

人文社会ビジネス科学学術院 ビジネス科学研究群 2023年度 春C

テキストマイニング

day 3

スケジュール

day 1

- 講義 – テキストマイニング概説 (津田先生)
- 講義 – 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
- 演習 – テキスト解析 (1)
- 演習 – データ理解

day 3

- 演習 – テキスト解析 (2)
- 講義&演習 – データ分析 (使い方編)

day 4

- TextMining Studio の紹介
- 講義&講義 – データ分析 (実践編)

day 5

- 講義&講義 – データ分析 (実践編)

緊急追加 ～ ChatGPT登場後のテキストマイニングのカタチ(2) ～

ChatGPT — Code Interpreter

● 7/11: Code Interpreter 機能が、全 ChatGPT Plus ユーザーで利用可能

- ChatGPT 上で自然言語を入力して、データ分析やグラフ化ができる
- 「誰でもデータアナリストになれる時代が来る」と話題

年代別と性別のクロス集計により、投稿者の傾向を分析してください
rakuten-1000-2022-2023.xlsx

承知しました。まずは、提供いただいたファイルを確認してみましょう。その後、年代別と性別のクロス集計を行います。

データが正しく読み込まれました。このデータセットには以下の列が含まれています：

- ・カテゴリー
- ・エリア
- ・施設番号
- ・施設名
- ・コメント
- ・総合
- ・サービス
- ・立地
- ・部屋
- ・設備・アメニティ
- ・風呂
- ・食事
- ・旅行の目的
- ・同伴者
- ・宿泊年月
- ・投稿者
- ・年代
- ・性別

入力したデータの項目名を列挙



Finished working Show work

'年代'と'性別'の列にはそれぞれ6419の'na'値が含まれています。これは大部分のデータが欠けていることを意味します。

それでも分析を進めてよろしいでしょうか？または別のアプローチをとりましょうか？

進めてください

Finished working Show work

年代別と性別のクロス集計結果は以下の通りです：

	年代	女性	男性
10代	1	7	
20代	117	88	
30代	237	228	
40代	354	519	
50代	412	783	
60代	205	488	
70代	31	100	
80代	2	9	

この表から、次のような傾向が見て取れます：

- ・40代、50代、60代の男性が最も多くいます。
- ・年齢が上がるにつれて投稿数は減少傾向
- ・全年代において男性の投稿数が女性よりも多い

以上の情報は、投稿者の年齢と性別によるものです。

欠損が多いことを指摘
「…これは大部分のデータが欠けていることを意味します。それでも分析を進めてよろしいでしょうか？」

クロス集計表を作成
(グラフをリクエストすればグラフ化も可能)

結果の読解も可能

- 40代、50代、60代の男性が最も多く投稿しています。特に50代の男性の投稿数が最も多いです。
- 年齢が上がるにつれて投稿数は減少傾向にあります。
- 全年代において男性の投稿数が女性よりも多いです。

参考: <https://openai.com/blog/chatgpt-plugins#code-interpreter>

Code Interpreter (実行例: データ理解)

次に、エリアごとに各数値評価項目(総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事)の平均値をクロス集計し、結果をヒートマップで可視化するとともに、エリアごとの各数値項目の平均値の傾向を考察してください。

Finished working Show work ▾

4.10 4.03 4.27 3.95 3.88 4.31 4.08
4.25 4.19 4.28 4.03 3.92 4.31 4.15
4.23 4.22 4.15 4.12 3.97 4.22 4.28
4.26 4.16 4.41 4.10 4.00 4.09 4.21
4.28 4.36 4.27 4.50 4.57
4.37
4.26
4.14 3.97 4.34
4.24 4.06 4.34
4.18 4.01 4.40

クリックするとコードを表示

Heatmap from below shows the following trends:

- 「05_湯布院」は全体的に高い評価を得ています。
- 温泉地として知られる「01_登別」、「02_琴似」は他のエリアに比べて高いです。
- 大都市エリアである「06_札幌」、「07_大阪」の評価が他のエリアに比べて高いです。
- 「03_箱根」は全体的な評価が平均的で、どの評価項目でも特に目立った評価はありません。
- 「09_大阪」は「立地」の評価が高い一方、「風呂」の評価が他のエリアに比べて低いです。

Regenerate response

クロス集計のヒートマップ

Finished working Hide work ^

python

```
# Create a heatmap with English labels
plt.figure(figsize=(12, 8))
sns.heatmap(mean_ratings, annot=True, fmt=".2f", cmap='YlGnBu', linewidths=.5)
plt.title('Average Ratings by Area')
plt.xlabel('Rating Item')
plt.ylabel('Area')

# Change labels to English
plt.xticks(ticks=[0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5], labels=['Overall', 'Service', 'Location', 'Room', 'Facilities', 'Bath', 'Meal'])
plt.yticks(ticks=[0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5], labels=['01_Noboribetsu', '02_Kusatsu', '03_Hakone', '04_Dogo', '05_Yubin', '06_Sapporo', '07_Osaka', '08_Tokyo', '09_Nagoya', '10_Fukuoka'])

plt.show()
```

プロットラベルを日本語にするようリクエストした

申し訳ありませんが、現在の環境では日本語フォントが利用できないため、プロットのラベルやタイトルを日本語で表示することはできません。

その代わり、エリアと評価項目の英語版を作成することは可能です。それでよろしければ、英語のラベルとタイトルでヒートマップを作成します。

お願いします

Finished working Show work ▾

Area

	Overall	Service	Location	Room	Facilities	Bath	Meal
01_Noboribetsu	4.10	4.03	4.27	3.95	3.88	4.31	4.08
02_Kusatsu	4.25	4.19	4.28	4.03	3.92	4.31	4.15
03_Hakone	4.23	4.22	4.15	4.12	3.97	4.22	4.28
04_Dogo	4.26	4.16	4.41	4.10	4.00	4.09	4.21
05_Yubin	4.54	4.52	4.28	4.36	4.27	4.50	4.57
06_Sapporo	4.20	3.99	4.37	4.09	3.92	3.81	4.17
07_Osaka	4.16	3.98	4.26	4.06	3.92	3.82	3.99
08_Tokyo	4.14	3.97	4.34	4.11	3.91	3.73	3.99
09_Nagoya	4.24	4.06	4.34	4.13	3.96	3.86	4.12
10_Fukuoka	4.18	4.01	4.40	4.11	3.89	3.85	4.02

ラベルを翻訳して英語表記

こちらはエリア別の各評価項目の平均値を表示したヒートマップで、ラベルとタイトルは英語で表示されています。このヒートマップから、各エリアがどの評価項目で高評価を得ているのか、または低評価であるのかが一目で分かります。これらの情報は、各エリアの強みや弱みを理解するための一助となるでしょう。

Code Interpreter (実行例: テキスト分析)

- 実習用データの「コメント」列を読んで、「エリア」ごとの特徴を要約して説明してください

The screenshot shows a conversation between a user and ChatGPT. The user asks about the differences in features between the 'category' and 'comment' columns. ChatGPT responds by suggesting various methods to analyze the data, including using the 'janome' library for Japanese text analysis or the 'sudachipy' library. It also mentions calculating word frequency and comparing words across regions. A red box highlights the text "色々とやりとりをしているうちに、それらしい分析結果を出力してくれることも…". Another red box highlights the text "データの説明と分析方法の提案". A third red box highlights the text "日本語形態素解析器など新たなライブラリはインストールできないらしい". A fourth red box highlights the text "以下に、各エリアのコメントを要約した文章を示します". A fifth red box highlights the numbered list of regional comments.

「カテゴリー」列の値が「A_レジャー」のデータについて、「コメント」列のテキストから各エリアの特徴の違いを説明してください。

rakuten-1000-2022-2023.xlsx

色々とやりとりをしているうちに、それらしい分析結果を出力してくれることも…

以下に、各エリアのコメントを要約した文章を示します：

- 登別：顧客は食事、特に夕食について多くコメントしており、部屋とお風呂の設備も注目されています。また、再訪を考えているという意見も多く見られました。
- 草津：再度利用したいという意見が多く、食事や部屋の設備についてのコメントが多いです。また、お世話になったと感謝の意を表すコメントも多く見られました。
- 箱根：箱根の顧客は食事や部屋、お風呂について特にコメントをしています。再訪を希望する声も多く、その地域を評価しています。
- 道後：道後の顧客は部屋の設備や立地条件について特にコメントをしています。また、再訪したいという意見も多く、特に夕食についての満足度が高いようです。
- 湯布院：湯布院の顧客は食事についてのコメントが多く、特に夕食に対する満足度が高いようです。また、「また行きたい」という意見も多く見られ、お世話になったという感謝の意も見受けられました。
- 札幌：札幌の顧客は朝食について特に満足しており、再訪したいという意見も多く見られました。立地についても評価が高いようです。
- 名古屋：名古屋の顧客は滞在が快適であったとコメントしており、立地やWi-Fiについても特に好評でした。また、再訪したいという意見も多く見られました。
- 東京：東京の顧客は施設の立地について特にコメントをしており、駅からの近さを評価しています。また、再訪したいという意見も多く見られました。
- 大阪：大阪の顧客は部屋や朝食、立地について満足度が高く、お風呂についても特に好評でした。また、再訪する意向が強いようです。
- 福岡：福岡の顧客は朝食について特に満足しており、再訪したいという意見も多く見られました。また、お世話になったという感謝の意も見受けられました。

これらの要約は、顧客が各エリアのどの部分を特に評価しているかを示しています。それは食事、部屋の設備、立地、お風呂など、宿泊施設の様々な側面に及んでいます。再訪したいという意見も各エリアで共通して見られ、施設の満足度が高いことを示しています。

データの説明と分析方法の提案

日本語形態素解析器など新たなライブラリはインストールできないらしい

Model: Code Interpreter

の結果をエリアごとに文章として要約してください

Finished working Show work ▾

提供されたデータは、カテゴリー、エリア、施設番号、施設名、コメント、総合、サービス、立地、部屋、設備・アメニティ、風呂、食事、旅行の目的、同伴者、宿泊年月、投稿者、年代、性別というカラムを含むレビューデータのようです。

「カテゴリー」列の値が「A_レジャー」のデータについて、「コメント」列のテキストから各エリアの特徴の違いを説明するためには、まず「A_レジャー」のカテゴリーに属するデータをフィルタリングし、それから各エリアのコメントを分析する必要があります。テキスト分析には様々な手法がありますが、一般的には頻出単語の分析や感情分析を行います。

まず、「カテゴリー」列の値が「A_レジャー」のデータをフィルタリングしましょう。その後、各エリアで最も頻繁に現れる単語を調べます。ただし、一部の一般的な単語（助詞、接続詞など）は頻繁に出現しますが、エリアの特徴を理解するのにはあまり役立たない可能性があります。そのため、これらの一般的な単語（ストップワード）は分析から除外します。そして、各エリアで最も頻繁に現れる単語を比較することで、エリア間の特徴の違いを理解する手がかりを得ることができます。

以上の手順で進めてよろしいでしょうか？

頻出単語を分析する方法で進めてください

Finished working Show work ▾ Regenerate response

申し訳ありませんが、この環境では「janome」という日本語のテキスト分析に必要なライブラリが利用できません。

しかし、代わりに「sudachipy」を使用して日本語のテキスト解析を行うことが可能です。「sudachipy」は形態素解析器の一つで、文を単語に分割するなどの処理を行います。

それでは、「sudachipy」を使ってテキスト解析を進めてみましょう。まずは「sudachipy」の形態素解析を使用して、各エリアのコメントから名詞を抽出し、その頻度を計算します。その結果を基に、各エリアの特徴を説明します。

申し訳ありませんが、この環境では「sudachipy」という日本語のテキスト分析に必要なライブラリも利用できません。

残念ながら、この環境では日本語のテキスト解析を行うためのライブラリが限られています。そのため、頻出単語の分析を行うことは難しい状況です。

一方で、それぞれのエリアについて、コメントの数や評価の平均値など、基本的な統計量を計算することは可能です。これにより、各エリアの評価の高さや、どのエリアが最もコメントが多いなどの情報を得ることができます。

この方法で進めてみましょうか？

Send a message

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

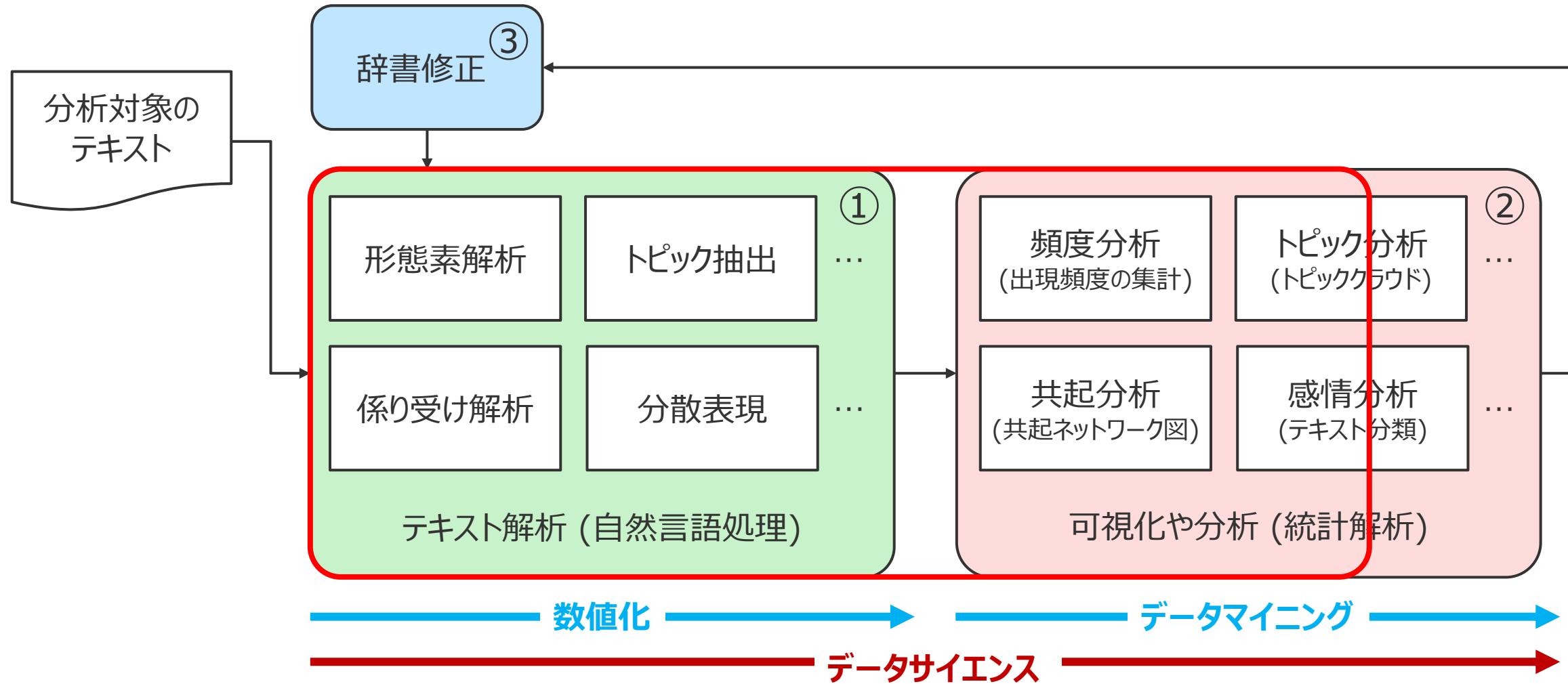
テキスト解析 (2)

