

テキストマイニングの実習

— 2日目 —

2018/7/11

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-201907>



前回: データの理解

	データセットの特徴	注意すべきバイアス等
年代別・性別	<ul style="list-style-type: none"> 約60%が年代や性別を表明していない 年代別では、目的によらず40~50代が多い 全体的に男性の投稿者が多い（女性の倍以上） レジャーに比べてビジネス方が男女差が大きい レジャーの中でも男女差が大きいのは道後 	<ul style="list-style-type: none"> 無回答(na)層がある年代や性別に偏っている可能性 コメントの観点が年代や性別によって偏っている可能性
目的別	<ul style="list-style-type: none"> レジャーは家族が多い、ビジネスは一人が多い（出張は単独） レジャーの中でも、道後は一人が多い（道後はもはや仕事で行く場所） 	<ul style="list-style-type: none"> コメントの観点が性別によって偏っている可能性 コメントの観点がカテゴリと一致していない可能性
数値評価 (総合)	<ul style="list-style-type: none"> 目的によらず評価は高め レジャーがビジネスより評価が高め レジャーの中で評価が高いのは湯布院、低いのは登別 ビジネスの中で評価が高いのは札幌と大阪、低いのは東京都と福岡だが僅差 	<ul style="list-style-type: none"> 好評価しか投稿しない→コメントが好評価に偏っている可能性 目的によって投稿の動機が異なっている可能性
数値評価 (項目ごと)	<ul style="list-style-type: none"> レジャーの評価は、風呂や食事 > 設備や部屋 ビジネスの評価は、立地 > その他 レジャーの中で湯布院は軒並み高評価 レジャーもビジネスも立地は高評価 	<ul style="list-style-type: none"> 目的によって評価の観点が異なっている可能性
全体	<ul style="list-style-type: none"> あくまでも、投稿者の傾向であって、旅行者の実態ではないことにも注意 	

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)

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：101 / 2695

[KH Coderを用いた研究事例のリスト 2695件](#)

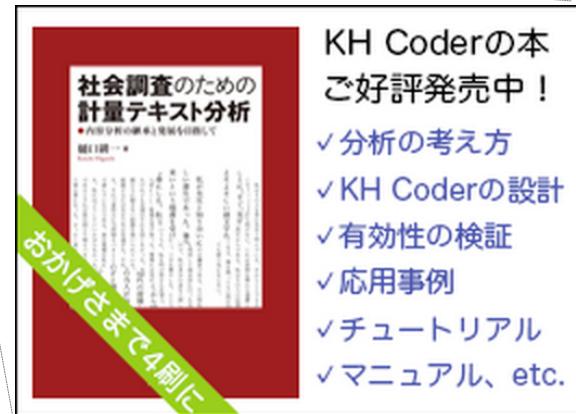
※ 2019/6/16 現在 (961件→1206件→昨年1646件→昨年2042件)

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 logo 'KH Coder' with a blue, wavy background. Below it is a navigation bar with Japanese and English language options. The main content area has a large 'Index' section with a message about an upcoming seminar. Below this are sections for '概要' (Overview), '機能紹介 (スクリーンショット)' (Function Introduction (Screenshot)), and 'KH Coderの入手' (How to Get KH Coder). A sidebar on the right contains links to various resources, including a tutorial for 'Anne of Green Gables'.

参考書



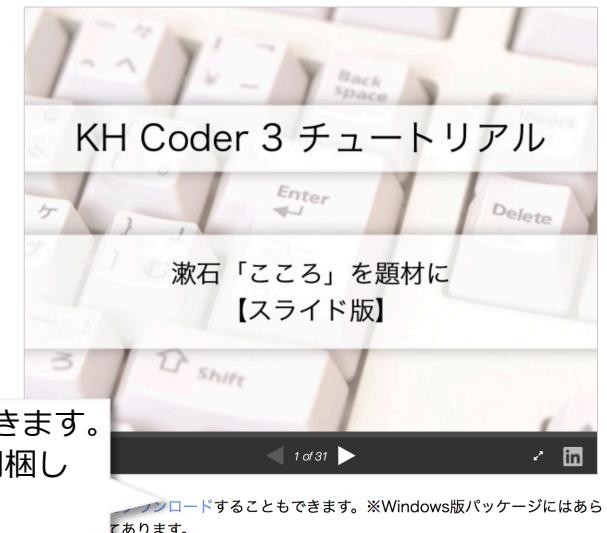
KH Coderの本 ご好評発売中！

- ✓ 分析の考え方
- ✓ KH Coderの設計
- ✓ 有効性の検証
- ✓ 応用事例
- ✓ チュートリアル
- ✓ マニュアル、etc.

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

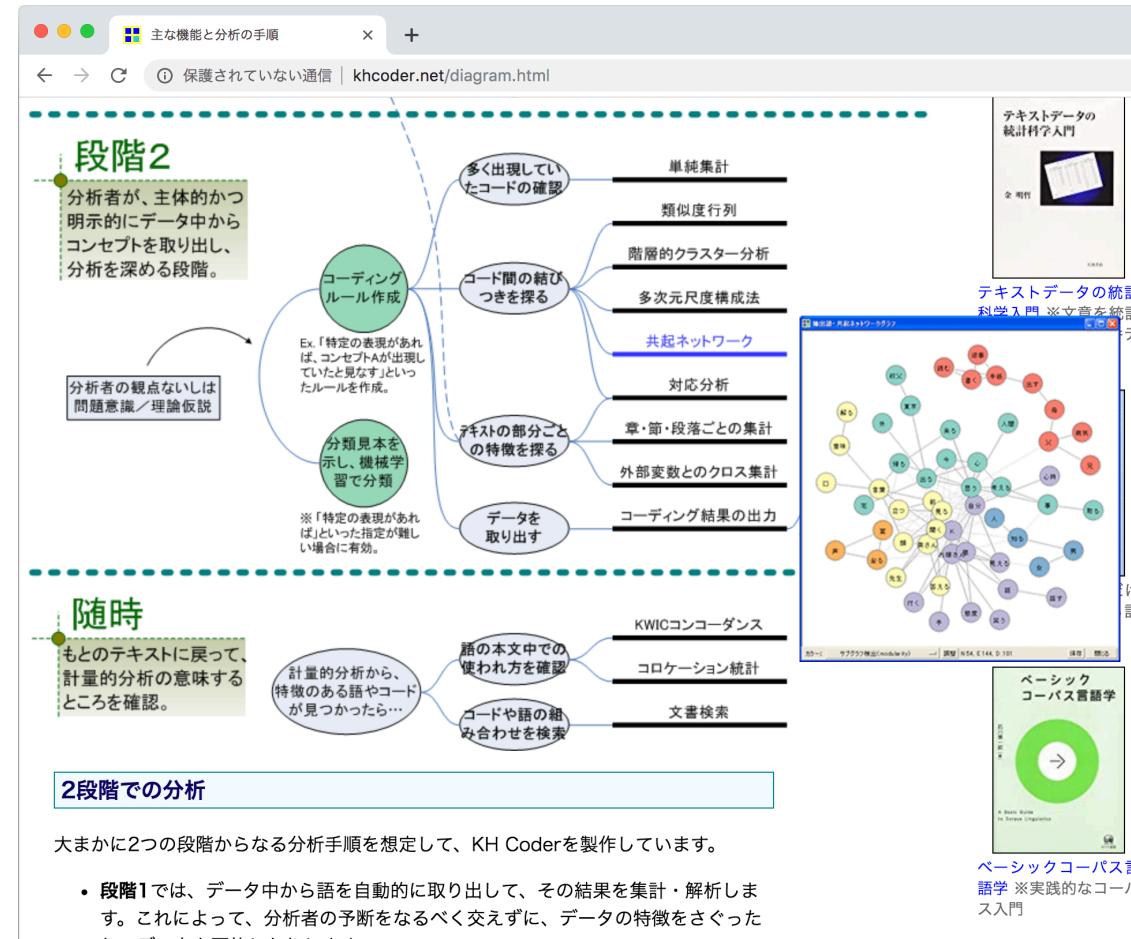
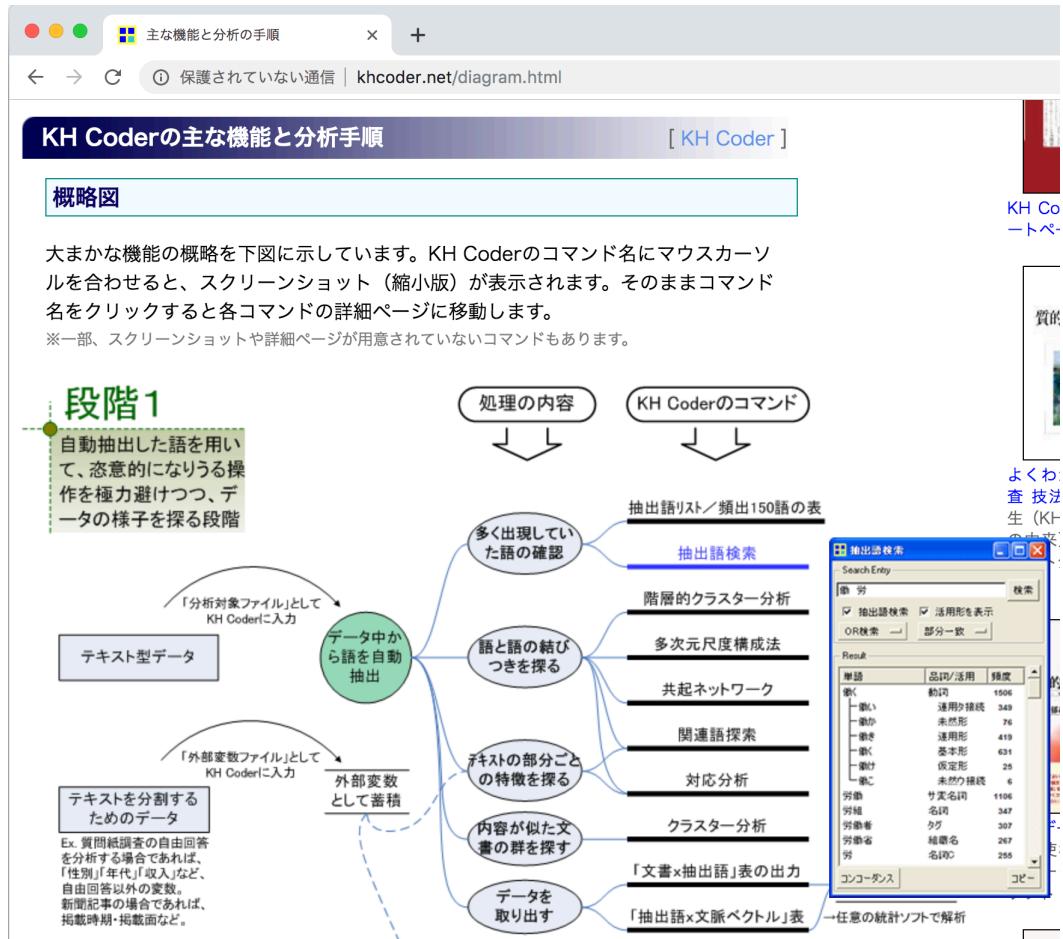


