉	.109		
4	温泉	.107	温泉	.136	露天風呂	.134	立地	.082		
5	美味しい	.094	風呂	.126	風呂	.116	最高	.066		
6	良い	.093	宿	.120	部屋	.109	広い	.063		
7	パーキング	.090	美味しい	.102	良い	.106	浴場	.059		
8	残念	.078	良い	.100	温泉	.102	よい	.058		
9	ない	.077	部屋	.096	宿	.097	フロント	.057		
10	夕食	.076	最高	.090	スタッフ	.096	大変	.057		
11	種類	.075	夕食	.085	夕食	.095	夕食	.055		
12	露天風呂	.074	ない	.074	ない	.083	便利	.055		
13	05_湯布院		06_札幌		07_名古屋		08_東京			
14	宿	.180	ホテル	.092	ホテル	.086	駅	.102		
15	美味しい	.144	立地	.077	便利	.072	ホテル	.086		
16	露天風呂	.135	便利	.077	駅	.070	便利	.078		
17	風呂	.127	綺麗	.071	綺麗	.069	立地	.077		
18	温泉	.124	浴場	.070	フロント	.066	近い	.071		
19	最高	.114	フロント	.065	立地	.065	綺麗	.064		
20	スタッフ	.110	広い	.063	近い	.059	快適	.063		
21	家族	.104	快適	.056	アニメティ	.056	コンビニ	.059		
22	部屋	.099	駅	.056	快適	.055	フロント	.055		
23	良い	.097	ベッド	.055	コンビニ	.051	アニメティ	.052		
24	09_大阪		10_福岡							
25	ホテル	.108	ホテル	.090						
26	駅	.096	便利	.087						
27	便利	.080	立地	.082						
28	立地	.074	駅	.074						
29	綺麗	.072	フロント							
30	フロント	.067	綺麗							
31	快適	.064	トイレ							
32	広い	.064	コンビ							
33	近い	.064	よい							
34	アニメティ	.054	快適							

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

● 共起ネットワークの作成(2)

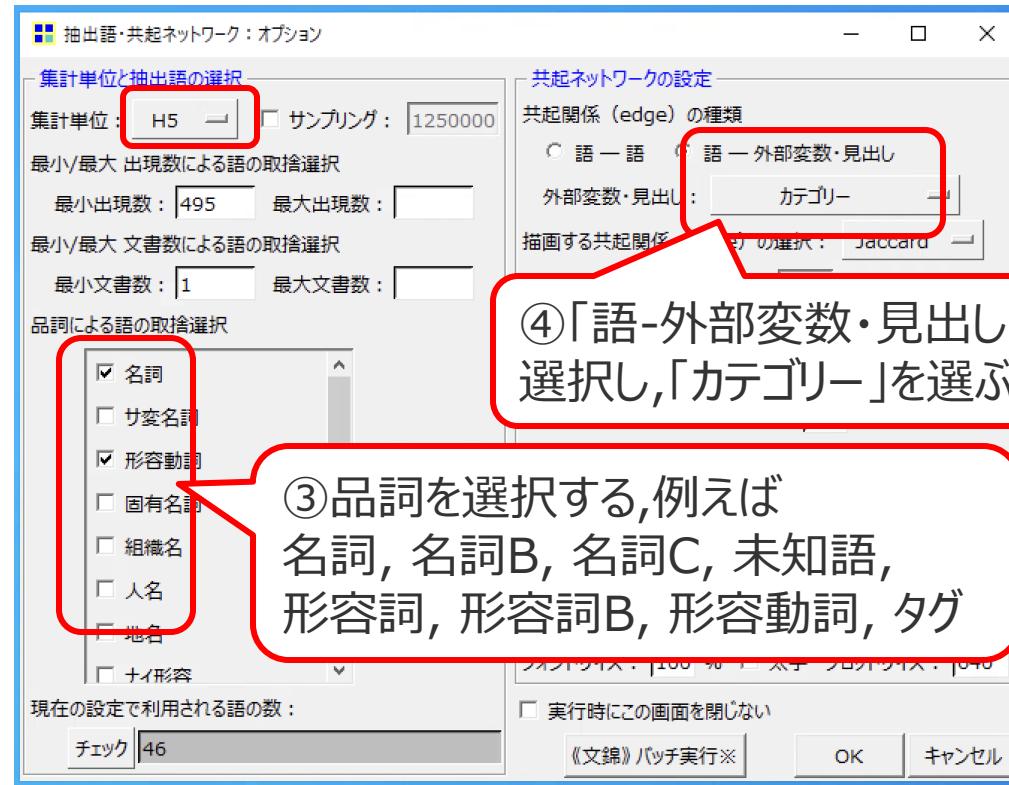
- ①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ
 - ②「集計単位」として「H5」を選んで「OK」をクリック



● 共起ネットワークの作成(3)

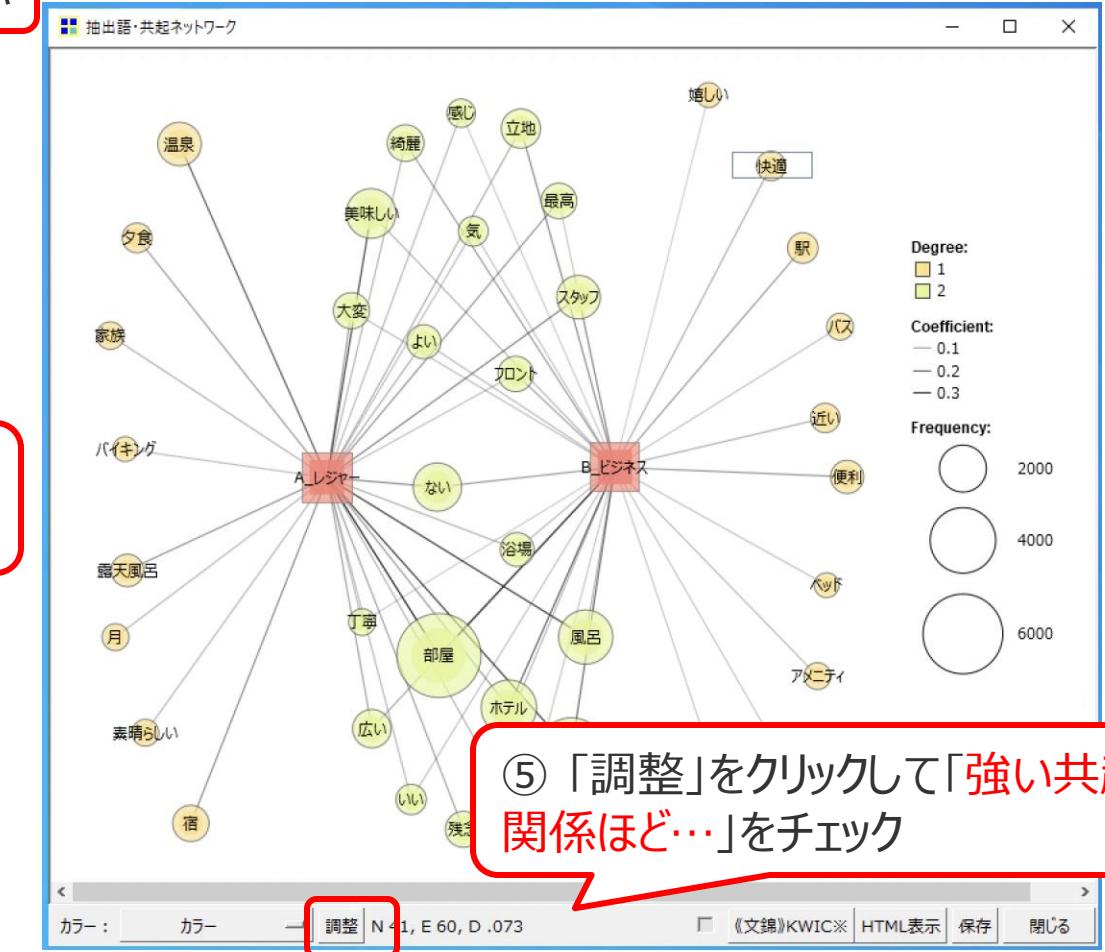
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「H5」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, 形容動詞, タグ