(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是?
 - 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

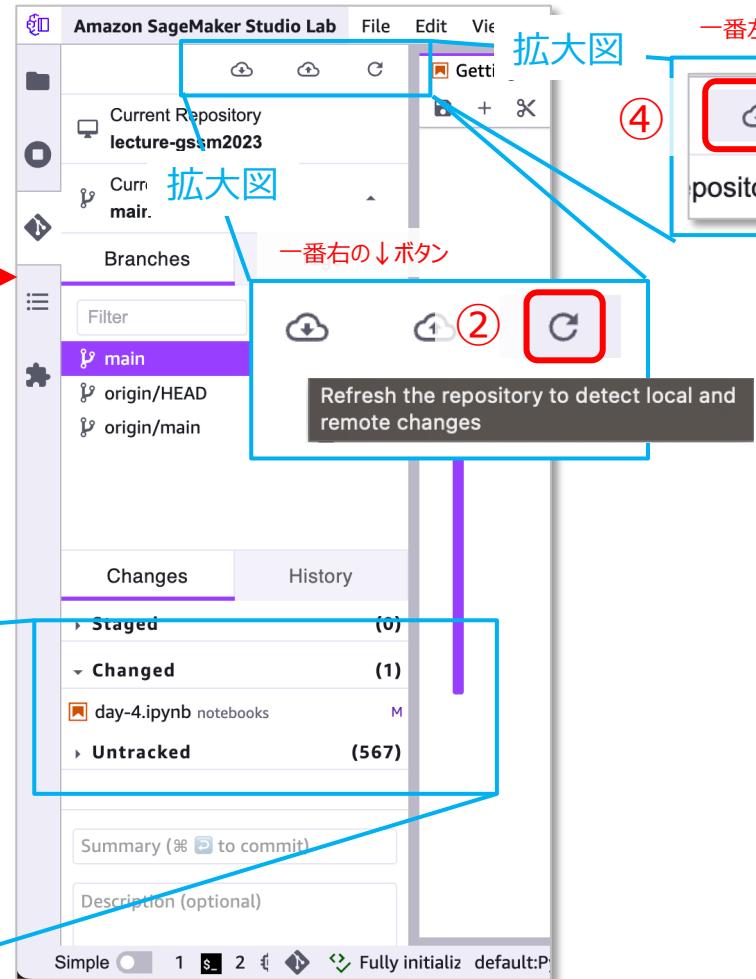
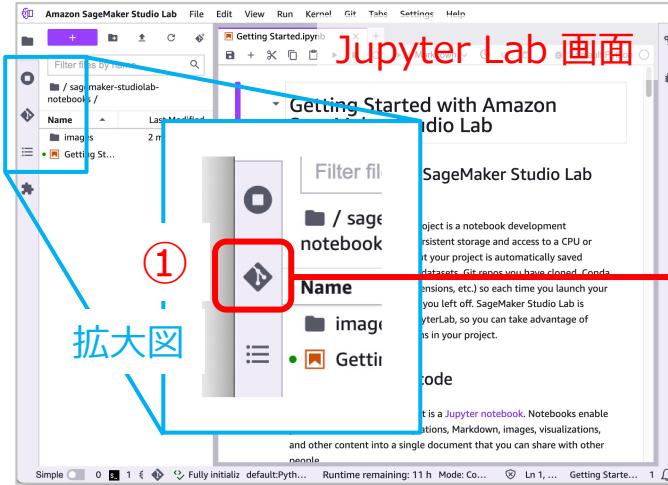
(再掲) テキスト分析の手順

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)

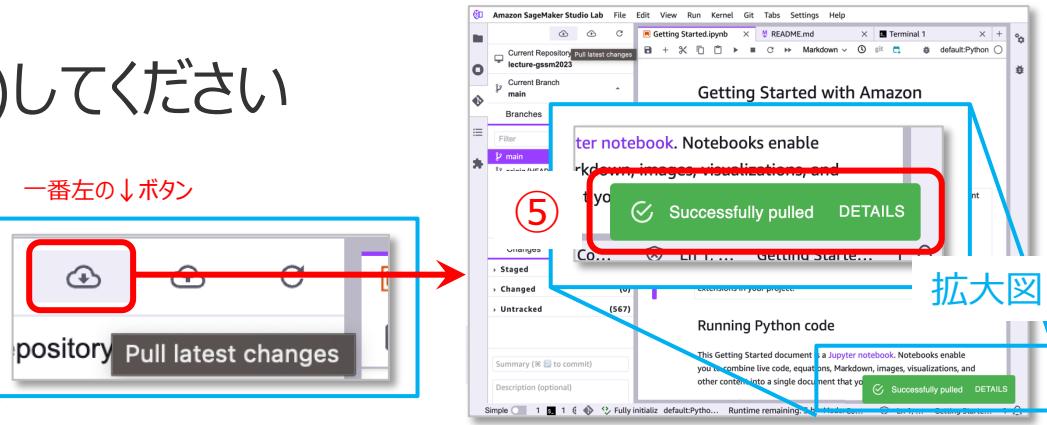
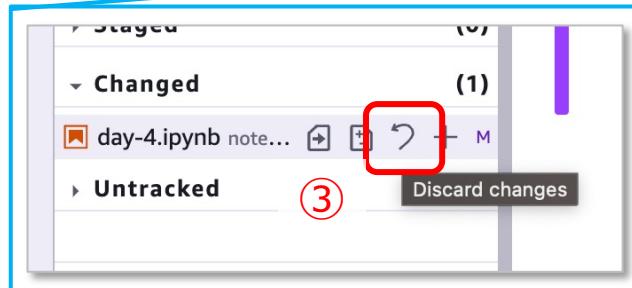


演習 — テキスト解析 (2)

● Jupyter Lab を起動して、教材を最新化(pull)してください



(例) 競合がある場合のみ 拡大図



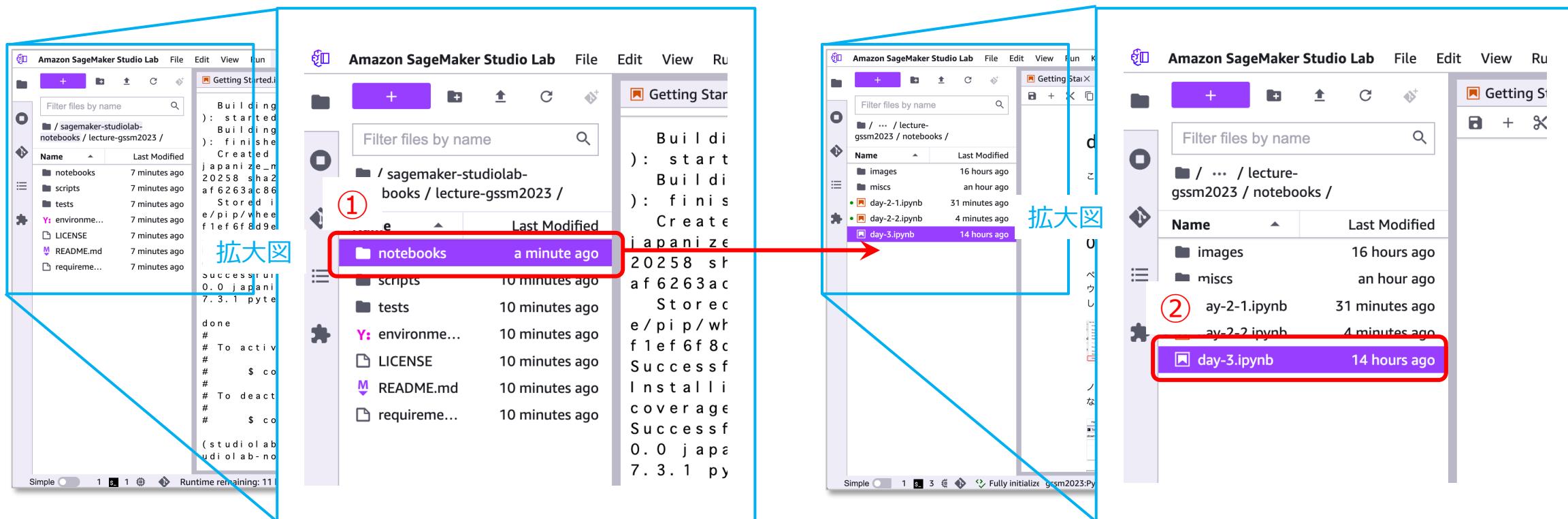
● 教材を最新化する

- ① 「Git」 ボタンを押す
- ② 「Refresh」 ボタンを押す
- ③ もし、競合がある場合(Changedが0でない場合)、
対象のファイルを手動で退避した後、
「Discard changes」 ボタンを押し
て変更を破棄する
- ④ 「Pull latest changes」 を押す
- ⑤ 画面の右下に「Successfully published」が表示されること確認する

演習 — テキスト解析 (2)

● clone したファイルを最新化(pull)してください

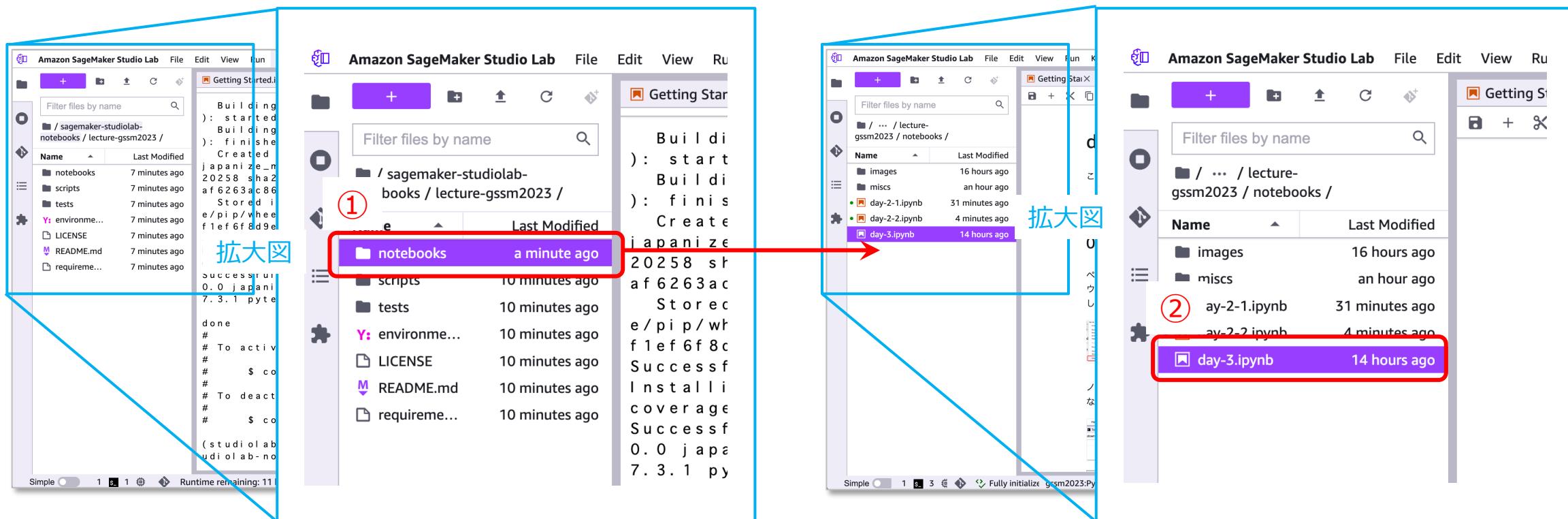
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-3.ipynb** ノートブックを開く



演習 — テキスト解析 (2)

