

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)

(前回) day 2 – レポート課題

- 以下を PDF ファイルで提出してください
 - データ集計により作成した「集計表」のキャプチャ (P.47~51) ※ページ番号は各スライド右下に記載
 - 作成した「集計結果の整理」の表 (P.52) ※ページ番号は各スライド右下に記載
- ※「集計表」のキャプチャは Jupyter の出力でも EXCEL でも構いません
- ※ 何らかの事情で上記2つを提出できない場合、本日の講義の感想を文章で記述してください

レポート形式	提出先	期限
PDF	manaba	次回～18:20

- 後述する KHCoder を各自の環境にインストールしてください ← このレポート提出は不要

(再掲) 実習用のデータ (Webサイトクローリング)

● ホテルのクチコミ数: 1,325万件 ※年間約60~80万

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 the number '13,246,463' displayed prominently. Below this, there is a search bar for reviews and a section for '新着！最新のクチコミ' (Latest reviews). On the right side, there is 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 thoughts from users who actually used the service). The overall layout includes various navigation links and promotional banners.

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 (2022)
→ **1,325万件 (今回)**
※ 2021/5/27現在

鴨川シーワールドホテルのクチコミ・お客様の声

[●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

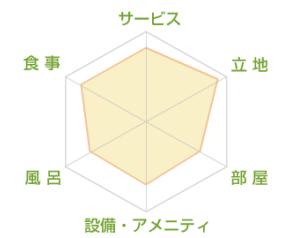
評価内訳

- 5点
- 4点
- 3点
- 2点
- 1点

236件
302件
47件
15件
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・情報)



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（品数が多く、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、それは思いませんでした。

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



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

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。

セラベーションアップに繋がる
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。

数値評価

(再掲) 実習用のデータ (Webサイトクローリング)

● 楽天トラベル のクチコミデータ

- 収集期間は 2019-2020 および 2022-2023(～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
ビジネス	5エリア	札幌、名古屋、東京、大阪、福岡	= 計10,000件

(再掲) 実習用のデータ (Webサイトクローリング)

● 楽天トラベル のクチコミデータ

- 収集期間は 2019-2020 および 2022-2023(～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
- 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) 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

(参考) データ理解 — 集計例

①件数 (エリア別)

行ラベル	個数 / コメン
■ A_レジャー	5000
01_登別	1000
02_草津	1000
03_箱根	1000
04_道後	1000
05_湯布院	1000
■ B_ビジネス	5000
06_札幌	1000
07_名古屋	1000
08_東京	1000
09_大阪	1000
10_福岡	1000
総計	10000

②投稿者の傾向 (年代別x性別)

行ラベル	個数 / コメン	列ラベル	男性	女性	na	総計
10代			0.07%	0.01%	0.00%	0.08%
20代			0.88%	1.17%	0.00%	2.05%
30代			2.28%	2.37%	0.00%	4.65%
40代			5.19%	3.54%	0.00%	8.73%
50代			7.83%	4.12%	0.00%	11.95%
60代			4.88%	2.05%	0.00%	6.93%
70代			1.00%	0.31%	0.00%	1.31%
80代			0.09%	0.02%	0.00%	0.11%
na			0.00%	0.00%	64.19%	64.19%
総計			22.22%	13.59%	64.19%	100.00%

③投稿者の傾向 (性別xカテゴリ別)

行ラベル	A_レジャー	B_ビジネス	総計
男性	22.26%	22.18%	22.22%
女性	15.22%	11.96%	13.59%
na	62.52%	65.86%	64.19%
総計	100.00%	100.00%	100.00%

- 男性の投稿者が多い(女性の倍程度) → 男性の観点によるコメントが多い

- 無回答(na)の中の分布が、表明した層と異なる(ある年代や性別に偏っている)可能性もある

(参考) データ理解 — 集計例

④投稿者の傾向 (性別xカテゴリーエリア別)

