

テキストマイニングの実践

— 2日目 —

2020/7/10

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 — データ分析の手順
 - 実習 — 環境構築, 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)
 - 説明 — データ分析の実践 (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 —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,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://khcoder.net/bib.html?year=2018&auth=all&key=>

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：630 / 3741

KH Coderを用いた研究事例のリスト 3741件

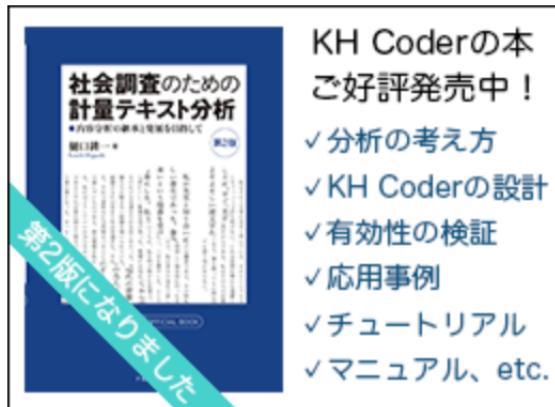
※ 2020/6/21 現在 (961→1206→1646→2042→昨年2695件)

KH Coder の情報

ホームページ <http://khcoder.net/>

The screenshot shows the official website for KH Coder. At the top is the header "KH Coder: 計量テキスト分析 / テキストマイニング". Below it is a banner with the text "KH Coder" in blue. The main content area has a "Index" section with a message about an upcoming seminar. A "概要" section provides a brief introduction to the software. A "機能紹介 (スクリーンショット)" section shows a screenshot of the software interface. A "KH Coderの入手" section lists download links for different versions. On the right side, there's a sidebar with a "参考書" section featuring a book cover for "社会調査のための計量テキスト分析 第2版になりました" and a "チュートリアル & ヒント" section with a link to "http://khcoder.net/tutorial.html".

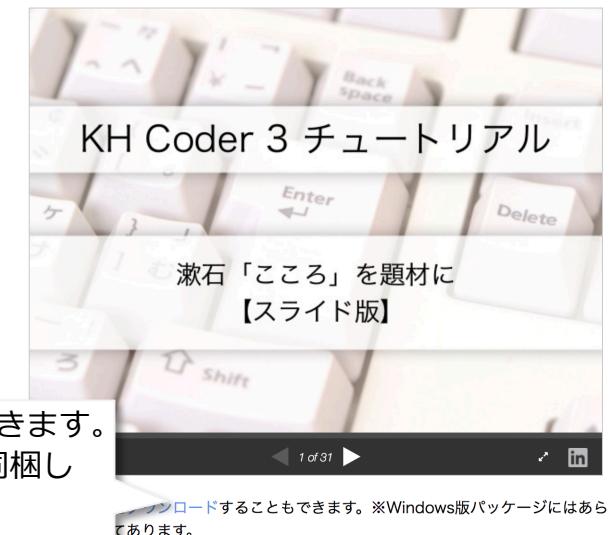
参考書



PDFファイルをダウンロードすることもできます。
※Windows版パッケージにはあらかじめ同梱してあります。

チュートリアル
<http://khcoder.net/tutorial.html>

チュートリアル & ヒント

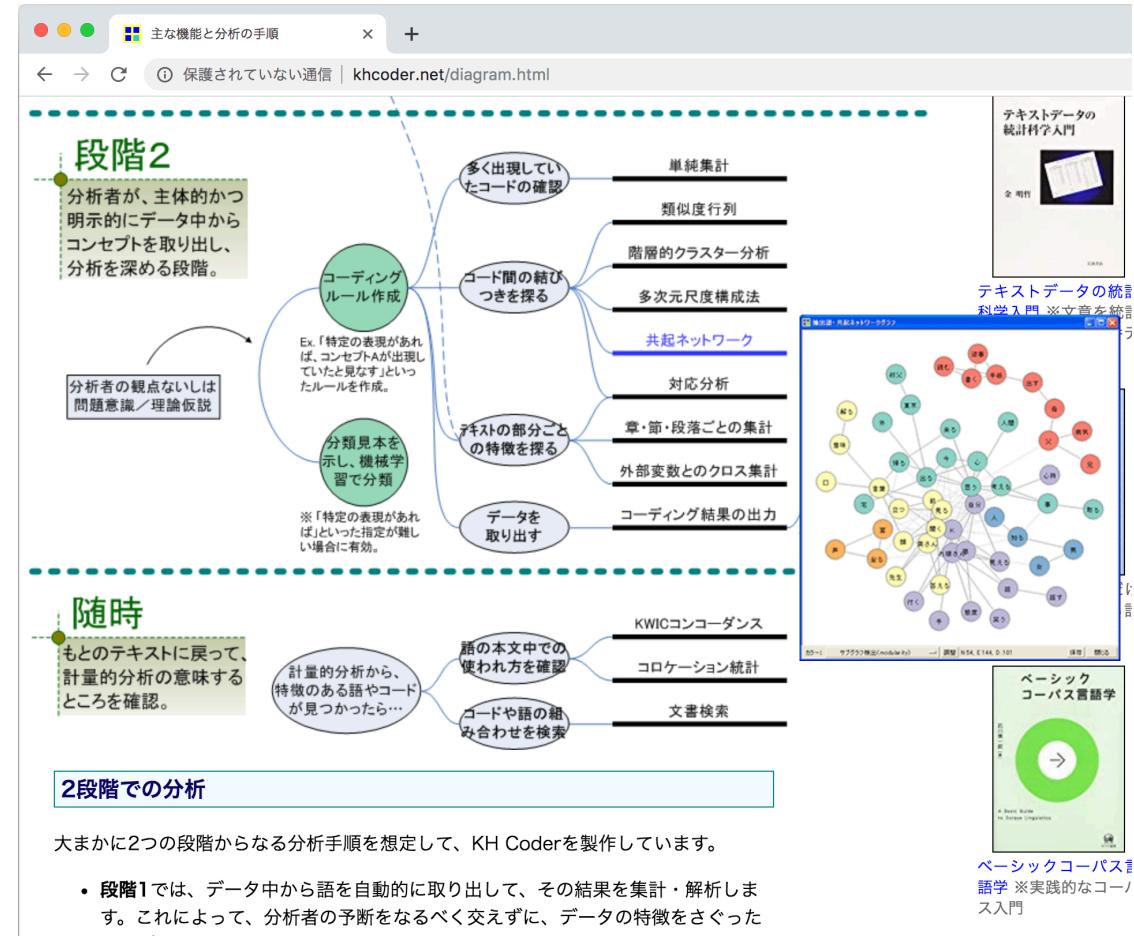
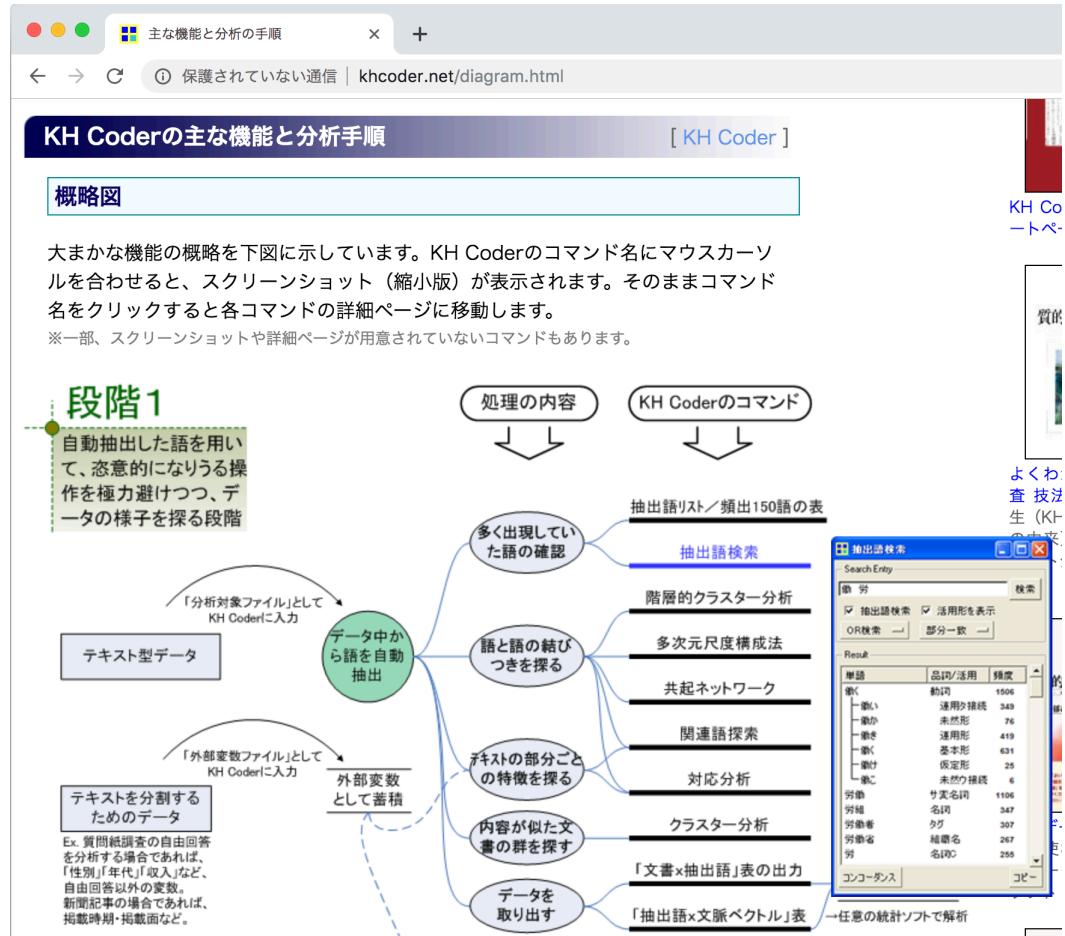