● day-3.ipynb を開いてください

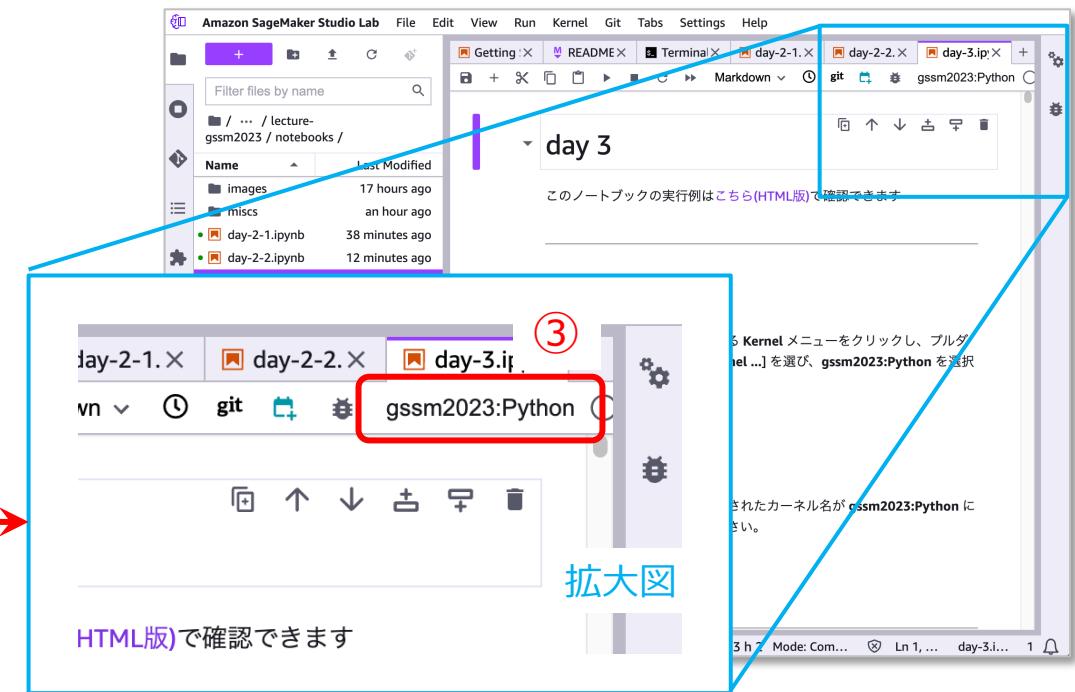
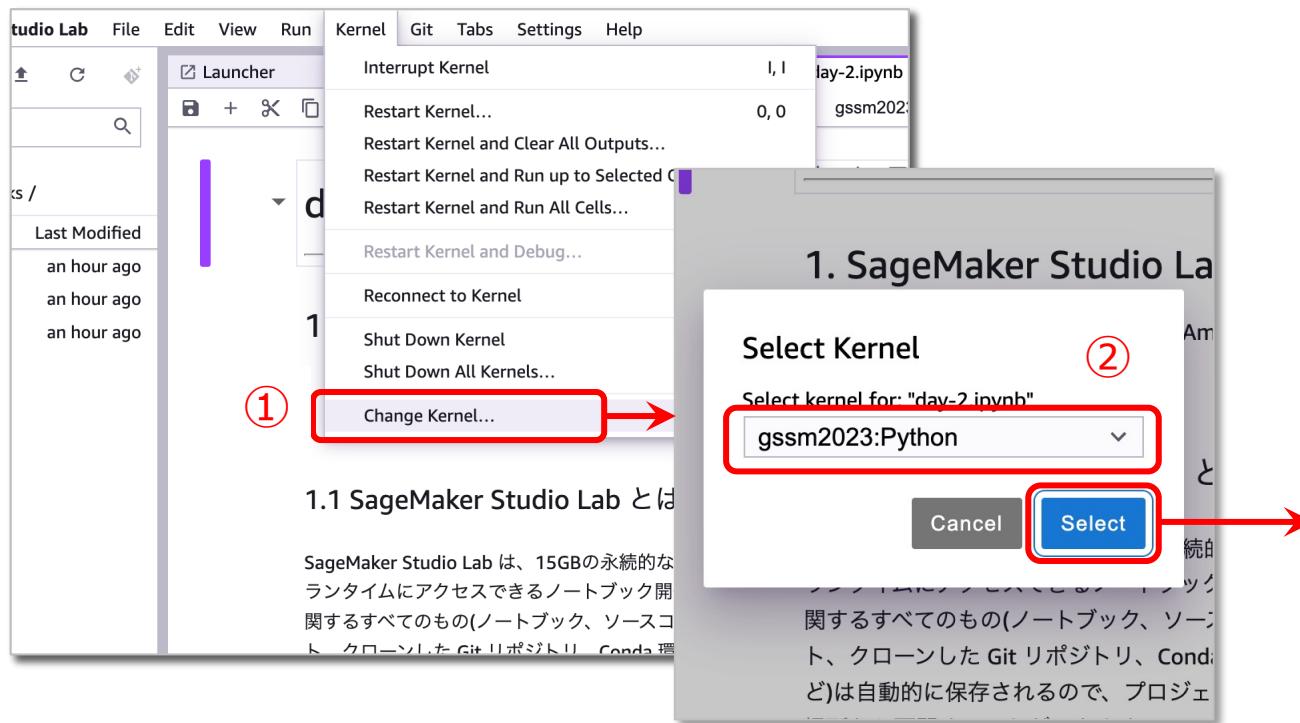
- ① 画面左の File Browser から ① notebooks をフォルダを開く (既に開いている場合はスキップ)
- ② 次に day-3.ipynb ノートブックを開く



演習 — テキスト解析 (2)

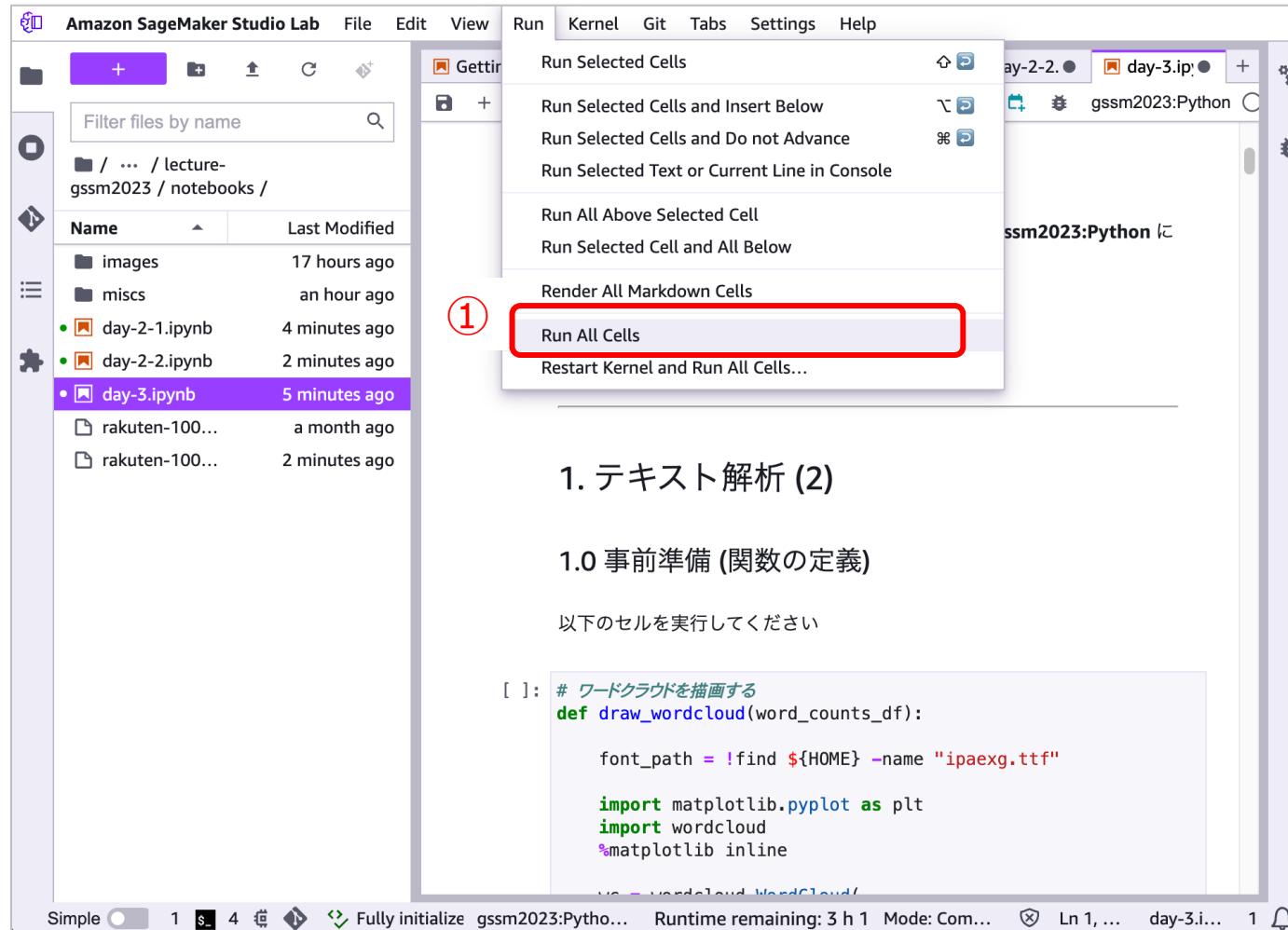
● カーネル gssm2023:Python を選択してください !重要!

- ① ページ上部の **Kernel** メニューから「Change Kernel ...」を選ぶ
- ② ポップアップ画面から「gssm2023:Python」を選択し、「Select」を押す
- ③ 右上隅にカーネル名「gssm2023:Python」が表示されていることを確認する



演習 — テキスト解析 (2)

● テキスト解析と可視化



The screenshot shows the Amazon SageMaker Studio Lab interface. On the left, there's a file browser with a list of notebooks. In the center, a code editor cell contains Python code for generating a word cloud. A context menu is open over the cell, with the 'Run All Cells' option highlighted by a red box and a circled number 1.

```
[ ]: # ワードクラウドを描画する
def draw_wordcloud(word_counts_df):
    font_path = !find ${HOME} -name "ipaexg.ttf"
    import matplotlib.pyplot as plt
    import wordcloud
    %matplotlib inline
```

演習:

- ① ページ上部の Run メニューから「Run All Cells」を選ぶ

この後、Step-by-step で解説します

テキストマイニングツール KHCoder の紹介

(再掲) 実習環境について

- 以降の実習では、**KHCoder** (フリーソフトのテキストマイニングツール) を使用します
- KHCoder の利用には Windows OS (10 or 11) が必要になります

PC の種類	Windows OSの有無	方法	備考
Windows	有り	Windows PC に KHCoder をインストールして使用する	最もオススメ
		全学計算機システムの Windows [※] に KHCoder をインストールして使用する	※ 利用手順: https://www.u.tsukuba.ac.jp/remote/#vm2win
Mac	仮想環境 [※] 上で動く Windows がある	仮想環境上の Windows に KHCoder をインストールして使用する	※ Vmware Fusion や Parallels Desktop などを想定
	なし	全学計算機システムの Windows [※] に KHCoder をインストールして使用する	※ 利用手順: https://www.u.tsukuba.ac.jp/remote/#mac2win
		SageMaker Studio Lab 上の Python スクリプト [※] を利用する	※ 救済的な措置で、一部の KHCoder 機能は未対応です

(再掲) KH Coder のインストール

- 前ページを参考に、各自で選択した環境(個人PC or 全学計算機システム)に **KH Coder** をインストールしてください
- 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. 任意の保存先を指定します (**全学計算機ではCドライブへの保存は禁止されています**)
例: 「Z:¥Desktop¥khcoder3」 (全学の場合)
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 とは

- 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>

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)

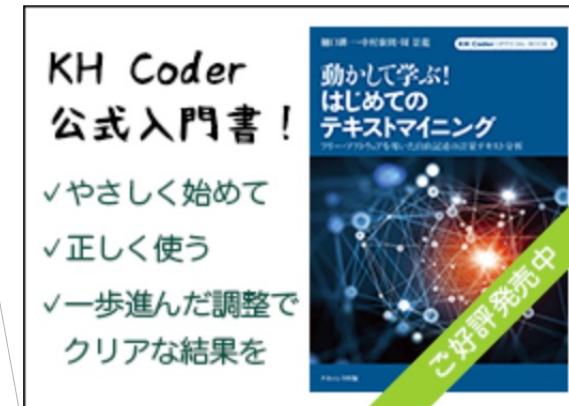
KH Coder の情報

ホームページ <http://khcoder.net/>

The screenshot shows the main website for KH Coder. It includes:

- Index:** A large banner with the text "KH Coder" in blue, stylized letters.
- 概要:** Text about the seminar and analysis methods.
- 機能紹介 (スクリーンショット):** Screenshots of the software interface.
- ダウンロードと使い方:** Download links and usage instructions.

参考書



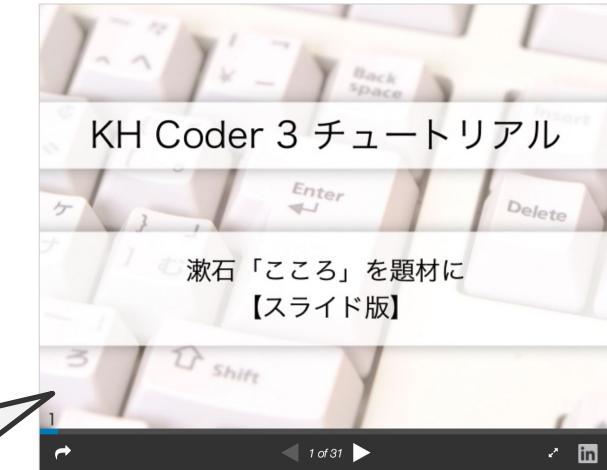
PDFファイルをダウンロードすることもできます。
※Windows版パッケージにはあらかじめ同梱してあります。

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。

チュートリアル

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

チュートリアル & ヒント



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。

チュートリアル用データ

[tutorial-data-3x.zip](#) (2018 04/25)

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。

KH Coder — 分析手法 (1)

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。



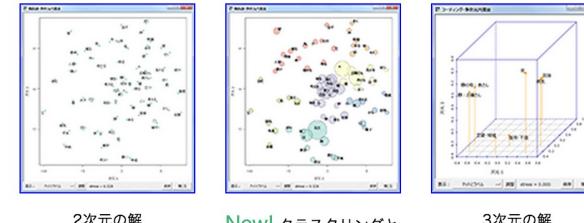
共起の程度が非常に強いものだけを線で結んだ図

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	説明
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)	<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 (コレスポンデンス分析)	<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、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

KH Coder – 分析手法 (2)

トピックモデル (LDA)

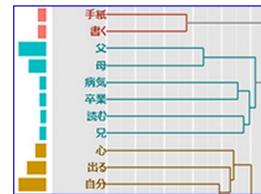
文書ごとにトピックの出現割合を表示したり、各トピックに高い確率で出現する語を表示できます。



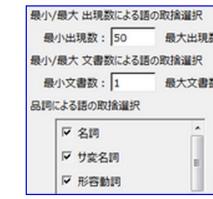
文書ごとのトピック比率

階層的クラスター分析

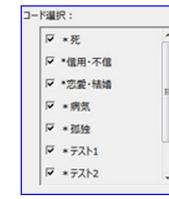
抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけではなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