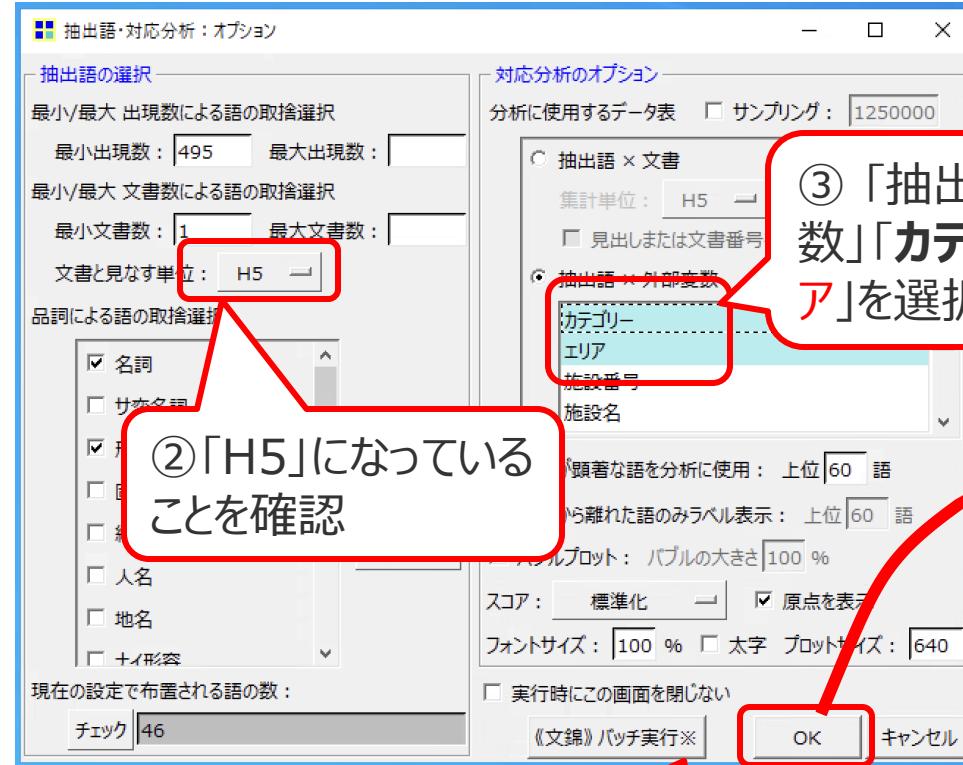
④「語-外部変数・見出し」を選択し、「カテゴリ」を選ぶ



⑤「調整」をクリックして「強い共起関係ほど…」をチェック

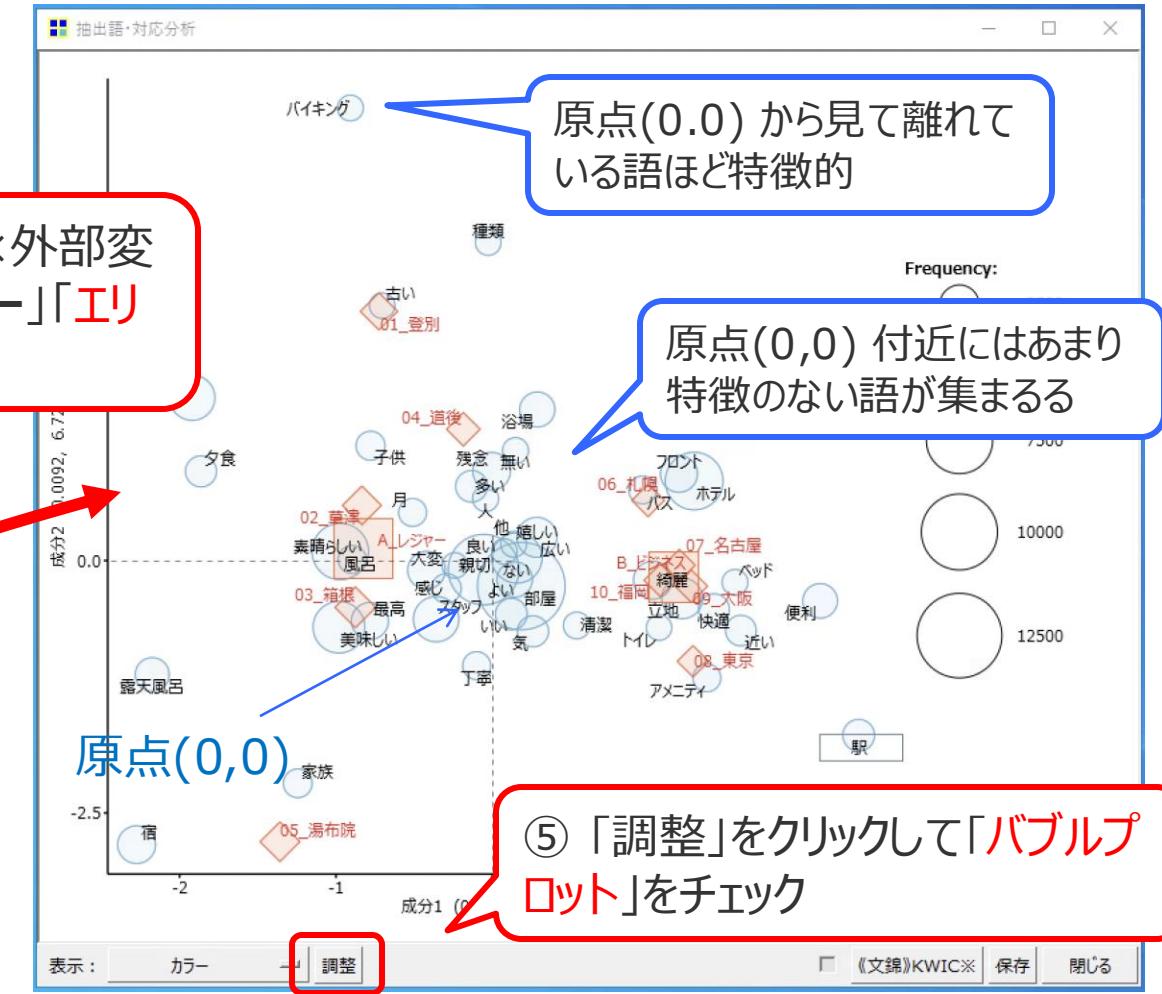
● 対応分析による探索(1)