チュートリアル用データ

チュートリアルの実行に必要なデータファイルです。
※Windows版パッケージには同梱してありますので、別途ダウンロードする必要はありません。

参考 — KH Coder の分析手順

<http://khcoder.net/diagram.html>



文の出現パターンと単語の出現パターン

【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)

h5	bun	部屋	ホテル	風呂	温泉	お部屋	スタッフ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイク	家族	場所	トイレ	子供	ペット	コンビ	良い
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	
3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

距離で「似てる」を図る

- Jaccard 距離: KHCoder で標準的な距離尺度
 - 1つ文書に含まれる語が少ないケースや,各語が一部の文書中にしか含まれていないケースに向いている →スパースなデータ分析向き

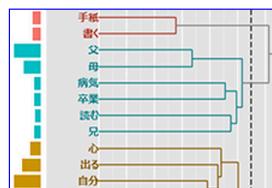
Jaccard 距離	ユークリッド距離	コサイン距離								
<ul style="list-style-type: none">• 1つの文書の中に語が1回出現した場合も10回出現した場合も単に「出現あり」と見なしてカウントした語と語の共起数を計算• 語Aと語Bのどちらも出現していない文書(0-0対)が沢山あっても語Aと語Bが類似しているとは見なさない	<ul style="list-style-type: none">• 文書中における語の出現回数(1,000語あたりの出現回数に調整)を計算• 1つひとつの文書が長く,多数の文書に含まれている語が多いデータ向き(各文書中の語の出現回数の大小が重要な場合)	<ul style="list-style-type: none">• サイズの差までも見る場合向き• 傾きが似ているかどうかだけを見る場合向き								
<table border="1"><tr><td>1</td><td>0</td></tr><tr><td>1</td><td>n_{11}</td><td>n_{10}</td></tr><tr><td>0</td><td>n_{01}</td><td>n_{00}</td></tr></table> $J^S = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$	1	0	1	n_{11}	n_{10}	0	n_{01}	n_{00}	$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$	$\cos S(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$
1	0									
1	n_{11}	n_{10}								
0	n_{01}	n_{00}								

<http://mjin.doshisha.ac.jp/R/68/68.html>

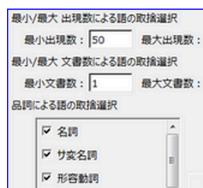
KH Coder—スクリーンショット

階層的クラスター分析

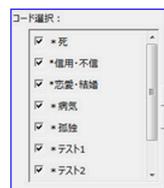
抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけでなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



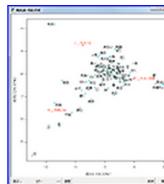
抽出語は出現数や品詞で選択



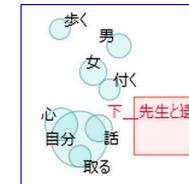
コードはチェックボックスで直接選択

対応分析

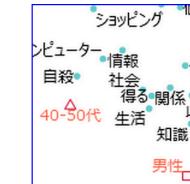
同じく抽出語またはコードを用いての、対応分析です。



同時布置図



New! バブルプロット



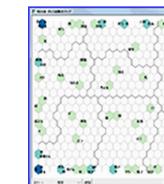
複数の外部変数を用いた多重対応分析

自己組織化マップ

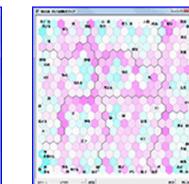
抽出語またはコードを用いての、自己組織化マップです。



クラスター色分け



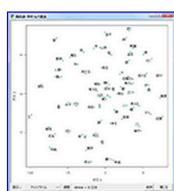
頻度のプロット



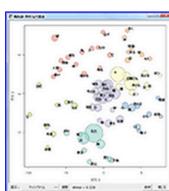
U-Matrix

多次元尺度構成法 (MDS)

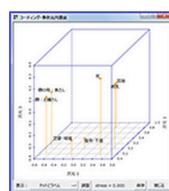
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



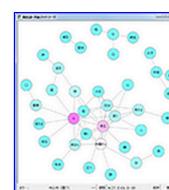
New! クラスタリングと
色分け



3次元の解

共起ネットワーク

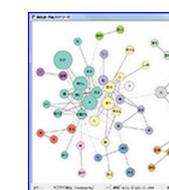
抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



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描画に含め、自動的にグループ分け（色分け）



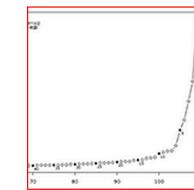
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

文書のクラスター分析

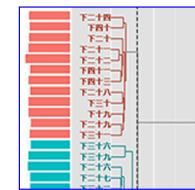
文書の分類を行うクラスター分析です。



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。



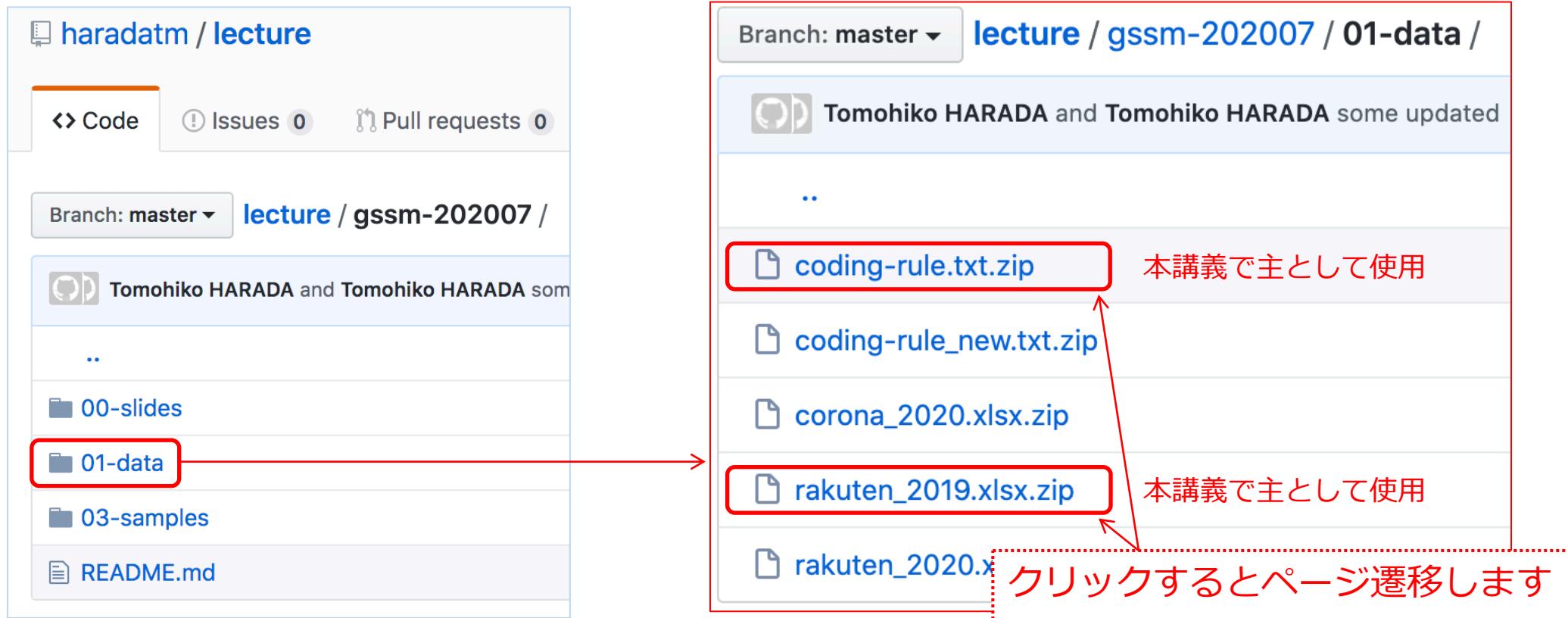
文書のデンドログラム。左の棒グラフは各文書の長さをあらわす。なお、文書数が500を超える場合、デンドログラムは表示不可。

KH Coder の主な分析手法

分析手法	解説
階層的クラスター分析	<ul style="list-style-type: none">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(MDS)	<ul style="list-style-type: none">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 (PCAが元の情報をそのまま可視化するのに対し, 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, 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	<ul style="list-style-type: none">出現パターンの似た単語同士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
文書のクラスター分析	<ul style="list-style-type: none">似た文書同士をグルーピング(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