抽出語は出現数や品詞で選択



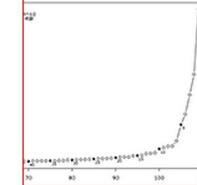
コードはチェックボックスで直接選択

文書のクラスター分析

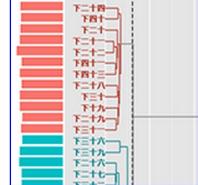
文書の分類を行うクラスター分析です。



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

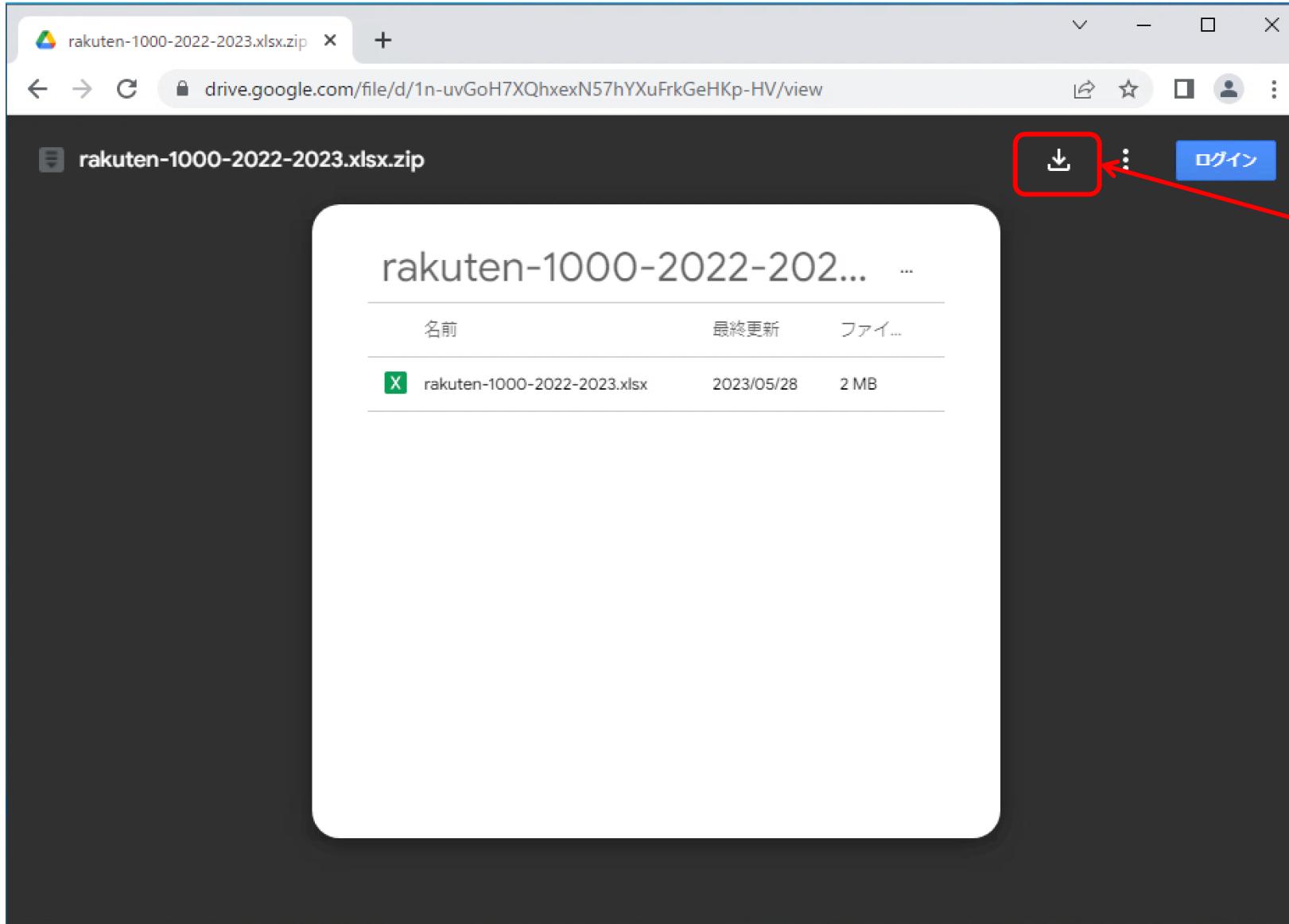
テキスト分析（使い方編）

(再掲) 実習用データ — ファイル一覧

● 実習用データは以下の通り → 主に「**rakuten-1000-2022-2023.xlsx**」を使用する

ファイル名	件数	データセット	備考
<u>rakuten-1000-2022-2023.xlsx.zip</u>	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: 2022/1~2023 GW明け	本講義の全体を通して使用する
<u>rakuten-1000-2020-2021.xlsx.zip</u>	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: 2020/1~2021/12	演習用 (年度で比較する場合など)
<u>rakuten-all-2022-2023-tsv.zip</u>	142,061	<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: 2022/1~2023 GW明け	参考用
<u>rakuten-all-2020-2021-tsv.zip</u>	198,885	<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: 2020/1~2021/12	参考用
<u>rakuten-all-tsv.zip</u>	1,659,396	<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	参考用

(参考) Google Drive ダウンロード画面

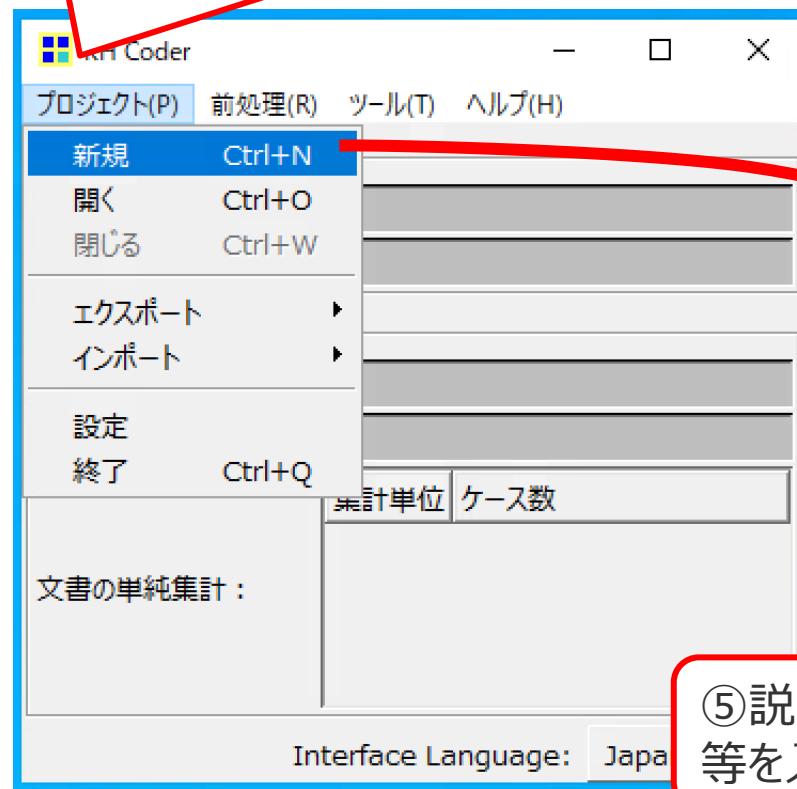


ここをクリックすると
ダウンロードが始ま
ります

KHCoder の使い方

● 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-1000-2022-2023.xlsx」を開く

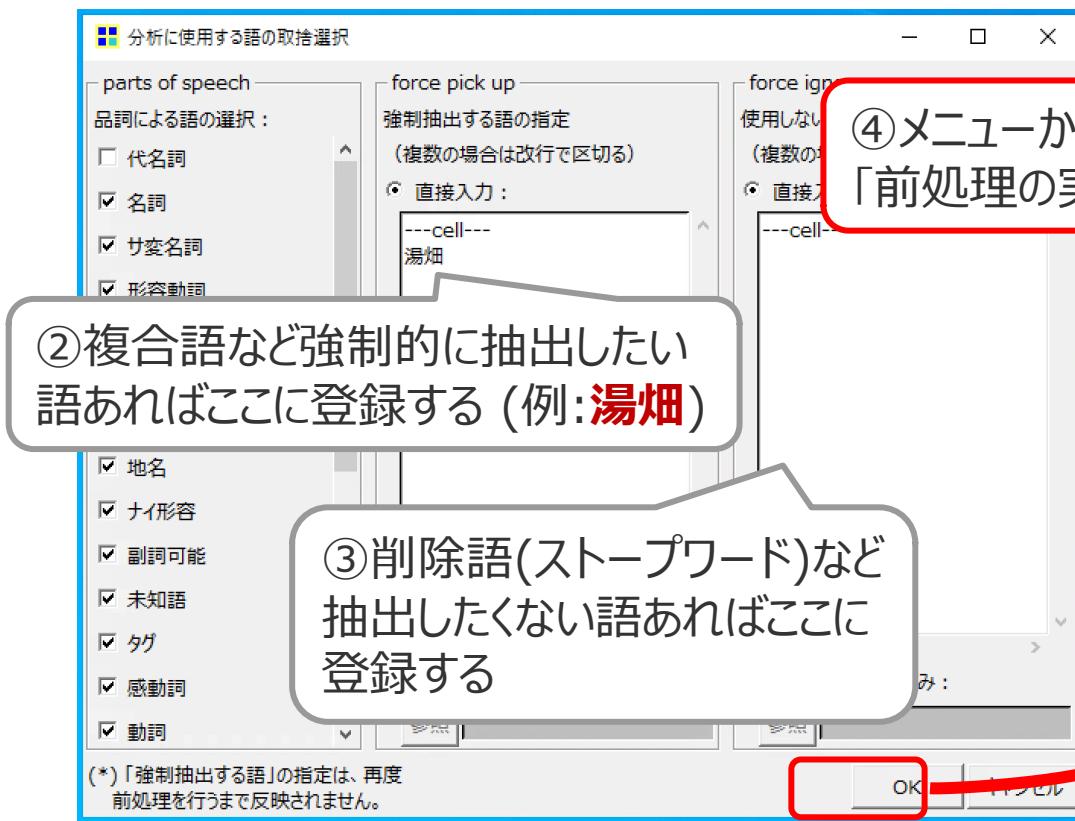


- ⑤説明「楽天トラベル」
等を入力

KHCoder の使い方

● 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



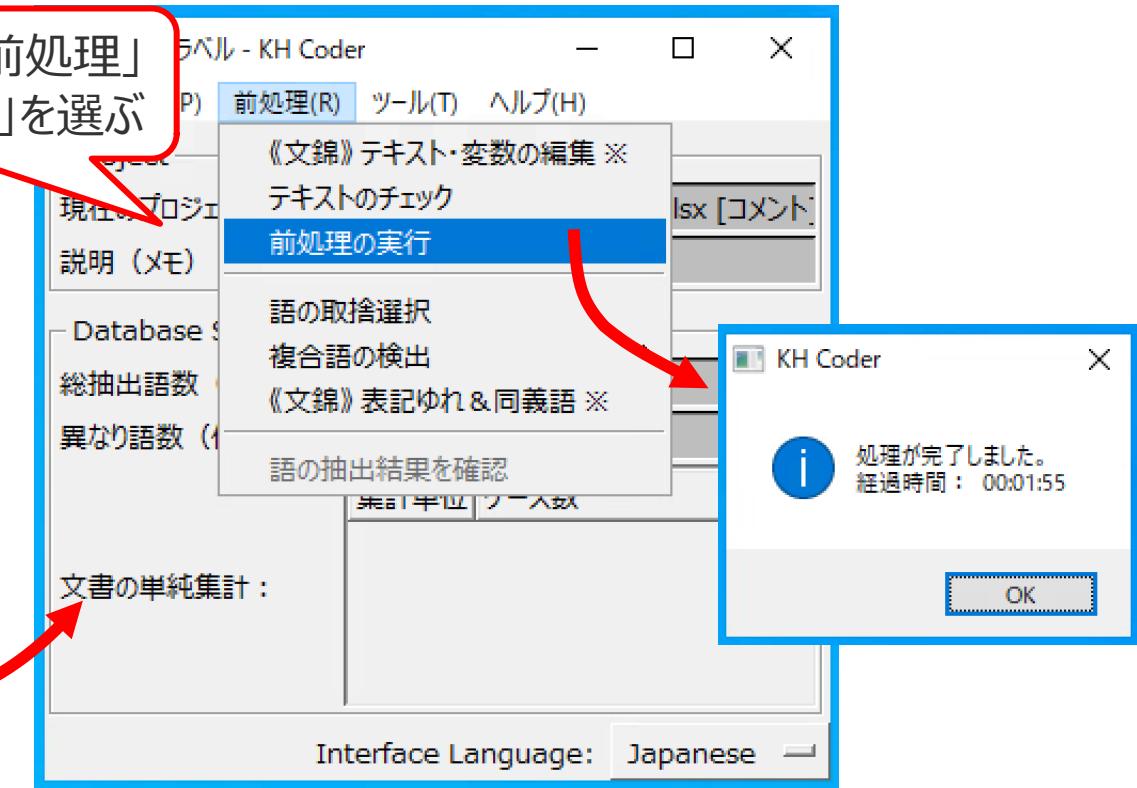
②複合語など強制的に抽出したい語あればここに登録する (例:湯畑)