個数 / コメント	列ラベル	A_レジャー 集計										B_ビジネス 集計			総計
		01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡				
男性	A_レジャー	26.70%	23.90%	16.30%	24.80%	19.60%	22.26%	24.70%	22.20%	20.50%	20.20%	23.30%	22.18%	22.22%	
女性	A_レジャー	13.10%	15.60%	16.00%	12.50%	18.90%	15.22%	12.60%	10.50%	12.70%	11.80%	12.20%	11.96%	13.59%	
na		60.20%	60.50%	67.70%	62.70%	61.50%	62.52%	62.70%	67.30%	66.80%	68.00%	64.50%	65.86%	64.19%	
総計		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

- 男女差は、レジャーに比べビジネスが大きい
- 男女差がレジャーで大きいのは道後(次いで登別や草津も大きい)

⑤投稿者の傾向 (年代別xカテゴリーエリア別)

個数 / コメント	列ラベル	A_レジャー 集計										B_ビジネス 集計			総計
		01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡				
10代	A_レジャー	0.00%	0.00%	0.30%	0.00%	0.00%	0.06%	0.20%	0.00%	0.20%	0.10%	0.00%	0.10%	0.08%	
20代	A_レジャー	1.60%	3.70%	2.80%	1.40%	2.70%	2.44%	1.30%	2.00%	1.40%	2.10%	1.50%	1.66%	2.05%	
30代	A_レジャー	5.10%	4.60%	5.60%	4.40%	5.80%	5.10%	6.40%	3.50%	3.40%	3.80%	3.90%	4.20%	4.65%	
40代	A_レジャー	9.80%	10.00%	6.20%	8.00%	8.60%	8.52%	9.20%	9.20%	9.80%	7.60%	8.90%	8.94%	8.73%	
50代	A_レジャー	13.20%	11.40%	8.90%	13.10%	11.80%	11.68%	12.40%	11.10%	11.60%	11.30%	14.70%	12.22%	11.95%	
60代	A_レジャー	7.70%	8.30%	5.70%	8.80%	8.60%	7.82%	7.00%	6.20%	5.70%	5.80%	5.50%	6.04%	6.93%	
70代	A_レジャー	2.30%	1.10%	2.80%	1.50%	0.90%	1.72%	0.70%	0.50%	1.00%	1.30%	1.00%	0.90%	1.31%	
80代	A_レジャー	0.10%	0.40%	0.00%	0.10%	0.10%	0.14%	0.10%	0.20%	0.10%	0.00%	0.00%	0.08%	0.11%	
na		60.20%	60.50%	67.70%	62.70%	61.50%	62.52%	62.70%	67.30%	66.80%	68.00%	64.50%	65.86%	64.19%	
総計		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

- 年代別では、目的によらず40~50代が多い

- あくまでも投稿者の傾向であって、旅行者の実態と一致するは限らない

(参考) データ理解 — 集計例

- レジャーの中で一人が多いのは道後 →道後はもはや仕事で行く場所 (性別でも男性が多い)

⑥投稿者の傾向 (同行者別)

個数 / コメント 行ラベル	A_レジャー 集計										B_ビジネス 集計				総計
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
一人	28.40%	15.90%	13.90%	47.90%	16.40%	24.50%	60.20%	65.20%	67.20%	55.70%	55.20%	60.70%		42.60%	
家族	59.10%	61.80%	68.00%	42.00%	65.90%	59.36%	27.90%	24.20%	21.50%	30.20%	32.70%	27.30%		43.33%	
恋人	5.40%	13.00%	10.20%	4.60%	10.40%	8.72%	5.50%	4.70%	4.10%	5.50%	4.10%	4.78%		6.75%	
友達	4.60%	8.30%	6.40%	3.50%	6.20%	5.80%	3.90%	3.50%	4.10%	6.50%	5.80%	4.76%		5.28%	
仕事仲間	2.00%	0.60%	1.00%	1.40%	0.40%	1.08%	1.80%	1.70%	1.90%	1.60%	1.80%	1.76%		1.42%	
その他	0.50%	0.40%	0.50%	0.60%	0.70%	0.54%	0.70%	0.70%	1.20%	0.50%	0.40%	0.70%		0.62%	
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%		100.00%	

⑦投稿者の傾向 (年代別x性別)