チュートリアル用データ

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>



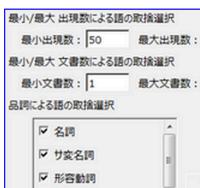
KH Coder –スクリーンショット

階層的クラスター分析

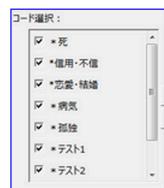
抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけでなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



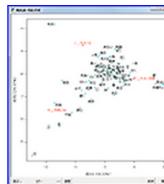
抽出語は出現数や品詞で選択



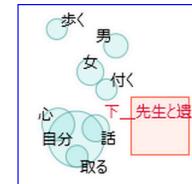
コードはチェックボックスで直接選択

対応分析

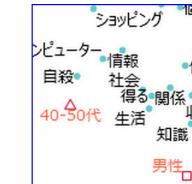
同じく抽出語またはコードを用いての、対応分析です。



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

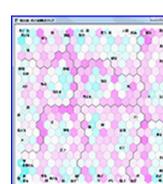
抽出語またはコードを用いての、自己組織化マップです。



クラスター色分け



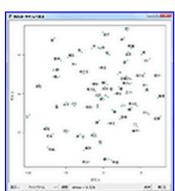
頻度のプロット



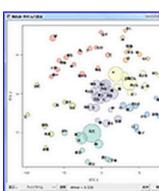
U-Matrix

多次元尺度構成法 (MDS)

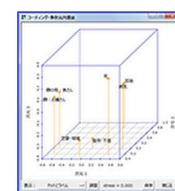
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



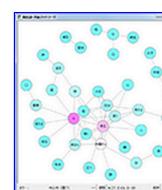
New! クラスタリングと色分け



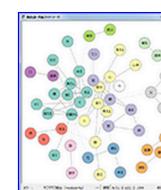
3次元の解

共起ネットワーク

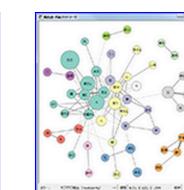
抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



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）



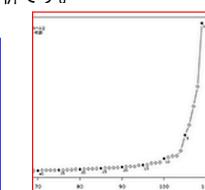
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

文書のクラスター分析

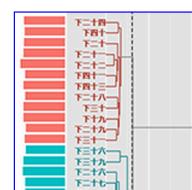
文書の分類を行うクラスター分析です。



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(MDS)	<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 (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	<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
文書のクラスター分析	<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

出現パターン – 「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

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	バイキ	家族	場所	トイレ	子供	ベット	コンビ	良い
1	1	1	0	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	1	0	0	0	0	0	1	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

KH Coder で使われる距離尺度

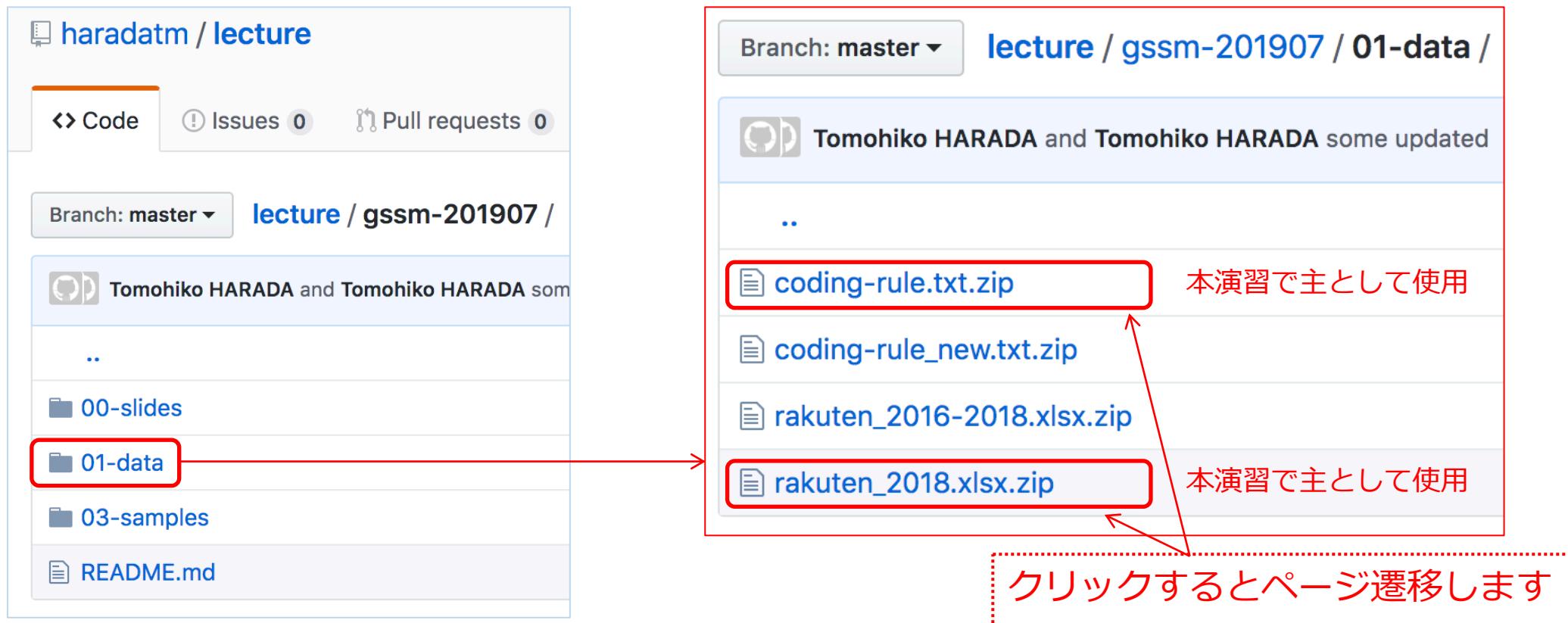
- KH Coder では Jaccard 距離を多用
 - 語Aと語Bのどちらも出現していない文書(0-0対)が沢山あっても語Aと語Bが類似しているとは見なさない → **スパースなデータ分析向き**

Jaccard 距離	コサイン距離	ユークリッド距離								
<ul style="list-style-type: none">• 1つ文書に含まれる語が少なく、各語が一部の文書中にしか含まれていないスパースデータ向き• 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	<ul style="list-style-type: none">• 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(各文書中での語の出現回数の大小が重要な場合)• 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	<ul style="list-style-type: none">• 増減傾向が似ているかどうかだけを見る場合向き• サイズの差までも見る場合向き								
<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> $J\text{S} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$	1	0	1	n_{11}	n_{10}	0	n_{01}	n_{00}	$\text{cos} s(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$	$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$
1	0									
1	n_{11}	n_{10}								
0	n_{01}	n_{00}								