- ① メニューから「ツール」「抽出語」「対応分析」を選ぶ



③ 「抽出語×外部変数」「カテゴリ」「エリア」を選択

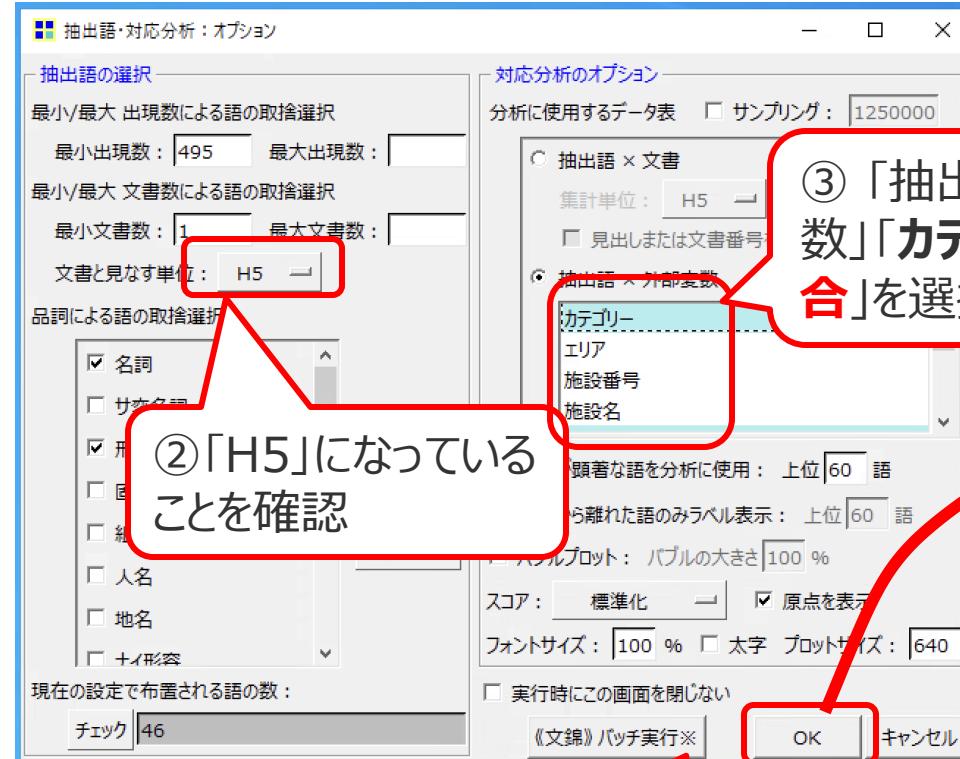
④ 「OK」をクリック



⑤ 「調整」をクリックして「バブルプロット」をチェック

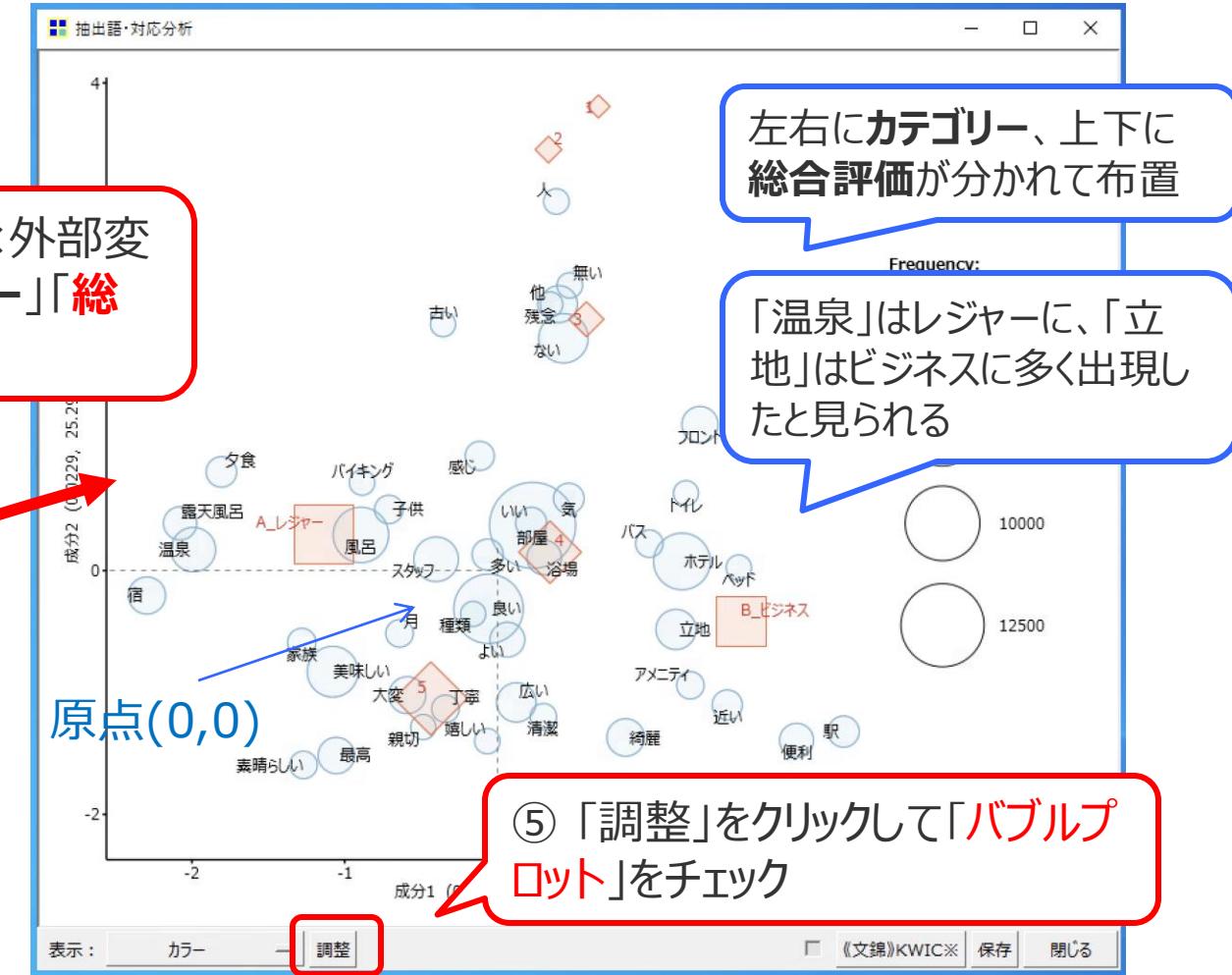
● 対応分析による探索(2)