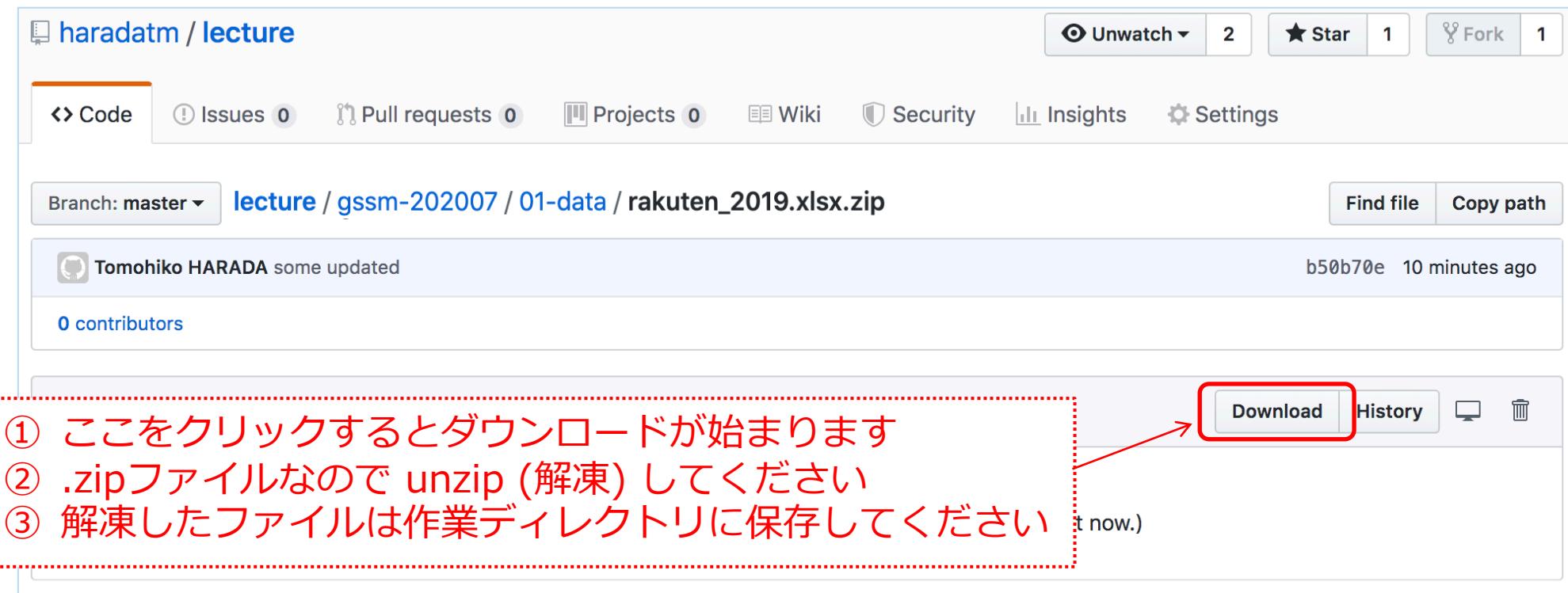
データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-202007>



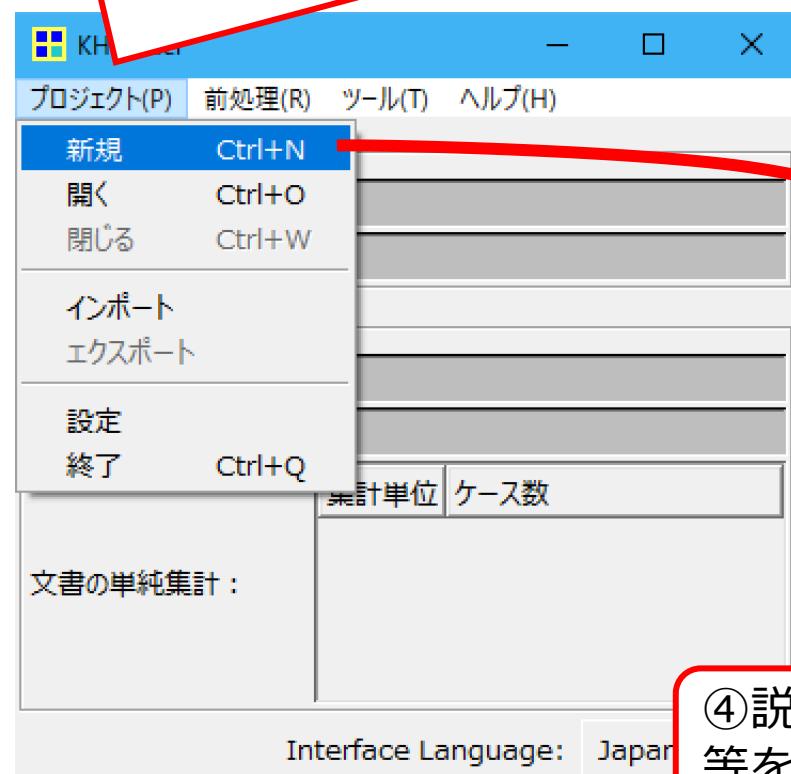
ダウンロード方法

- Download ボタンをクリックするとダウンロードを開始



操作説明 — プロジェクトの作成

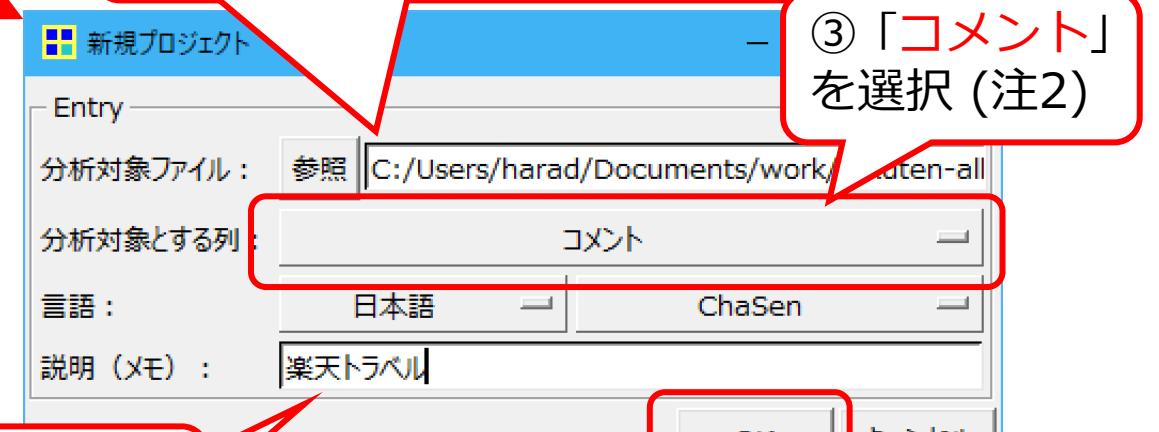
①メニューから「プロジェクト」「新規」を選択 (注1)



