

テキストマイニング

— Part 3 —

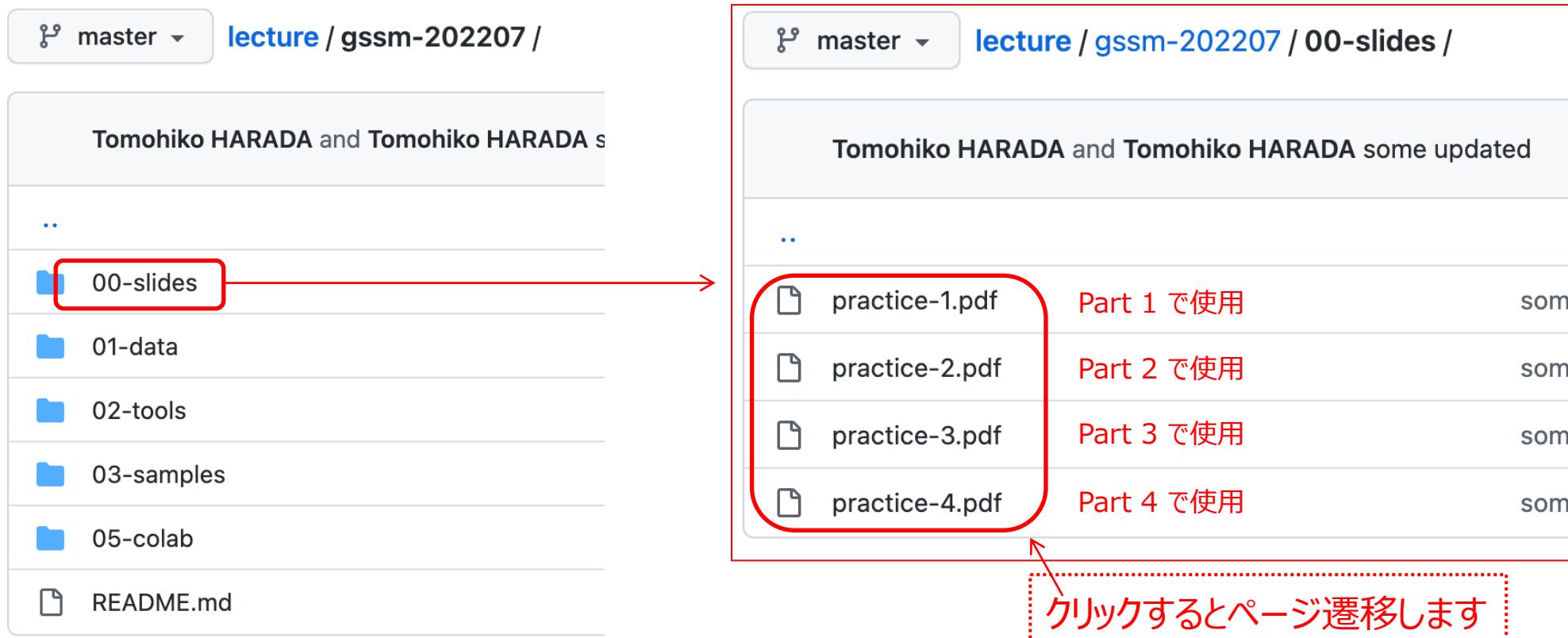
R04年度
人文社会ビジネス科学学術院
ビジネス科学研究群

スケジュール

- Part 1
 - 説明 — 自然言語処理のトレンド
 - 説明 — 環境説明
- Part 2
 - 説明 — テキストマイニングの手順
 - 説明 — データ理解
 - 実習 — データ理解 (Excel)
- Part 3
 - 説明 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)
- Part 4
 - 実習 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)

講義スライド

- <https://github.com/haradatm/lecture/tree/master/gssm-202207>



(復習) 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?

・データ分析に取り組む

- ・これら課題を解決するために、テキスト分析を実施

(復習) クチコミサイトの例



- ホテルのクチコミ数: 1,237万件 ※年間約60~70万

The screenshot shows the Rakuten Travel website at <https://travel.rakuten.co.jp/review/>. The main heading is 'お客様の声' (Customer Reviews) with a count of 'ホテルのクチコミ数No.1 12,369,840'. Below this, there's a search bar for reviews and a section for new reviews. On the right, there's a summary box stating '「お客様の声」には、実際にご利用になった方のご意見・ご感想が満載です。' (There are many opinions and feelings from customers who actually used the service). The overall theme is travel and accommodation reviews.

経年変化:

780万件 (2015)
→ 836万件 (2016)
→ 900万件 (2017)
→ 973万件 (2018)
→ 1,042万件 (2019)
→ 1,098万件 (2020)
→ 1,165万件 (2021)
→ **1,237万件 (今回)**
※ 2021/6/4現在

鴨川シーワールドホテルのクチコミ・お客さまの声

[●ホテル・旅行のクチコミTOPへ](#)

総合評価

4.12

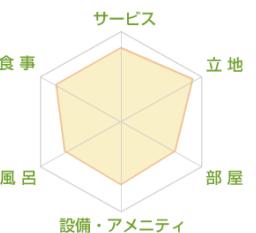
アンケート件数：886件

評価内訳

- 5点 ■■■■■ 236件
- 4点 ■■■■ 302件
- 3点 ■■ 47件
- 2点 ■ 15件
- 1点 ■ 9件

項目別の評価

サービス	4.11
立地	4.61
部屋	3.53
設備・アメニティ	3.62
風呂	3.53
食事	4.10



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、それは思ひませんでした。

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

... 総合 2

サービス 2

立地 4

部屋 4

設備・アメニティ 2

風呂 2

食事 4

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す。

客室内清掃の件、大変申し訳ござい

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

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。

テキストデータ

数値評価

(復習) 使用データ

楽天トラベル のクチコミデータ

- ・収集期間は **2019-2020** および **2021-2022(～GW明け)** の **2セット**
- ・以下の **10 エリアごと** 同数に **1,000件ずつ** ランダムサンプリング
- ・データ件数は **1万件** × 2セット

レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件 × 10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

(復習) データ項目

楽天トラベル のクチコミデータ

- データ項目は **18項目** (テキスト1項目+その他の属性**17項目**)

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント (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

数値評価で違いを見るのは難しい

- ユーザーの8割が4~5の評価,
1~2をつけない→本音が見えない

数値評価の平均 (エリア別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.29	4.29	4.18	4.07	4.34	4.29	4.34
01_登別	4.08	4.20	3.96	3.87	4.33	4.13	4.17
02_草津	4.29	4.27	4.13	4.04	4.38	4.18	4.33
03_箱根	4.26	4.16	4.18	4.05	4.28	4.25	4.27
04_道後	4.26	4.42	4.21	4.05	4.28	4.25	4.36
05_湯布院	4.58	4.39	4.40	4.05	4.28	4.25	4.58
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.29	4.32
06_札幌	4.17	4.42	4.26	4.07	3.96	4.15	4.35
07_名古屋	4.07	4.29	4.17	3.99	3.91	4.03	4.24
08_東京	4.13	4.43	4.20	4.04	3.88	4.21	4.32
09_大阪	4.16	4.42	4.20	4.04	3.88	4.17	4.37
10_福岡	4.17	4.43	4.20	4.04	3.94	4.25	4.32

- 同じ点数でもテキストを見れば差異があるかも

- すべての項目に回答する→どこに注目しているかよくわからない

数値評価の平均 (レジャー, ビジネス別)

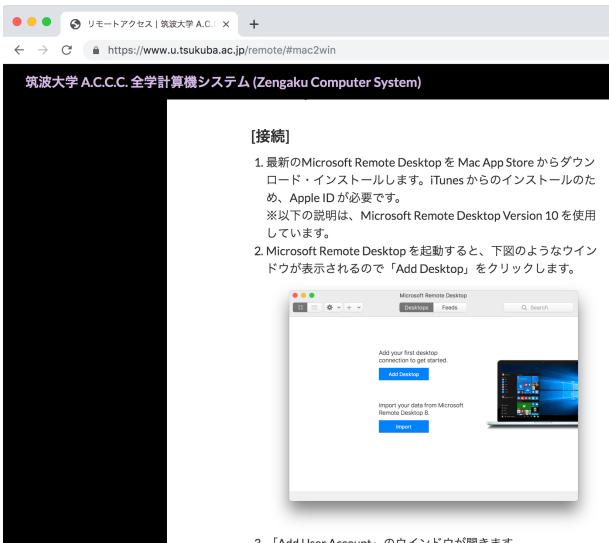
行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.29	4.29	4.18	4.07	4.34	4.29	4.34
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32

実習環境について

- 実習では,以下のツールを使用します
 - Part 2 では **Microsoft EXCEL** (用途: データの加工や修正)
 - Part 3 以降では **KHCoder** (用途: テキストマイニング) ※フリーソフト
- **KHCoder** は,全学計算機システムのリモートデスクトップでも動作します
 - 【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
 - 【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>
- 個人のPCで **KHCoder** を使用しても構いません
 - ただし, **Windows OS (11, 10, 8.1)** を搭載したPCが必要です

(参考) 全学計算機システムのリモートデスクトップ

- 事前に、下記のページの説明に従い全学計算機システム(Windows)へログインができるることを確認しておいてください
 - 【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
 - 【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>



Mac の場合:

左記のページにある説明に従って、事前にツール [Microsoft Remote Desktop](#) のインストールが必要です

KH Coder インストール時の注意:

全学の Windows では C ドライブへのファイル保存は禁止されています。ダウンロードした KH Coder を解凍する場合は、保存先を「**C ドライブ以外**」に変更してください。例)「**Z:¥Desktop¥khcoder3**」

KH Coder のインストール (Part 3 以降で使用)

- ・ダウンロードとインストール <https://khcoder.net/dl3.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 ダウンロードしたファイルを実行（ダブルクリックし、開いた画面上の「Unzip」ボタンをクリックします。）
- ③ 保存先を「**Cドライブ以外**」(**Cドライブへの保存は禁止されています**)に変更します。例)
'Z:\¥Desktop¥khcoder3'
- ④ 指定した保存先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。

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目的で開発されたフリー(無料)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?year=2022&auth=all>

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：247 / 4554

[KH Coderを用いた研究事例のリスト ◀5355件](#)

※2022/6/11 現在 (→1646→2042→2695→3741件 →昨年4554件→5355件)

KH Coder の情報

ホームページ <http://khcoder.net/>

The screenshot shows the homepage of khcoder.net. At the top, there's a navigation bar with Japanese and English language options. Below it is a large banner featuring the text "KH Coder 公式入門書!" and "ご好評発売中". To the left, there's a sidebar with sections for "Index", "【お知らせ】", "概要", "機能紹介 (スクリーンショット)", and "ダウンロードと使い方". On the right, there are several social media posts from users like "khcoderさん" and "JSS 日本社会学会 (非公式)".