- ① メニューから「ツール」「抽出語」「対応分析」を選ぶ



③ 「抽出語×外部変数」「カテゴリー」「総合評価」を選択

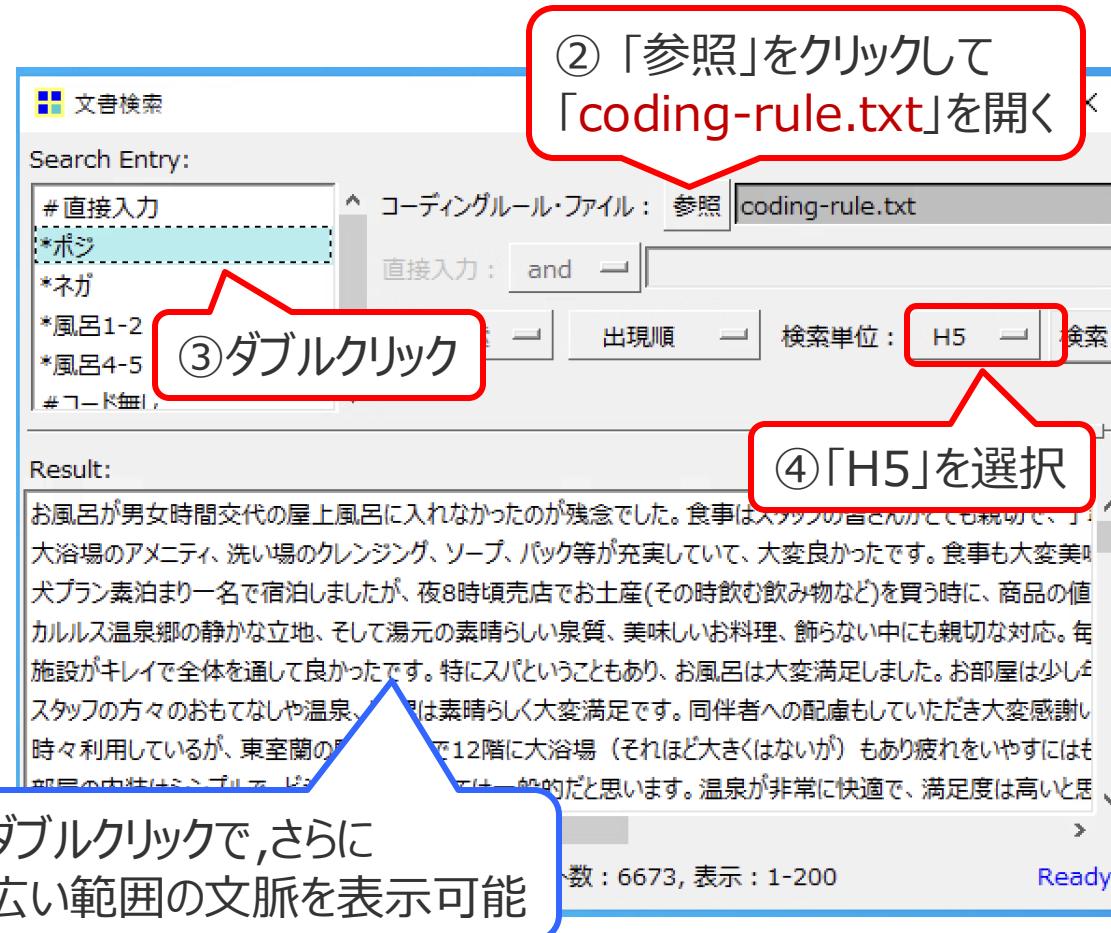
④ 「OK」をクリック



⑤ 「調整」をクリックして「バブルプロット」をチェック

●コーディングルール（語ではなくコンセプトを数える方法）

①メニューから「ツール」「文書」「文書検索」を選ぶ



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

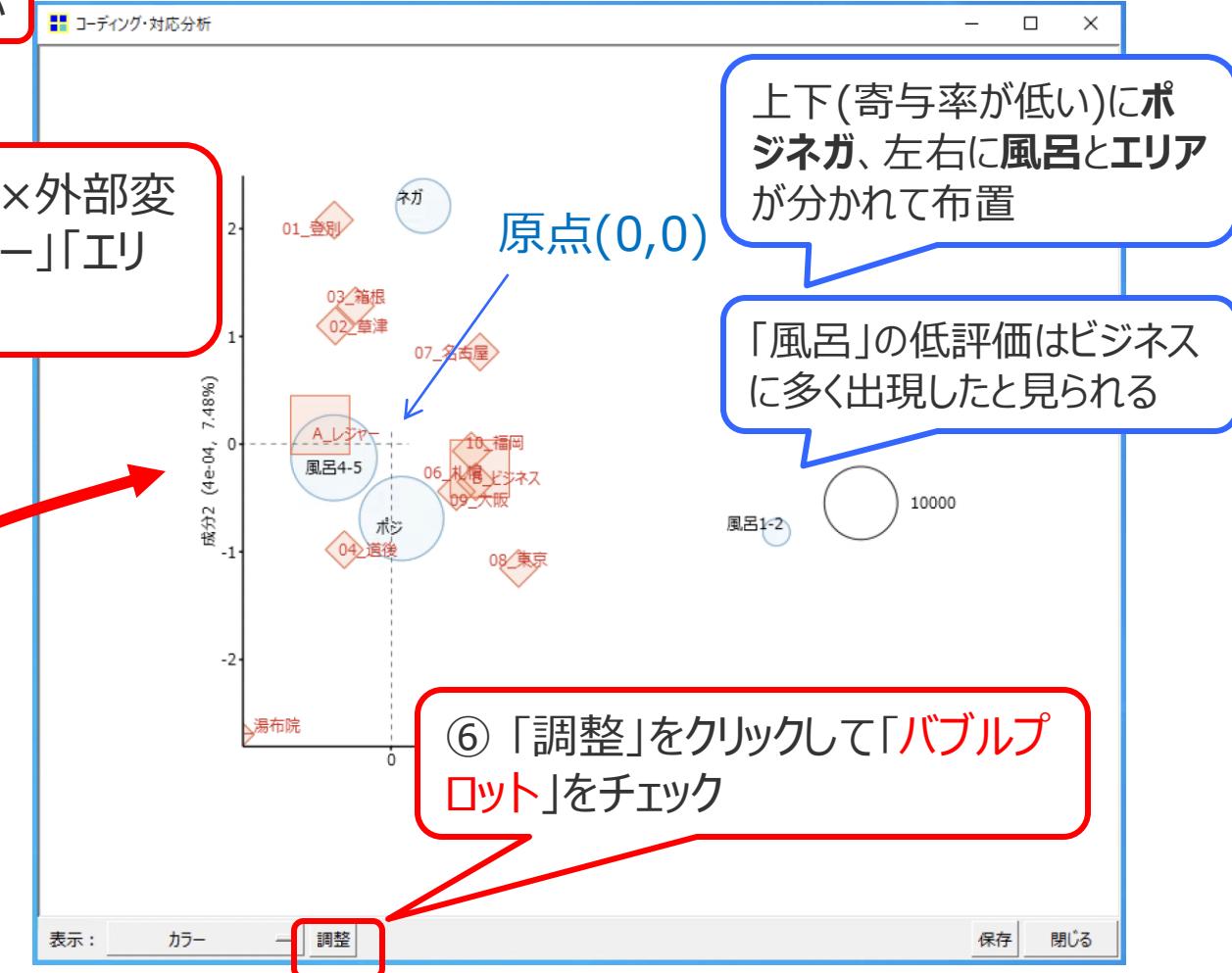
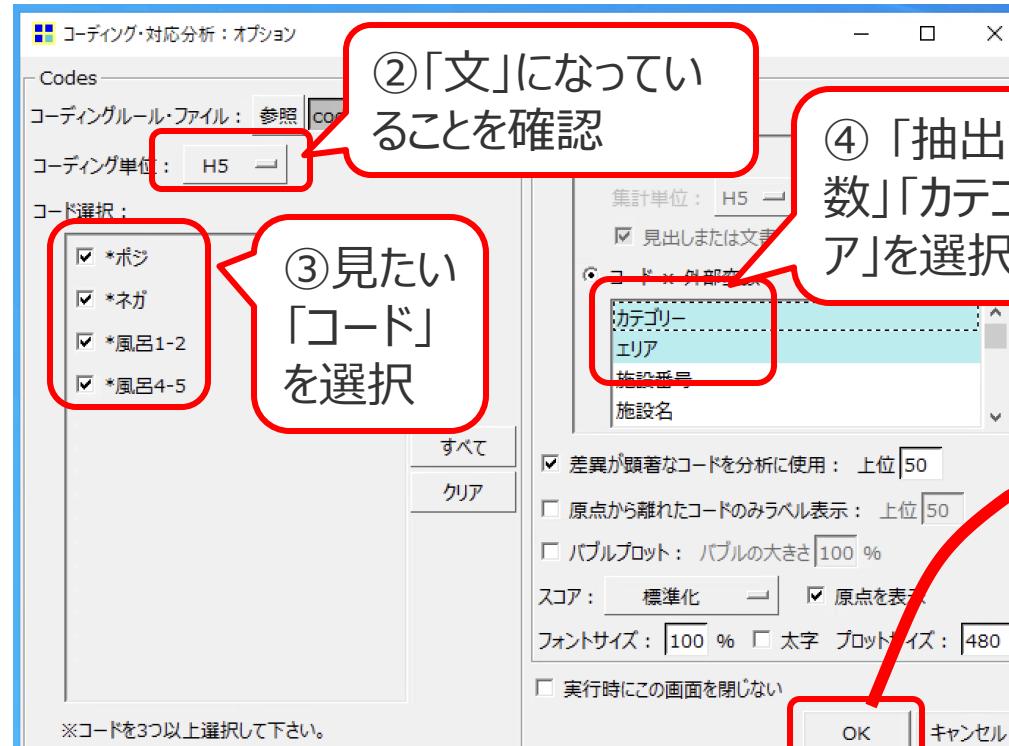
*風呂4-5

<>風呂-->4 | <>風呂-->5

外部変数

● 対応分析による探索(3)