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_2019.xlsx」を開く

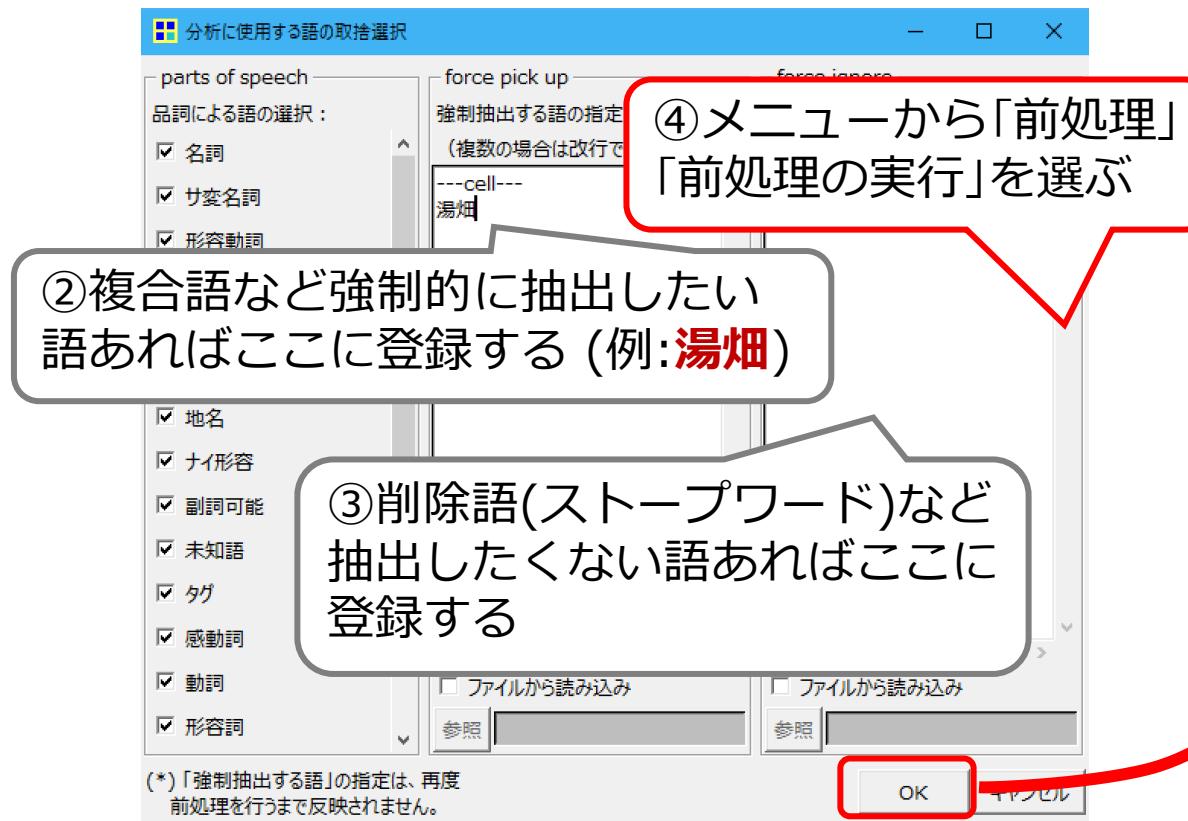


④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

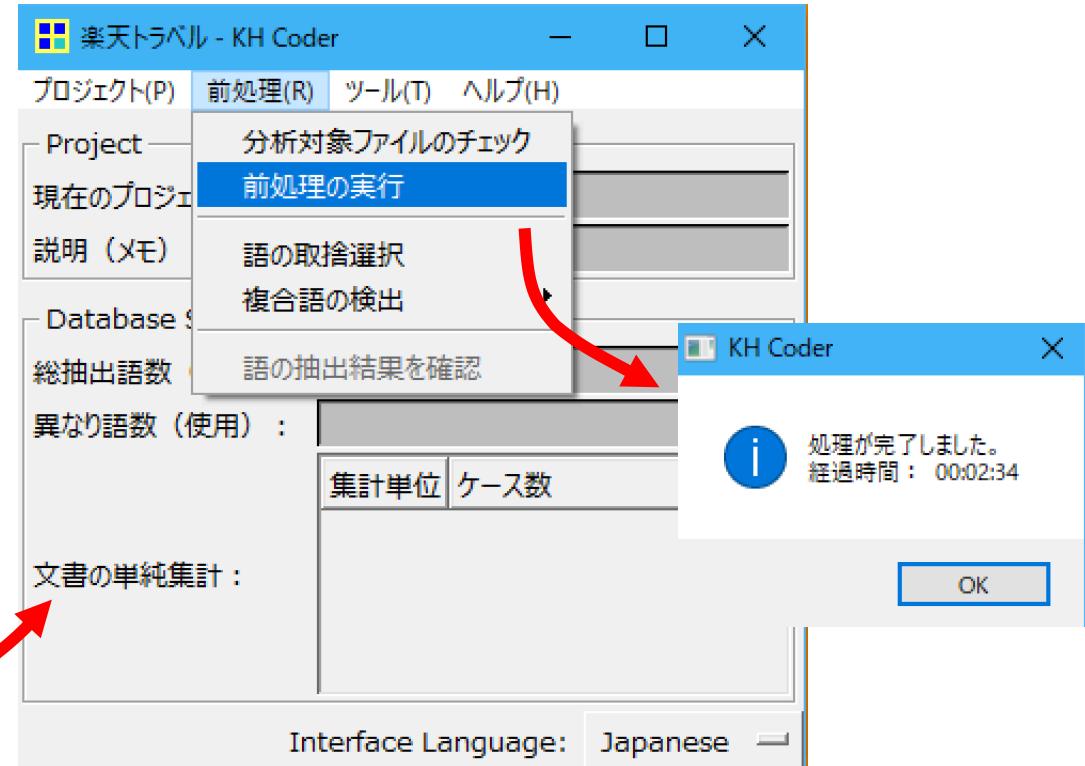
操作説明 — 前処理 (形態素解析)

①メニューから「前処理」「語の取捨選択」を選ぶ



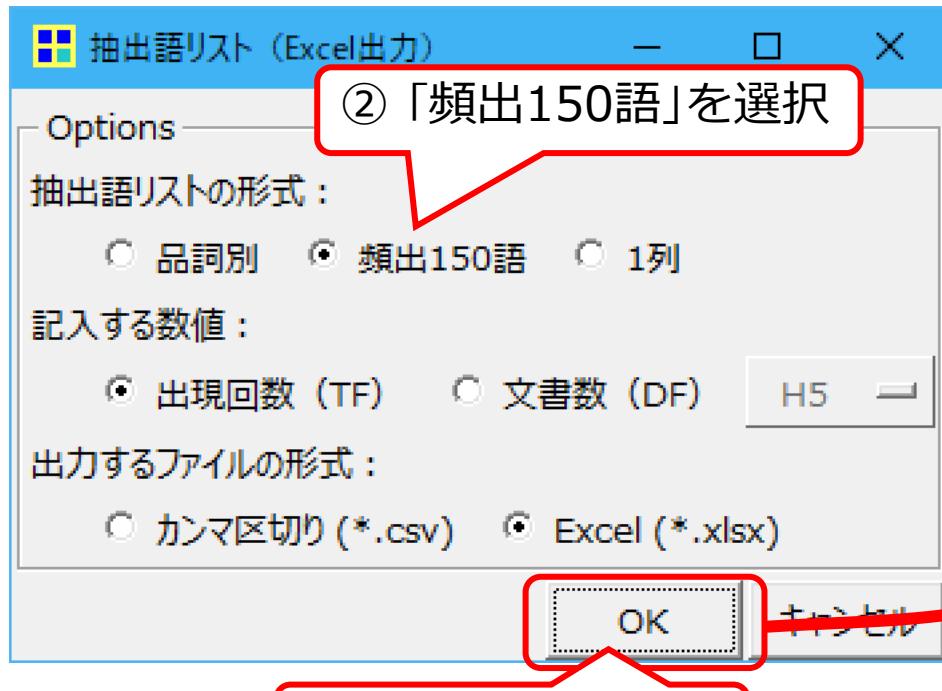
注1: EXCELファイルを読み込んで分析する場合、あらかじめ「---cell---」が入力されています

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



操作説明 — 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

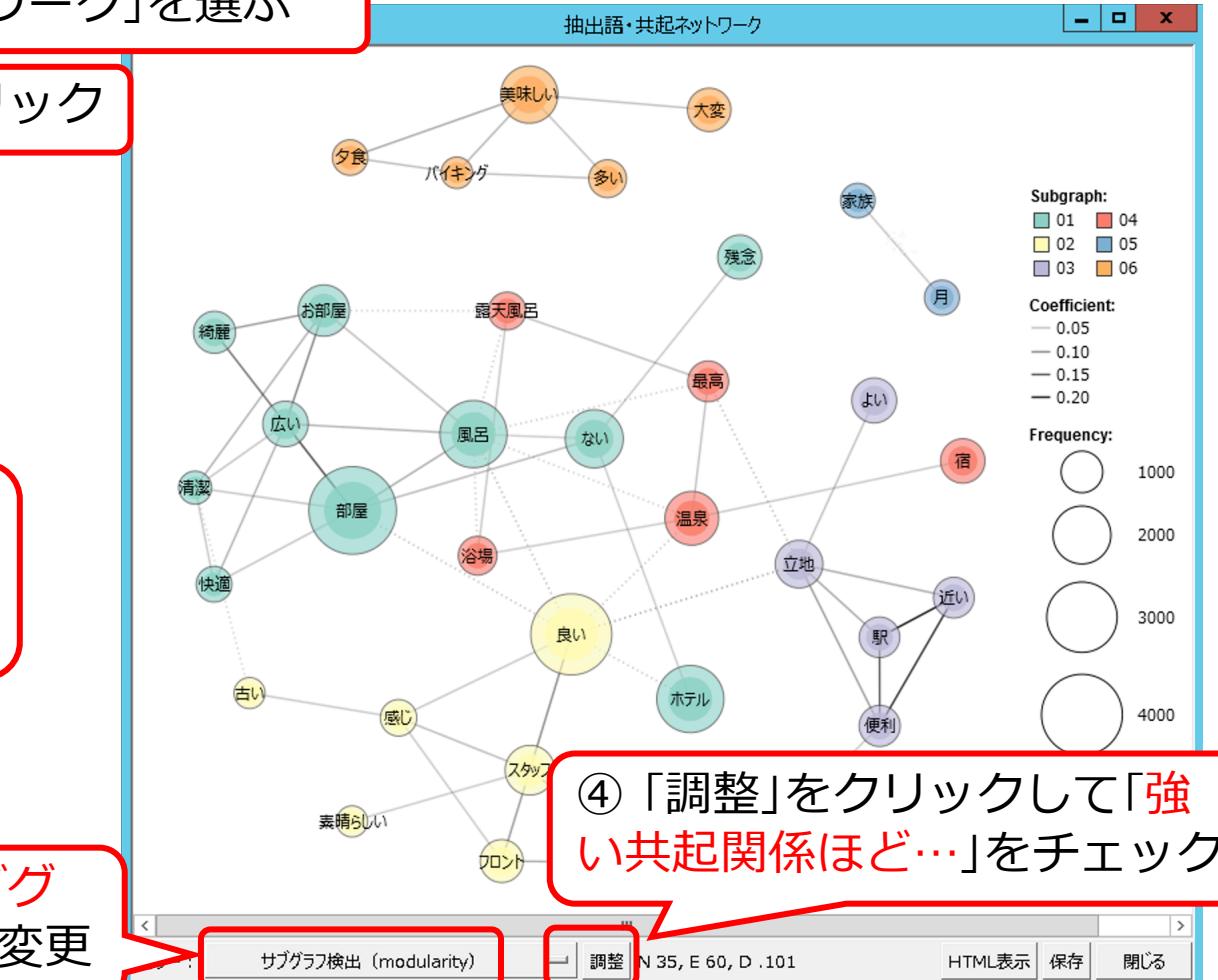
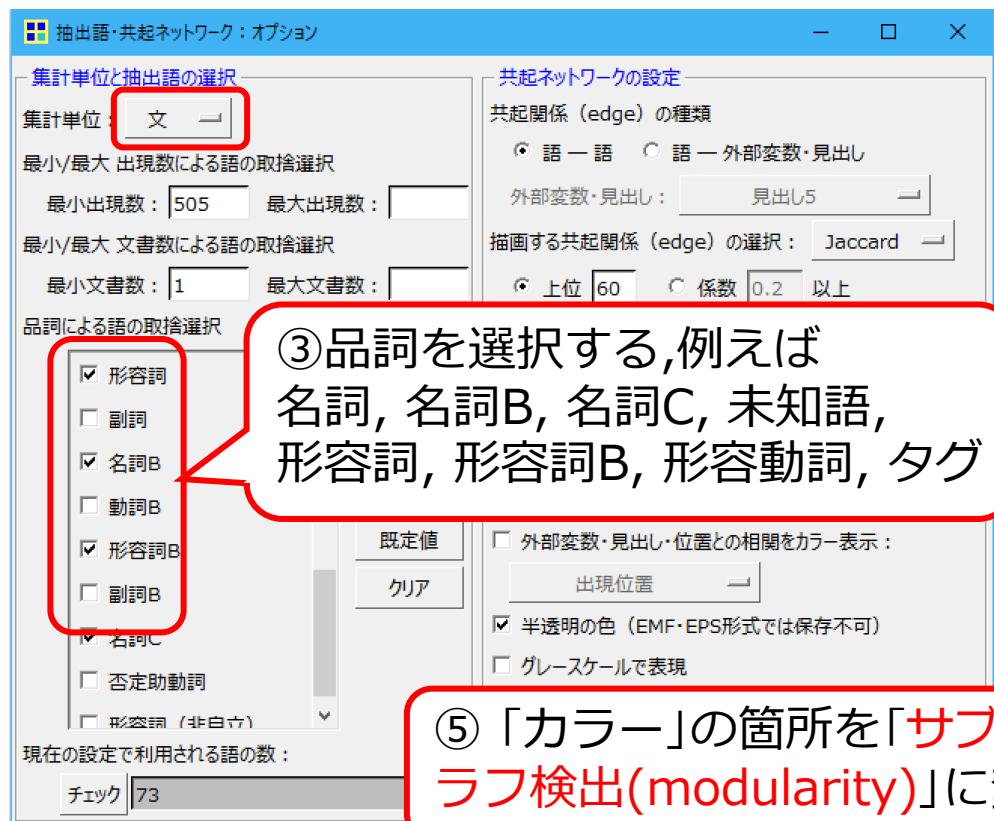