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

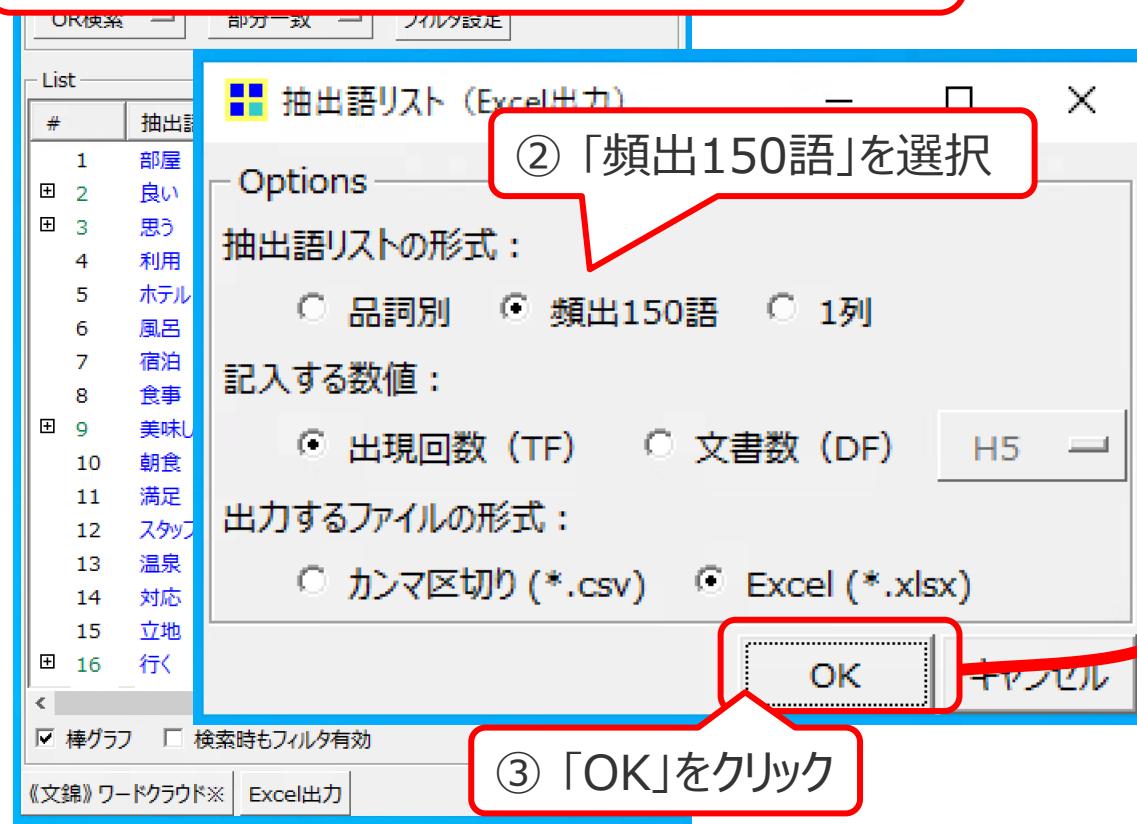
注2: メニューから「前処理」「複合語の検出」を選ぶと,複合語候補の一覧を出力できます



KHCoder の使い方

● 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



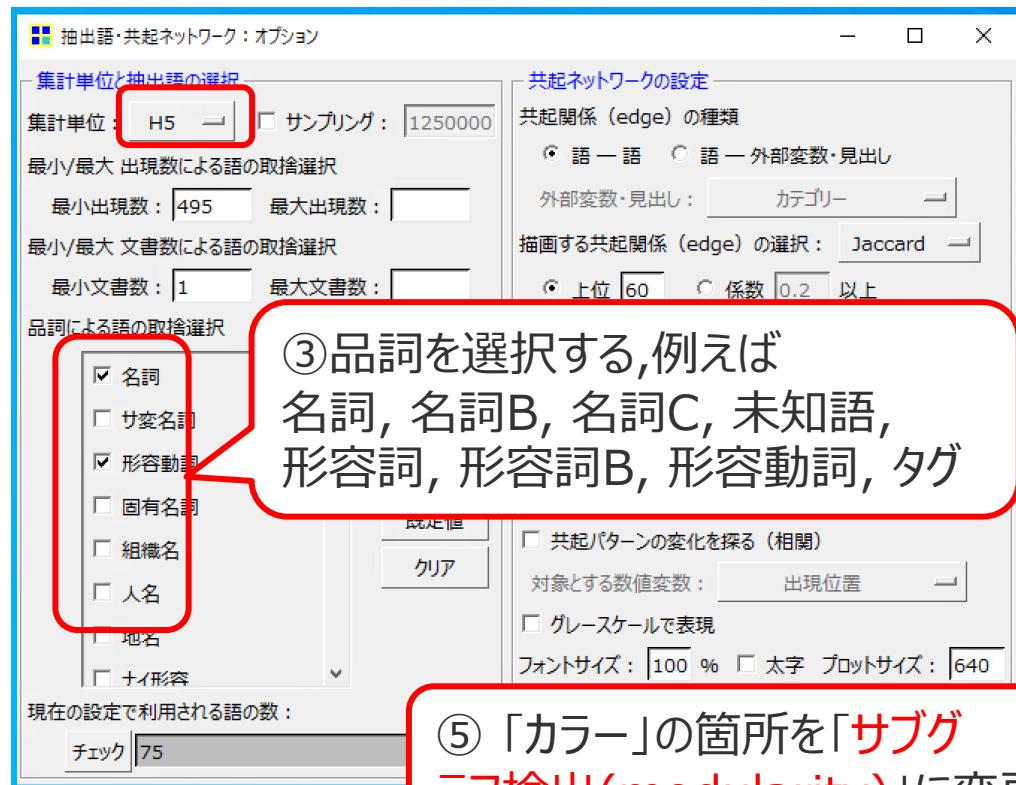
A	B	C	D	E	F	G	H
1	抽出語	出現回数	抽出語	出現回数	抽出語	出現回数	
2	部屋	6689	子供	661	プラン	389	
3	良い	4302	過ごす	657	見える	388	
4	思う	3976	家族	648	機会	387	
5	利用	3481	予約	636	設備	387	
6	ホテル	2831	過ごせる	626	旅館	386	
7	風呂	2702	駐車	613	置く	384	
8	宿泊	2649	素晴らしい	612	きれい	377	
9	食事	2447	月	611	歩く	368	
10	美味しい	2249	バス	610	湯	359	
11	朝食	2172	丁寧	610	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

KHCoder の使い方

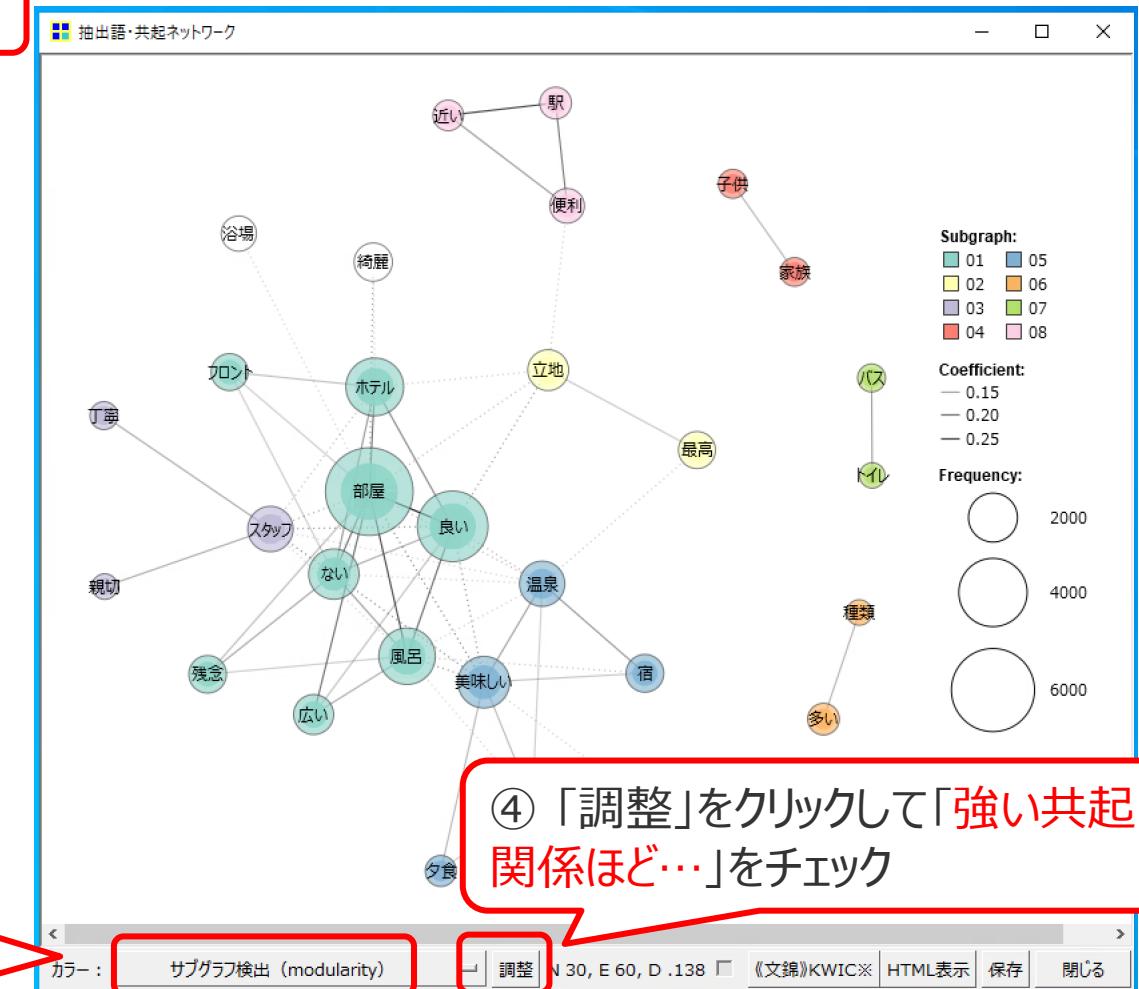
● 共起ネットワークの作成(1)

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「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



(参考) 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

出典: KH Coder 3 リファレンス・マニュアル

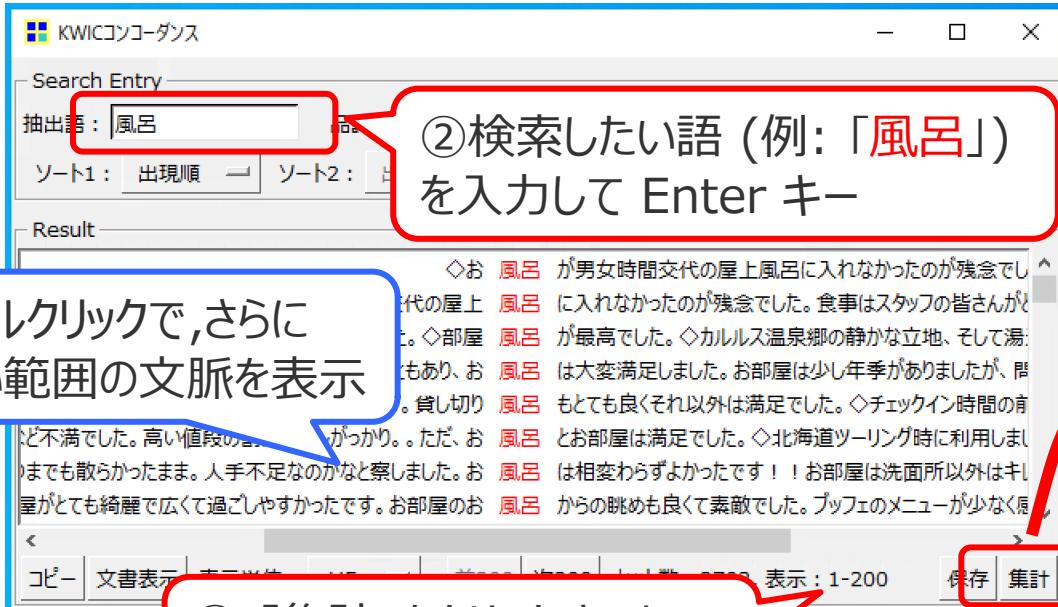
注: どの品詞を選択すべきかは、分析対象のデータや分析目的により異なります。

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。

KHCoder の使い方

● 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



「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 104 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, 形容動詞)

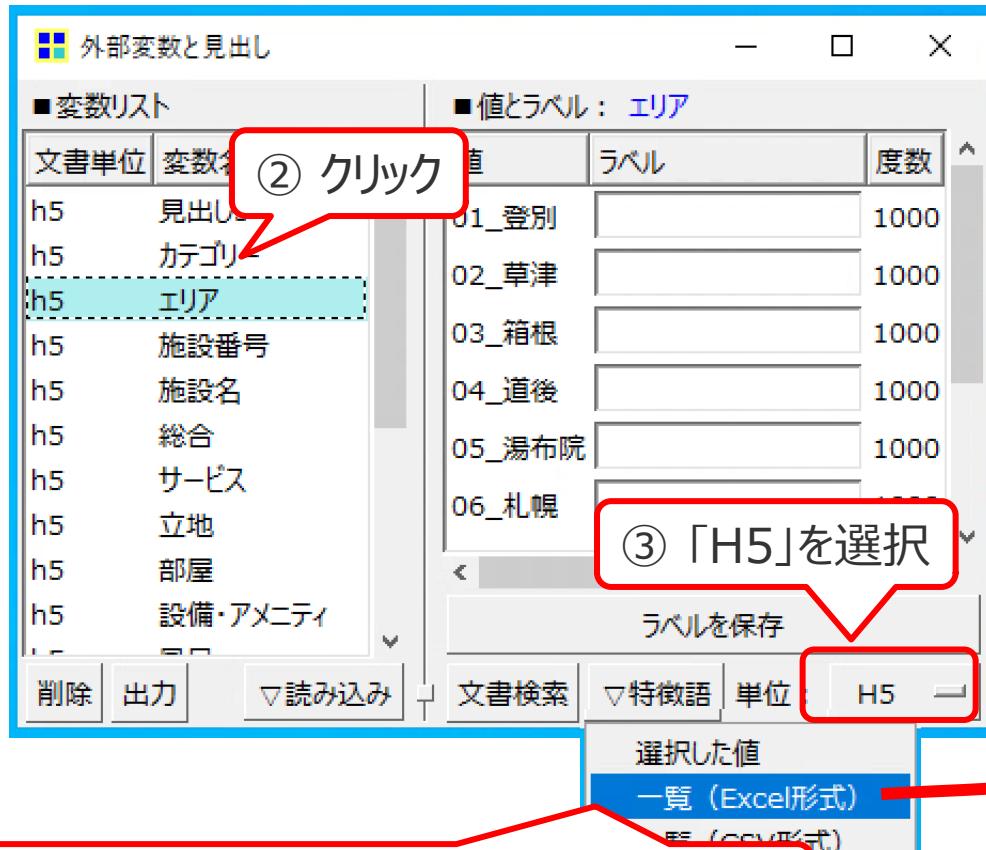
⑤「右合計」でソート

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	広い	形容詞	190	41	149	8	5	13	13	2	1	104	20	18	6	81.0
2	良い	形容詞	222	77	145	40	9	14	13	1	6	48	43	21	27	77.4
3	最高	名詞	105	13	92	5	3	3	2			44	12	22	8	42.8
4	部屋	名詞	439									22	20	18	172.7	
5	トイレ	名詞	117									1	4	4	60.2	
6	露天風呂	名詞	87									9	13	13	30.6	
7	風呂	名詞	130									31	14	20	35.6	