- ① メニューから「ツール」「コーディング」「対応分析」を選択



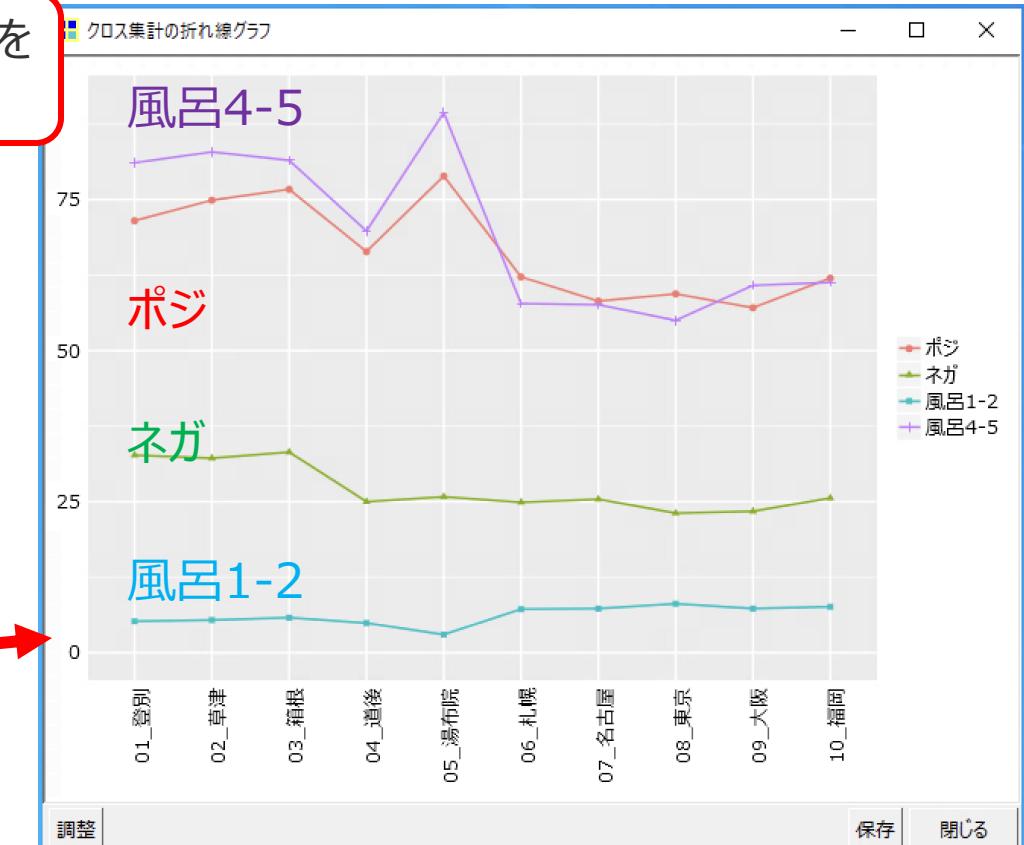
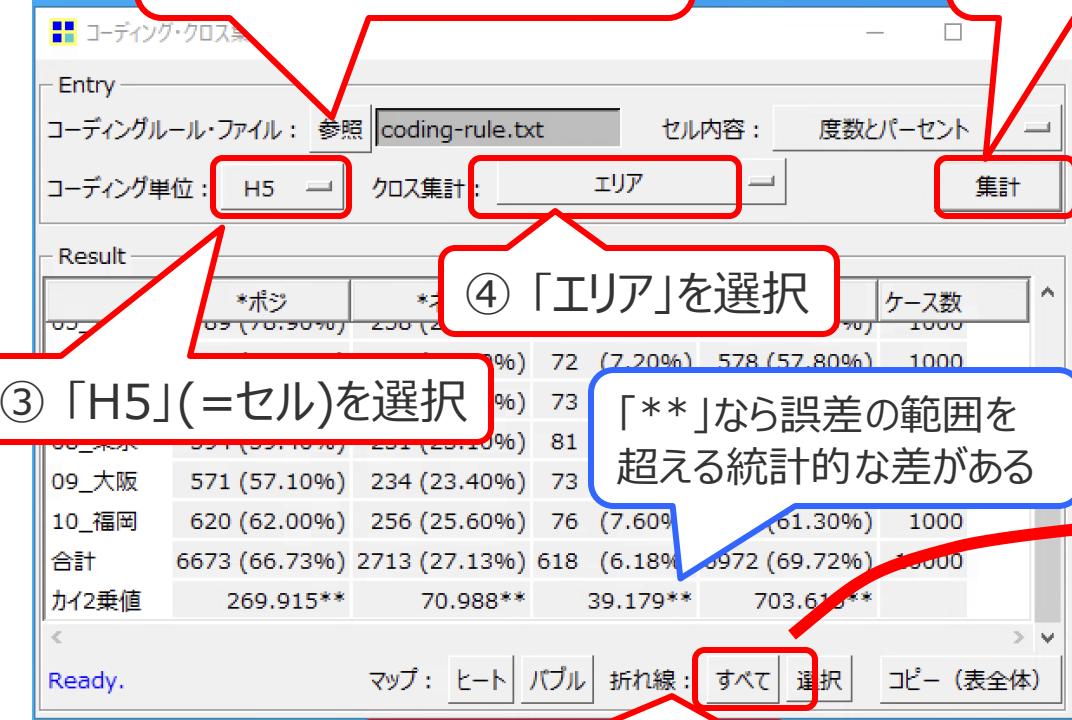
● クロス集計

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

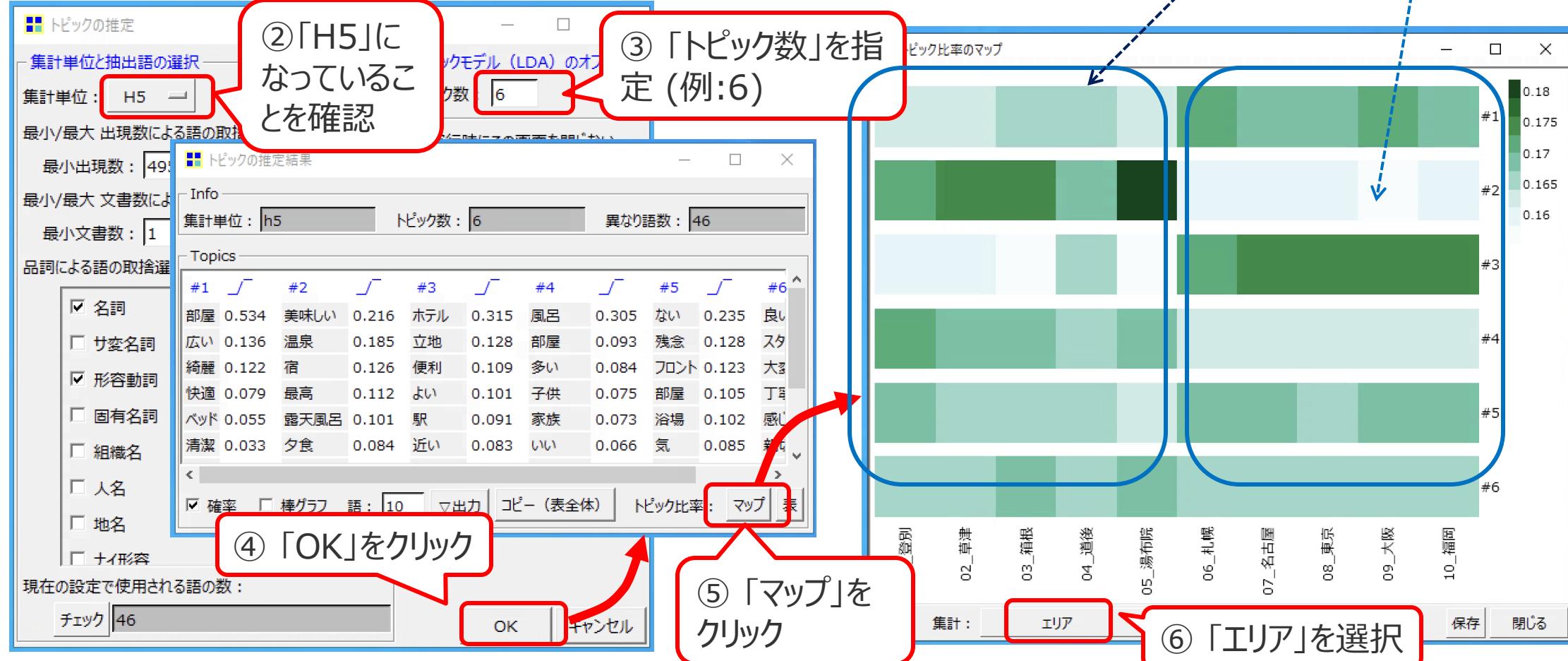
注: プロット左側のラベルは表示されません



⑥ 「すべて」をクリック

● トピックモデルによる分析

- ① メニューから「ツール」「文書」「トピックモデル」「トピックの推定」を選ぶ



KH Coder の解析・分析手法

- KH Coder は「単語と単語」、「外部変数と単語」の関係に注目した分析が得意

- 特徴的な単語を見つける

- 特定の文書に特徴的な単語を見つける → TF・IDF
→ その文書に特に頻出するが、他の文書ではそれほどではない

- 特徴的な関係を見つける

- 関係性の強い単語と単語を見つける → **共起ネットワーク(Jaccard係数)**
例) 単語:「風呂」と 単語:「広い」に関係が強そう
 - 関係性の強い単語と外部変数を見つける → **対応分析(カイ2乗値)**
例) 外部変数:「レジヤー」と 単語:「風呂」は関係が強そう