参考書 **New**

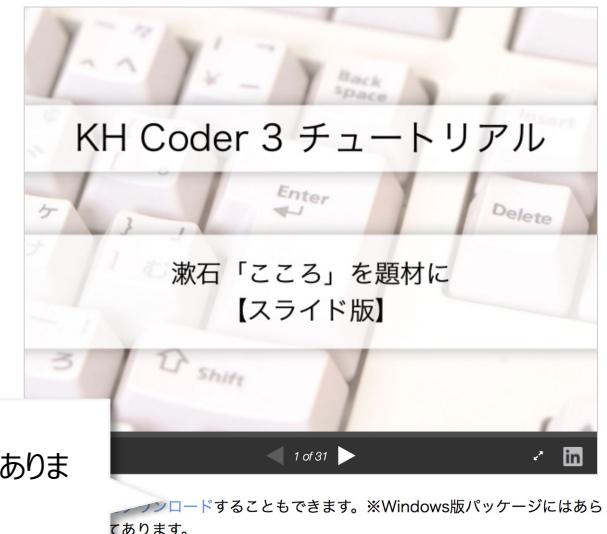


PDFファイルをダウンロードすることもできます。
※Windows版パッケージにはあらかじめ同梱してあります。

チュートリアル

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

チュートリアル & ヒント

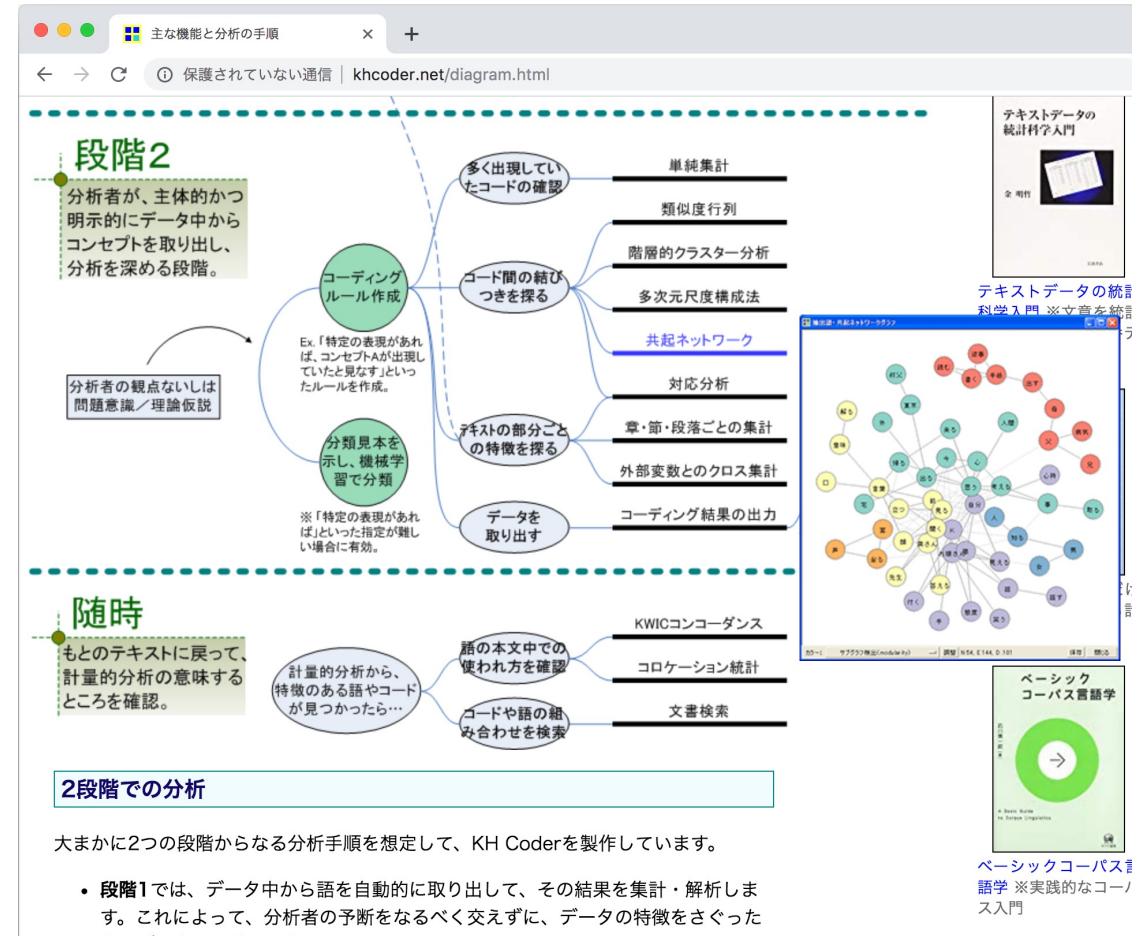
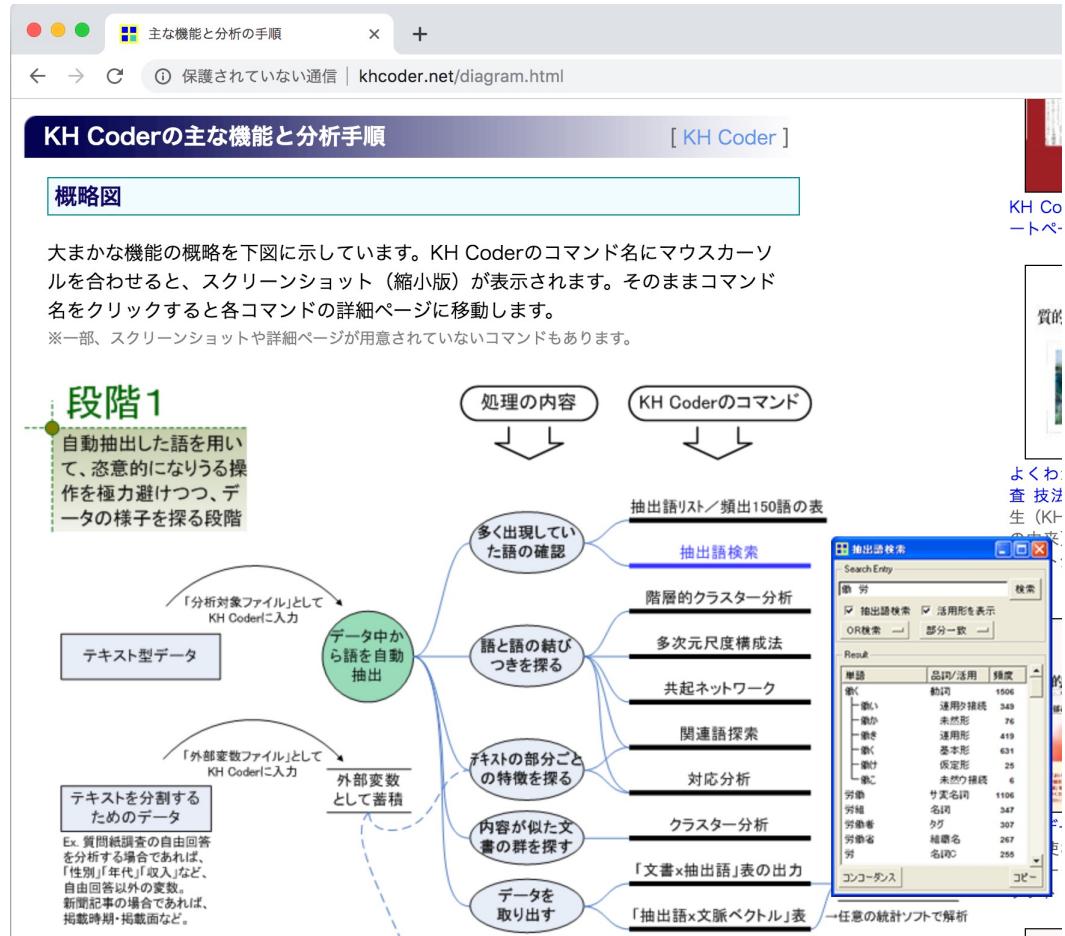


チュートリアル用データ

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>



KHCoder の基本:

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	

KHCoder の基本: 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」(2値)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(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"> 文中(1,000語あたり)における語の出現回数を計算 1つひとつの文の長さが長く,多数の文に含まれている語が多いデータ向き(各文中での語の出現回数の大小が重要な場合も) 	<ul style="list-style-type: none"> 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 						
<table border="1"> <tr> <td>1</td><td>0</td></tr> <tr> <td>1</td><td>n_{11}</td></tr> <tr> <td>0</td><td>n_{01}</td></tr> </table> $J^S = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$	1	0	1	n_{11}	0	n_{01}	$E d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$	$cos S(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$
1	0							
1	n_{11}							
0	n_{01}							

n_{11} : 2つの語が同時に出現した文書の数

\mathbf{x}, \mathbf{y} : はそれぞれの単語ベクトル

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

KH Coder — 分析手法

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



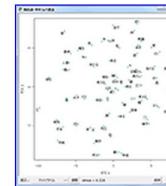
やや弱い共起関係も描画に含め、自動的にグループ分け（色分け）



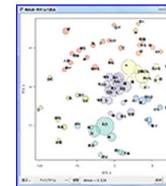
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

多次元尺度構成法 (MDS)

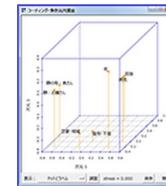
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



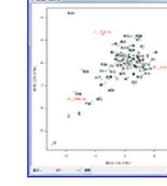
New! クラスタリングと
色分け



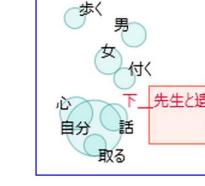
3次元の解

対応分析

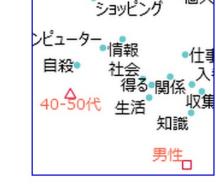
同じく抽出語またはコードを用いての、対応分析です。



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

分析手法

解説

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、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

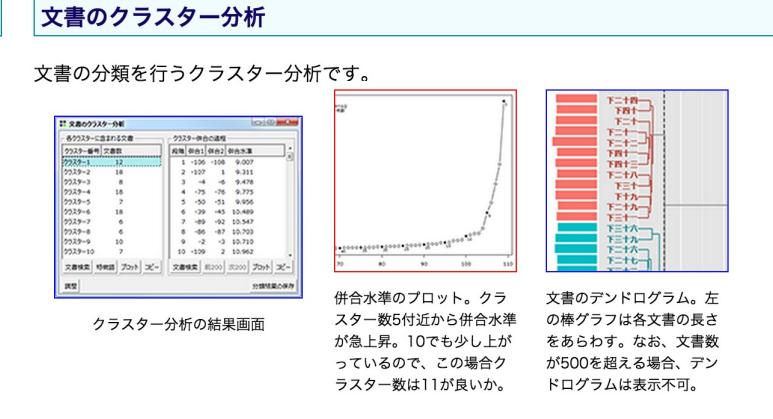
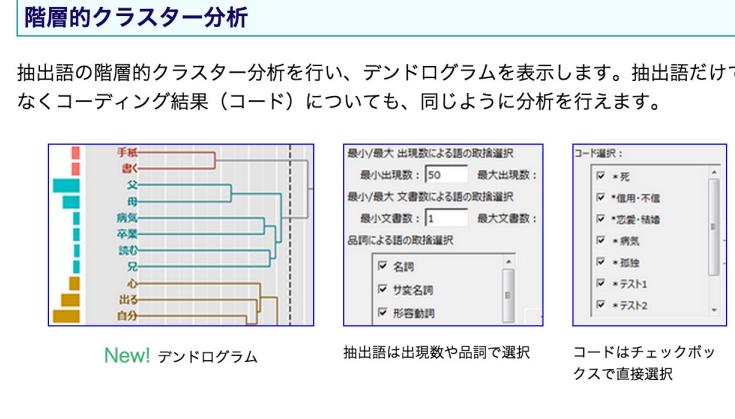
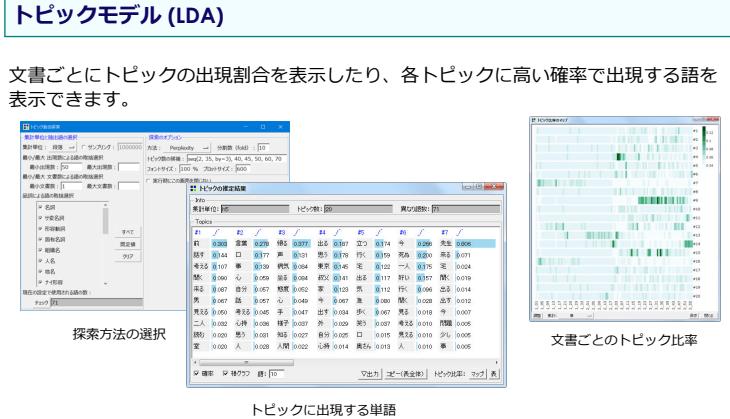
多次元尺度構成法 (MDS)

- 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 (コレスポンデンス分析)

- 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、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

KH Coder — 分析手法



分析手法	解説
トピックモデル (LDA)	<ul style="list-style-type: none"> 文書が複数のトピックを持つと仮定し、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 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、と乱数のシードを固定した以外はデフォルト設定 コーディングルールが専門家による単語の集約であるのに対し、トピックモデルは教師なし学習のため客観性が高まる
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
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

(再掲) 実習で使用するデータ

楽天トラベル のクチコミデータ

- ・収集期間は **2019-2020** および **2021-2022(～GW明け)** の **2セット**
- ・以下の **10 エリアごと** 同数に **1,000件ずつ** ランダムサンプリング
- ・データ件数は **1万件** × 2セット

レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件 × 10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

(再掲) 実習で使用するデータ

楽天トラベル のクチコミデータ

- データ項目は **18項目** (テキスト1項目+その他の属性**17項目**)

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント (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

(再掲) データ一覧

データファイル名	件数	データセット	備考
rakuten-1000-2021-2022.xlsx	10,000	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアエリアごと 1,000件 (ランダムサンプリング)期間: 2020/1~2022/5/15	<ul style="list-style-type: none">本講義の全体を通して利用する
rakuten-1000-2019-2020.xlsx	10,000	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアエリアごと 1,000件 (ランダムサンプリング)期間: 2019/1~2020/12	<ul style="list-style-type: none">実習用 (期間で比較する等)
rakuten-all-2021-2022-tsv.zip	142,475	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアサンプリング前の全データ (宿泊年月naを除く)期間: 2020/1~2022/5/15	<ul style="list-style-type: none">その他 (Python や R を使って分析したい人向け)
rakuten-all-2019-2020-tsv.zip	162,433	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアサンプリング前の全データ (宿泊年月naを除く)期間: 2019/1~2020/12	
rakuten-all-tsv.zip	1,593,525	<ul style="list-style-type: none">レジャー+ビジネスの 10エリアサンプリング前の全データ期間: 2009/3~2020/12	

データの取得方法 – 必ず再ダウンロードする

- <https://github.com/haradatm/lecture/tree/master/gssm-202207>

The image shows two screenshots of a GitHub repository. The left screenshot shows the main directory structure of 'gssm-202207' with a red box around the '01-data' folder. An arrow points from this folder to the right screenshot, which shows a detailed view of the '01-data' folder. In the right screenshot, a red box highlights the 'rakuten-1000-2021-2022.xlsx.zip' file. A red dashed box surrounds the text '本講義で主として使用' (Used primarily in this lecture) next to the file name. Below the file list, a red box highlights the 'Download data (to be used in the exercise)' button.

Tomohiko HARADA and Tomohiko HARADA s

..

00-slides

01-data

02-tools

03-samples

05-colab

README.md

master lecture / gssm-202207 /

Tomohiko HARADA and Tomohiko HARADA some updated

..

README.md

some updated

..