- 数値評価は、目的によらず高め →好評価しか投稿しない偏りがあるの可能性にも注意

- レジャーは家族が多く、ビジネスは一人が多い →出張は複数より単独が多い

個数 / コメント 行ラベル	A_レジャー 集計										B_ビジネス 集計				総計
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
5	41.30%	49.60%	48.50%	46.50%	67.50%	50.68%	44.50%	40.40%	41.10%	47.30%	42.60%	43.18%		46.93%	
4	39.20%	34.20%	36.00%	39.40%	23.70%	34.50%	39.50%	43.60%	41.20%	36.90%	39.80%	40.20%		37.35%	
3	11.30%	9.70%	8.20%	9.10%	5.50%	8.76%	9.90%	10.40%	11.30%	10.50%	12.90%	11.00%		9.88%	
2	5.00%	4.10%	4.60%	3.60%	2.40%	3.94%	3.80%	2.80%	3.60%	3.00%	2.80%	3.20%		3.57%	
1	3.20%	2.40%	2.70%	1.40%	0.90%	2.12%	2.30%	2.80%	2.80%	2.30%	1.90%	2.42%		2.27%	
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%		100.00%	

- レジャーの高評価は、湯布院が多く、登別が少ない

- ビジネスの高評価は、大阪が多い、福岡・東京・名古屋がやや少ない

(参考) データ理解 — 集計例

⑧-a 数値評価の平均 (エリア別×数値評価別)

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		
■ A_レジャー	4.22	4.28	4.11	4.01	4.29	4.26	4.28
01_登別	4.03	4.27	3.95	3.88	4.31	4.08	4.10
02_草津	4.19	4.28	4.03	3.92	4.31	4.15	4.25
03_箱根	4.22	4.15	4.12	3.97	4.22	4.28	4.23
04_道後	4.16	4.41	4.10	4.00	4.09	4.21	4.26
05_湯布院	4.52	4.28	4.36	4.27	4.50	4.57	4.55
■ B_ビジネス	4.00	4.34	4.10	3.92	3.82	4.06	4.19
06_札幌	3.99	4.37	4.09	3.92	3.82	4.06	4.19
07_名古屋	3.98	4.26	4.06	3.92	3.82	4.06	4.19
08_東京	3.97	4.34	4.11	3.91	3.73	3.99	4.14
09_大阪	4.06	4.34	4.14	3.96	3.86	4.12	4.24
10_福岡	4.01	4.40	4.11	3.89	3.85	4.02	4.18

- レジャーは、風呂や食事が、設備や部屋に比べて高評価

- 湯布院は、レジャーの中で、軒並み高評価が多い

⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

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	4.22	4.28	4.11	4.01	4.29	4.26	4.28
B_ビジネス	4.00	4.34	4.10	3.92	3.82	4.06	4.19

- レジャーもビジネスも立地が評価される
- ビジネスは、立地がその他に比べて高評価

(参考) データ理解 — 集計例

⑨-a 数値評価の平均 (20~30代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニティ			
□ A_レジャー	4.39	4.34	4.26	4.17			
男性	4.30	4.28	4.20	4.15	4.35	4.27	4.34
女性	4.48	4.40	4.32	4.19	4.43	4.46	4.43
□ B_ビジネス	4.16	4.32	4.09	4.01	3.95	4.20	4.19
男性	3.90	4.17	3.86	3.80	3.73	4.13	3.99
女性	4.38	4.45	4.28	4.18	4.14	4.25	4.35

- 20~50代はレジャーに対するサービスや風呂、食事の評価が概ね高い

⑨-b 数値評価の平均 (40~50代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニティ	平均 / 風呂	平均 / 食事	平均 / 総合
□ A_レジャー	4.29	4.35	4.17	4.05	4.35	4.28	4.35
男性	4.23	4.34	4.13	3.99	4.31	4.25	4.30
女性	4.36	4.37	4.23	4.15	4.41	4.33	4.43
□ B_ビジネス	4.03	4.35	4.10	3.92	4.00	4.22	4.16
男性	3.92	4.30	4.01	3.82	3.99	4.16	4.16
女性	4.25	4.46	4.29	4.14	4.33	4.22	4.16