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

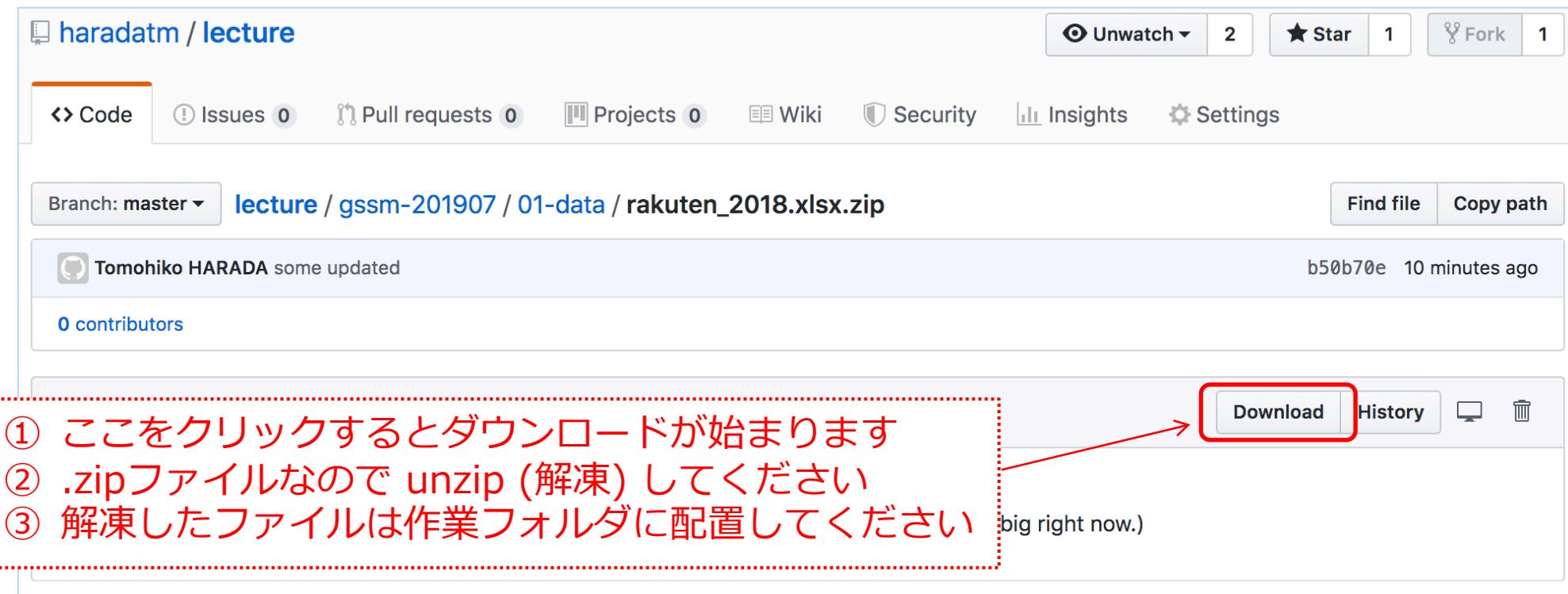
データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-201907>



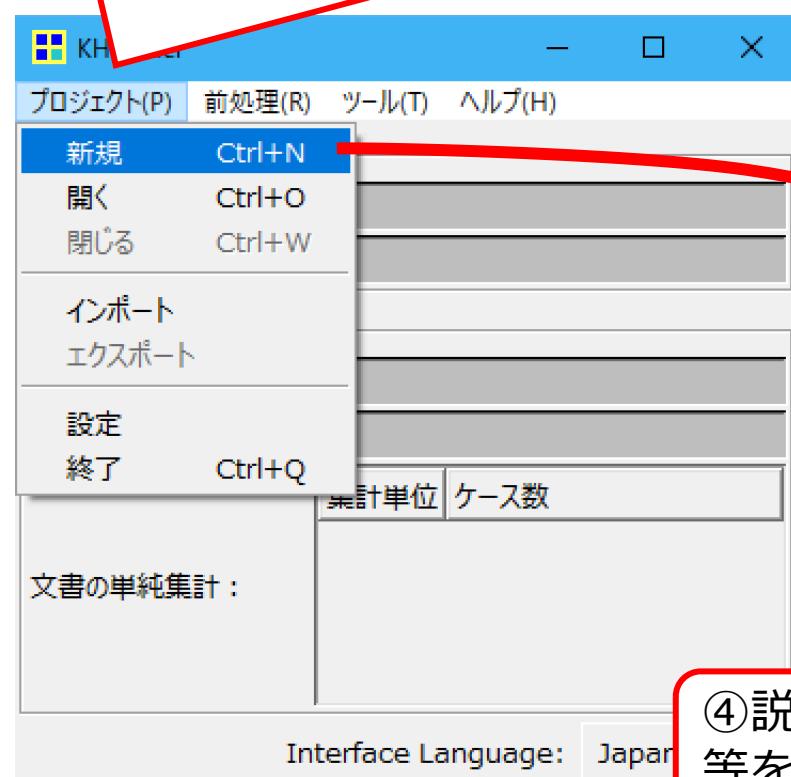
ダウンロード方法

- 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

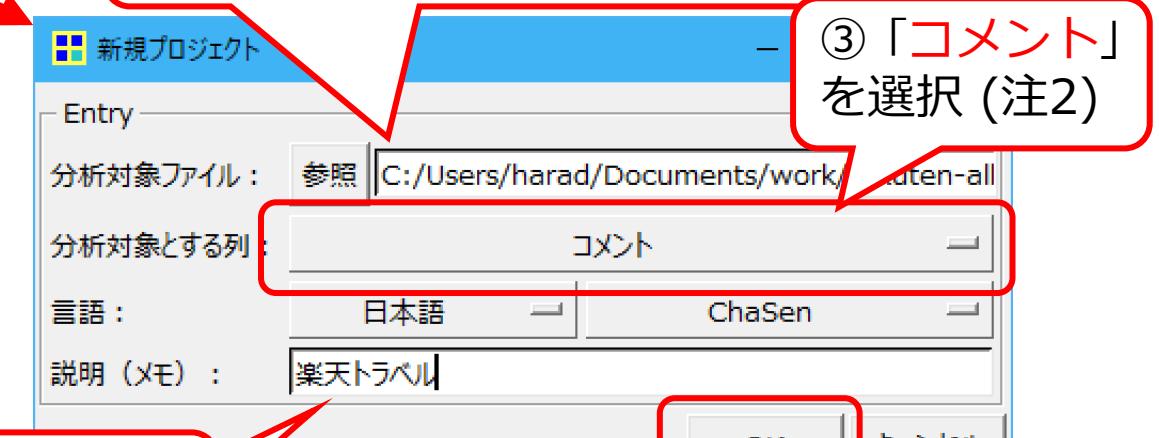
①メニューから「プロジェクト」「新規」を選択 (注1)