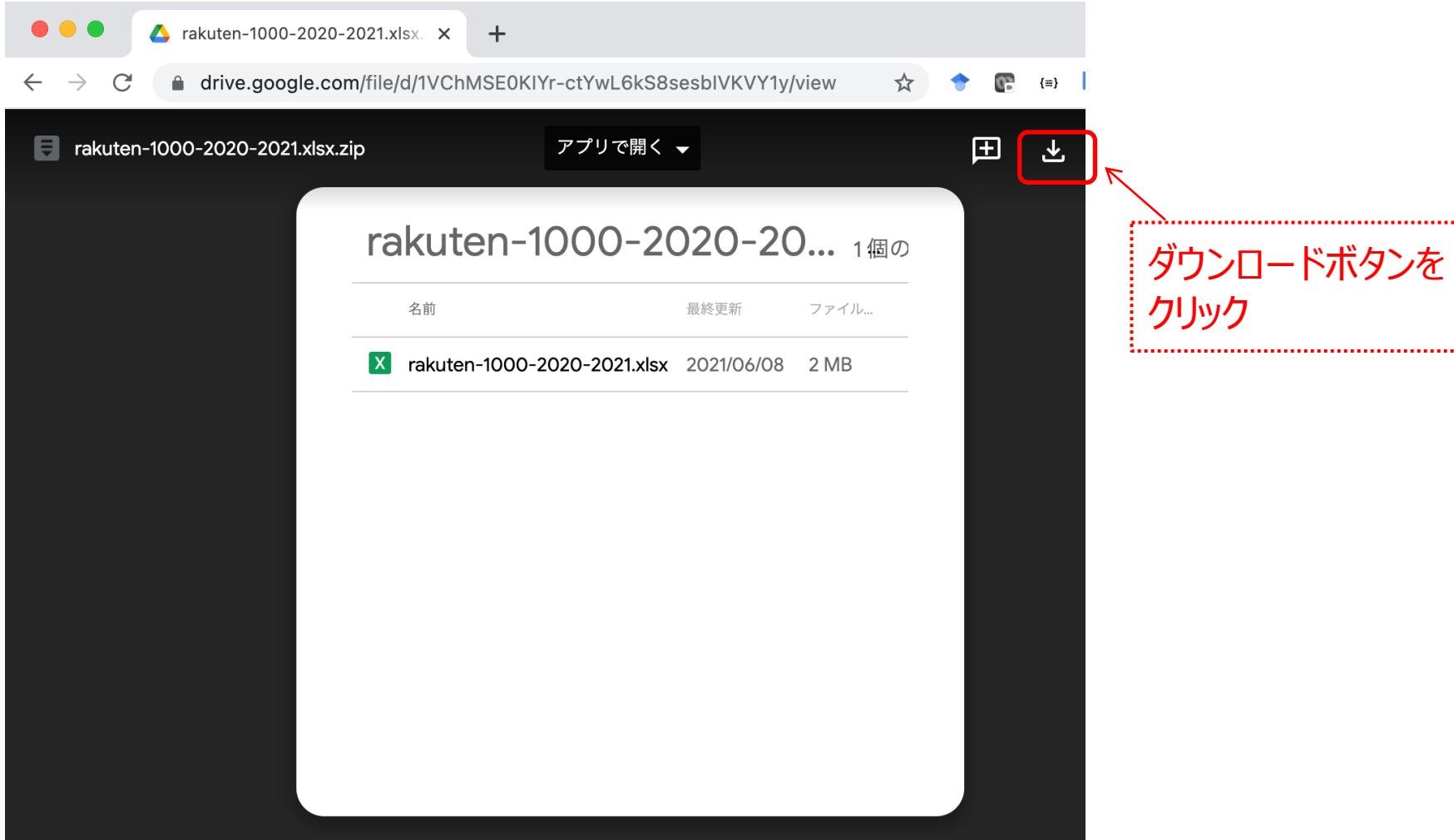
README.md

Download data (to be used in the exercise)

file name	# records	size (zipped)	period
rakuten-1000-2021-2022.xlsx.zip	10,000	2.4 MB	2021/1/1~2022/1/1
rakuten-1000-2019-2020.xlsx.zip	10,000	2.4 MB	2019/1/1~2020/1/1

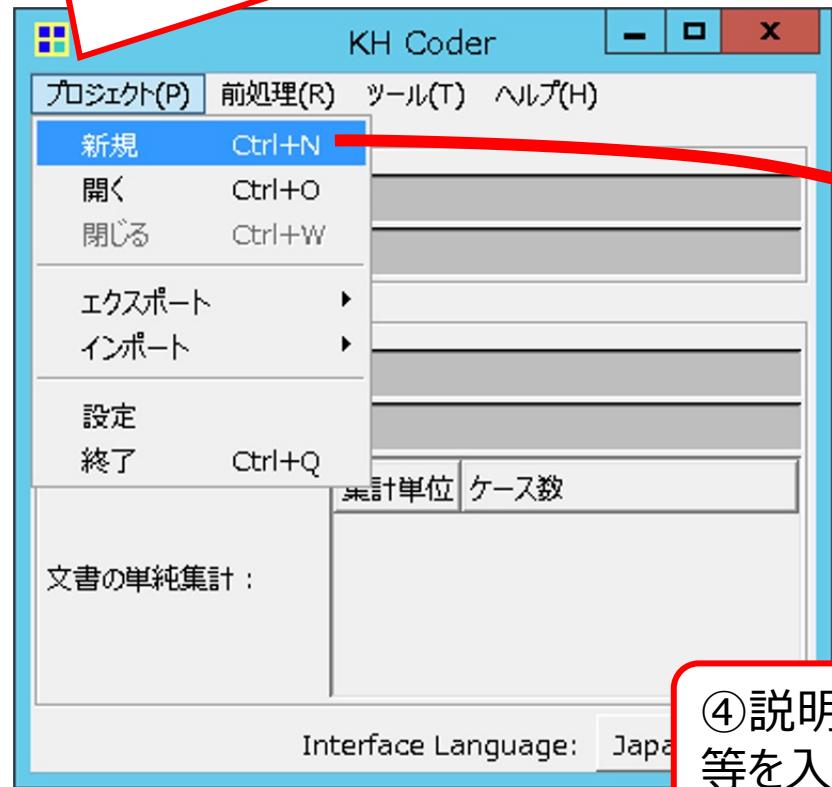
本講義で主として使用

ダウンロード方法 — 必ず再ダウンロードする



使い方 — プロジェクトの作成

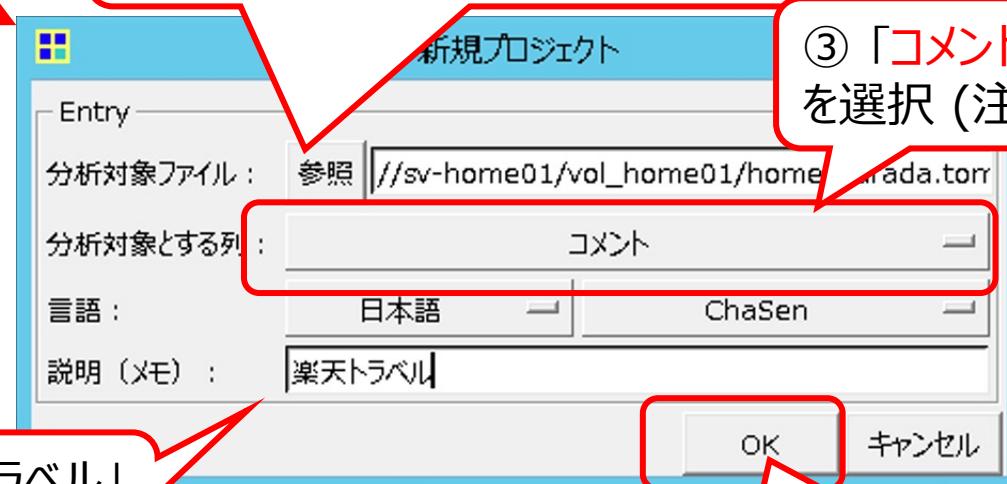
①メニューから「プロジェクト」「新規」を選択 (注1)