- 「文書-抽出語」表
 - 【行】ある文中に出現する単語の数を要素とする文ベクトル
 - 【列】全文中に出現する単語の数を要素とする単語ベクトル

「文書-抽出語」one-hot ベクトル表

カテゴリー	エリア	文書ID	部屋	良い	ホテル	風呂	美味しい	温泉	スタッフ	立地	よい	広い	綺麗だ	最高	宿	大変だ	便利だ	残念だ	浴場	フロン	駅	多い	露天風	近い	快適だ	いい	気	夕食	感じ	丁寧だ	家族	バス	アメニ	嬉しい	やすい	コンビ	素晴らしい	トイレ
A_レジヤー	01_登別	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
A_レジヤー	01_登別	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0			
A_レジヤー	01_登別	3	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
A_レジヤー	01_登別	4	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
A_レジヤー	01_登別	5	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0				
A_レジヤー	01_登別	6	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0				
A_レジヤー	01_登別	7	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0				
A_レジヤー	01_登別	8	1	1	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0				
A_レジヤー	01_登別	9	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0				
A_レジヤー	01_登別	10	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0				

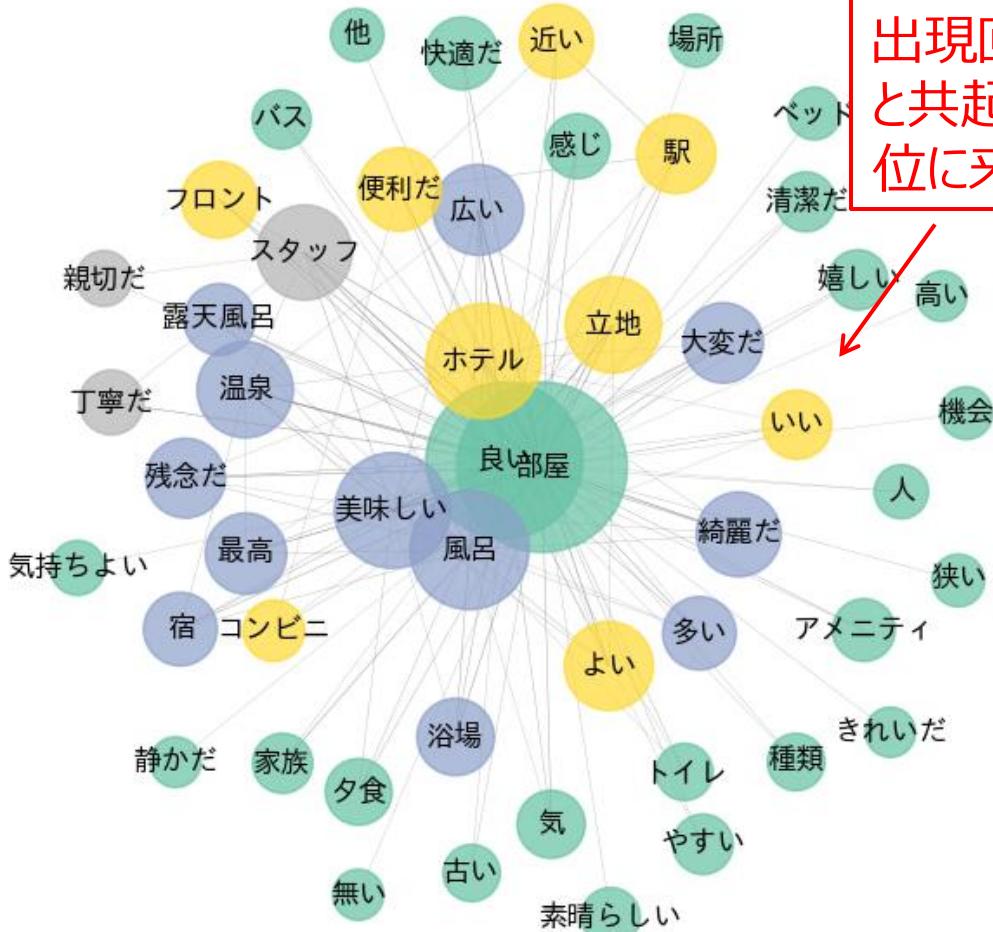
「抽出語-抽出語」共起頻度表 (共起ネットワーク)

表層	部屋	良い	ホテル	風呂	美味しい	温泉	スタッフ	立地	よい	広い	綺麗だ	最高	宿	大変だ	便利だ	残念だ
部屋	4390	1877	1032	1267	1055	708	777	666	655	973	790	479	463	508	478	571
良い	0	3703	873	1000	889	682	694	758	329	573	480	350	388	431	369	437
ホテル	0	0	1996	438	380	277	358	353	257	310	280	173	87	234	312	240
風呂	0	0	0	2142	634	366	363	301	344	445	309	315	302	260	189	317
美味しい	0	0	0	0	2023	469	420	247	281	341	282	282	309	304	192	242
温泉	0	0	0	0	0	1400	261	197	202	233	157	256	289	183	110	167
スタッフ	0	0	0	0	0	0	1371	200	207	215	184	180	181	203	125	201
立地	0	0	0	0	0	0	0	1394	263	179	178	227	92	160	259	138
よい	0	0	0	0	0	0	0	1193	201	135	108	124	127	134	145	
広い	0	0	0	0	0	0	0	0	1243	201	133	126	158	161	159	
綺麗だ	0	0	0	0	0	0	0	0	0	1086	125	100	126	126	98	
最高	0	0	0	0	0	0	0	0	0	0	997	148	101	56	91	

「外部変数-抽出語」クロス集計表 (対応分析)

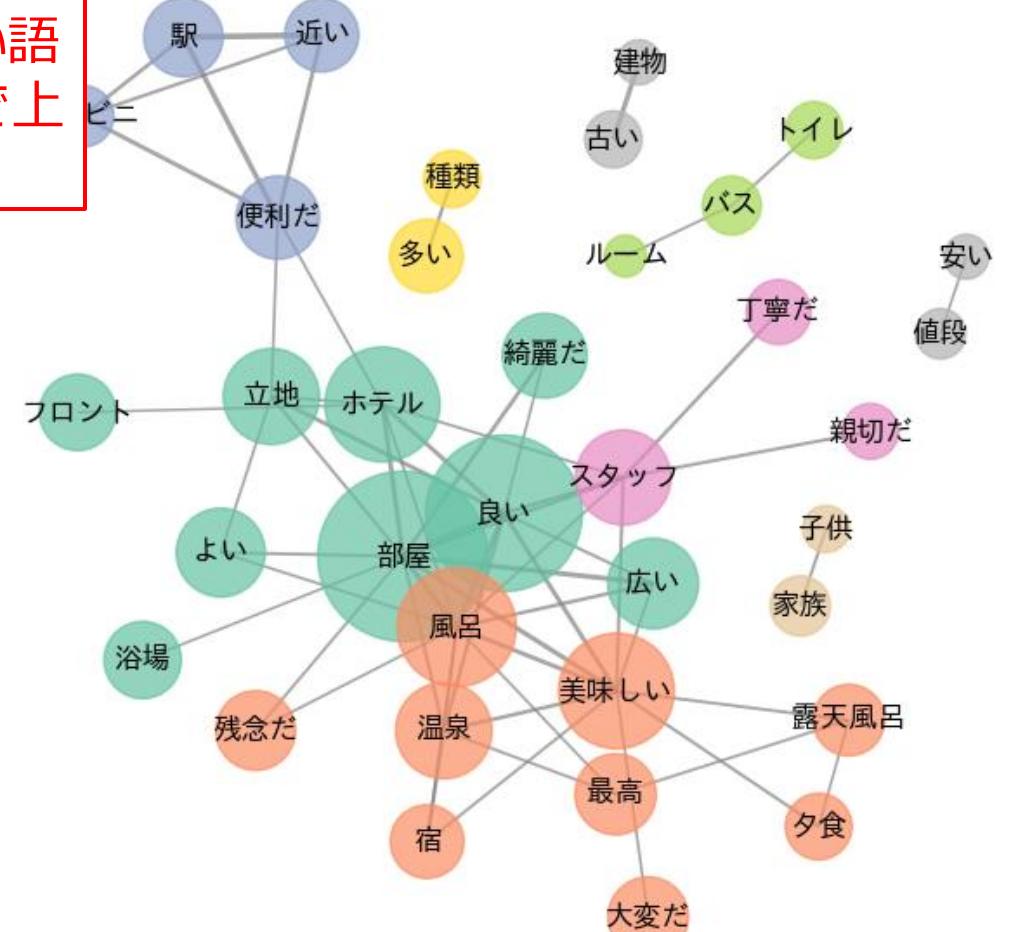
	部屋	良い	ホテル	風呂	美味しい	温泉	スタッフ	立地	よい	広い	綺麗だ	最高	宿	大変だ	便利だ	残念だ
A_レジヤー	2325	2054	753	1474	1504	1292	856	550	648	676	520	696	749	623	294	582
B_ビジネス	2065	1649	1243	668	519	108	515	844	545	567	566	301	70	353	753	379
01_登別	435	398	157	301	269	267	118	57	109	142	92	132	93	111	40	146
02_草津	481	441	168	362	289	287	193	161	141	130	123	155	182	130	72	126
03_箱根	510	450	166	287	351	220	195	57	142	144	121	117	152	156	35	146
04_道後	380	366	218	190	202	257	103	164	136	113	88	99	64	98	106	82
05_湯布院	519	399	44	334	393	261	247	111	120	147	96	193	258	128	41	82
06_札幌	399	324	263	118	145	27	102	170	105	125	121	72	15	83	157	86
07_名古屋	388	334	243	147	95	26	125	145	115	106	117	47	15	63	133	76
08_東京	416	333	224	133	80	15	99	170	96	86	105	63	16	75	155	65
09_大阪	423	305	252	126	88	19	96	171	118	132	117	57	16	62	152	71
10_福岡	439	353	261	144	111	21	93	188	111	118	106	62	8	70	156	81