KHCoder の使い方

● 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



	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	ピング	.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	.072							
30	フロント	.067	綺麗	.067							
31	快適	.064	トイレ	.064							
32	広い	.064	コンビ	.064							
33	近い	.064	よい	.064							
34	アニメティ	.054	快適	.054							

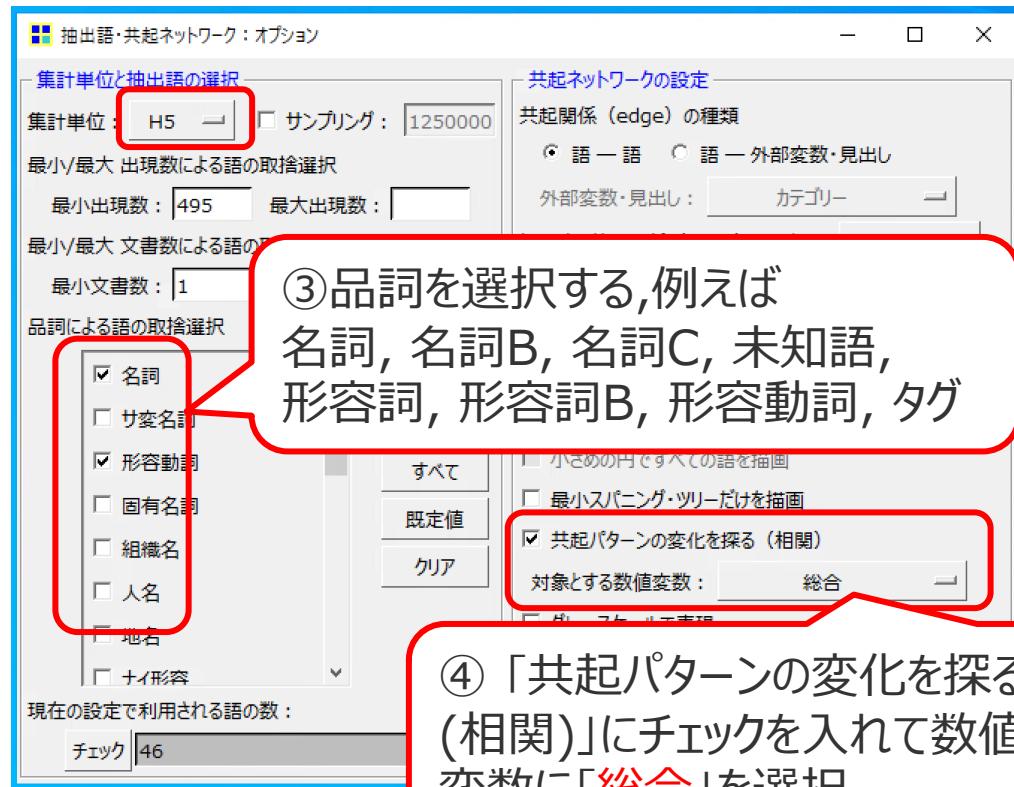
各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

KHCoder の使い方

● 共起ネットワークの作成(2)

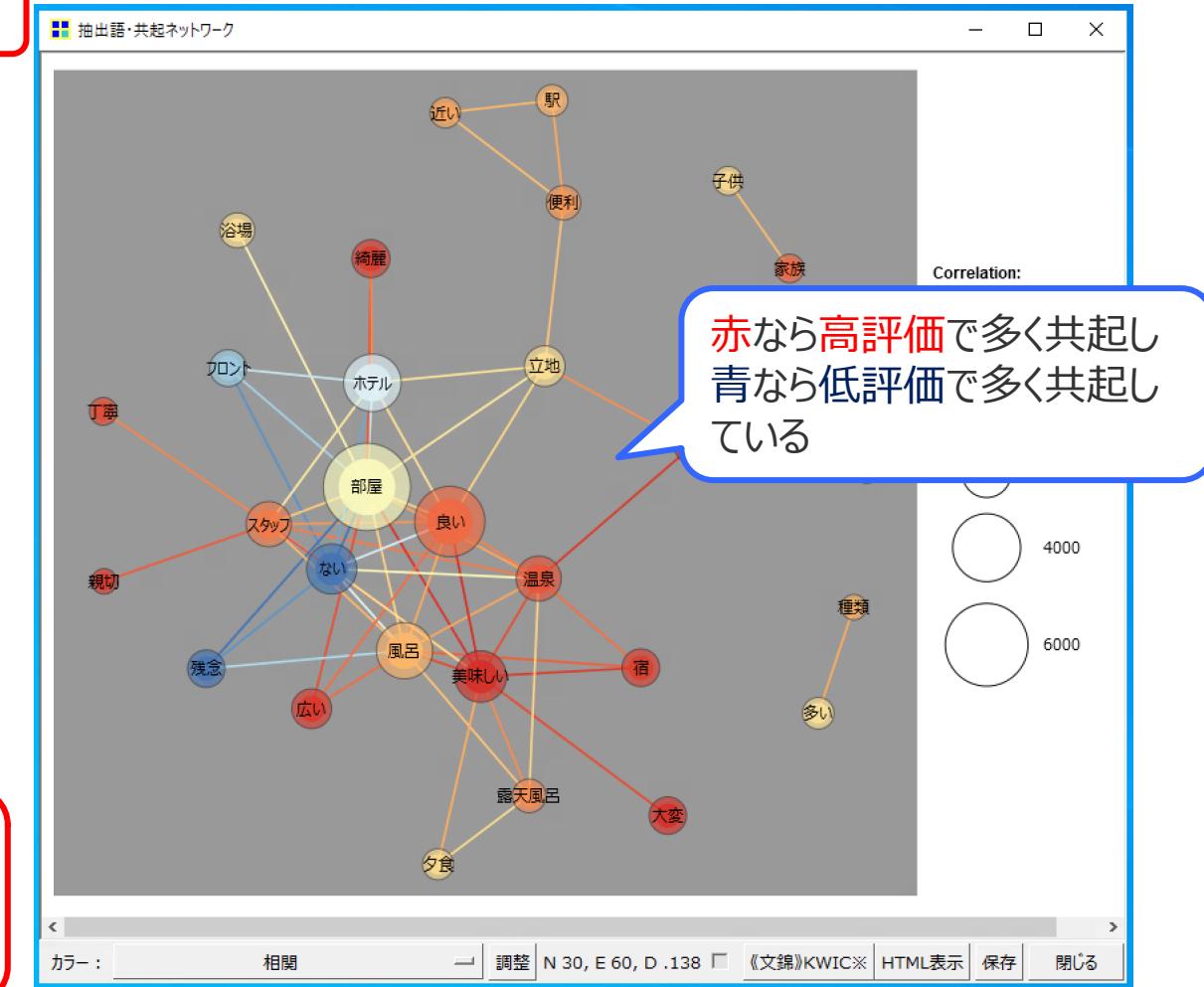
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「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「共起パターンの変化を探る
(相関)」にチェックを入れて数値
変数に「総合」を選択

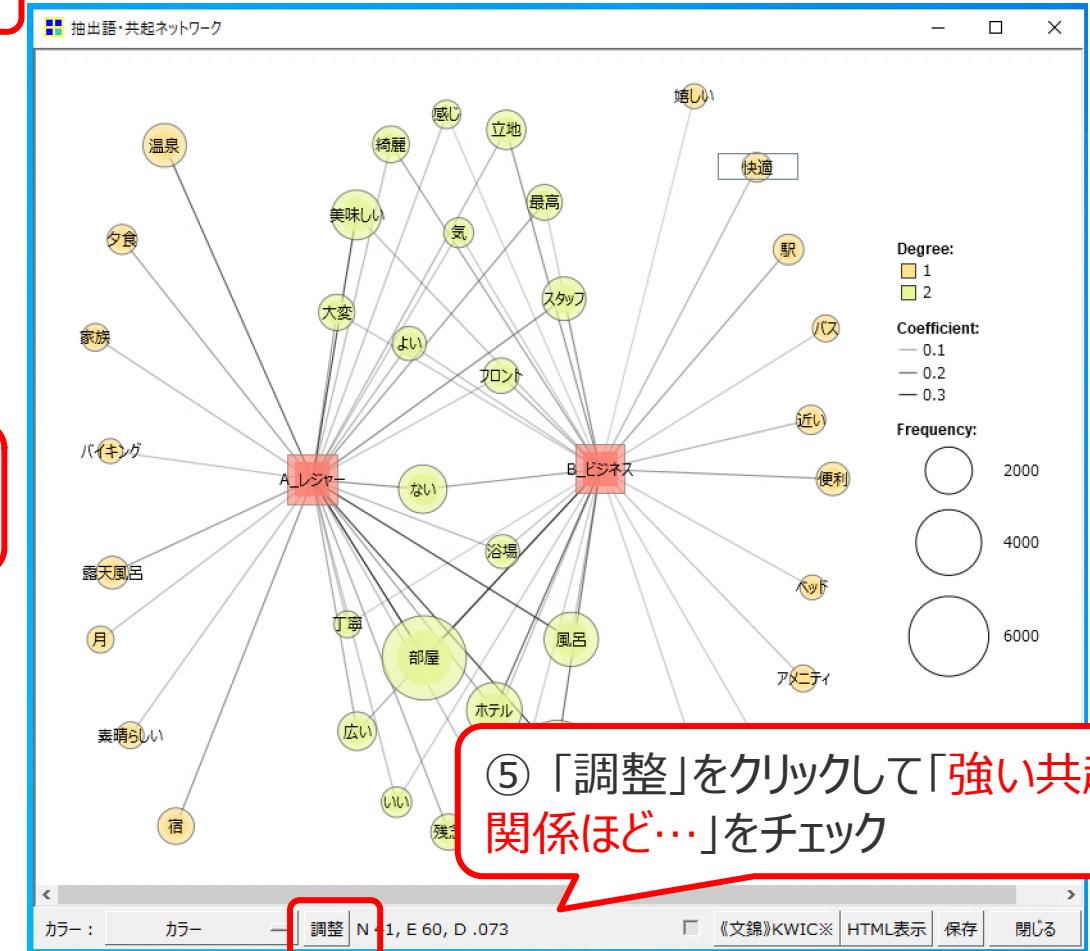
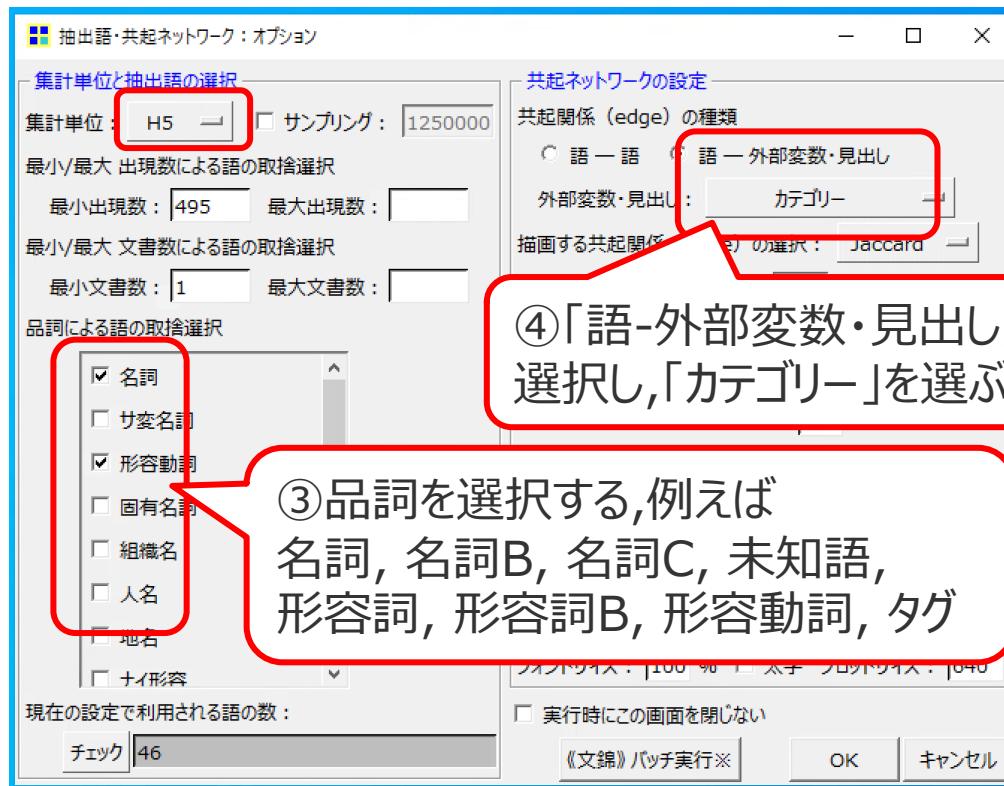