A	B	C	D	E	F	G	H
1 抽出語	出現回数		抽出語	出現回数		抽出語	出現回数
2 部屋	4713		家族	677		非常	420
3 思う	4080		バス	661		設備	418
4 良い	4013		予約	638		湯	407
5 利用	3550		旅行	629		接客	403
6 宿泊	2731		清潔	618		高い	401
7 風呂	2727		子供	616		静か	398
8 ホテル	2726		初めて	613		掃除	397
9 食事	2329		駐車	580		無料	391
10 朝食	2126		過ごせる	571		新しい	372
11 満足	2005		入れる	559		問題	368
12 美味しい	1954		お世話	556		お湯	366
13 温泉	1719		バイキング	555		施設	364
14 お部屋	1553		丁寧	551		置く	361
15 スタッフ	1454		過ごす	545		お願ひ	359
16 対応	1407		素晴らしい	541		女性	357
17 行く	1394		人	540		音	354
18 立地	1334		古い	539		草津	353

操作説明 — 共起ネットワークの作成

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「文」を選んで「OK」をクリック



KH Coder の品詞体系

表 A.1 KH Coder の品詞体系

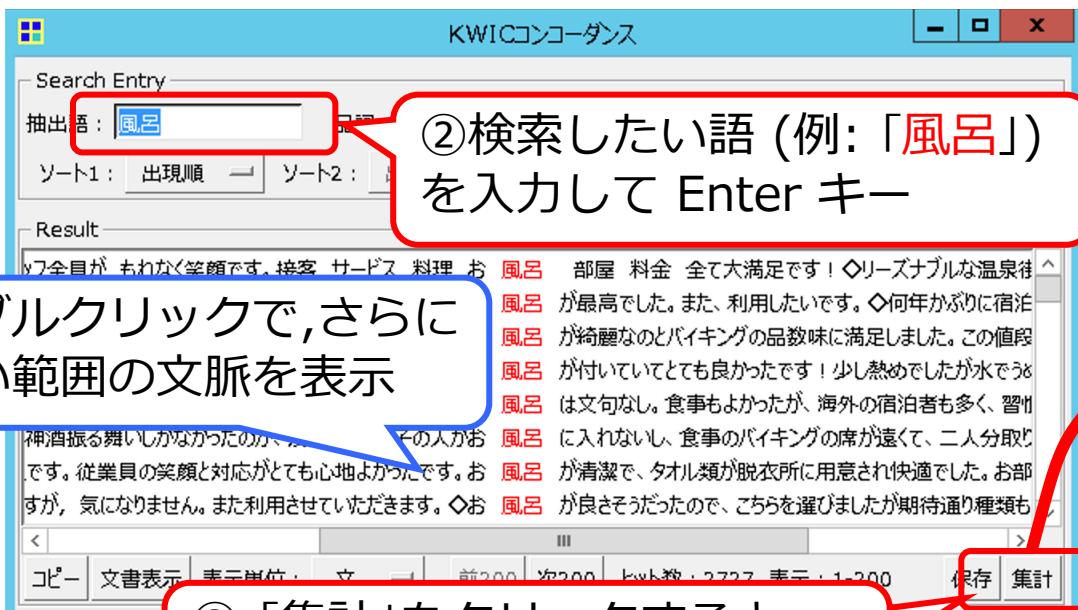
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般（漢字を含む 2 文字以上の語）
名詞 B	名詞一般（平仮名のみの語）
名詞 C	名詞一般（漢字 1 文字の語）
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立（漢字を含む語）
動詞 B	動詞-自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平仮名のみの語）
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

「KH Coder 3 リファレンス・マニュアル」
P.14 より

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

操作説明 — 語句の前後文脈を表示する

①メニューから「ツール」「抽出語」「KWICコンコーダンス」を選ぶ



②検索したい語（例：「風呂」）
を入力して Enter キー

ダブルクリックで、さらに広い範囲の文脈を表示

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)