- 年齢が高くなるに連れてレジャーに対する評価が厳しくなる

⑨-c 数値評価の平均 (60~90代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニティ	平均 / 風呂	平均 / 食事	総合
□ A_レジャー	4.18	4.21	4.05	3.95	4.25	4.29	4.26
男性	4.11	4.21	4.07	3.95	4.26	4.24	4.24
女性	4.33	4.20	4.07	3.95	4.35	4.30	4.30
□ B_ビジネス	3.93	4.30	4.07	3.85	3.73	3.85	4.15
男性	3.90	4.28	4.07	3.85	3.73	3.85	4.15
女性	4.02	4.36	4.12	3.98	4.06	4.00	4.18

- 60~90代は立地に対する評価も厳しい(→期待が高い)

(参考) データ理解 — 集計結果の整理

観点	データの特徴	テキスト分析時に注意すべき点
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40~60代が多い 全体的に男性の投稿者が多い（女性の倍程度） 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"> レビュー観点がある年代や性別に偏っている可能性 無回答("na")中が、ある年代や性別に偏っている可能性 <p style="border: 1px solid black; padding: 5px;">• 無回答が多い場合は、属性に引っ張られない解釈を心がける</p>
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（出張は単独で宿泊するケースが多い） レジャーの中でも、道後は男性の一人客が多い（道後はもやは仕事で行く場所か） 	<ul style="list-style-type: none"> レビ ーの観点が属性と一致しない可能性 レビューの観点がカテゴリと一致していない可能性（道後→仕事）
数値評価 (総合)	<ul style="list-style-type: none"> 旅行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 レジャーの中で高評価が多いのは湯布院 小さいのは 屋がやや少ない ビジネスの中で高評価が多い 	<p>• 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にも注意</p> <p>旅行目的によって投稿の動機が異なっている可能性</p> <p>• 属性に偏りがある場合は、乱暴に一般化しようとせず、属性ごとの分析を心がける</p>
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 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（重視している） 	目的によって評価の観点や重みが異なっている可能性
全体	<ul style="list-style-type: none"> あくまでも楽天トラベルの特性であるので、旅行者の傾向として主張するためには別途裏付けが必要 	

(再掲) 数値評価で違いを見るのは難しい

【再掲】⑧-a 数値評価の平均 (エリア別×数値評価別)

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			
■ A_レジャー	4.22	4.28	4.11	4.01	4.29	4.26	4.28	
01_登別	4.03	4.27	3.95	3.88	4.31	4.08	4.10	
02_草津	4.19	4.28	4.03	3.92	4.31	4.15	4.25	
03_箱根	4.22	4.15	4.12	3.97	4.22	4.28	4.23	
04_道後	4.16	4.41	4.10	4.00	4.09	4.21	4.26	
05_湯布院	4.52	4.28	4.36				4.55	
■ B_ビジネス	4.00	4.34	4.10				4.19	
06_札幌	3.99	4.37	4.09				4.20	
07_名古屋	3.98	4.26	4.06	3.92	3.82		4.16	
08_東京	3.97	4.34	4.11	3.91	3.73	3.99	4.14	
09_大阪	4.06	4.34	4.14	3.96	3.86	4.12	4.24	
10_福岡	4.01	4.40				4.02	4.18	

- 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

- 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

【再掲】⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

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	4.22	4.28	4.11	4.01	4.29	4.26	4.28
B_ビジネス	4.00	4.34	4.10	3.92	3.82	4.06	4.19