KHCoder の使い方

● 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

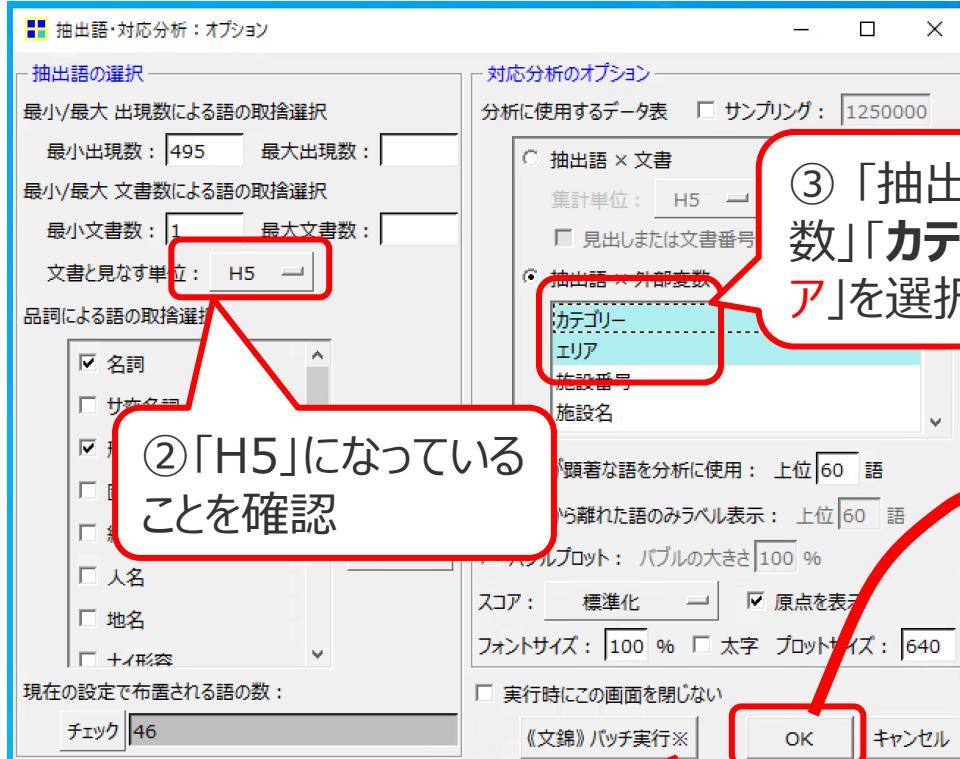
②「集計単位」として「H5」を選んで「OK」をクリック



KHCoder の使い方

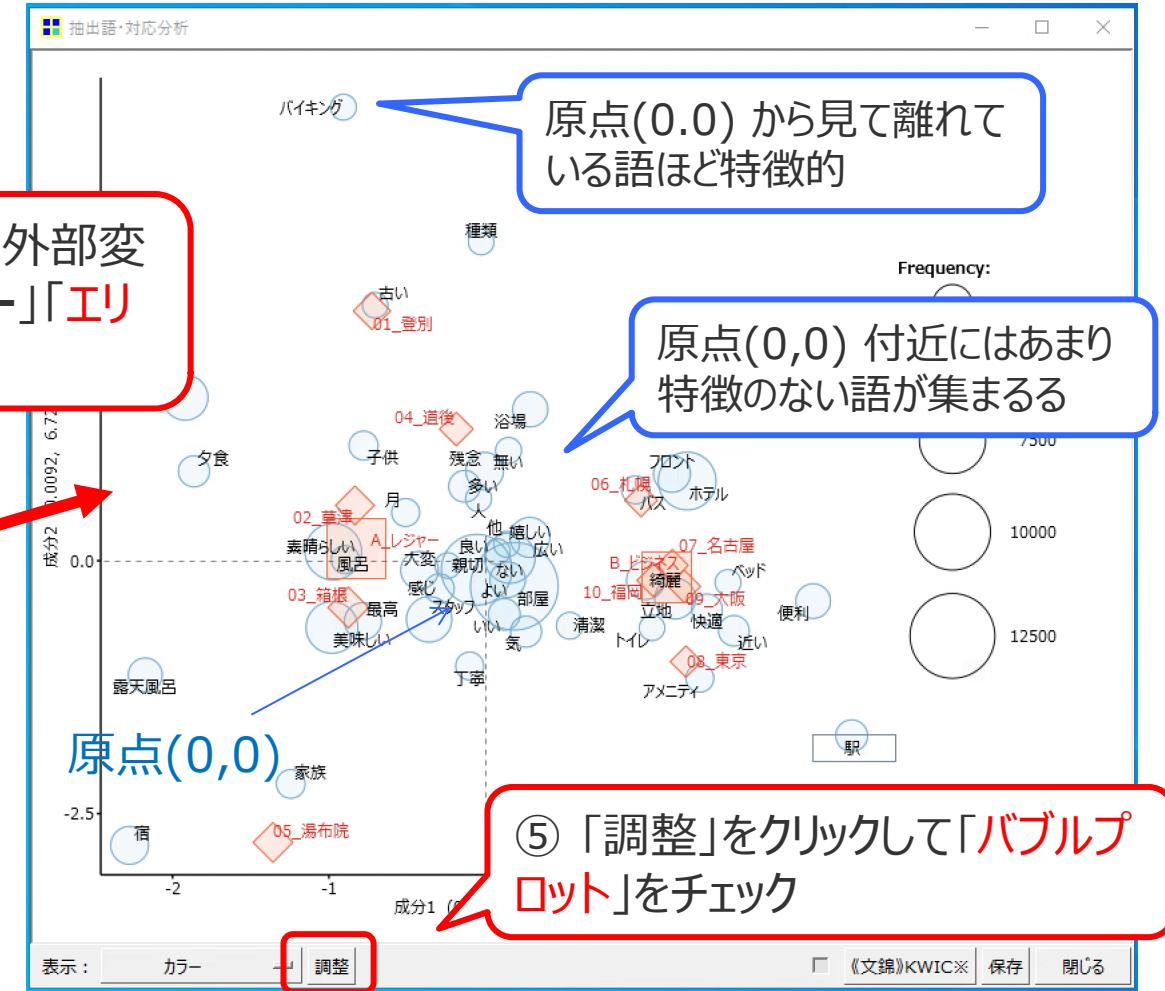
● 対応分析による探索(1)

- ① メニューから「ツール」「抽出語」「対応分析」を選ぶ



④ 「OK」をクリック

③ 「抽出語×外部変数」「カテゴリー」「エリア」を選択

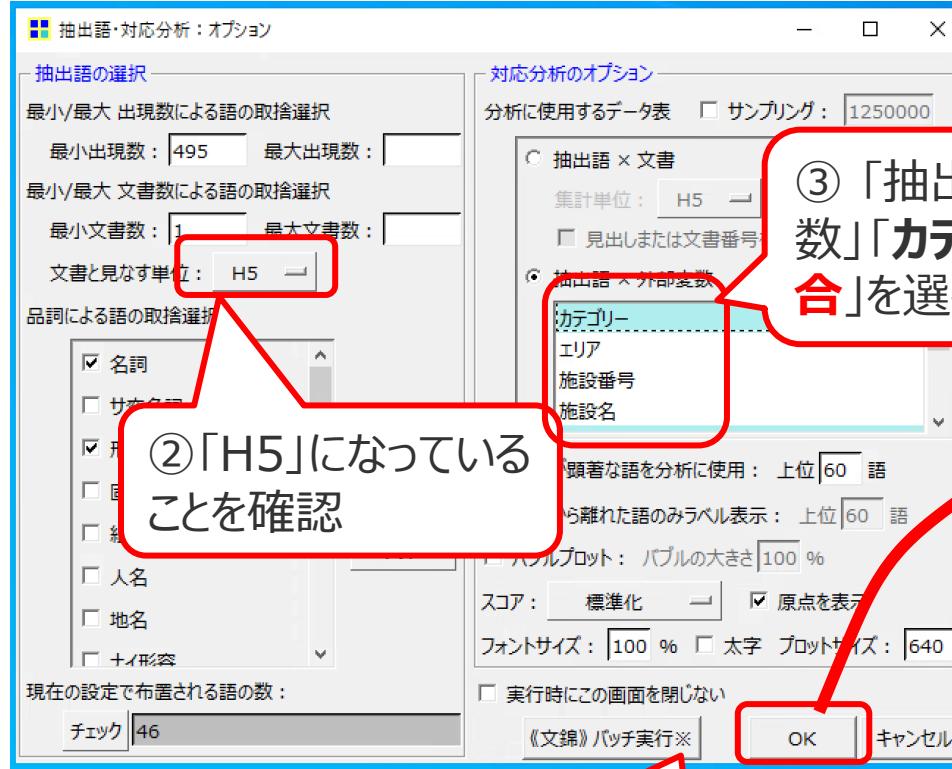


⑤ 「調整」をクリックして「バブルプロット」をチェック

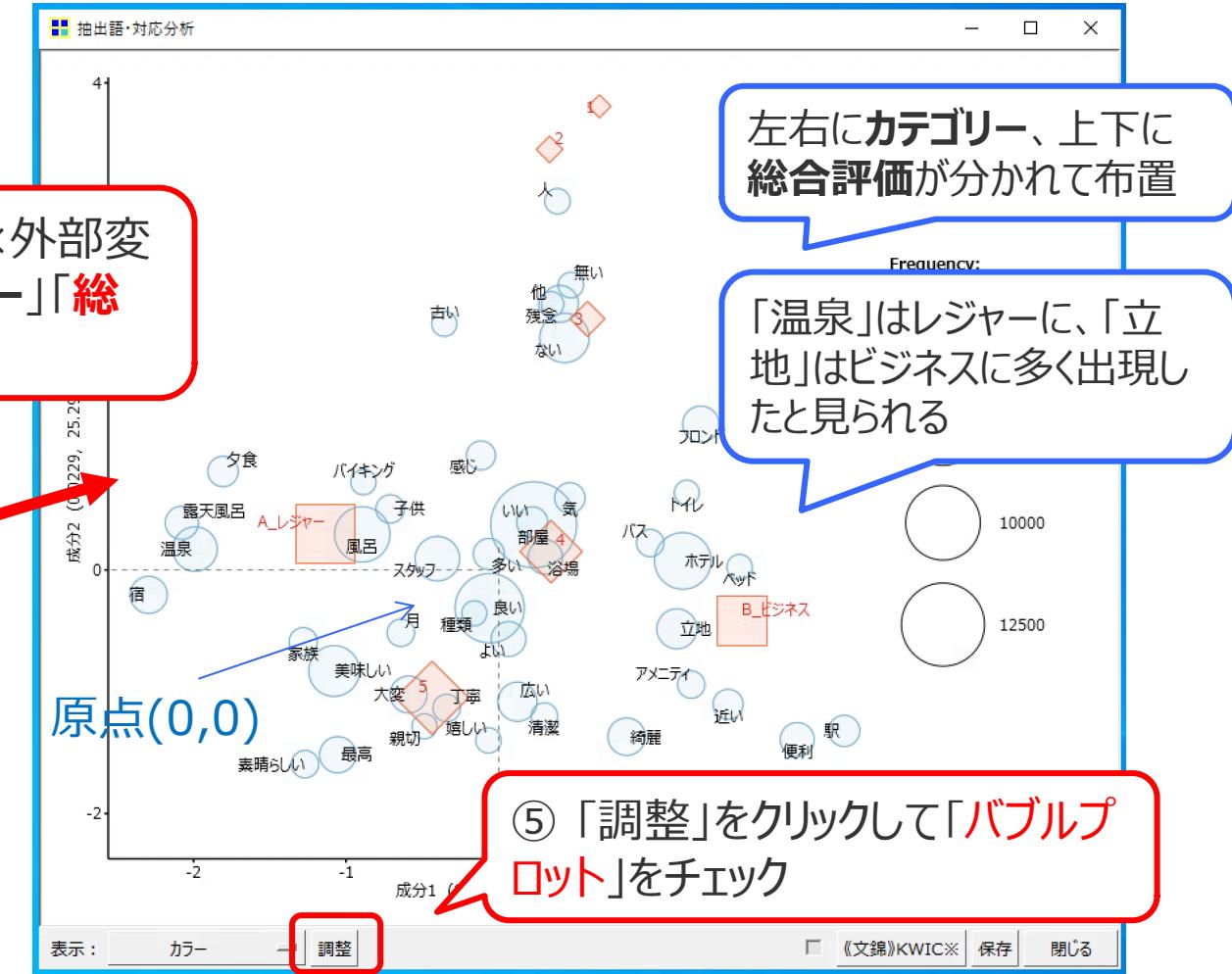
KHCoder の使い方

● 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



- ④ 「OK」をクリック

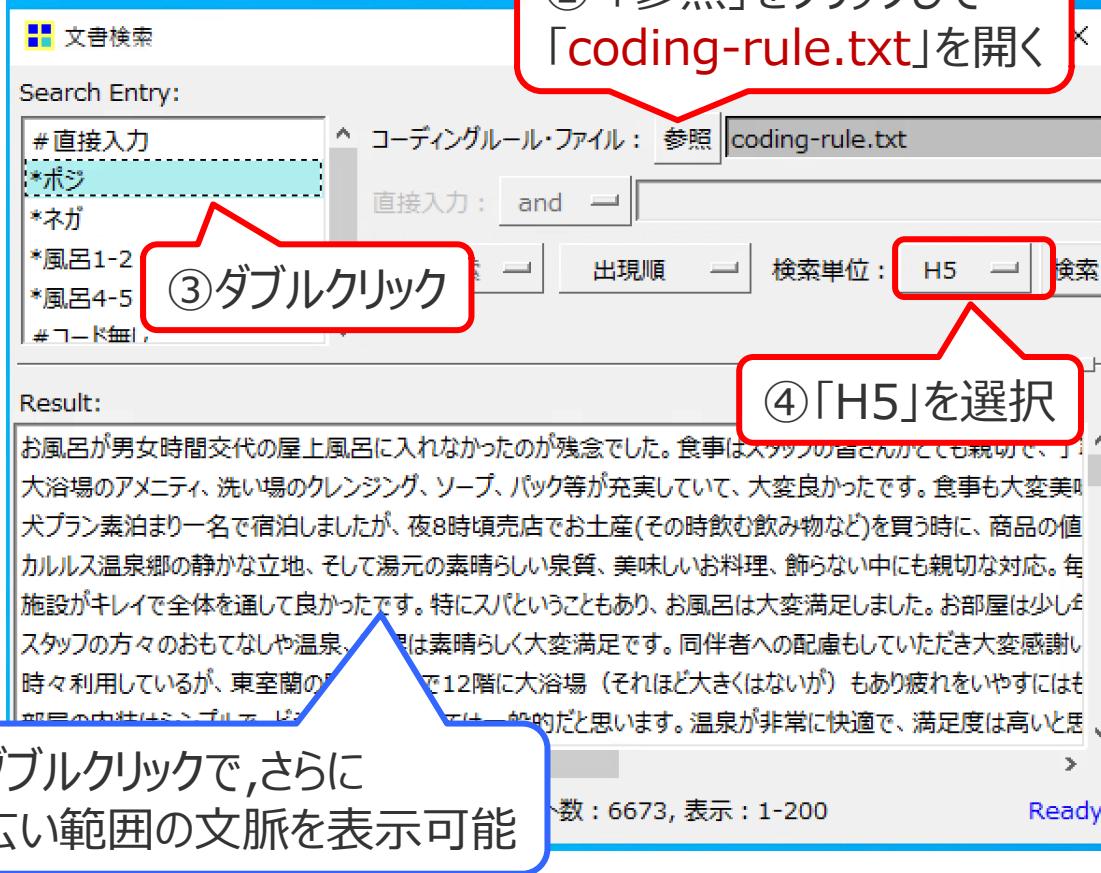


- ⑤ 「調整」をクリックして「バブルプロット」をチェック

KHCoder の使い方

●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 気持ちはいい or 楽しい or 近い or 大きい or 気持ち良い or 温かい or 早い or 優しい or 新しい or 暖かい or 快い or 明るい or 美しい or 可愛い