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-1000-2020-2021.xlsx」を開く

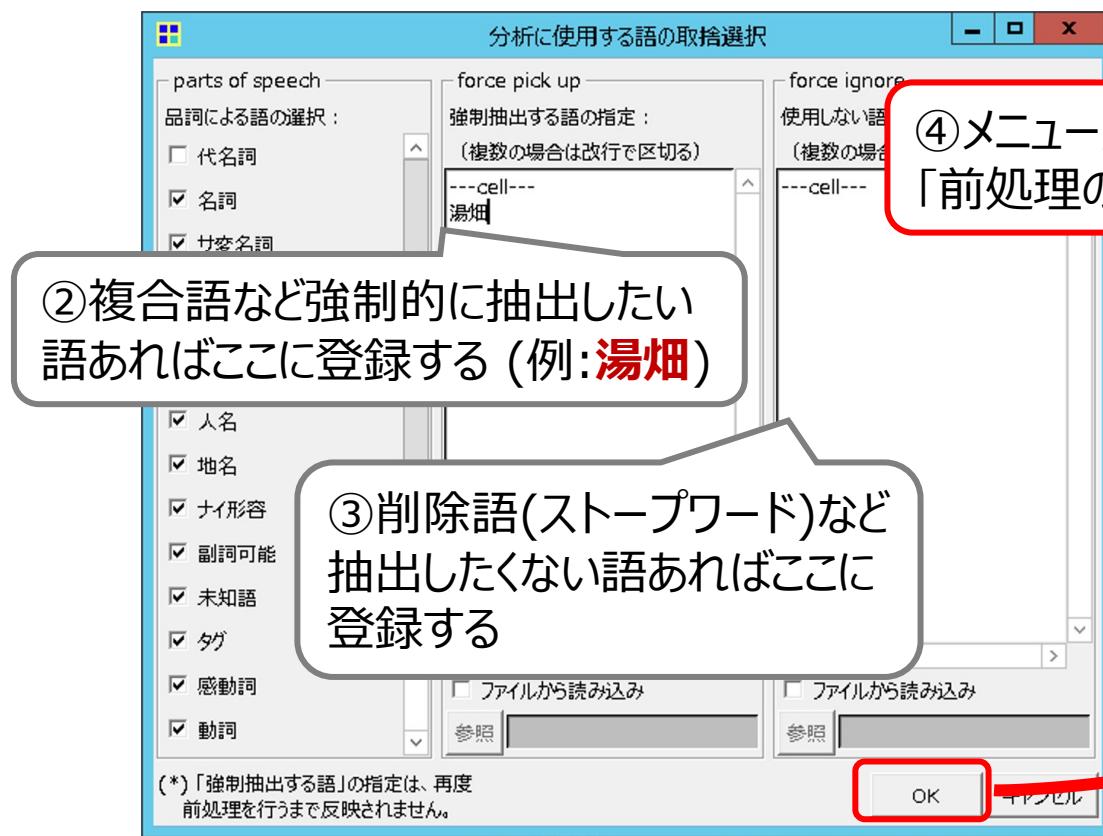


④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

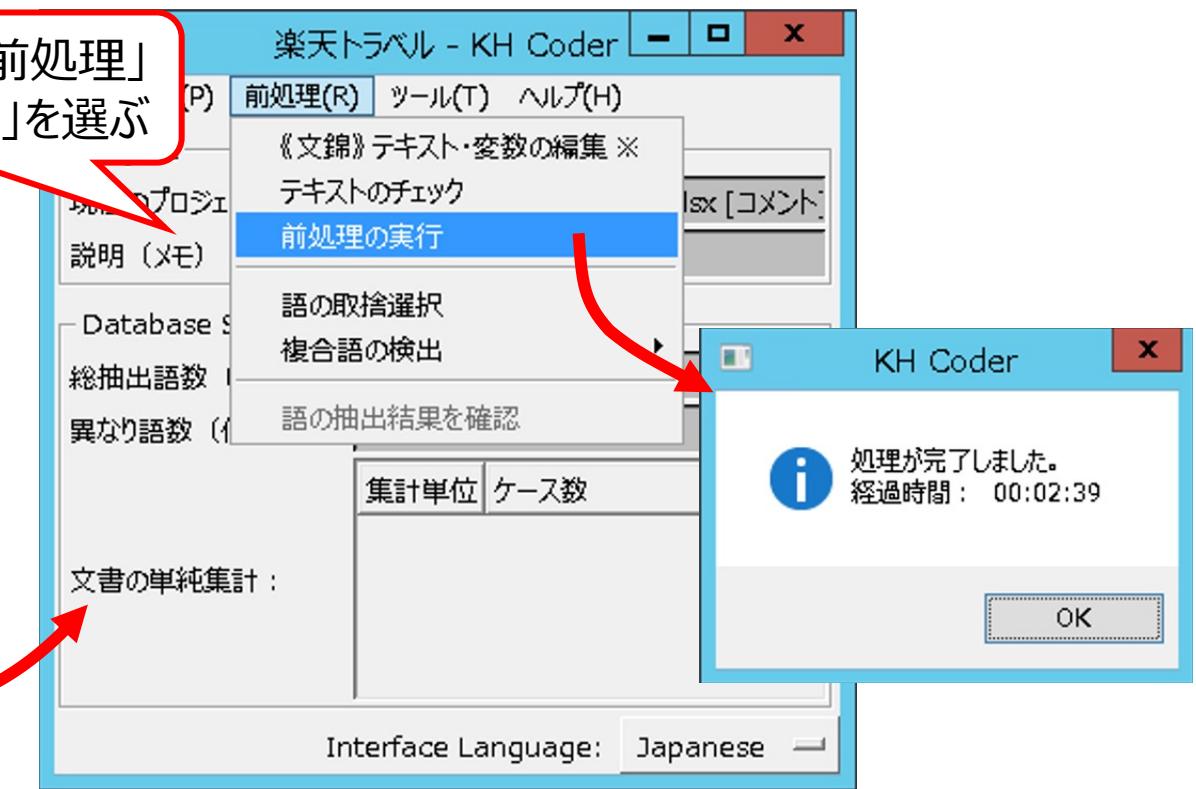
使い方 – 前処理（形態素解析）

①メニューから「前処理」「語の取捨選択」を選ぶ

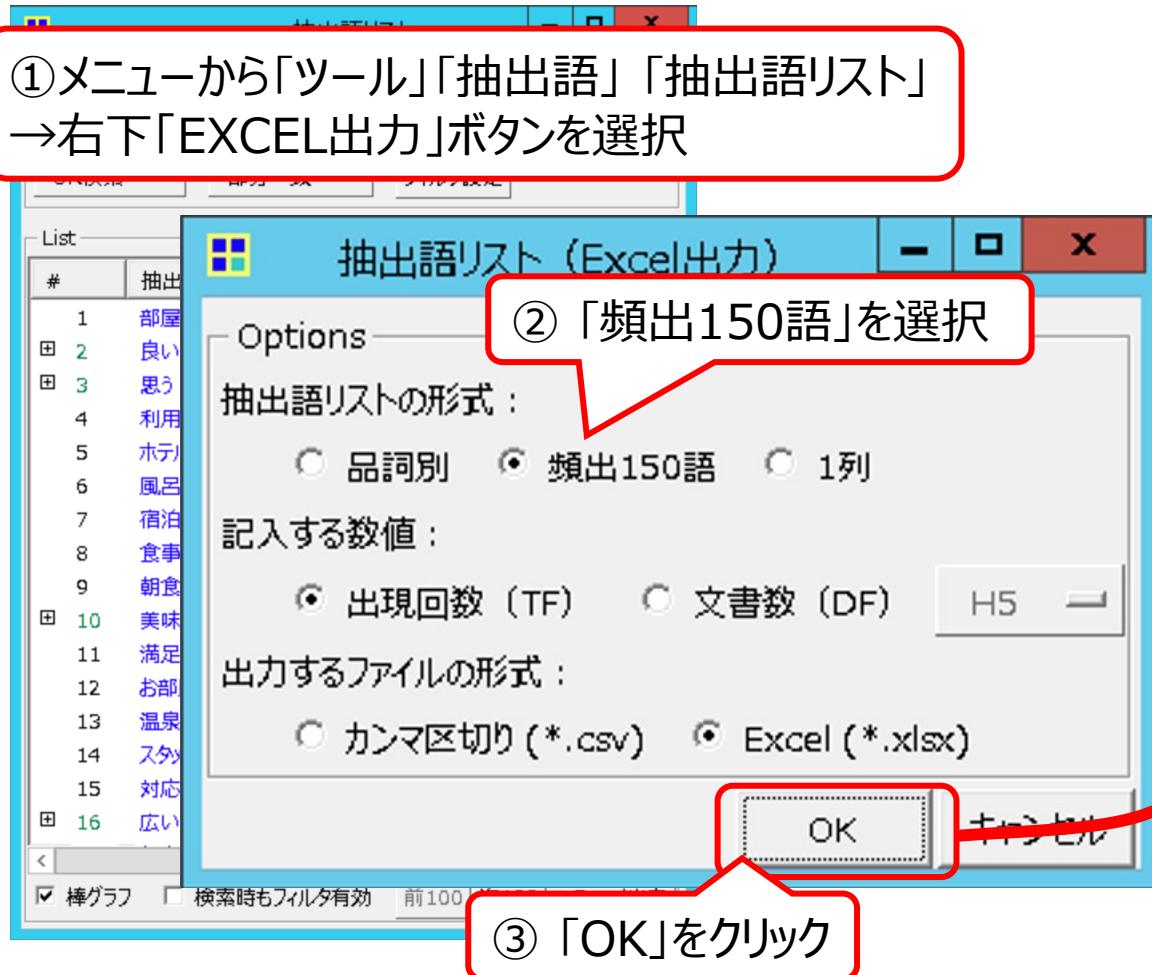


④メニューから「前処理」「前処理の実行」を選ぶ

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



使い方 — 頻出語を確認する

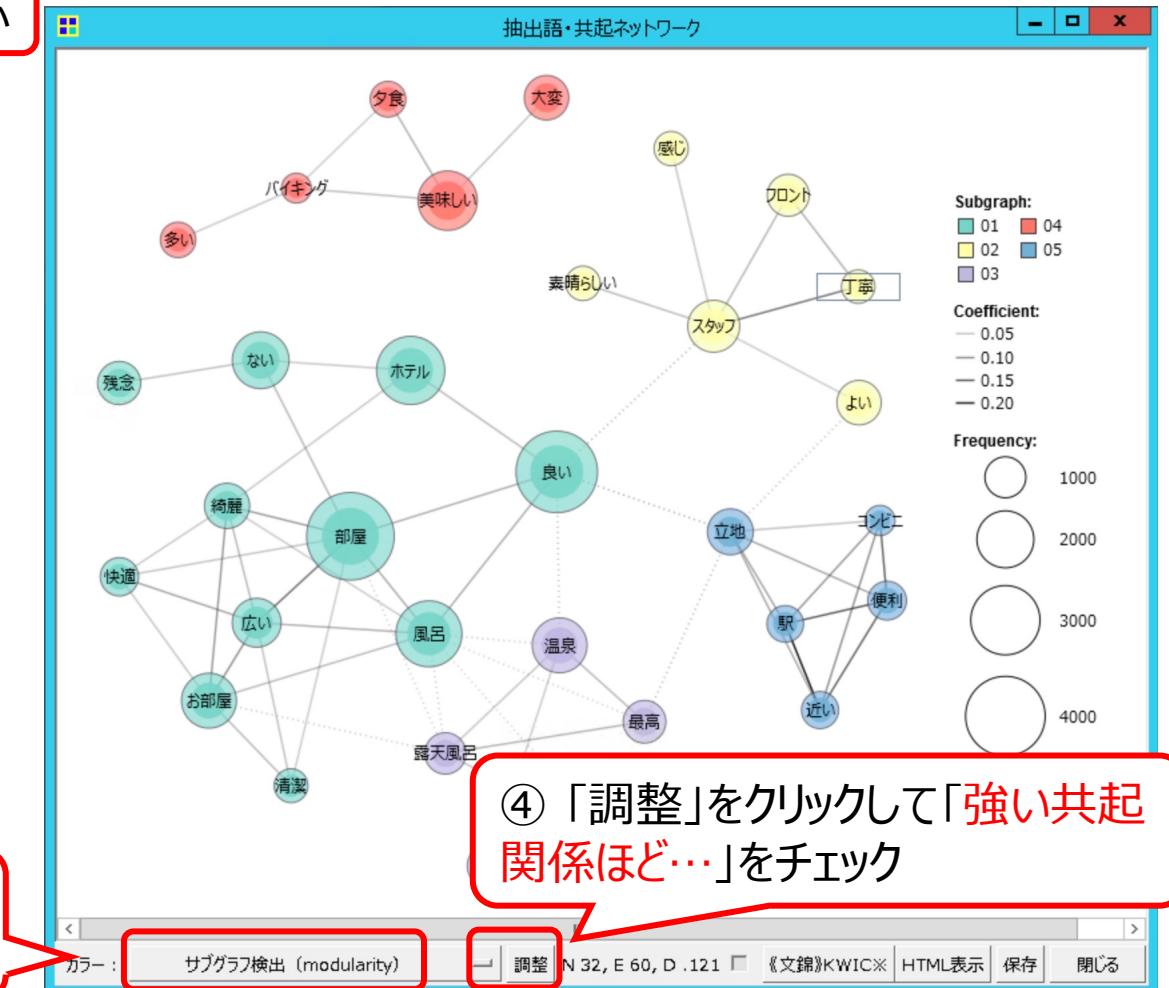
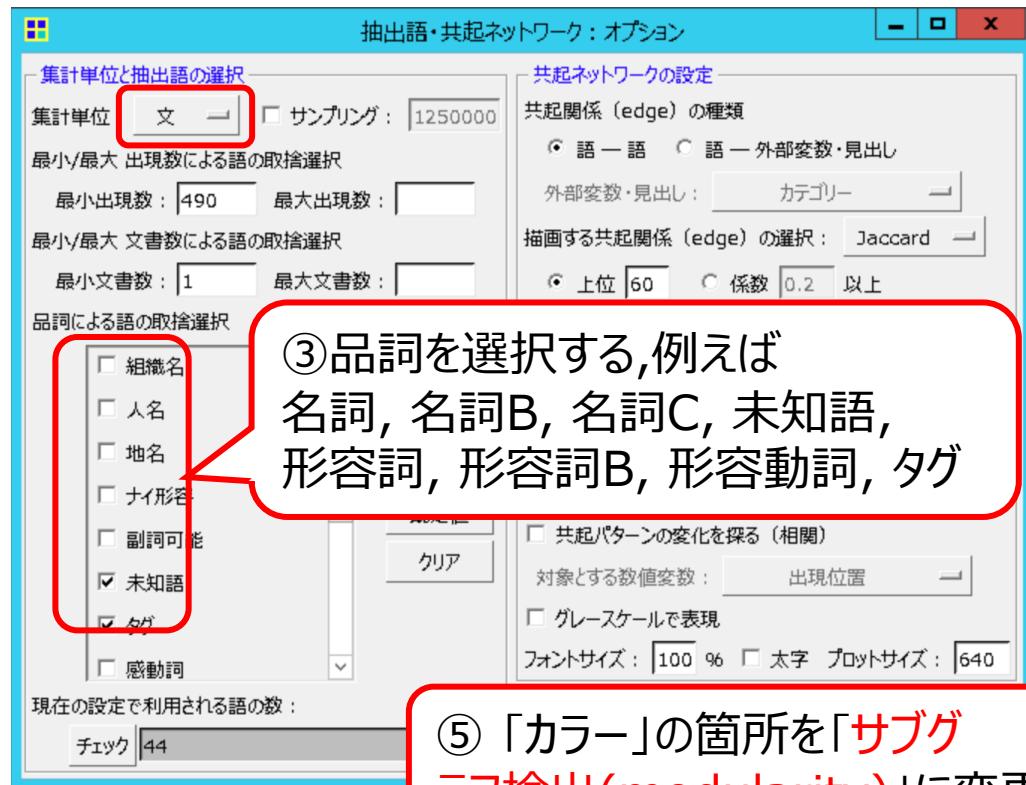


A	B	C	D	E	F	G	H
1 抽出語	出現回数	抽出語	出現回数	抽出語	出現回数		
2 部屋	4876	素晴らしい	696	ご飯	405		
3 良い	4213	過ごす	694	清掃	403		
4 思う	4126	感じ	686	高い	397		
5 利用	3554	清潔	673	無料	396		
6 ホテル	2915	過ごせる	669	悪い	395		
7 風呂	2724	丁寧	656	新しい	387		
8 宿泊	2723	バス	640	設備	383		
9 食事	2385	家族	635	安心	380		
10 朝食	2221	月	619	旅館	372		
11 美味しい	2184	コロナ	594	パ	366		
12 満足	2171	アメニティ	589	楽しめる	366		
13 お部屋	1861	初めて	573	見える	362		
14 温泉	1852	使う	554	狭い	358		
15 スタッフ	1645	入れる	552	対策	350		
16 対応	1569	泊	549	シャワー	349		
17 広い	1458	駐車	542	お願い	345		
18 行く	1387	子供	534	お湯	345		
19 立地	1279	旅行	533	全て	343		
20 綺麗	1236	コンビニ	525	湯畑	343		
21 大変	1195	夜	523	少ない	342		
22 サービス	1130	バイキング	514	置く	339		
23 残念	1122	プラン	513	用意	332		
24 料理	1119	値段	504	問題	330		

使い方 – 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「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

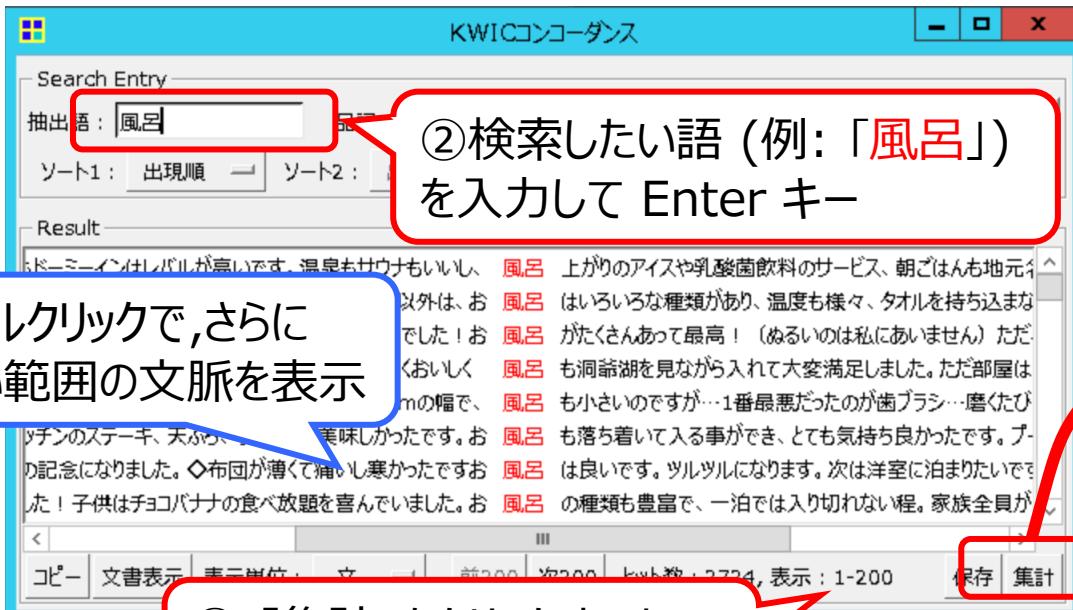
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般（漢字を含む 2 文字以上の語）
名詞 B	名詞一般（平仮名のみの語）
名詞 C	名詞一般（漢字 1 文字の語）
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立（漢字を含む語）
動詞 B	動詞-自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平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

使い方 — 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



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

The screenshot shows the 'Node Word' section of the KWIC Concordance tool. It displays a table of words and their collocation statistics. The table includes columns for the word itself, part of speech, total count, and counts for various contexts (left 5, left 4, left 3, left 2, left 1, right 1, right 2, right 3, right 4, right 5, score). A red box highlights the '抽出語' field, and another red box highlights the text '「広い」は「風呂」の2語後に91回出現' (The word '広い' appears 91 times two words after '風呂'). A blue box with a red border encloses the text '「右合計」でソート' (Sort by Right Total). A red box highlights the '右合計' button in the toolbar.