「右1」は右側の1つ目(=直後)
に出現していた回数

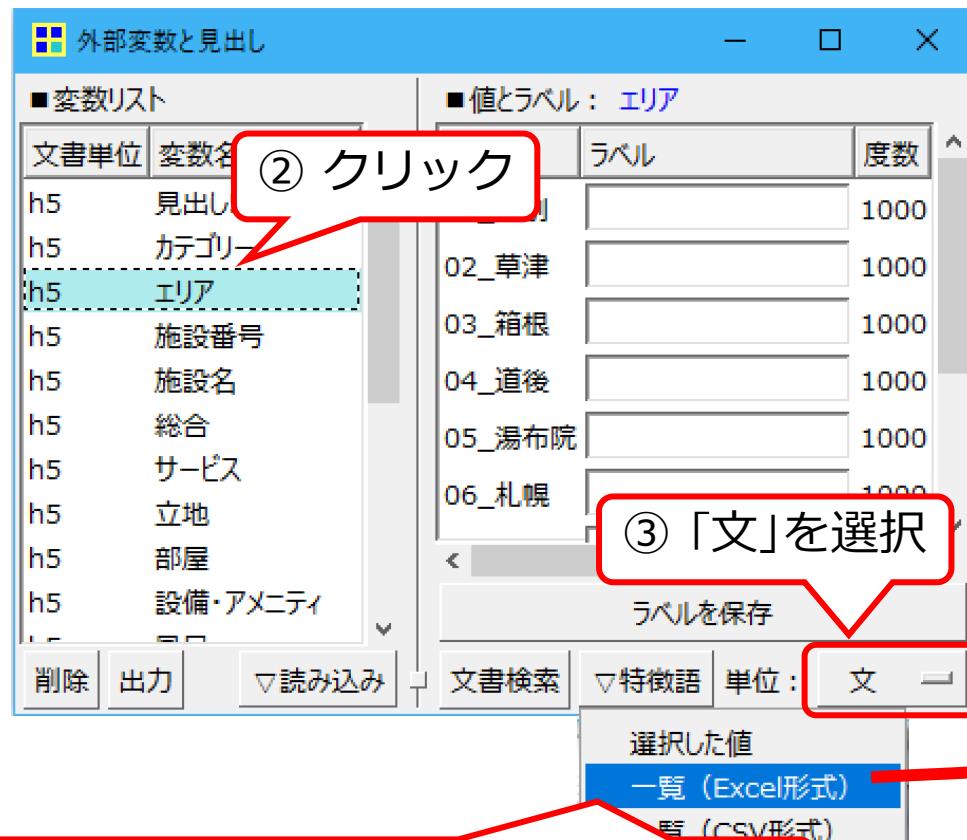
「広い」は「風呂」の
2語後に 86 回出現

④表示する語の品詞を選択（例：形容詞、形容詞B、形容動詞）

⑤ 「右合計」でソート

操作説明 — 外部変数(エリア)を利用する

①メニューから「ツール」「外部変数と見出し」「リスト」を開く

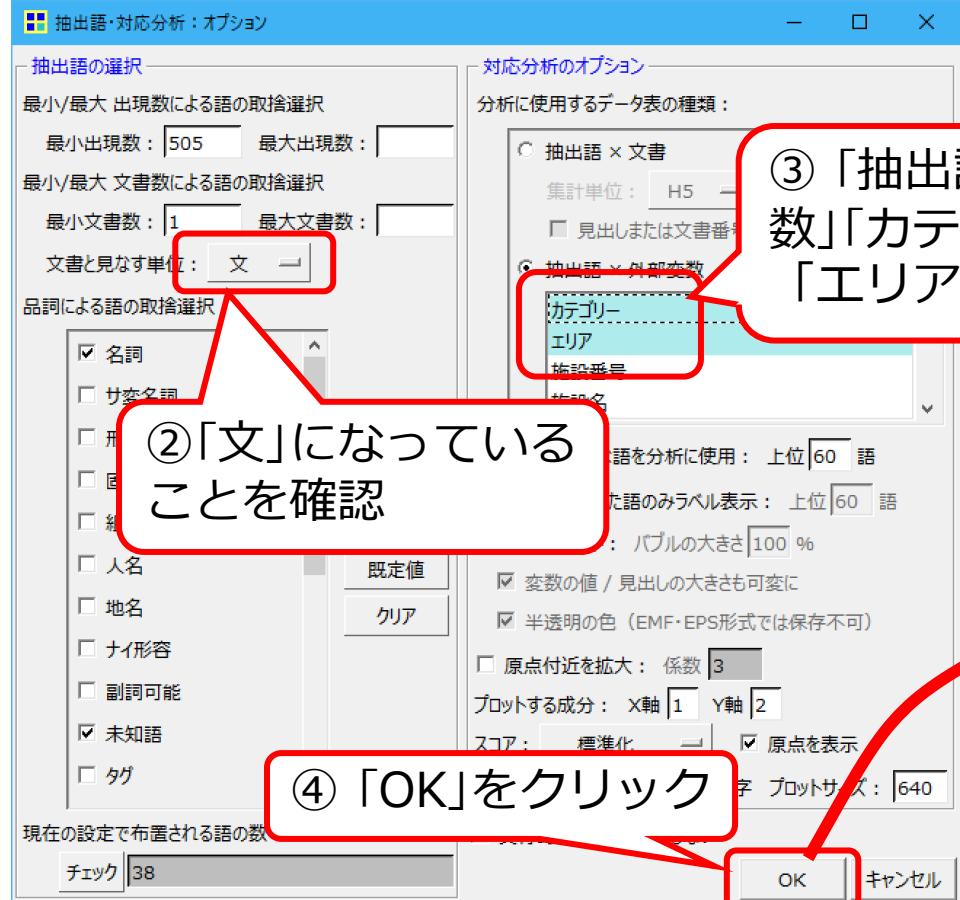


	B	C	D	E	F	G	H	I	J	K
2	01_登別		02_草津		03_箱根		04_道後			
3	食事	.059	湯畑	.081	食事	.070	温泉	.058		
4	部屋	.058	草津	.066	良い	.064	良い	.053		
5	良い	.055	温泉	.066	風呂	.056	利用	.052		
6	風呂	.053	良い	.064	美味しい	.053	ホテル	.045		
7	宿泊	.045	風呂	.064	お部屋	.045	朝食	.044		
8	温泉	.043	食事	.056	満足	.044	道後	.042		
9	美味しい	.039	宿泊	.046	スタッフ	.041	宿泊	.041		
10	満足	.035	満足	.041	温泉	.040	満足	.034		
11	残念	.035	宿	.040	宿	.038	松山	.033		
12	行く	.031	美味しい	.040	露天風呂	.036	美味しい	.033		
13	05_湯布院		06_札幌		07_名古屋		08_東京			
14	食事	.070	札幌	.058	名古屋	.060	利用	.062		
15	美味しい	.067	思う	.056	利用	.056	部屋	.057		
16	宿	.064	部屋	.055	朝食	.053	ホテル	.056		
17	風呂	.063	ホテル	.053	ホテル	.052	駅	.046		
18	思う	.060	利用	.051	部屋	.048	便利	.044		
19	満足	.048	朝食	.051	駅	.038	朝食	.038		
20	料理	.045	宿泊	.046	便利	.031	立地	.035		
21	スタッフ	.043	立地	.039	立地	.029	近い	.034		
22	露天風呂	.042	広い	.031	フロント	.028	フロント	.032		
23	温泉	.042	近い	.031	近い	.027	近く	.026		
24	09_大阪		10_福岡							
25	利用	.068	利用	.059						
26	ホテル	.060	部屋	.056						
27	部屋	.054	ホテル	.049						
28	思う	.051	博多	.046						
29	大阪	.047	朝食							
30	朝食	.043	立地							
31	便利	.040								
32	立地	.040								
33	駅	.037								
34	近い	.031								

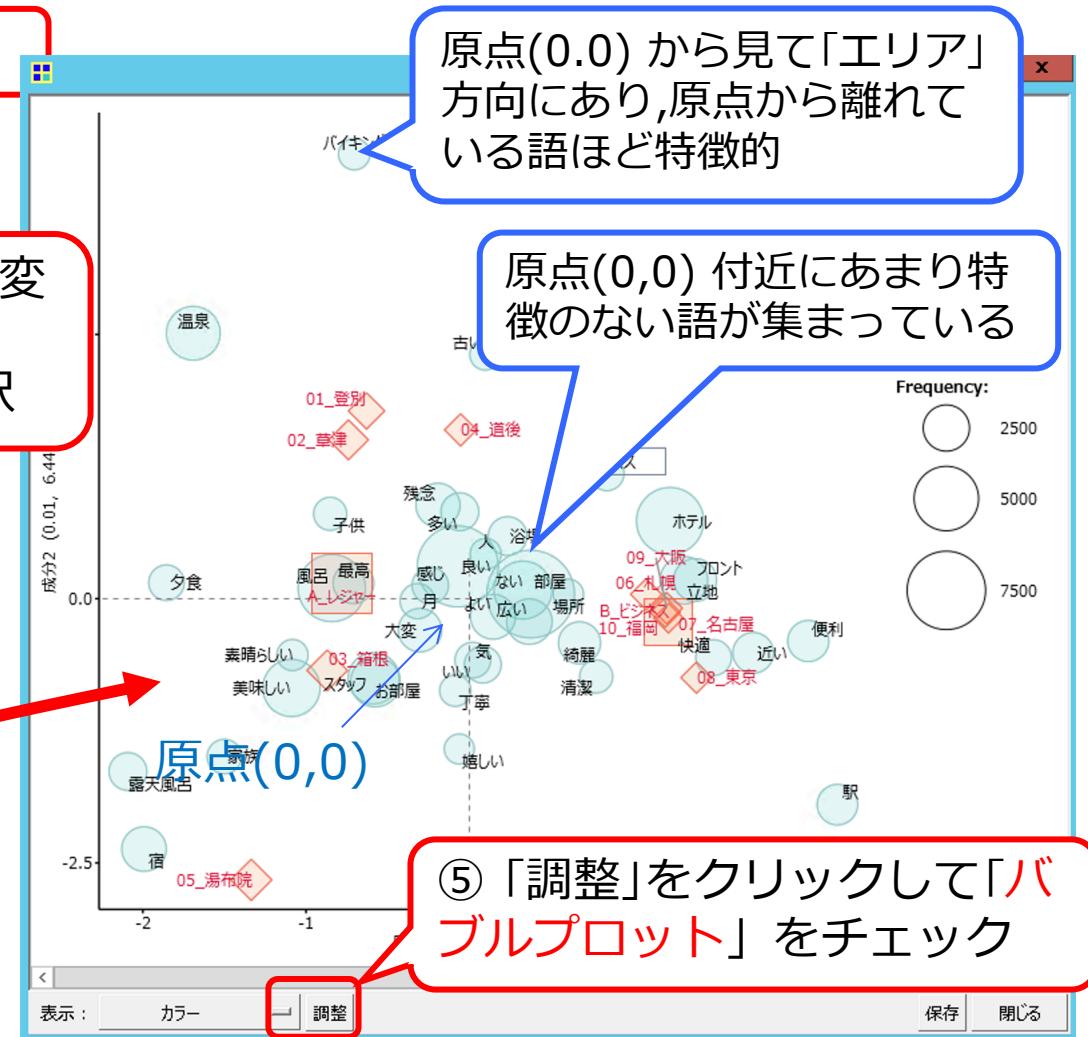
各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