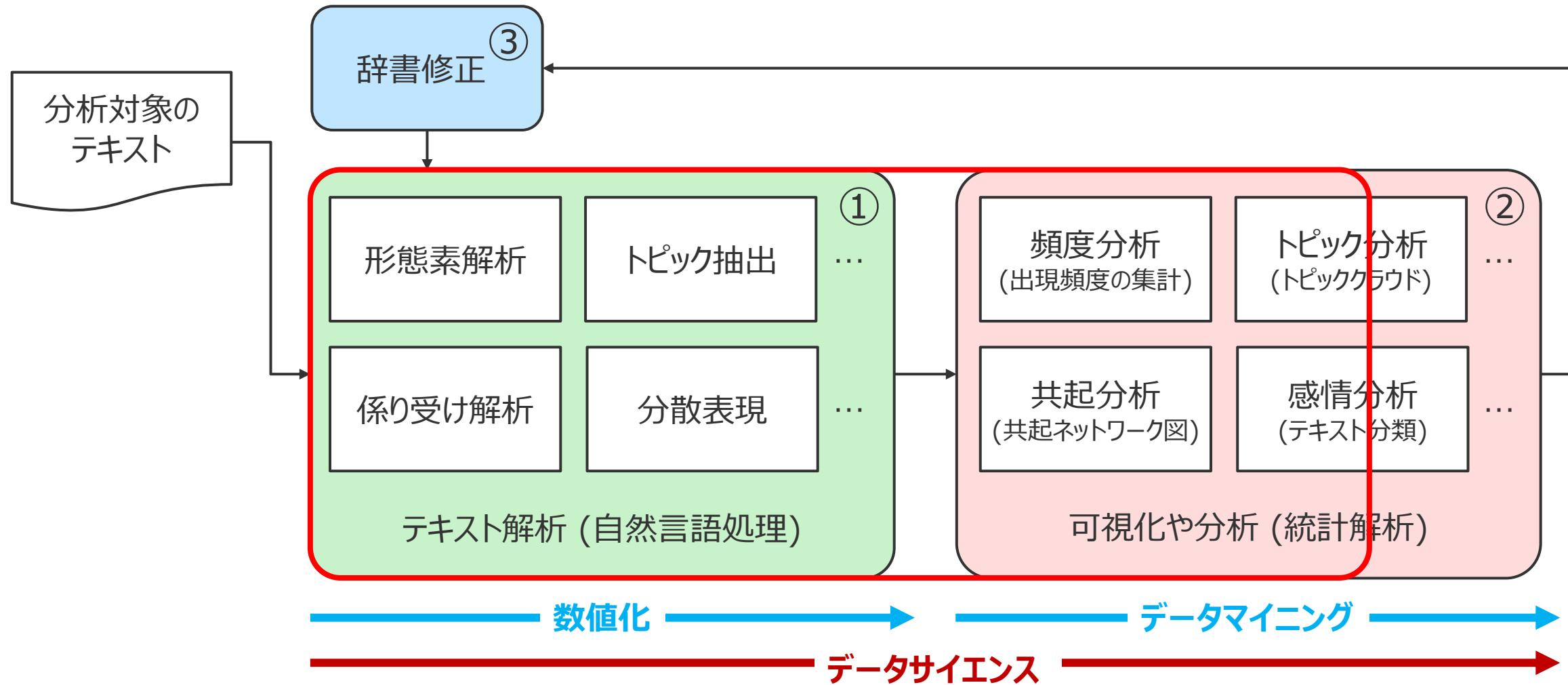
テキスト解析 (2)