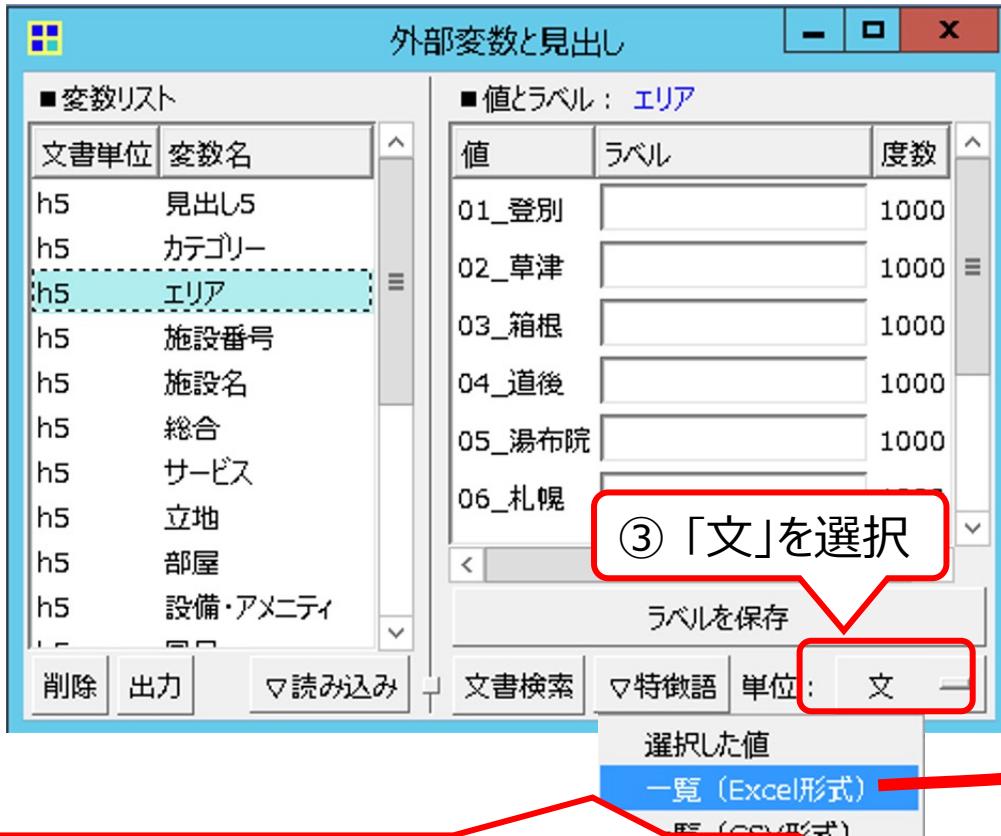
N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	スコア
1	良い	形容詞	223	74	149	42	15	13	4	0	1	51	39	26	32	70.88
2	広い	形容詞	189	48	141	10	8	20	8	2	1	91	24	17	8	77.01
3	綺麗	形容動詞	100	41	59	9	12	17	3	0	34	8	11	6	35.58	
4	よい	形容詞B	76									15	13	10	9	23.50
5	ない	形容詞B	56									9	5	11	12	18.20
6	清潔	形容動詞	46									2	11	7	4	15.23
7	気持ちよい	形容詞	37									2	7	7	7	12.28

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

使い方 — 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」を開く



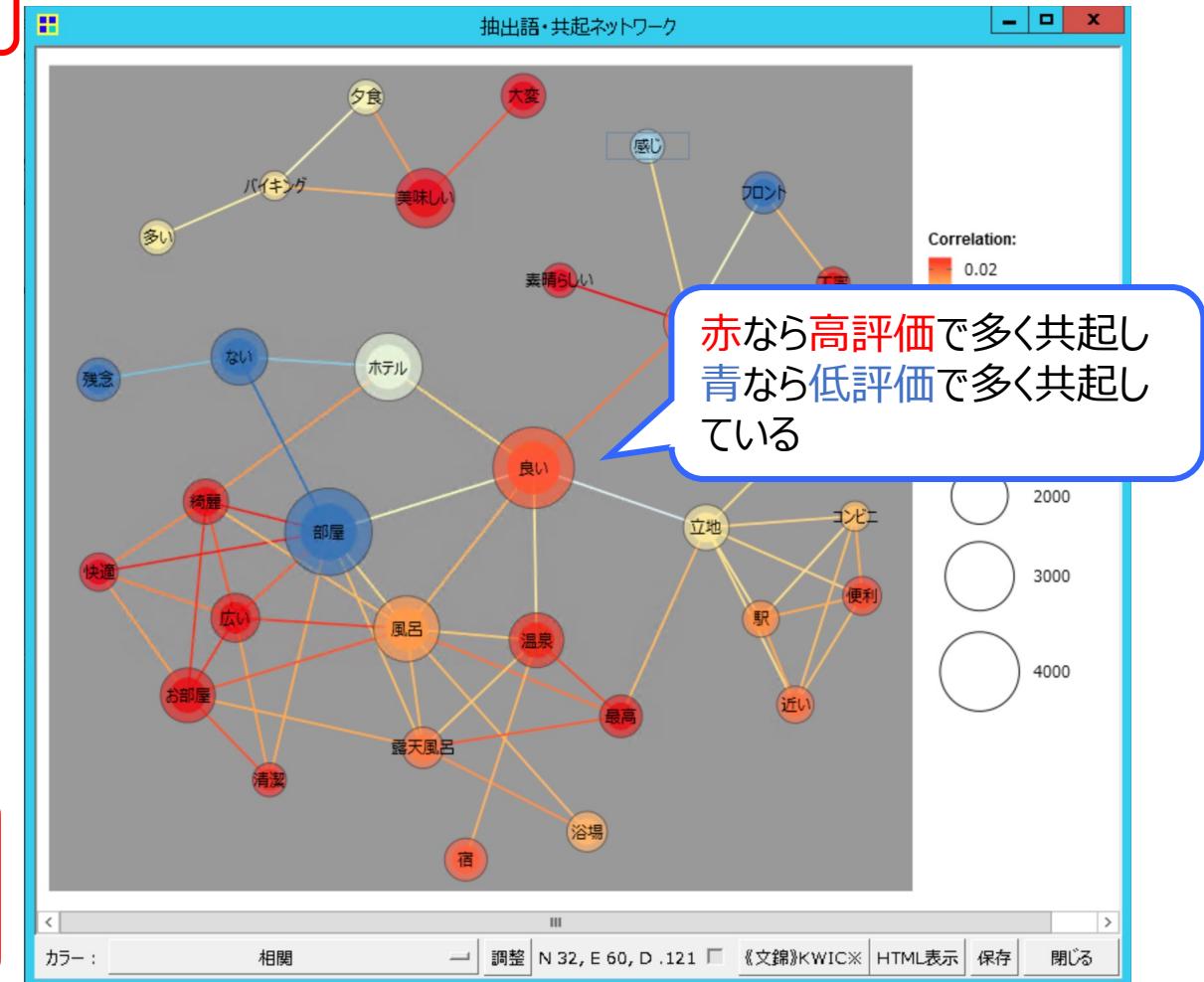
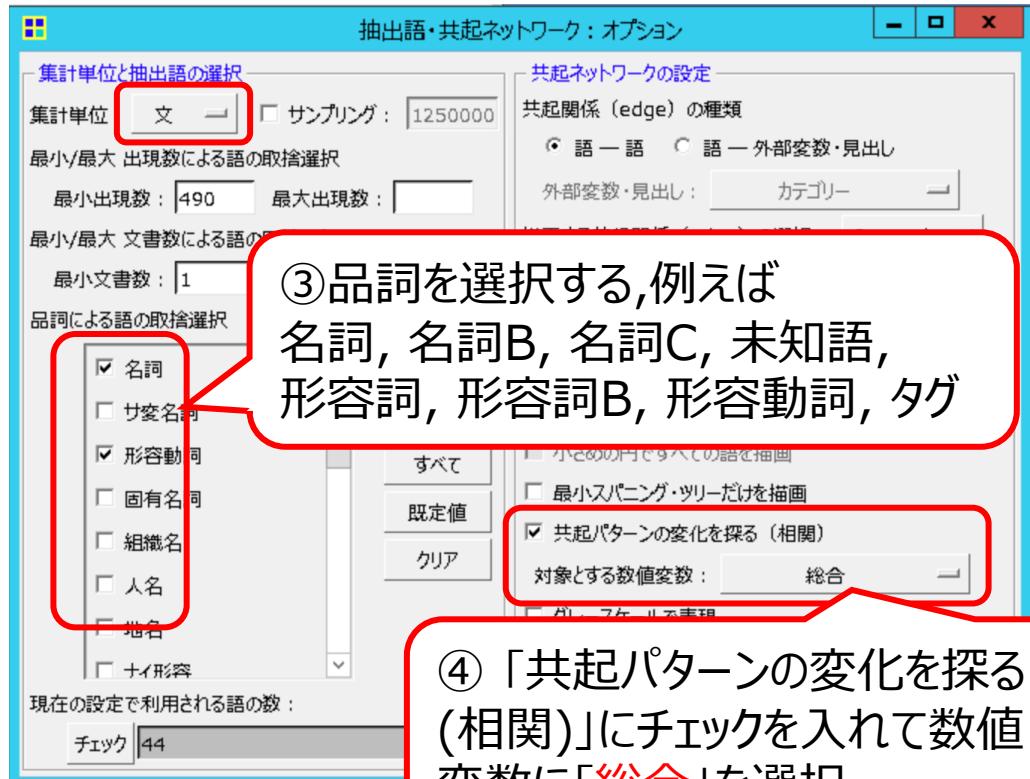
④「特徴語」「一覧(Excel形式)」を選択

	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	食事	.059	温泉	.068	思う	.066	温泉	.054			
4	良い	.058	湯畑	.064	食事	.064	良い	.051			
5	風呂	.057	風呂	.062	良い	.060	朝食	.045			
6	思う	.054	良い	.061	風呂	.053	ホテル	.042			
7	温泉	.049	食事	.056	美味しい	.049	美味しい	.042			
8	美味しい	.044	草津	.055	露天風呂	.048	道後	.041			
9	宿泊	.043	満足	.042	お部屋	.045	対応	.028			
10	満足	.041	美味しい	.042	温泉	.043	松山	.028			
11	料理	.033	宿	.041	満足	.043	立地	.026			
12	行く	.032	行く	.037	料理	.034	大変	.023			
13	05_湯布院		06_札幌		07_名古屋		08_東京				
14	食事	.072	ホテル	.061	ホテル	.063	利用	.060			
15	美味しい	.062	部屋	.058	名古屋	.059	部屋	.057			
16	宿	.061	朝食	.057	朝食	.058	ホテル	.054			
17	風呂	.059	利用	.055	利用	.055	宿泊	.039			
18	露天風呂	.050	札幌	.055	部屋	.055	朝食	.035			
19	料理	.049	良い	.052	思う	.047	快適	.034			
20	満足	.048	宿泊	.043	フロント	.035	お部屋	.034			
21	宿泊	.044	対応	.034	綺麗	.032	駅	.034			
22	温泉	.043	広い	.033	駅	.030	立地	.034			
23	お部屋	.042	立地	.031	対応	.029	フロント	.032			
24	09_大阪		10_福岡								
25	ホテル	.061	ホテル	.060							
26	利用	.056	利用	.060							
27	部屋	.050	部屋	.058							
28	宿泊	.040	朝食	.040							
29	立地	.039	博多	.039							
30	朝食	.039	立地	.039							
31	駅	.036	宿泊	.036							
32	綺麗	.033	便利	.033							
33	便利	.031	広い	.031							
34	フロント	.030	駅	.030							