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_2018.xlsx」を開く

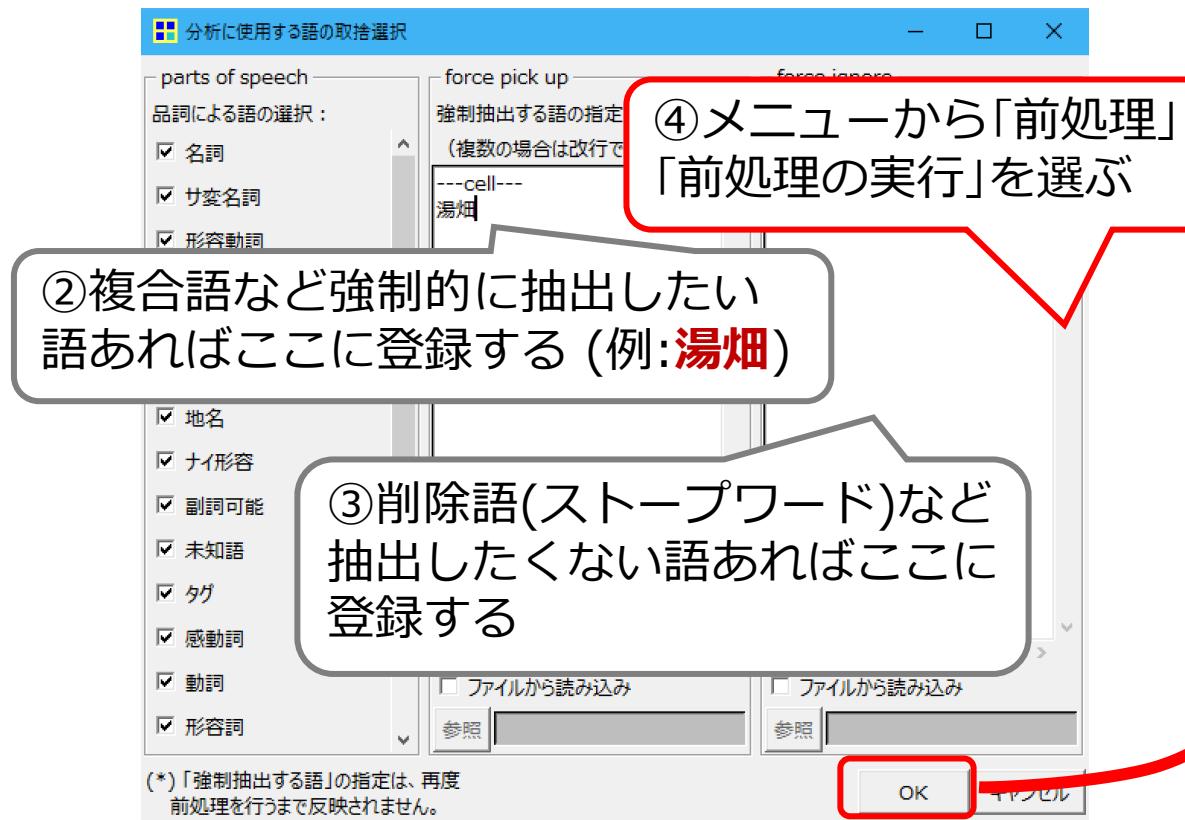


④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

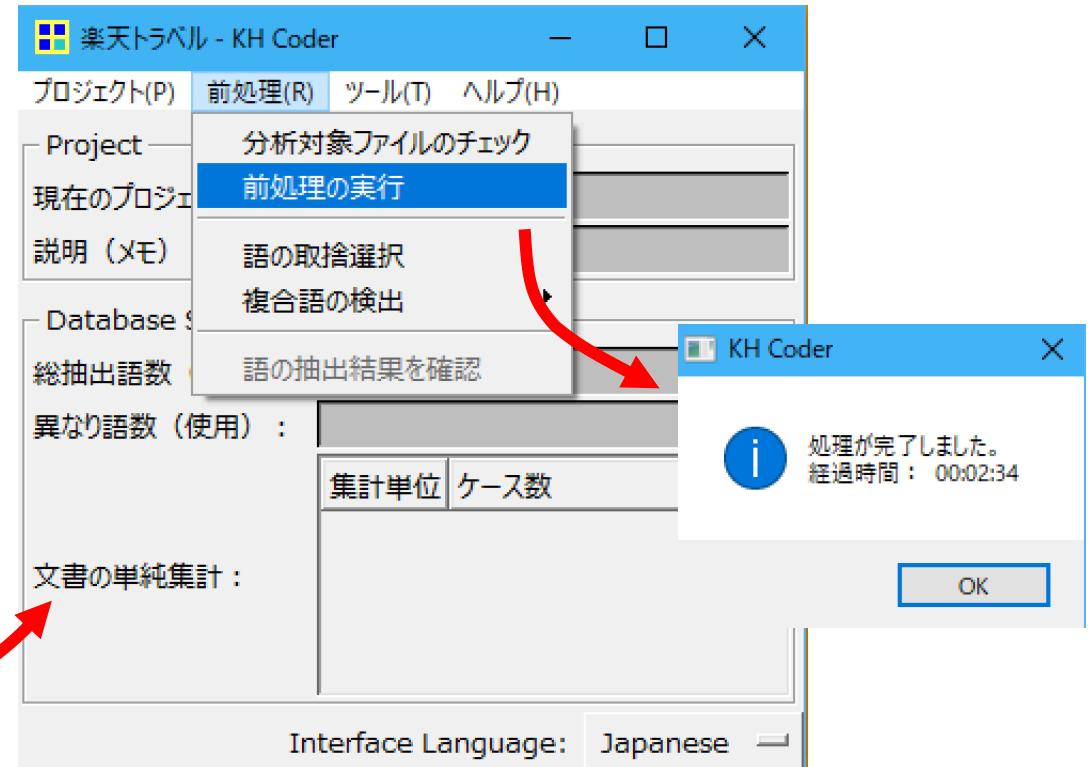
操作説明 – 前処理 (形態素解析)

①メニューから「前処理」「語の取捨選択」を選ぶ



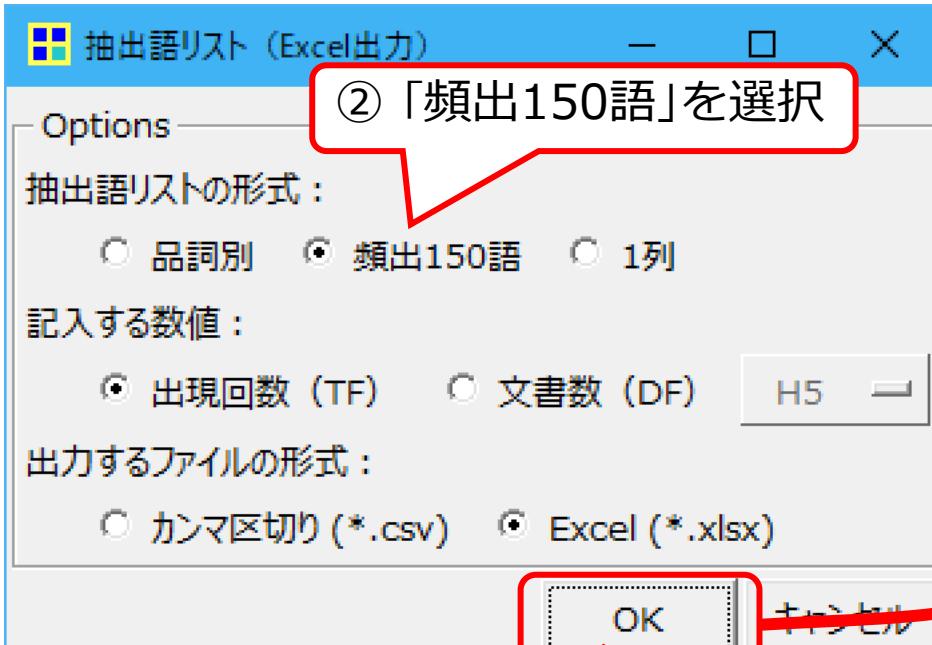
注1: EXCELファイルを読み込んで分析する場合、あらかじめ「---cell---」が入力されています

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」を選択



- ③「OK」をクリック

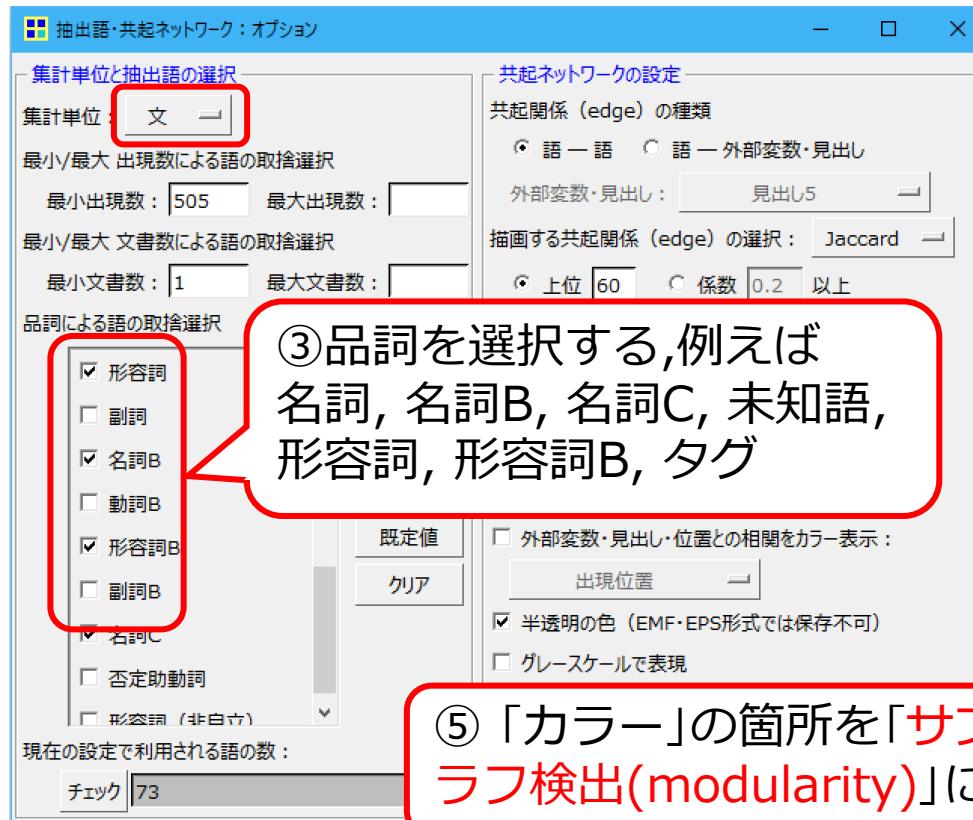
	A	B	C	D	E	F	G	H
1	抽出語	出現回数		抽出語	出現回数		抽出語	出現回数
2	部屋	5063	前	716			湯畑	413
3	思う	4284	月	674			歩く	413
4	良い	4101	清潔	655			気持ちよい	412
5	利用	3535	バイキング	636			シャワー	409
6	ホテル	2996	初めて	633			問題	406
7	宿泊	2880	旅行	612			施設	400
8	風呂	2781	使う	600			機会	399
9	食事	2466	家族	599			従業	399
10	朝食	2214	過ごせる	597			女性	395
11	満足	2123	人	584			掃除	393
12	美味しい	1853	夜	574			お願い	384
13	温泉	1742	素晴らしい	572			接客	379
14	対応	1607	古い	568			旅館	377
15	お部屋	1514	場所	568			タオル	376
16	スタッフ	1469	トイレ	567			静か	376
17	行く	1398	入れる	559			新しい	373
18	広い	1347	子供	548			置く	373
19	立地	1343	過ごす	543			清掃	367