操作説明 — 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



- ②「文」になっていることを確認
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



操作説明 — コーディングルール

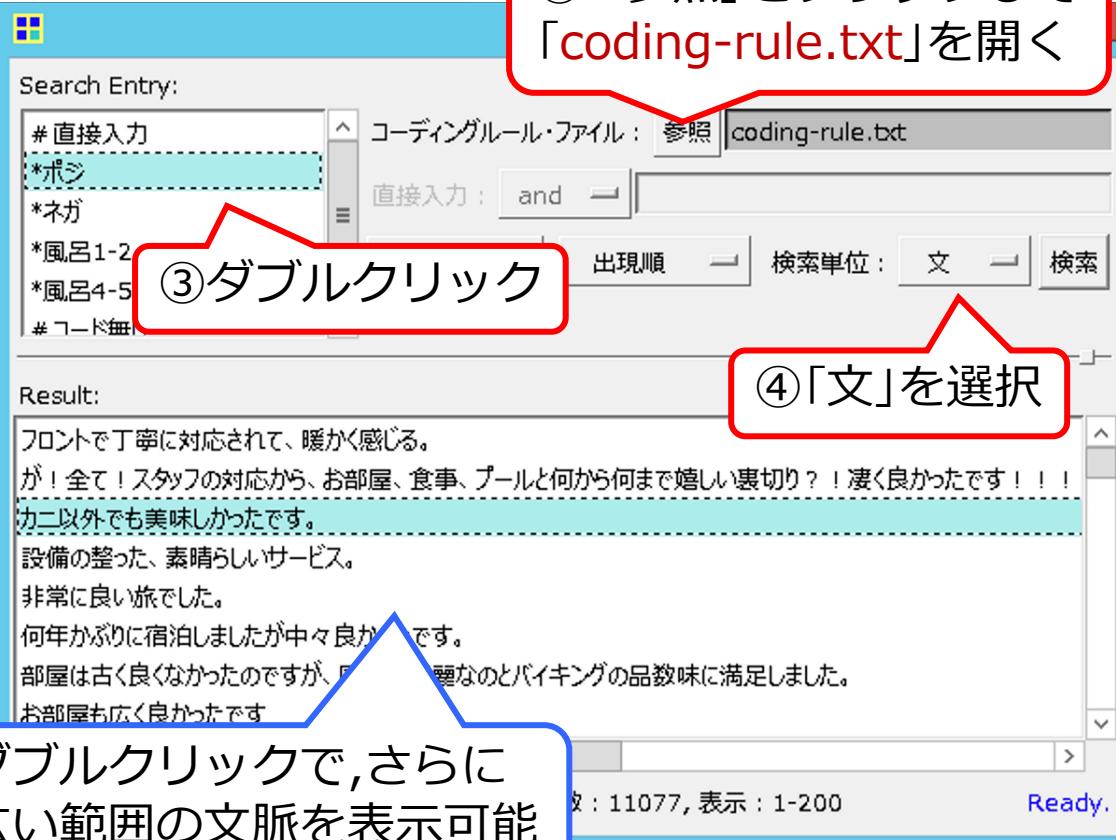
①メニューから「ツール」「文書」「文書検索」を選ぶ

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

ダブルクリックで、さらに
広い範囲の文脈を表示可能



※ コーディングルール: 語ではなくコンセプトを
数えるための方法

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

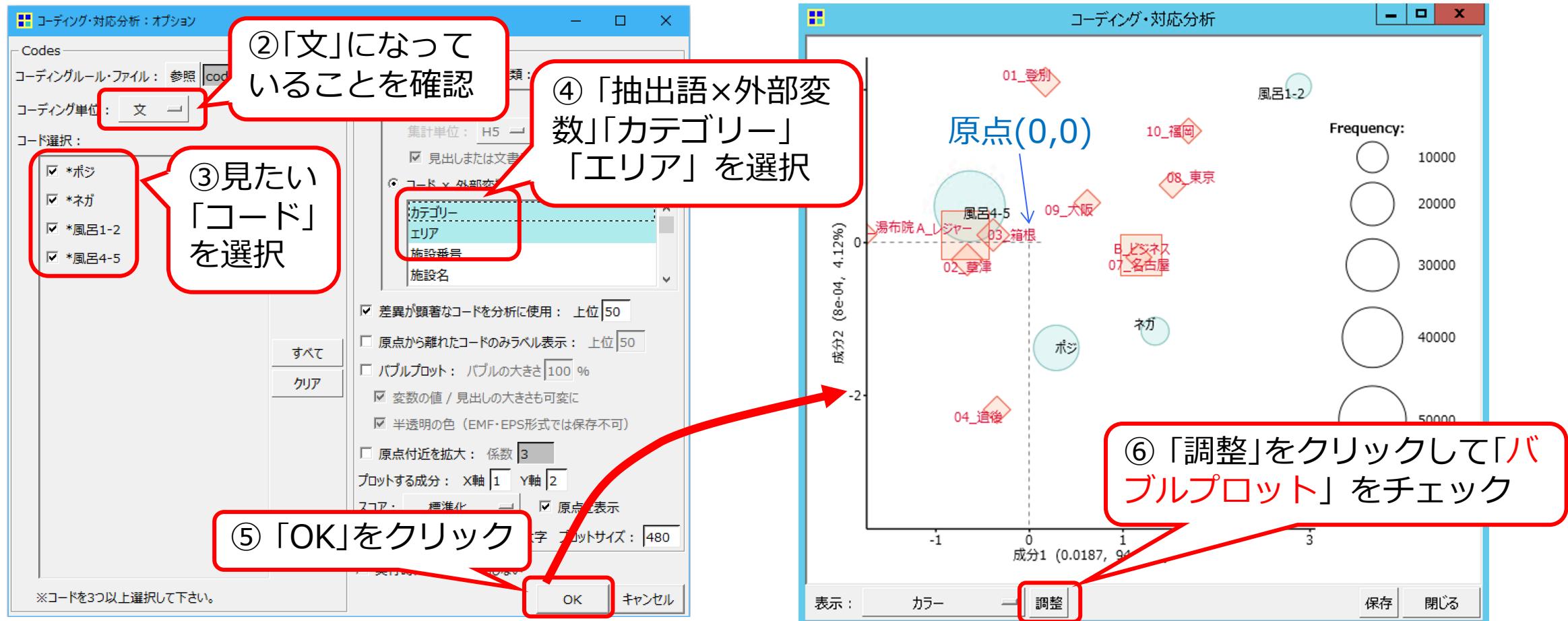
*風呂4-5

<>風呂-->4 | <>風呂-->5

外部変数

操作説明 — 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



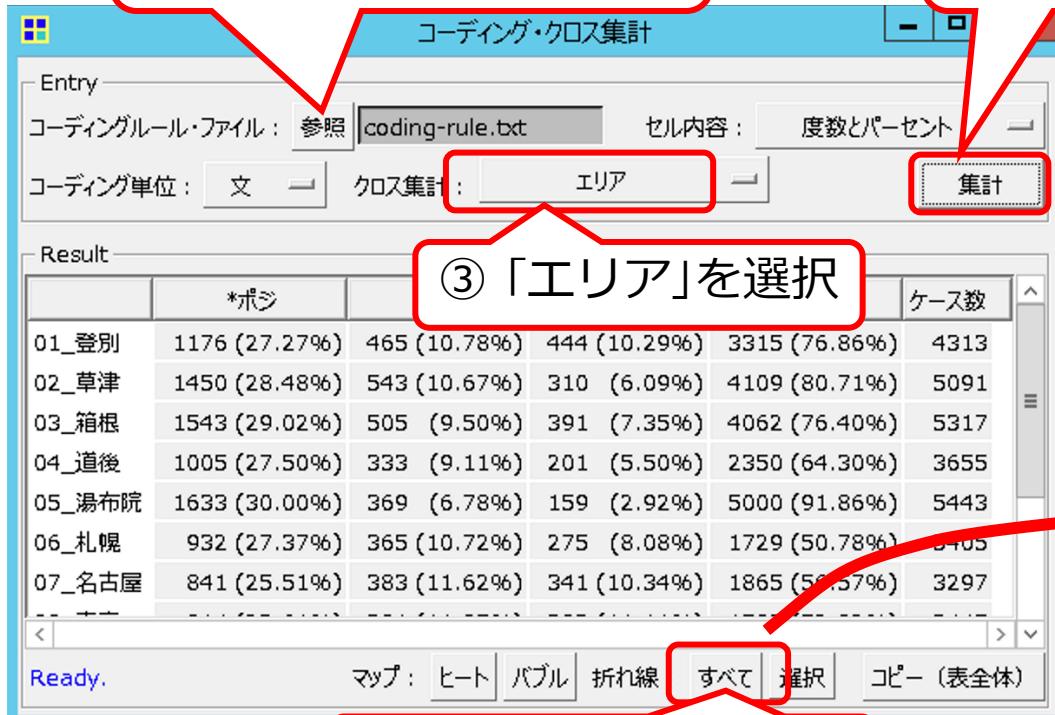
操作説明 — クロス集計

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