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

使い方 – 共起ネットワークの作成2

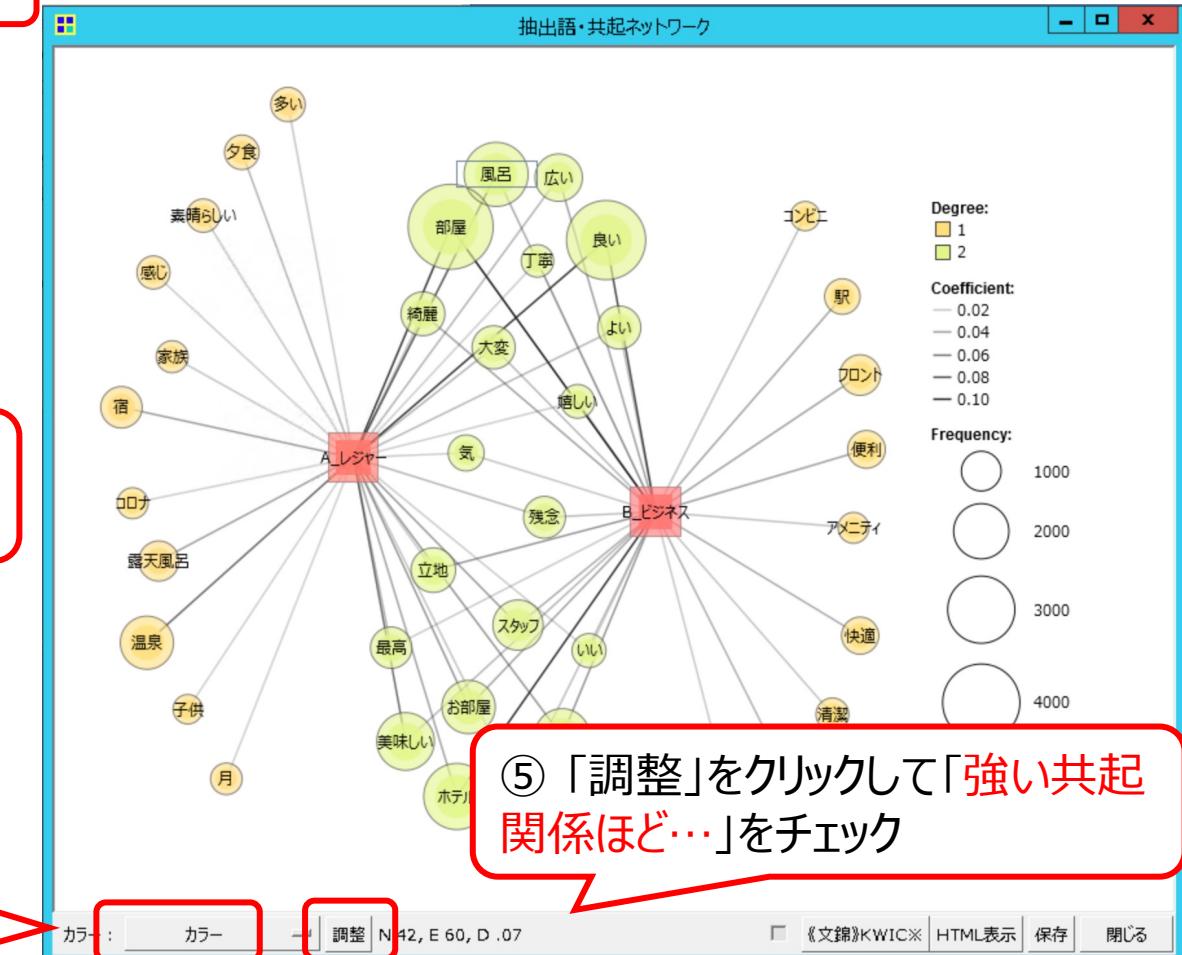
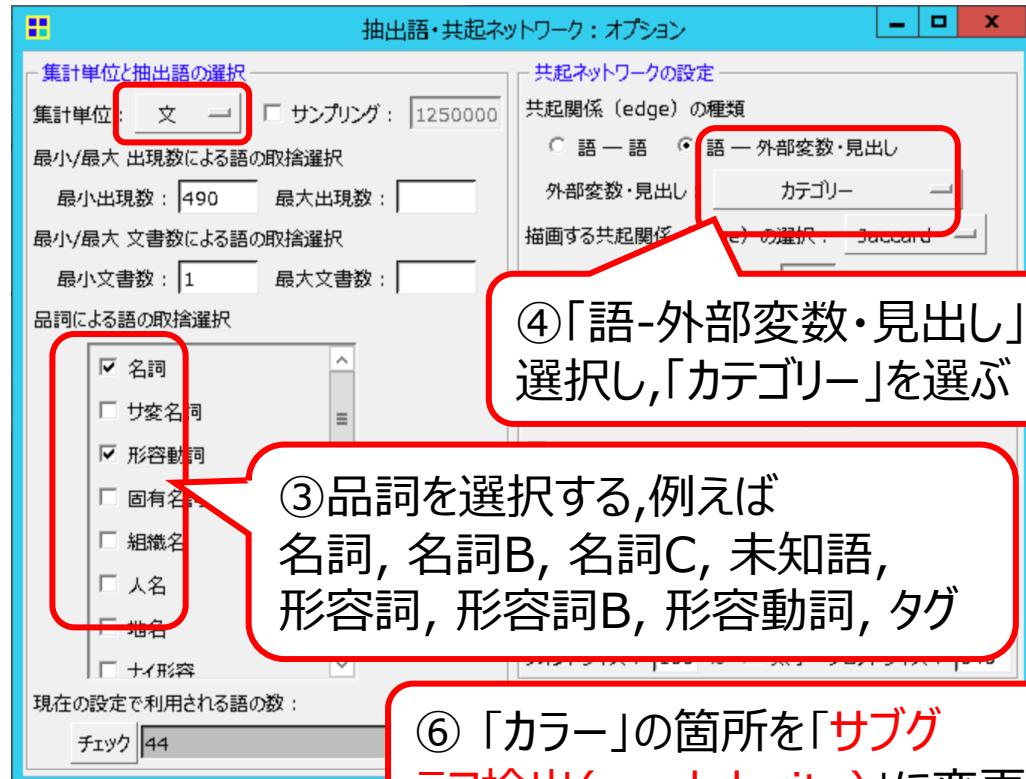
- ①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ
 - ②「集計単位」として「文」を選んで「OK」をクリック



使い方 – 共起ネットワークの作成3

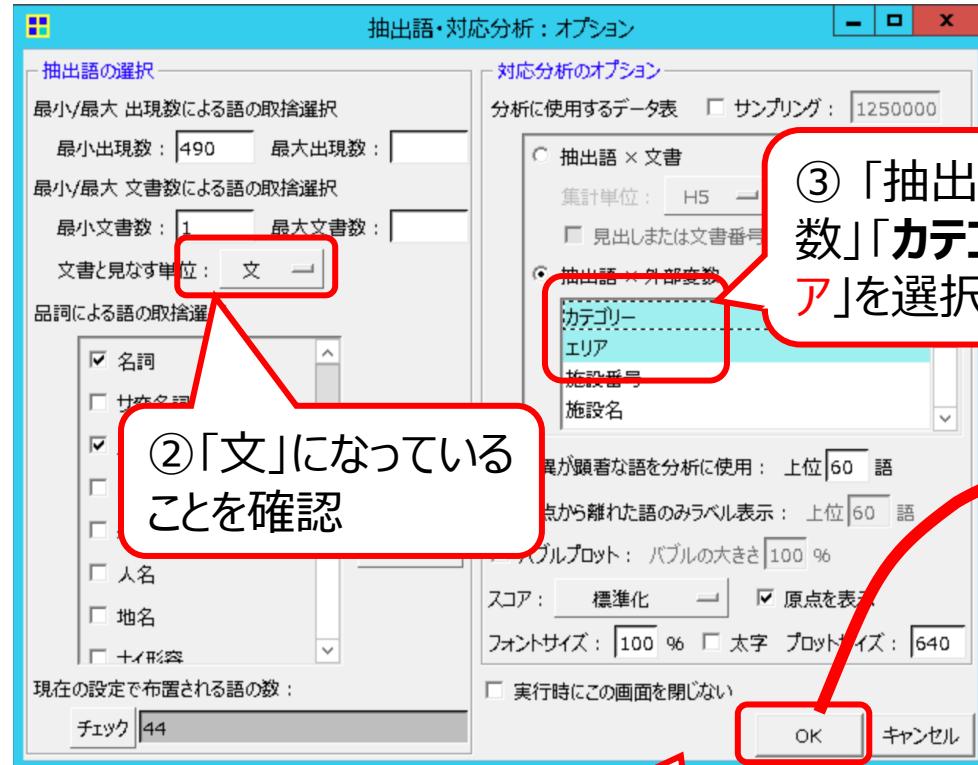
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「文」を選んで「OK」をクリック