操作説明 —共起ネットワークの作成

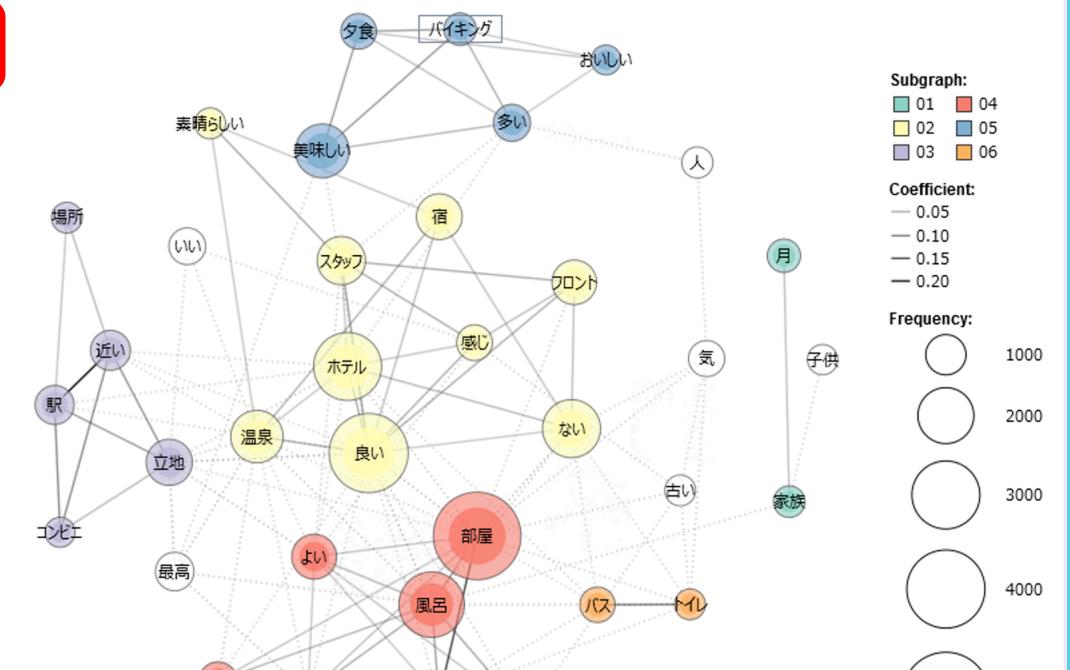
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

抽出語・共起ネットワーク

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, タグ



④「調整」をクリックして「上位」に120を
入力し、「強い共起関係ほど…」をチェック

サブグラフ検出 (modularity) 調整 N 38, E 120, D .171 HTML表示 保存 閉じる

⑤「カラー」の箇所を「サブグラフ
ラフ検出(modularity)」に変更

KH Coder の品詞体系

表 A.1 KH Coder の品詞体系

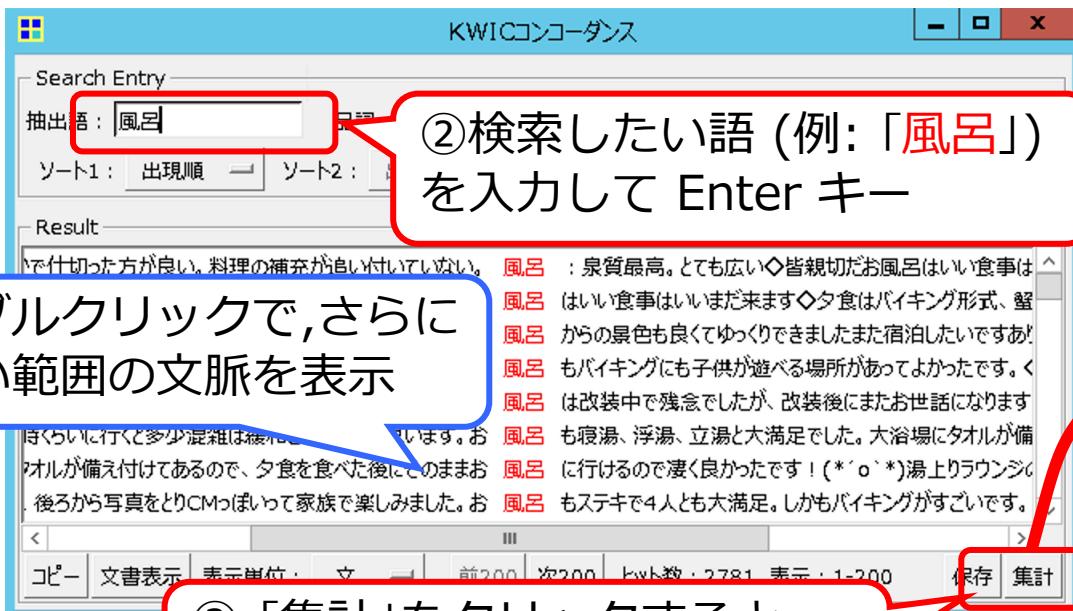
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般 (漢字を含む 2 文字以上の語)
名詞 B	名詞一般 (平仮名のみの語)
名詞 C	名詞一般 (漢字 1 文字の語)
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立 (漢字を含む語)
動詞 B	動詞-自立 (平仮名のみの語)
形容詞	形容詞 (漢字を含む語)
形容詞 B	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.11 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

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	スコア
1	良い	形容詞	207	67	140	35	14	8	9	1	5	50	38	21	26	71.78
2	広い	形容詞	176	42	134	7	12	10	10	3	0	91	18	21	4	73.28
3	よい	形容詞B	71	20	51	11	4	3	2	0	19	12	9	10	23.95	
4	狭い	形容詞	53	11	42	11	10	10	10	1	1	8	4	4	21.61	
5	ない	形容詞B	54	11	43	11	10	10	10	1	1	8	9	10	16.40	
6	気持ちよい	形容詞	38	11	27	11	10	10	10	1	6	6	3	13.85		
7	熱い	形容詞	27	11	16	11	10	10	10	1	5	6	3	10.26		

「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91回出現

③「集計」をクリックするとコロケーション統計(右)を開く

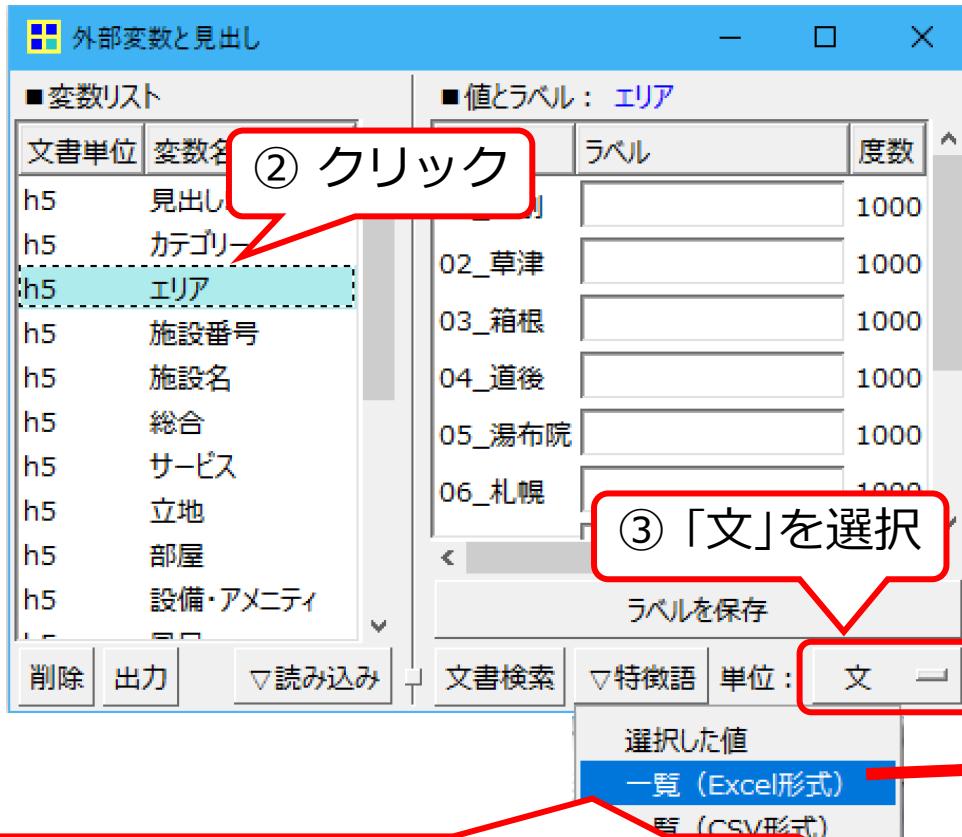
④表示する語の品詞を選択(例: 形容詞, 形容詞B)