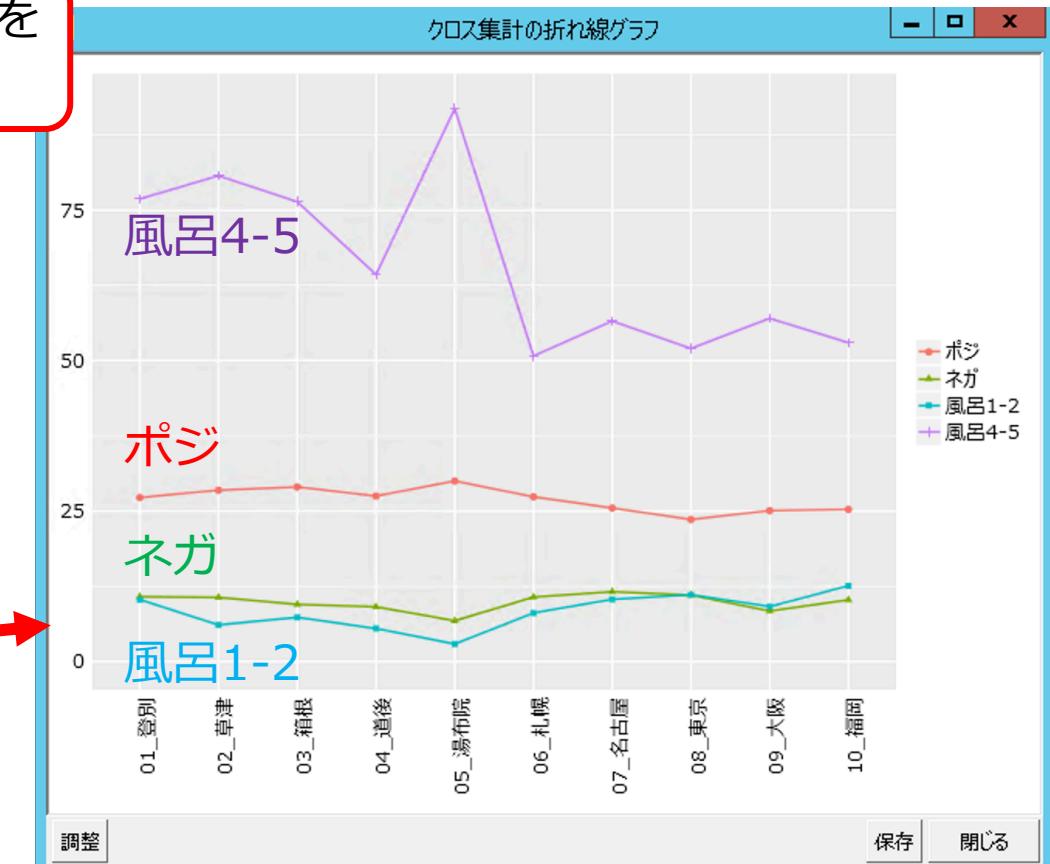
②「参照」をクリックして
「coding-rule.txt」を開く

④「集計」を
クリック

③「エリア」を選択



⑤「すべて」をクリック



練習 —数値評価と口コミの傾向比較

- ・コーディングルール「coding-rule.txt」中の「風呂1-2」「風呂4-5」を参考に、「総合1-2」「総合4-5」のルールを定義したコーディングルール「coding-rule_new.txt」を作成してください
- ・前ページで紹介したクロス集計を用いて,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,違いについて考察してください

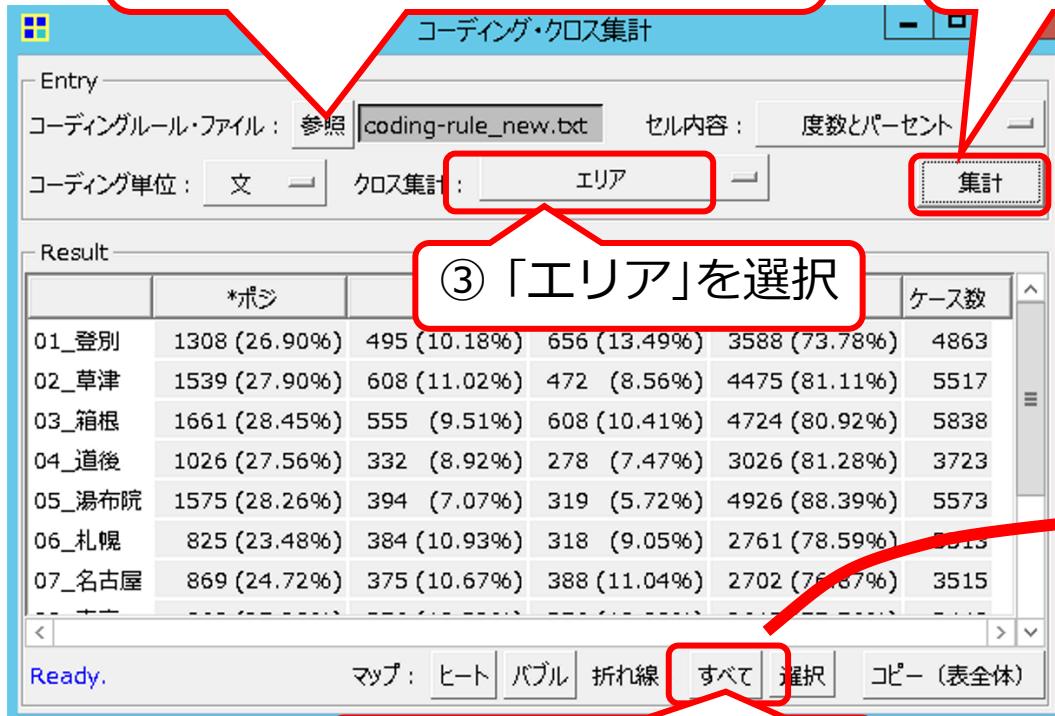
練習 — コーディングルール

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

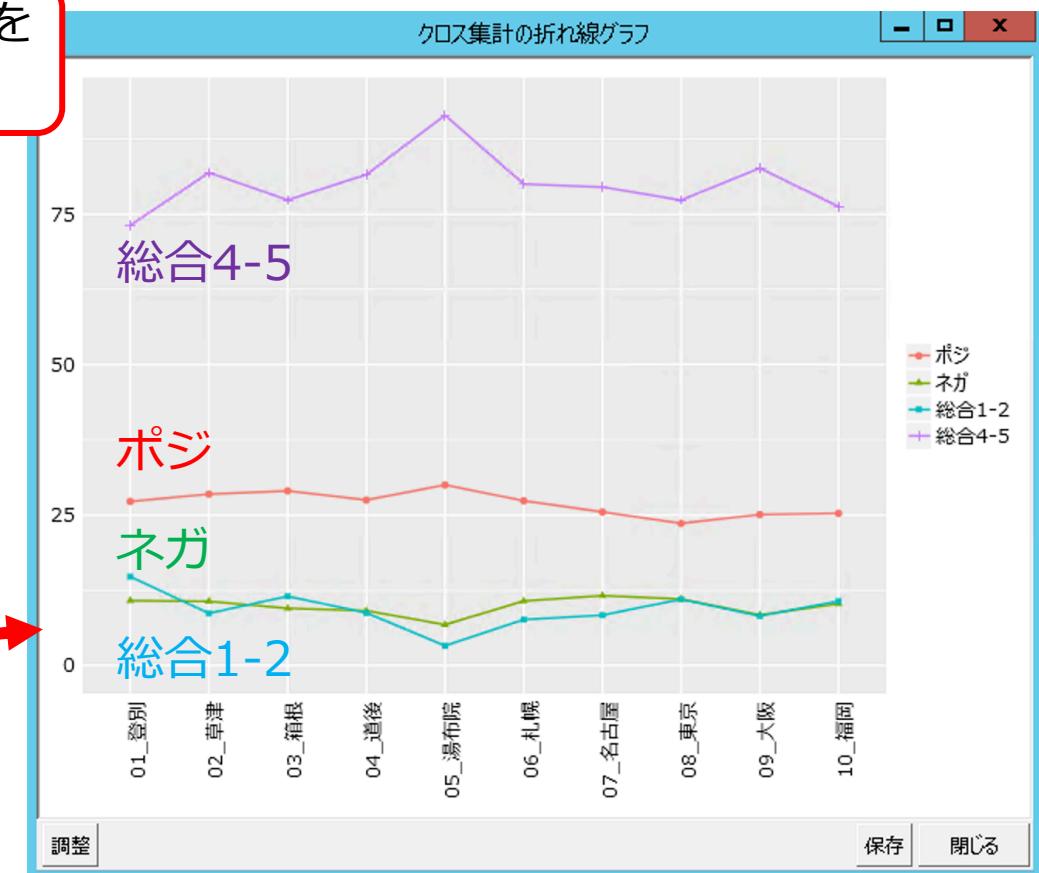
②「参照」をクリックして
「coding-rule_new.txt」を開く

④「集計」を
クリック

③「エリア」を選択



⑤「すべて」をクリック



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.