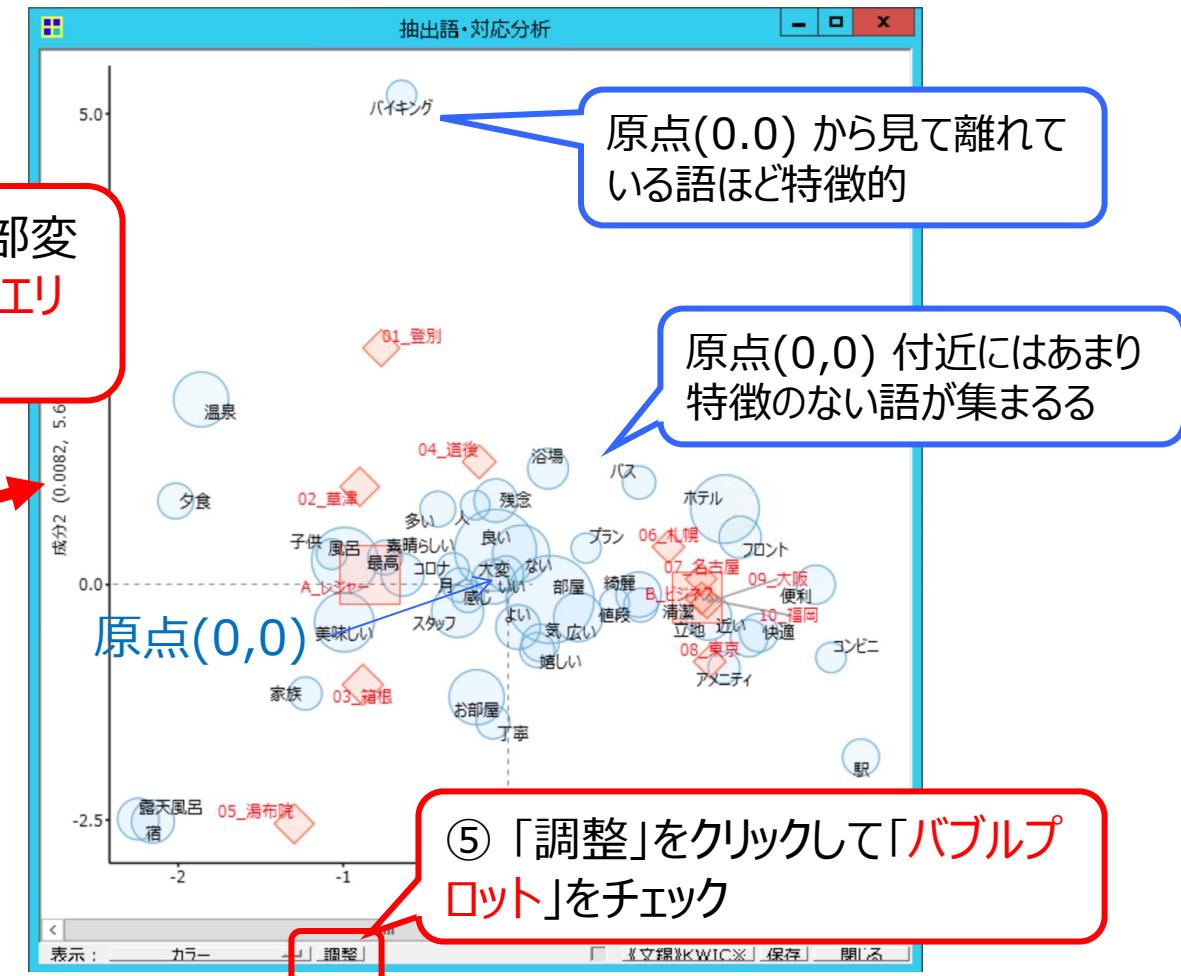
使い方 — 対応分析による探索1

①メニューから「ツール」「抽出語」「対応分析」を選ぶ



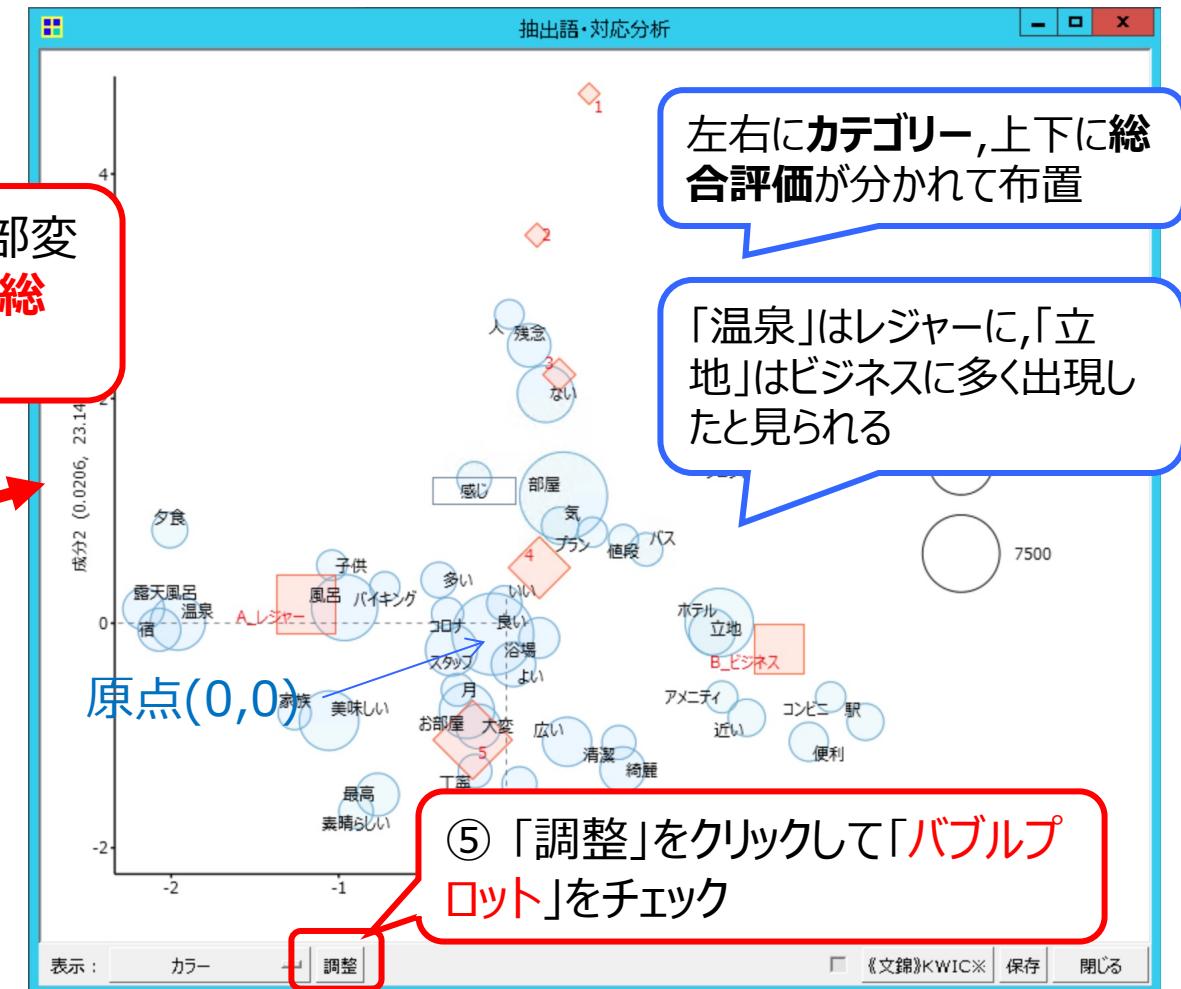
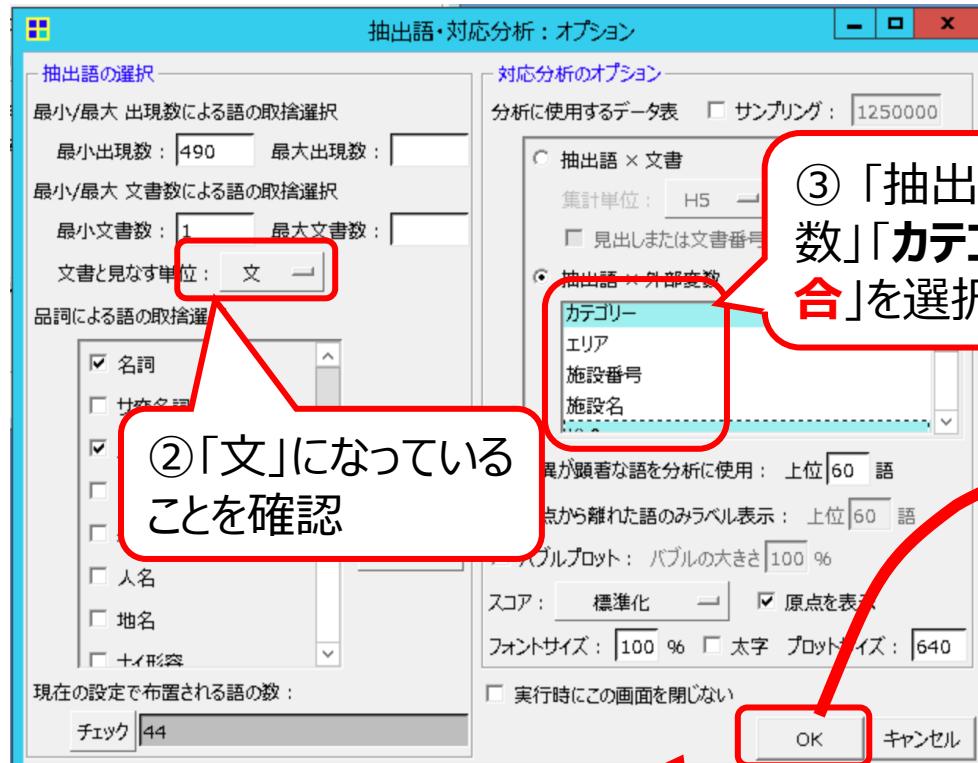
③「抽出語×外部変数」「カテゴリ」「エリア」を選択