⑤「右合計」でソート

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)

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



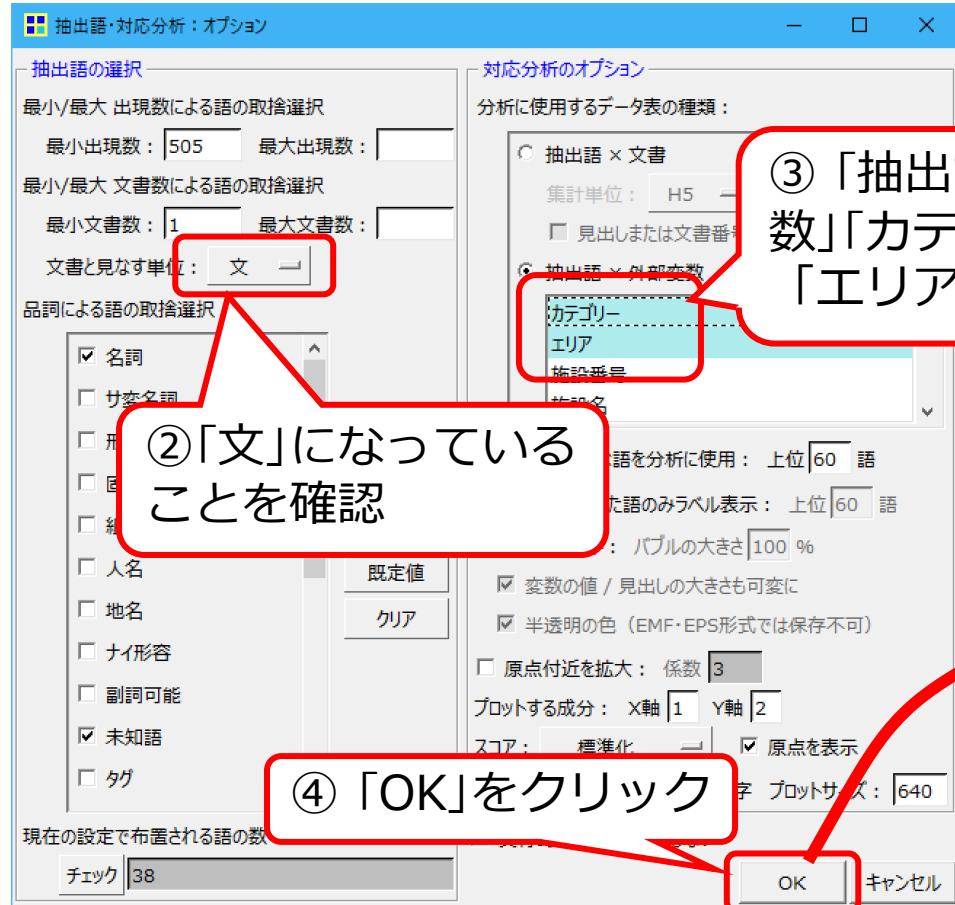
注: Jaccard係数は共起尺度のひとつで、
共通要素の数を少なくとも一方にある数で割ったもの

	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	食事	.062	湯畑	.071	食事	.072	温泉	.056			
4	風呂	.056	温泉	.065	思う	.064	部屋	.053			
5	良い	.056	風呂	.060	良い	.059	良い	.047			
6	思う	.056	良い	.059	風呂	.054	道後	.047			
7	宿泊	.044	草津	.058	美味しい	.051	利用	.046			
8	温泉	.043	食事	.056	満足	.044	朝食	.045			
9	満足	.038	満足	.047	温泉	.040	ホテル	.042			
10	美味しい	.036	宿	.044	露天風呂	.038	宿泊	.041			
11	スタッフ	.030	美味しい	.040	お部屋	.038	松山	.034			
12	行く	.030	お部屋	.033	スタッフ	.037	立地	.034			
13	05_湯布院		06_札幌		07_名古屋		08_東京				
14	宿	.069	部屋	.059	名古屋	.063	利用	.062			
15	食事	.068	ホテル	.055	ホテル	.057	ホテル	.054			
16	美味しい	.058	利用	.055	部屋	.055	部屋	.052			
17	風呂	.057	朝食	.054	朝食	.053	駅	.048			
18	宿泊	.045	札幌	.054	利用	.052	宿泊	.039			
19	満足	.043	思う	.047	思う	.046	便利	.038			
20	料理	.043	宿泊	.040	良い	.045	立地	.034			
21	露天風呂	.040	立地	.036	駅	.034	朝食	.034			
22	温泉	.040	フロント	.033	フロント	.034	フロント	.033			
23	お部屋	.039	便利	.032	便利	.031	近い	.031			
24	09_大阪		10_福岡								
25	利用	.059	ホテル	.064							
26	部屋	.055	部屋	.057							
27	ホテル	.053	博多	.055							
28	思う	.049	利用	.052							
29	朝食	.041	朝食	.041							
30	大阪	.039	立地	.036							
31	宿泊	.038									
32	フロント	.037									
33	便利	.036									
34	立地	.036									

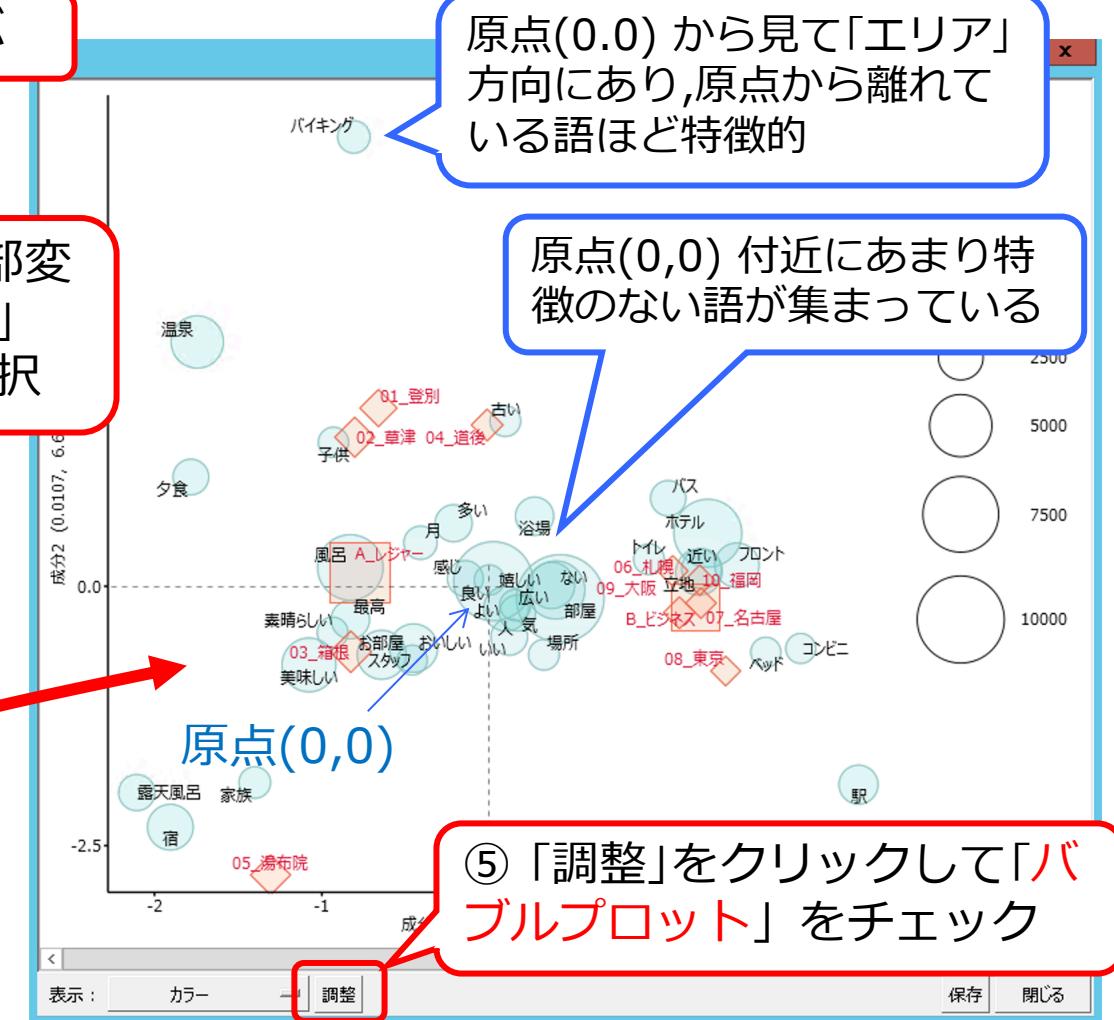
各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