*ネガ

古い or 無い or 高い or 悪い or 小さい or 狹い or 少ない or 寒い or 遅い or 熱い or 欲しい or 暑い or 冷たい or 遠い or 臭い or 暗い

*風呂1-2

<>風呂-->1 | <>風呂-->2

外部
変
数

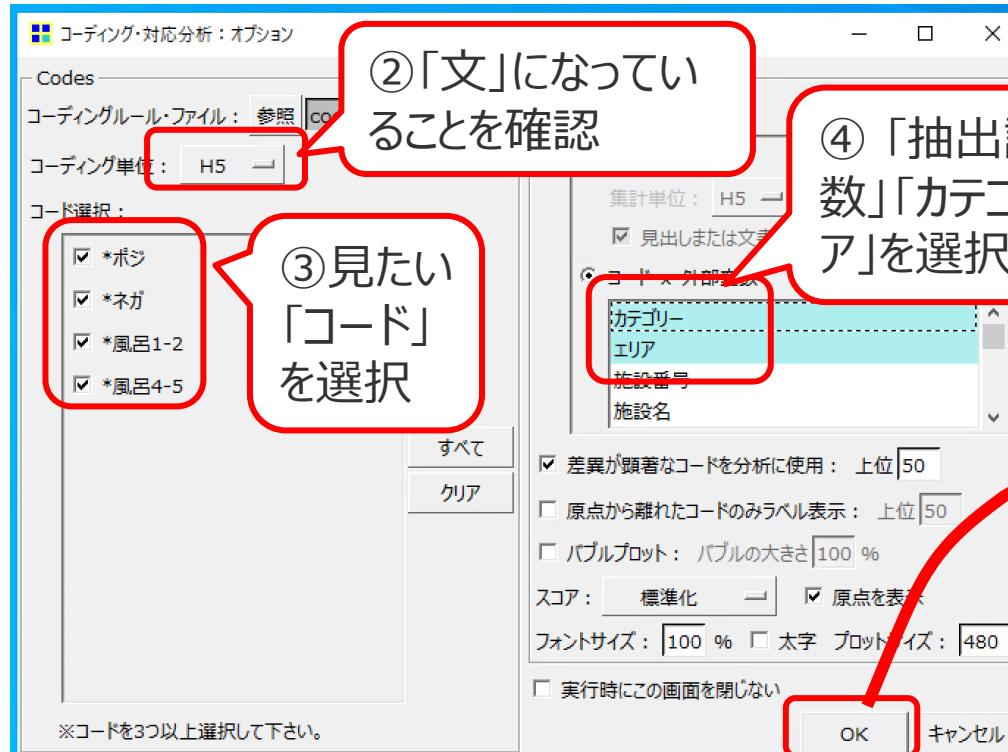
*風呂4-5

<>風呂-->4 | <>風呂-->5

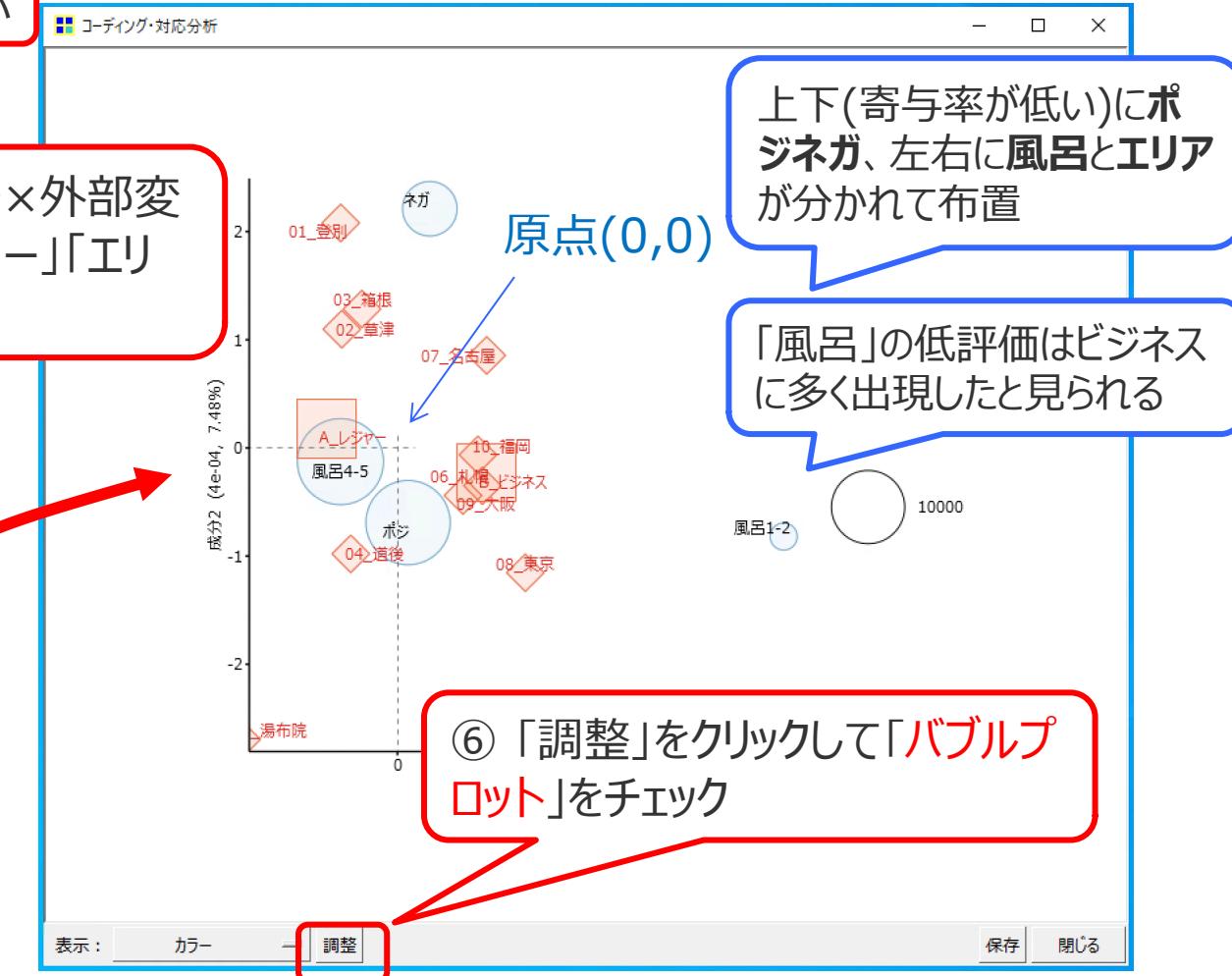
KHCoder の使い方

● 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



⑤ 「OK」をクリック



KHCoder の使い方

● 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

Entry

コーディングルール・ファイル：参照 coding-rule.txt

セル内容：度数とパーセント

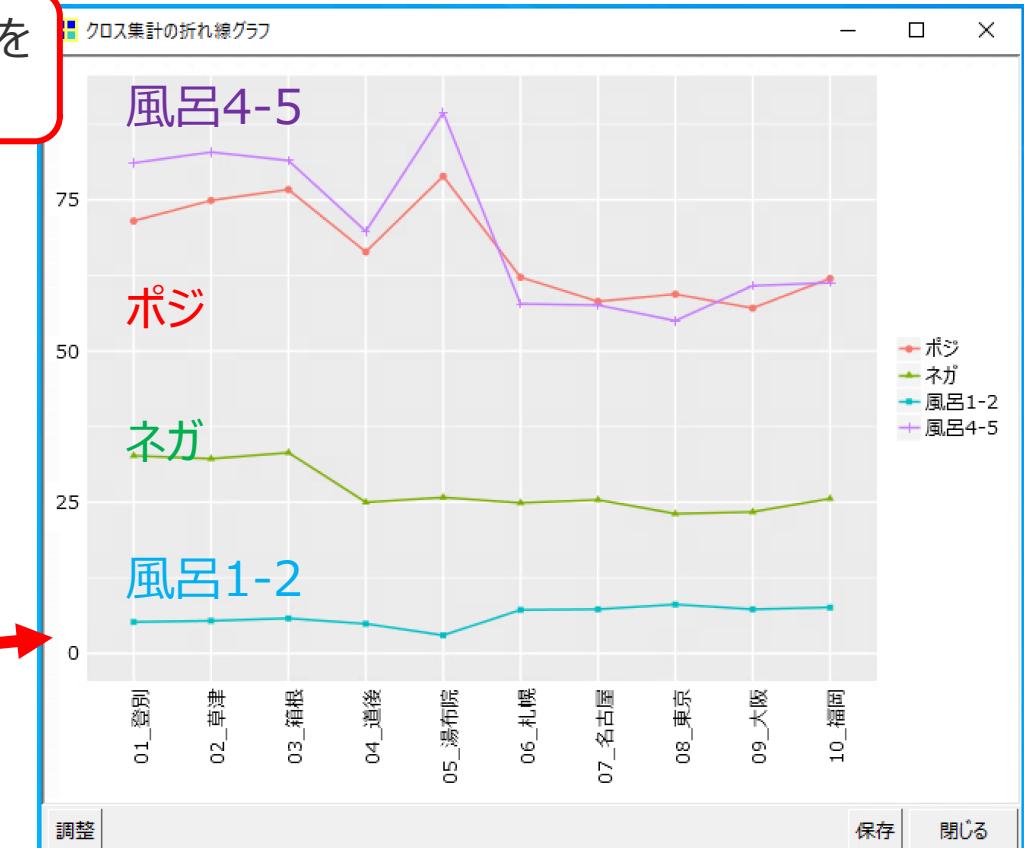
コーディング単位：H5 クロス集計：エリア 集計

Result

	*ポジ	*ネガ	ケース数
01_登別	72 (7.20%)	578 (57.80%)	1000
02_草津	73		
03_箱根	81		
04_道後			
05_湯布院			
06_札幌			
07_名古屋			
08_東京			
09_大阪	571 (57.10%)	234 (23.40%)	73
10_福岡	620 (62.00%)	256 (25.60%)	76 (7.60%)
合計	6673 (66.73%)	2713 (27.13%)	618 (6.18%)
カイ2乗値	269.915**	70.988**	39.179**
			703.613**

Ready. マップ：ヒート バブル 折れ線：すべて 選択 コピー（表全体）

注：プロット左側のラベルは表示されません

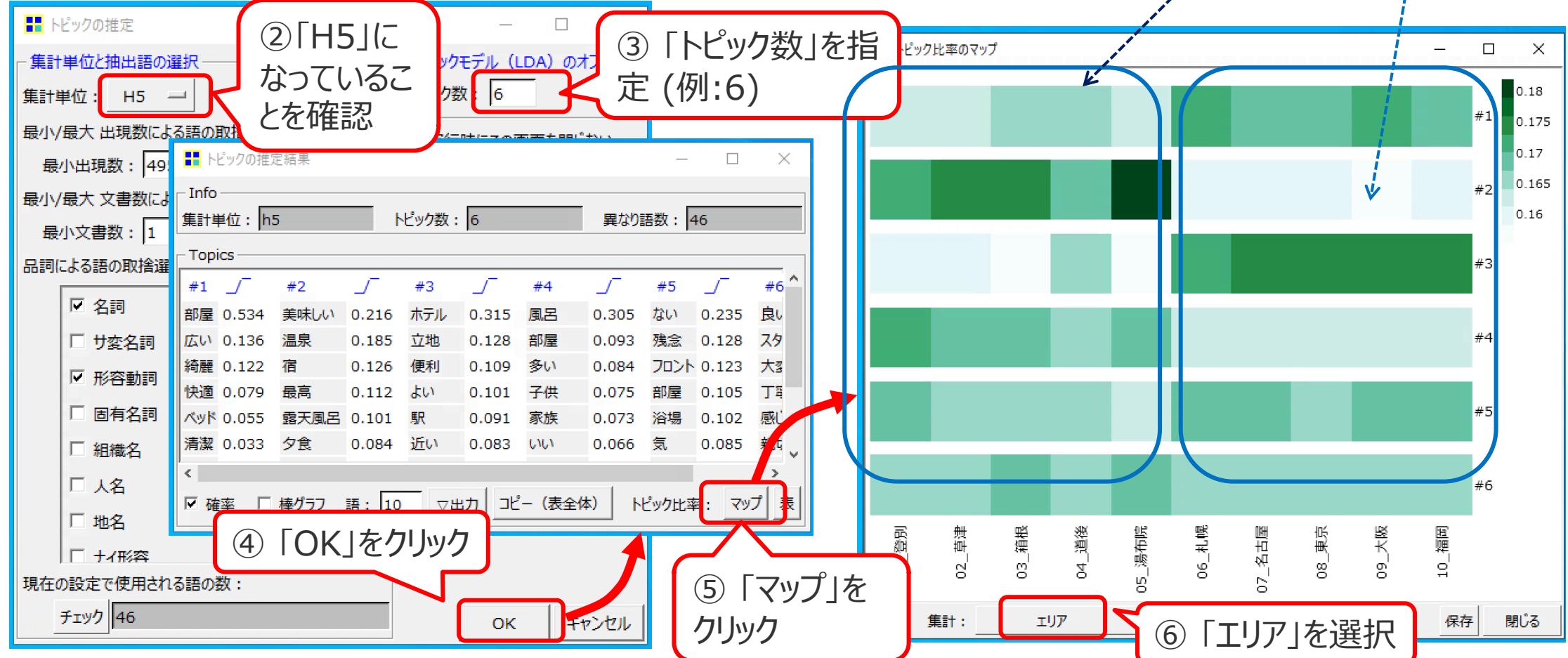


⑥ 「すべて」をクリック

KHCoder の使い方

● 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



day 3 – レポート課題

- 以下を 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」に倣って「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し、クロス集計を行って作成した「プロット」のキャプチャ (P.50)

※ 何らかの事情で上記のキャプチャ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

Q&A

参考資料

● KH Coder

- ・ 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.
- ・ 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一. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.
- ・ 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
- ・ 横口耕一. 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

- ・ 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. [[発表スライド](#)]

● Rを使った参考書

- ・ 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- ・ 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

- ・ 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.
- ・ 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

- ・ 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.