④「OK」をクリック



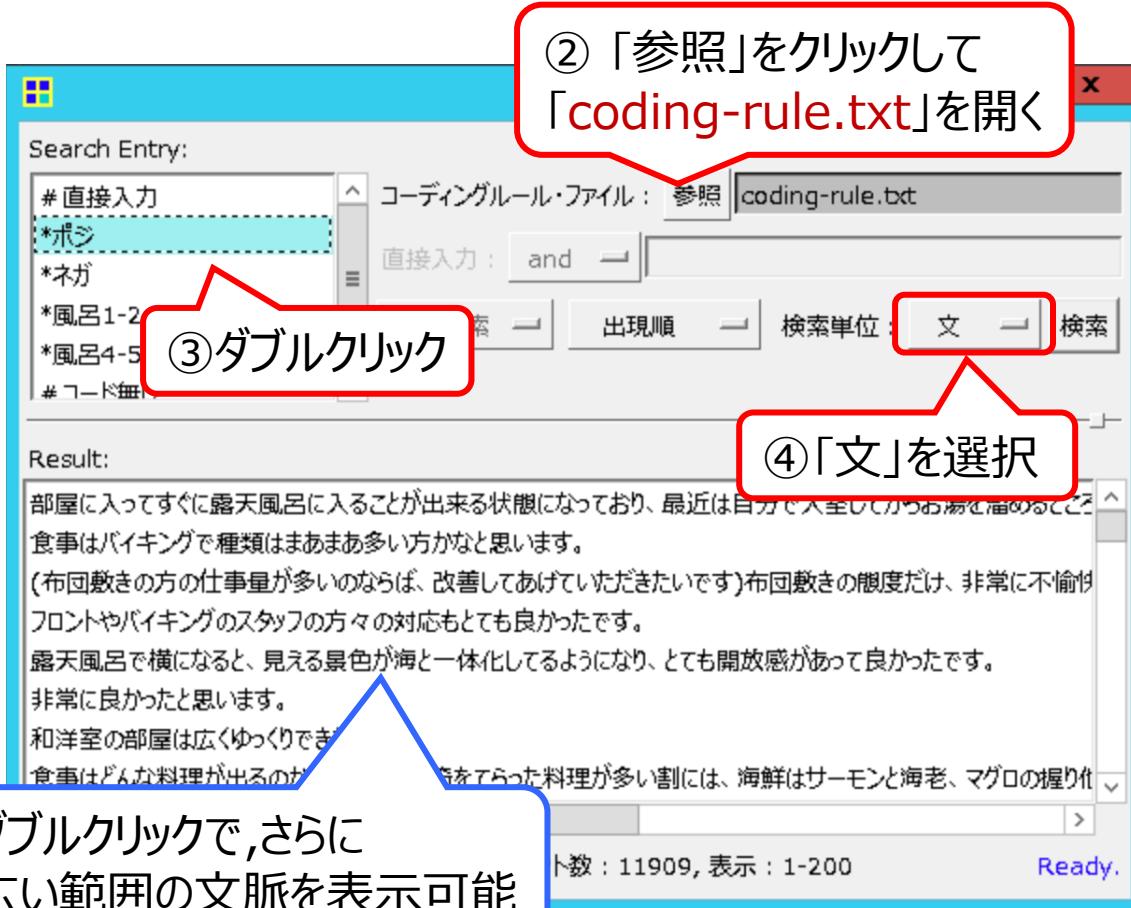
使い方 — 対応分析による探索2

①メニューから「ツール」「抽出語」「対応分析」を選ぶ



使い方 — 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



ダブルクリックで、さらに
広い範囲の文脈を表示可能

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

※ 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

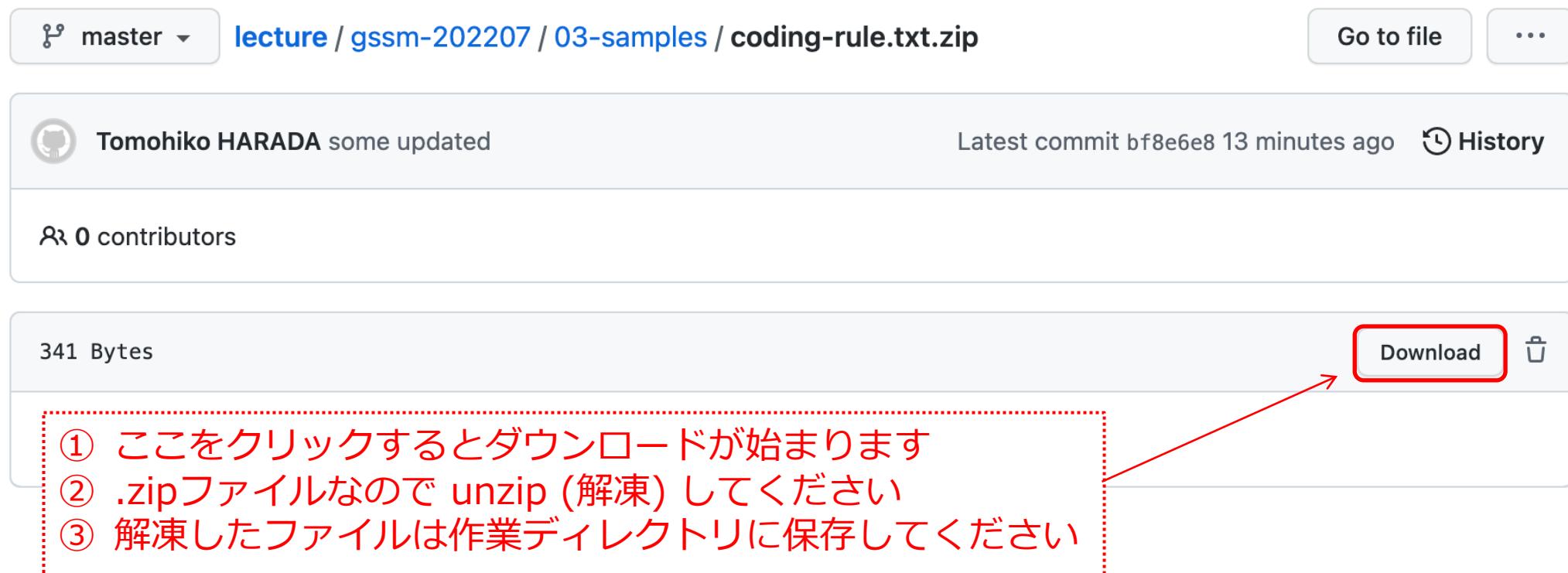
<>風呂-->4 | <>風呂-->5

外部変数

(参考) コーディングルールのサンプル

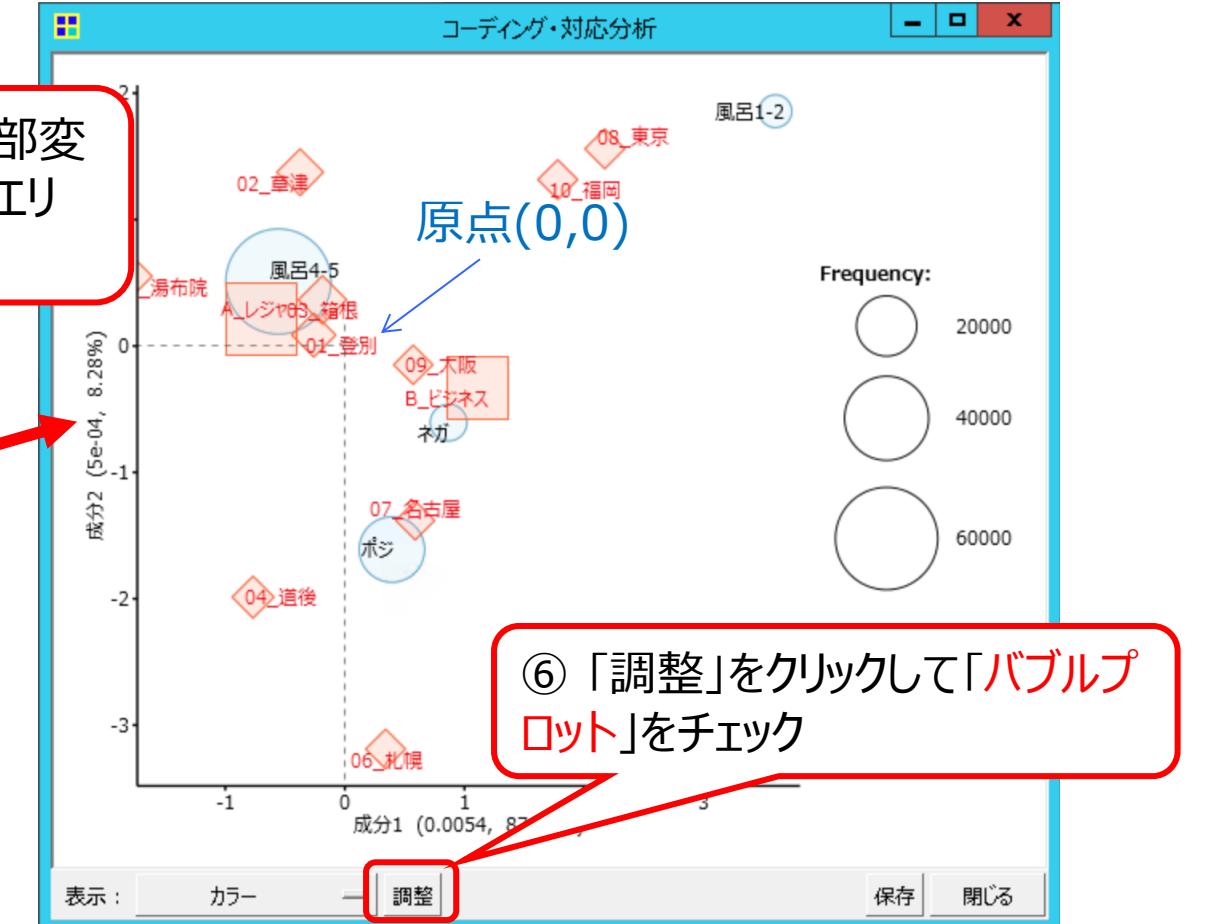
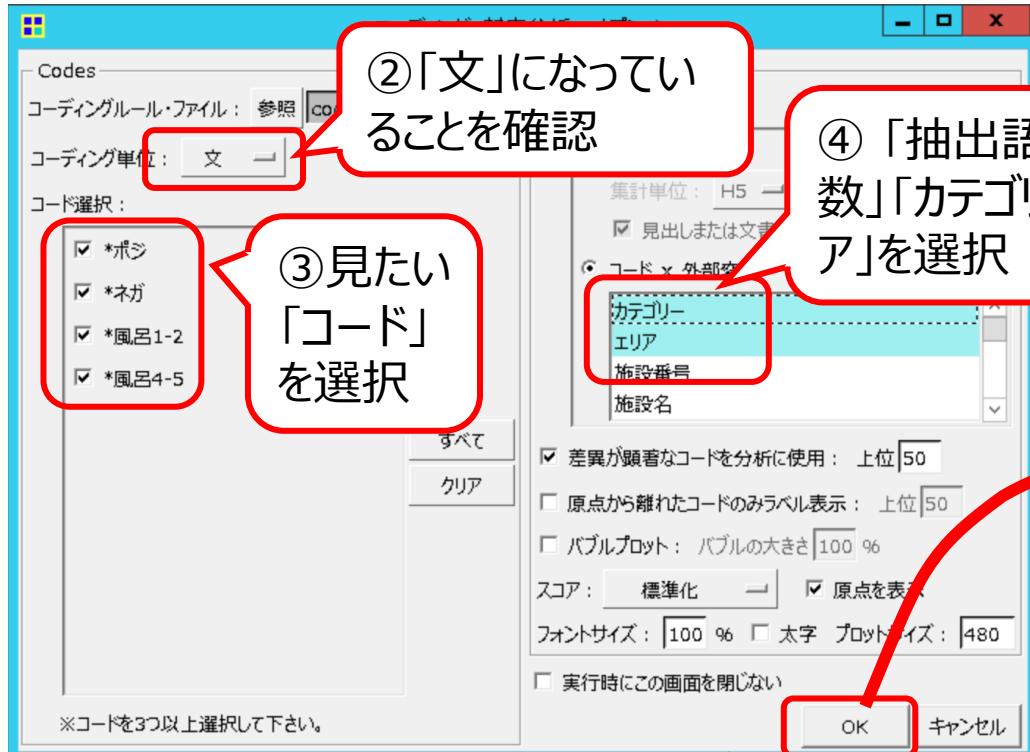
<https://github.com/haradatm/lecture/blob/master/gssm-202207/03-samples/coding-rule.txt.zip>

- Download ボタンをクリックするとダウンロードを開始



使い方 – 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」を選ぶ

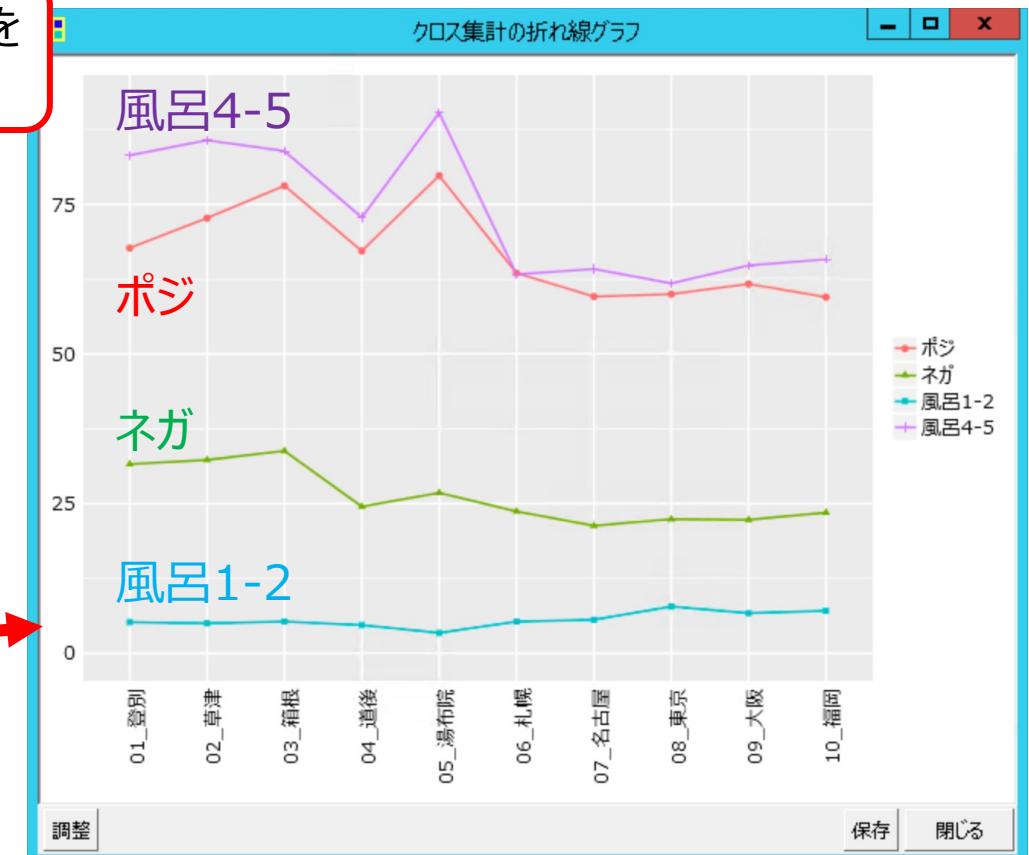
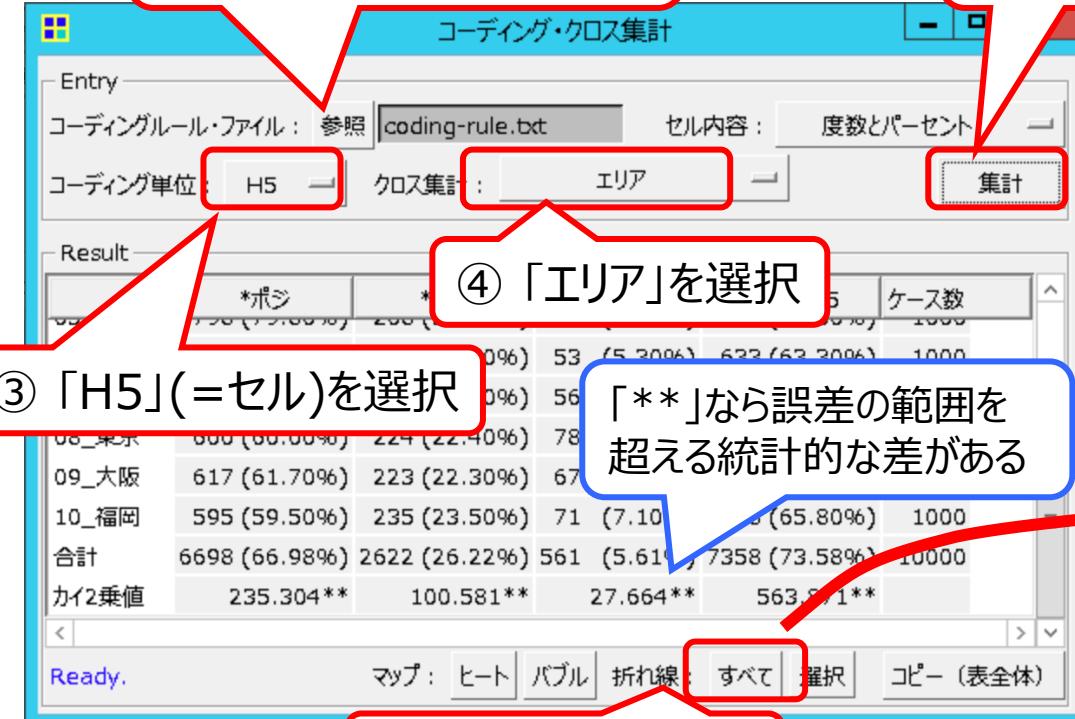


使い方 – クロス集計

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

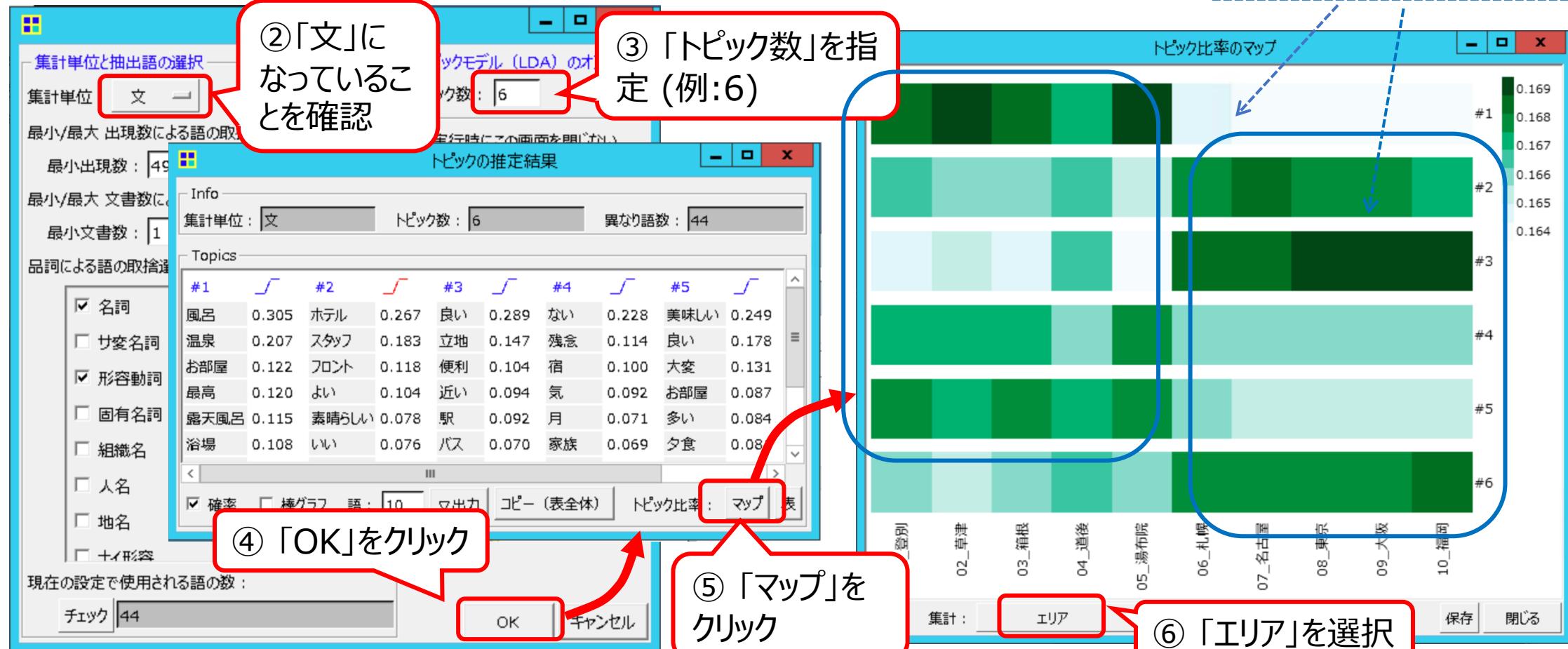
②「参照」をクリックして
「coding-rule.txt」を開く

⑤「集計」を
クリック



使い方 – トピックモデル

- 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



レジャーとビジネスで注目している観点(=トピック)が異なる

課題 — 数値評価と口コミの傾向比較

- 以下の 1点を **PDF ファイルで提出** してください
 - コーデ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を作成
 - 前ページで紹介したクロス集計を行い,**作成したプロットをPDFで提出**

形式: PDF, 提出先: manaba, 期限: 次週開始時刻(～18:20)

Q&A

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【第2版】 KH Coder オフィシャルブック. ナカニシヤ出版, 2020.
- [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一. 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
- New** [4] 樋口耕一. 動かして学ぶ! はじめてのテキストマイニング: フリー・ソフトウェアを用いた自由記述の計量テキスト分析 KH Coder オフィシャルブック II.ナカニシヤ出版, 2022.

(Windows環境によるデータ収集方法の参考に)

- [5] 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)

- [6] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [7] 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)

- [8] 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.
- [9] 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)

- [10] 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.