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

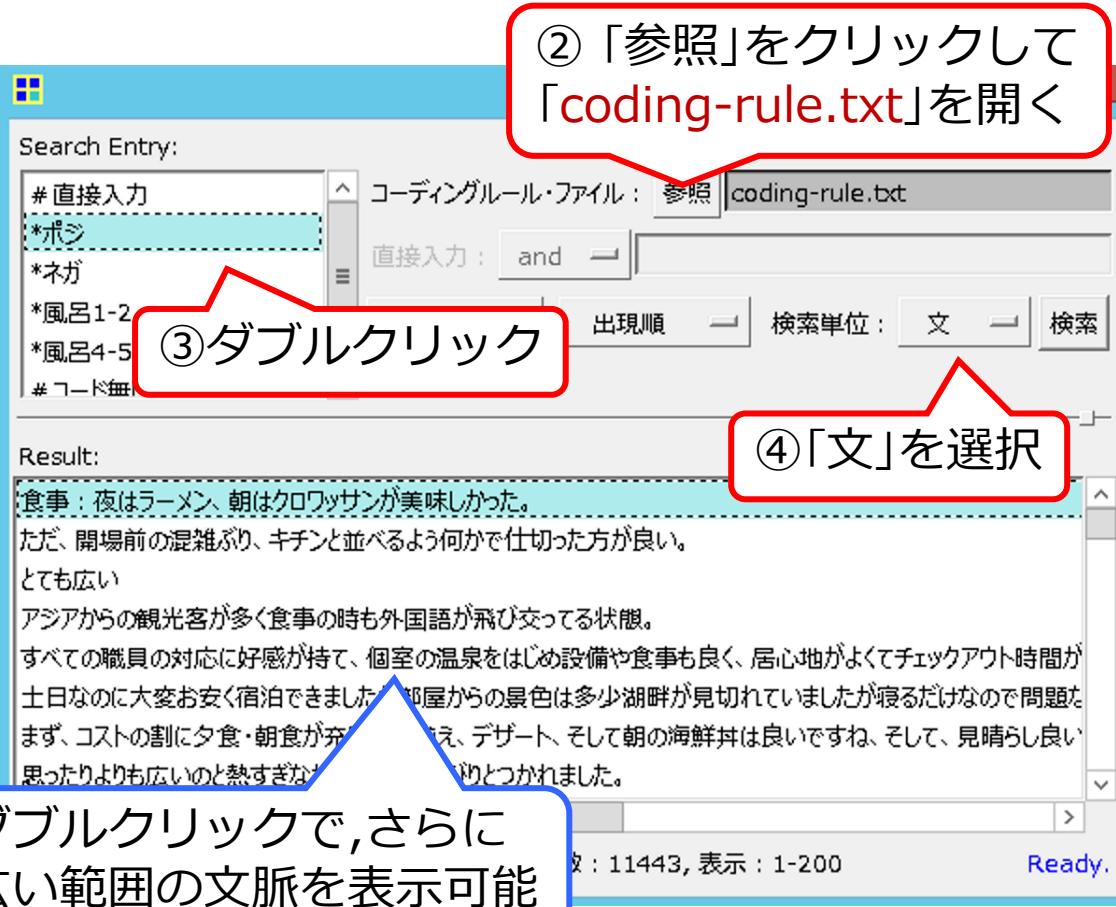


原点(0,0)から見て「エリア」方向にあり、原点から離れている語ほど特徴的



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



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

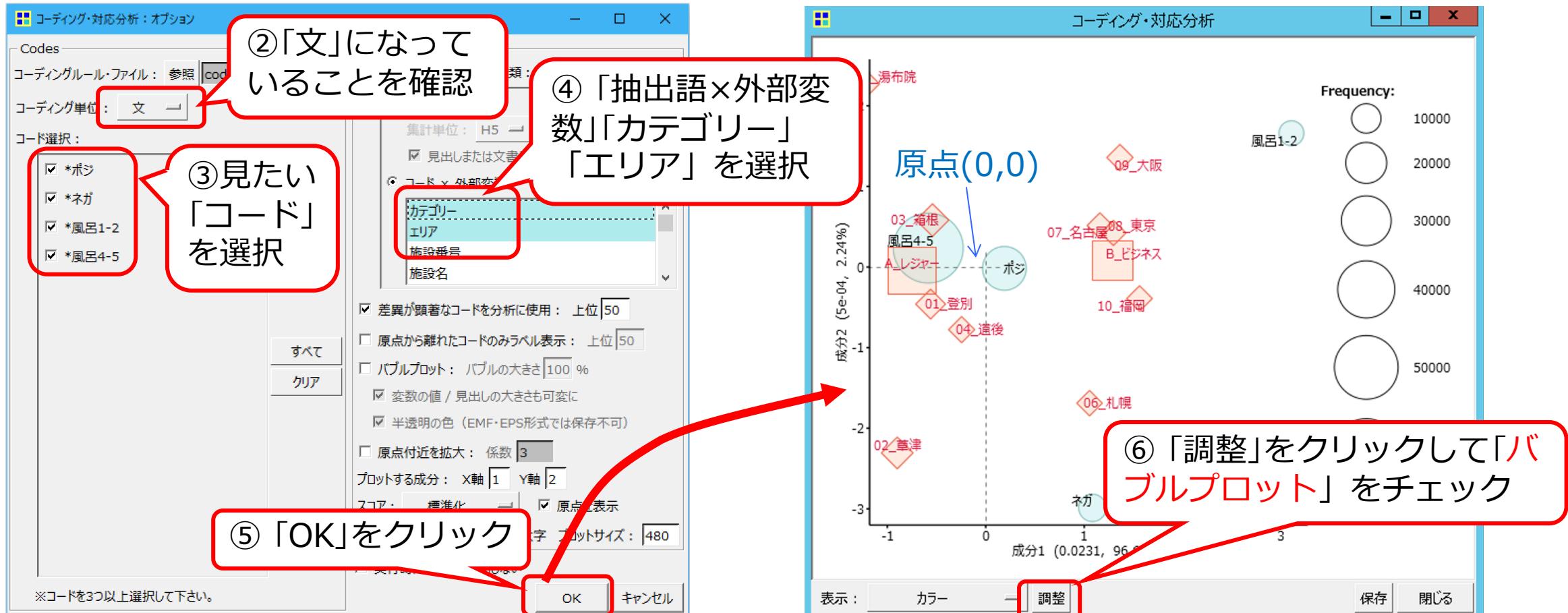
*風呂4-5

<>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

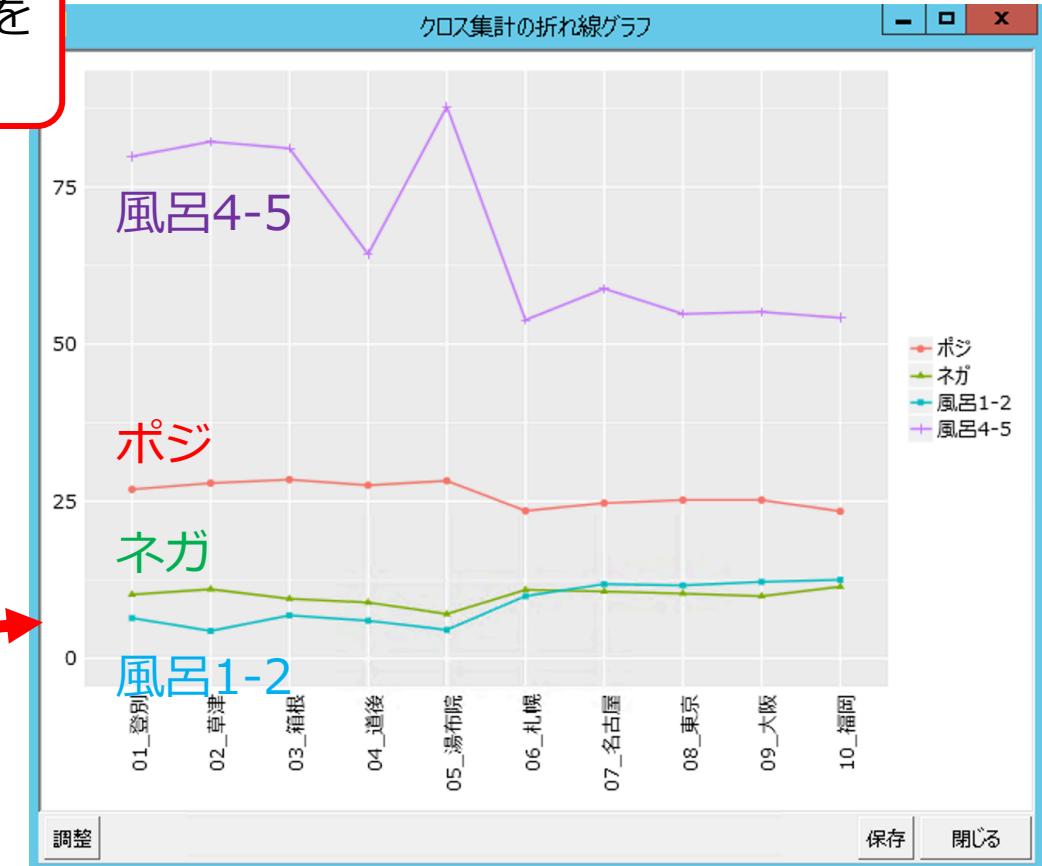
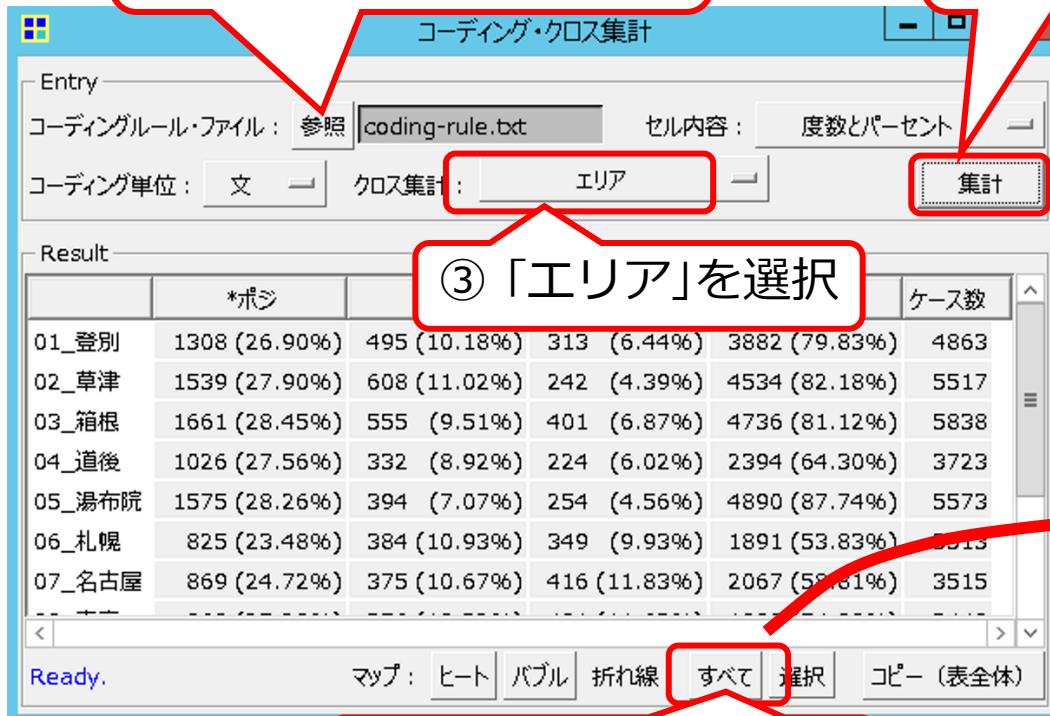
①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