- 共起頻度の高いエッジを残す



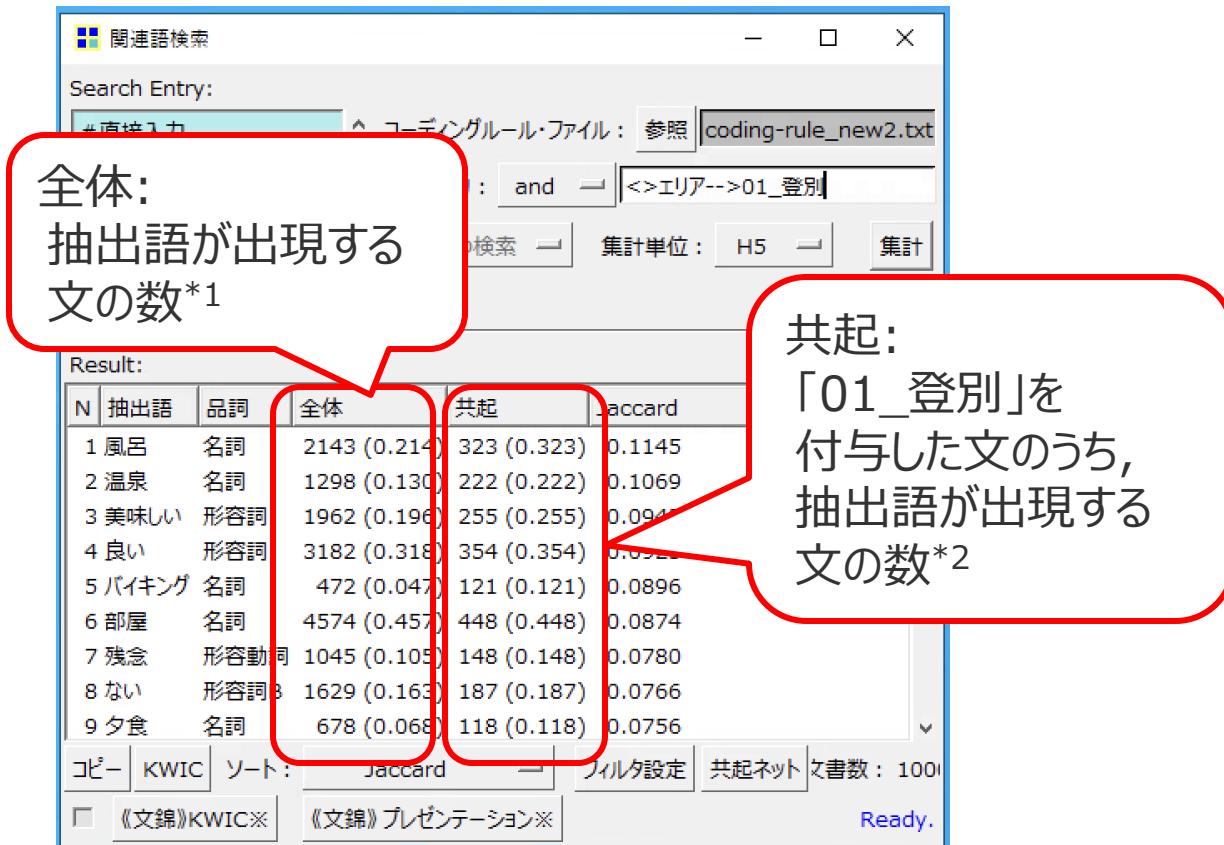
出現回数の高い語
と共に起するだけで上
位に来てしまう

- Jaccard 係数が上位のエッジを残す

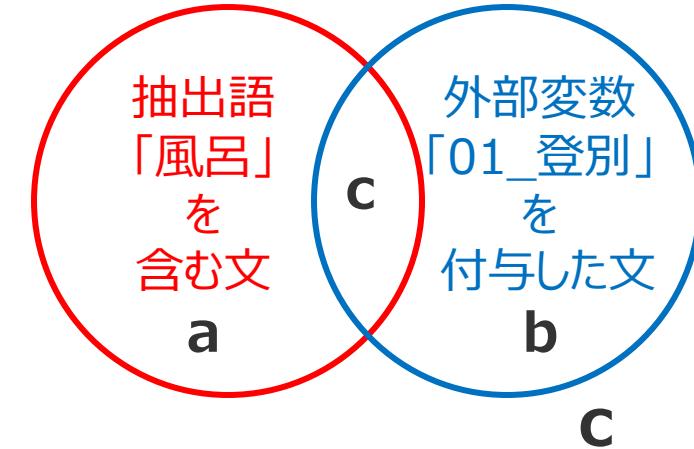


Jaccard 系数

- Jaccard 系数は、共起の強さを測る尺度 (KH Coderで標準的に使用)
 - どちらの語も含まない文書を無視 → 言語のようなスパースデータ分析に向いている



*1 括弧内はデータ全体に対する割合(前提確率) *2 括弧内は「01_登別」を付与したデータに対する割合(条件付き確率)



$$\text{Jaccard 系数} = \frac{c}{a+b+c}$$

抽出語「風呂」の場合:
c = 323 ("共起"列の値)
a = 2143 ("全体"列の値) - 323 = 1820
b = (323 / 0.323) - 323 = 677

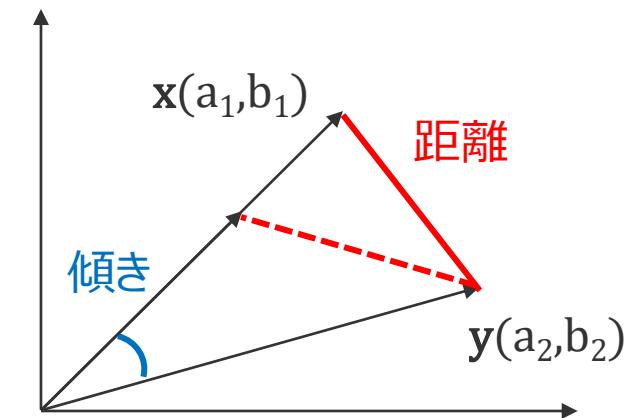
その他の尺度

- 出現パターンが似てる を測る = ユークリッド距離、コサイン距離

- 1つひとつの文が長く、各文中での語の出現回数の大小が重要なケースに向く
(語が1回出現したか、10回出現したかを区別したい)

ユークリッド距離	コサイン距離
サイズ(出現回数の大小)の差まで見る場合向き	傾きが似ているかどうかだけを見る場合向き
$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$	$d(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$

※ x, y はそれぞれの単語ベクトル (単語の出現パターン)