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 一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」 を作成する
- ・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,違いについて考察する

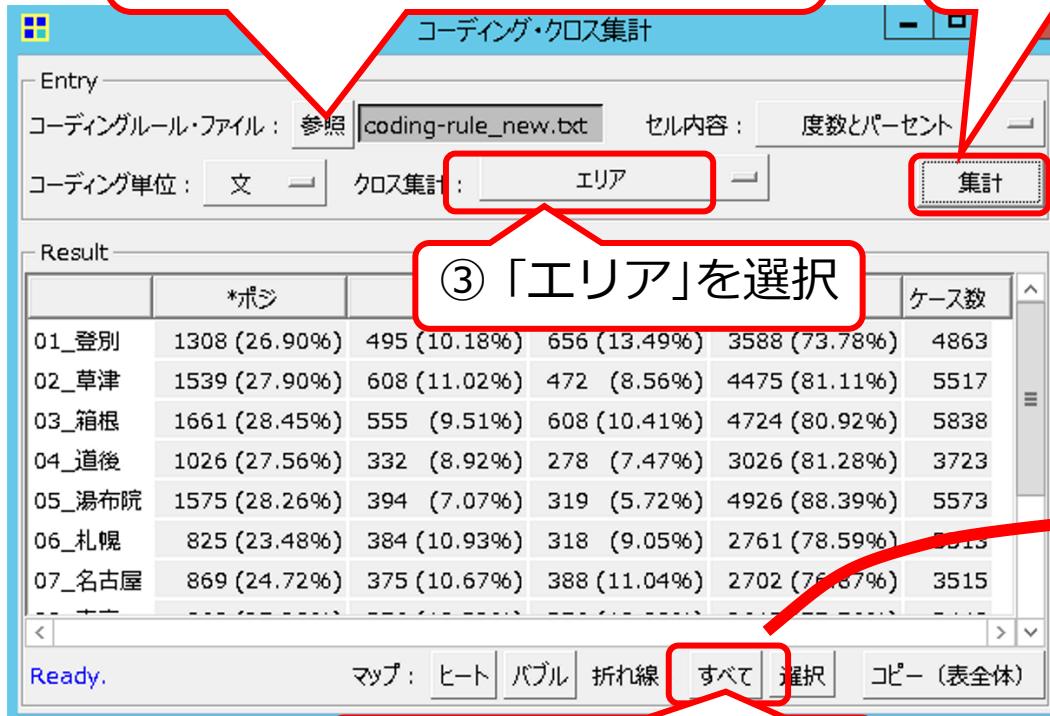
操作説明－クロス集計2

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

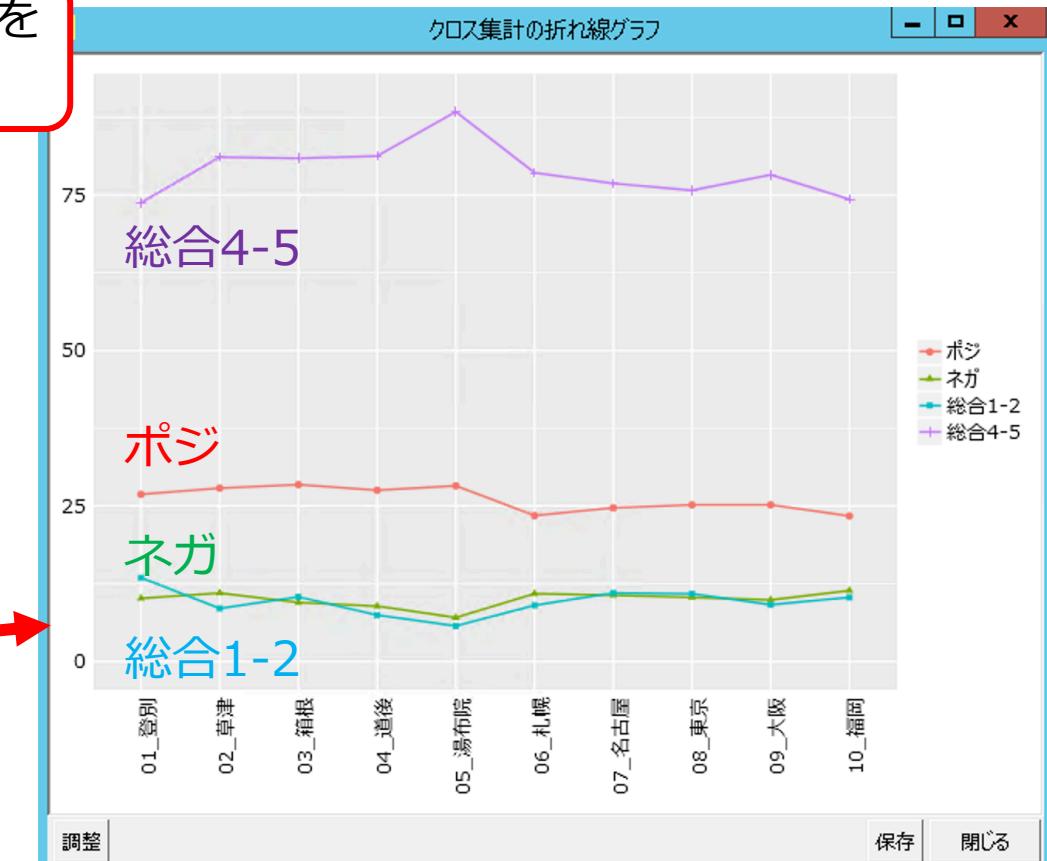
②「参照」をクリックして
「coding-rule_new.txt」を開く

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] 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.