カイ2乗値

- カイ2乗値は「無関係でない」度合いを測る尺度 → カテゴリと変数間の関連性を測定

$$\text{カイ2乗値} = \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

「観測度数」: カテゴリと変数に従ってクロス集計された度数

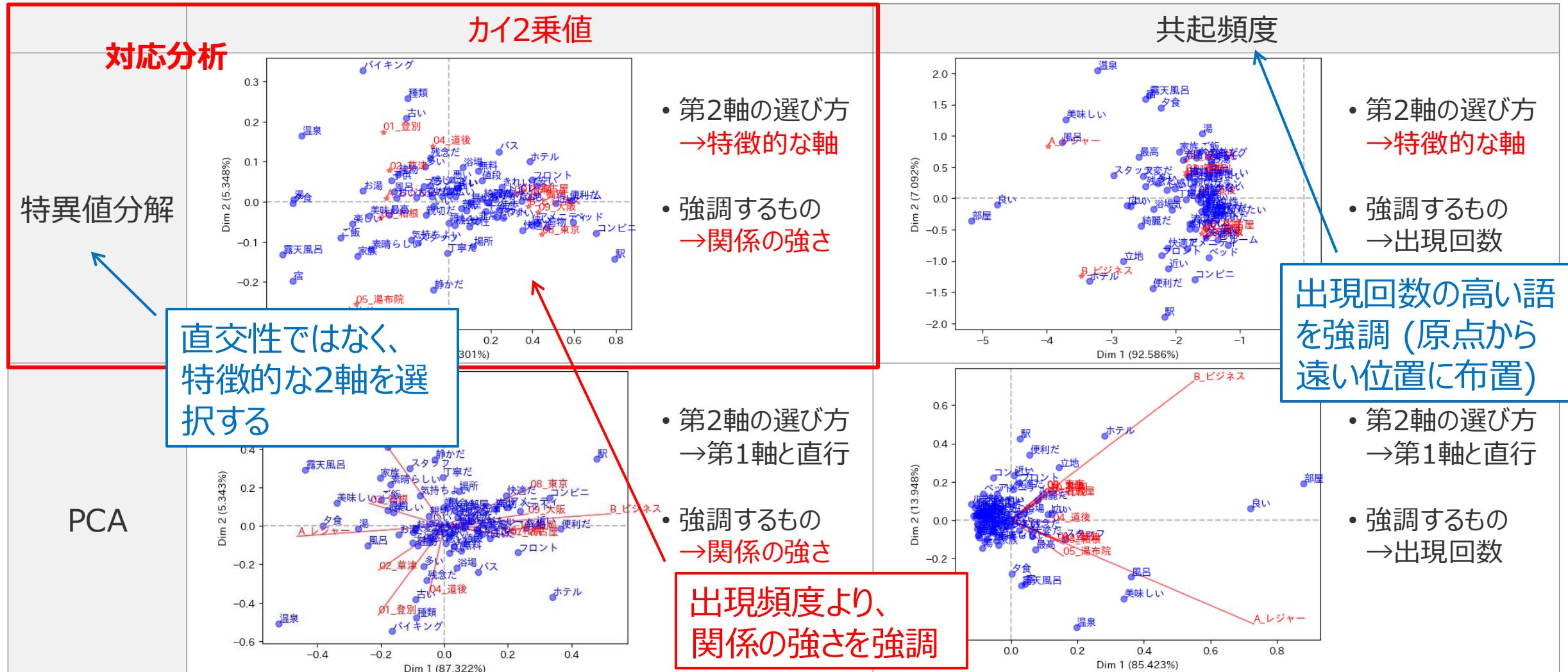
「期待度数」: 変数が互いに独立している場合に期待される度数

「観測度数 - 期待度数」: 実際の度数と独立と期待される度数の差

- カイ2乗値も大きい → カテゴリと変数間の関係が**期待より強い**を示す

クロス集計表 (観測度数)					期待度数					観測度数-期待度数					カイ2乗値												
	部屋	良い	ホテル	風呂	美味しい	合計		部屋	良い	ホテル	風呂	美味しい	合計		部屋	良い	ホテル	風呂	美味しい	合計		部屋	良い	ホテル	風呂	美味しい	合計
01_登別	435	398	157	301	269	1560	01_登別	480.45	405.27	218.45	234.43	221.40	1560.00	01_登別	-45.45	-7.27	-61.45	66.57	47.60	0.00	01_登別	4.30	0.13	17.28	18.91	10.23	50.85
02_草津	481	441	168	362	289	1741	02_草津	536.20	452.29	243.79	261.63	247.09	1741.00	02_草津	-55.20	-11.29	-75.79	100.37	41.91	0.00	02_草津	5.68	0.28	23.56	38.51	7.11	75.14
03_箱根	510	450	166	287	351	1764	03_箱根	543.28	458.26	247.01	265.08	250.36	1764.00	03_箱根	-33.28	-8.26	-81.01	21.92	100.64	0.00	03_箱根	2.04	0.15	26.57	1.81	40.46	71.03
04_道後	380	366	218	190	202	1356	04_道後	417.63	352.27	189.88	203.77	192.45	1356.00	04_道後	-37.63	13.73	28.12	-13.77	9.55	0.00	04_道後	3.39	0.54	4.16	0.93	0.47	9.49
05_湯布院	519	399	44	334	393	1689	05_湯布院	520.18	438.78	236.51	253.81	239.71	1689.00	05_湯布院	-1.18	-39.78	-192.51	80.19	153.29	0.00	05_湯布院	0.00	3.61	156.70	25.33	98.02	283.66
06_札幌	399	324	263	118	145	1249	06_札幌	384.67	324.47	174.90	187.69	177.26	1249.00	06_札幌	14.33	-0.47	88.10	-69.69	-32.26	0.00	06_札幌	0.53	0.00	44.38	25.88	5.87	76.66
07_名古屋	388	334	243	147	95	1207	07_名古屋	371.74	313.56	169.02	181.38	171.30	1207.00	07_名古屋	16.26	20.44	73.98	-34.38	-76.30	0.00	07_名古屋	0.71	1.33	32.38	6.52	33.99	74.93
08_東京	416	333	224	133	80	1186	08_東京	365.27	308.11	166.08	178.22	168.32	1186.00	08_東京	50.73	24.89	57.92	-45.22	-88.32	0.00	08_東京	7.05	2.01	20.20	11.48	46.35	87.08
09_大阪	423	305	252	126	88	1194	09_大阪	367.73	310.19	167.20	179.43	169.46	1194.00	09_大阪	55.27	-5.19	84.80	-53.43	-81.46	0.00	09_大阪	8.31	0.09	43.01	15.91	39.16	106.47
10_福岡	439	353	261	144	111	1308	10_福岡	402.84	339.80	183.16	196.56	185.64	1308.00	10_福岡	36.16	13.20	77.84	-52.56	-74.64	0.00	10_福岡	3.25	0.51	33.08	14.05	30.01	80.90
合計	4390	3703	1996	2142	2023	14254	合計	4390.00	3703.00	1996.00	2142.00	2023.00	14254.00	合計	0.00	0.00	0.00	0.00	0.00	0.00	合計	35.26	8.65	401.34	159.32	311.67	916.23

- 対応分析は、カイ2乗値を利用して、関係の強さを強調して表現できる

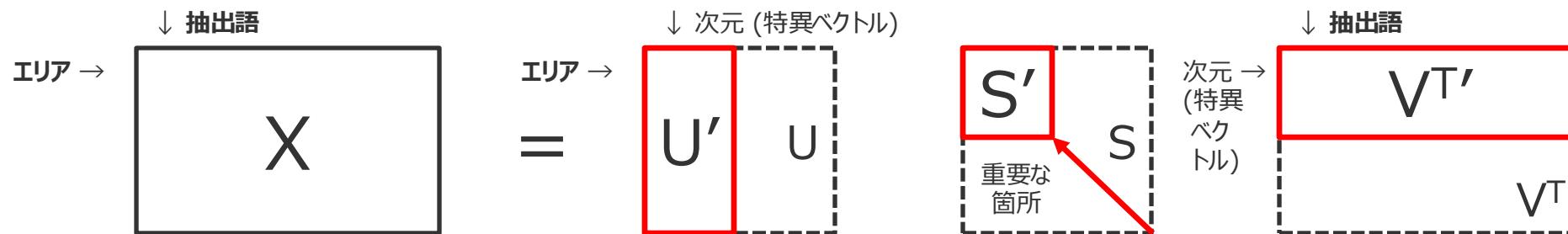


特異値分解

- 特異値分解 $X = USV^T$



- S の特異値が小さいものを削る



演習用の教材 – Google Colab ノートブック

- URL: <https://github.com/haradatm/lecture-gssm2025>