(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

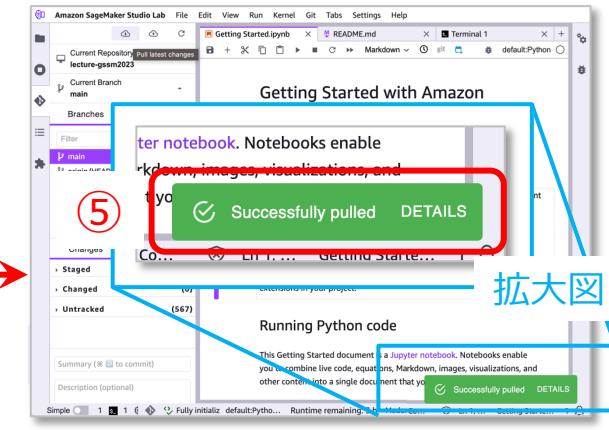
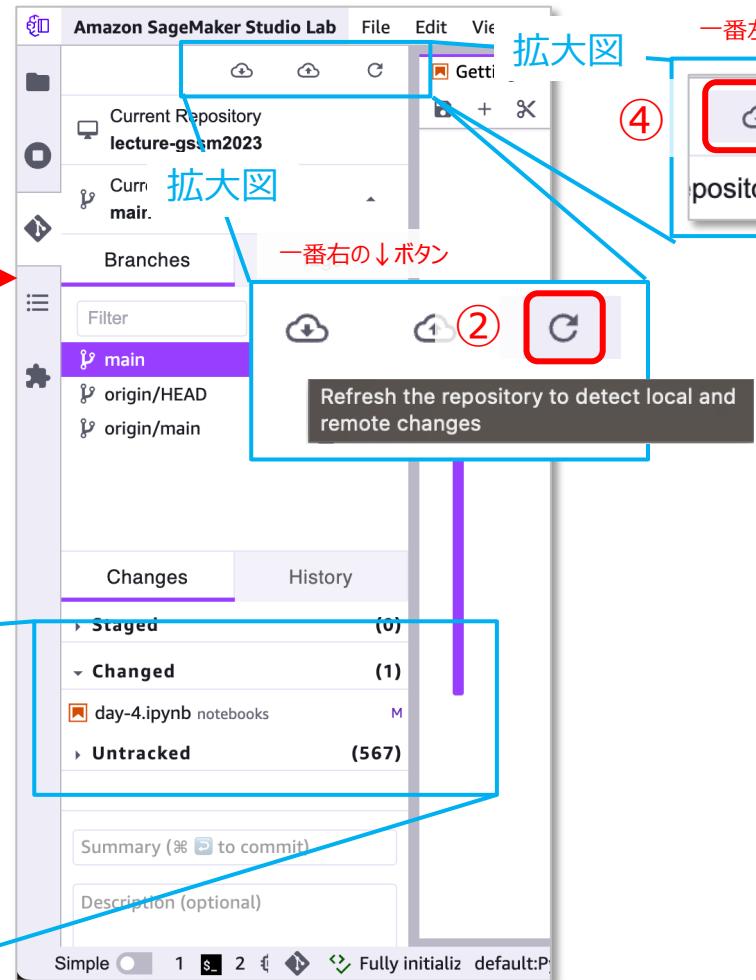
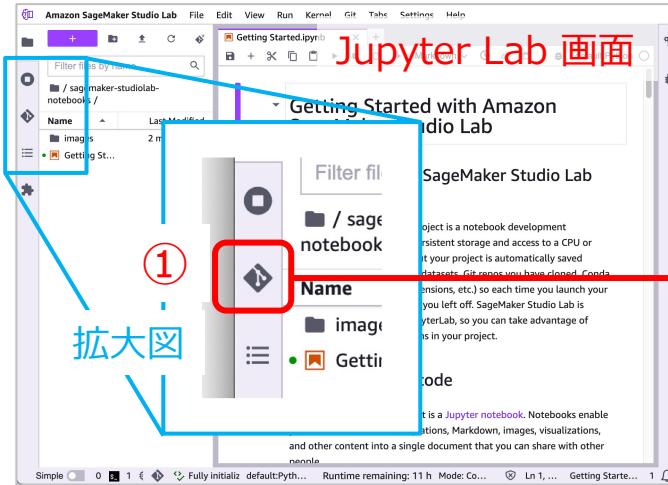
(再掲) テキスト分析の手順

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)

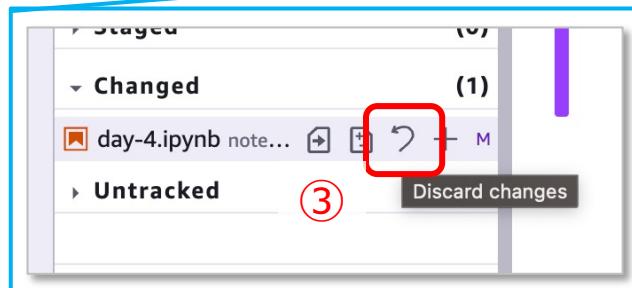


演習 — テキスト解析 (2)

● 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



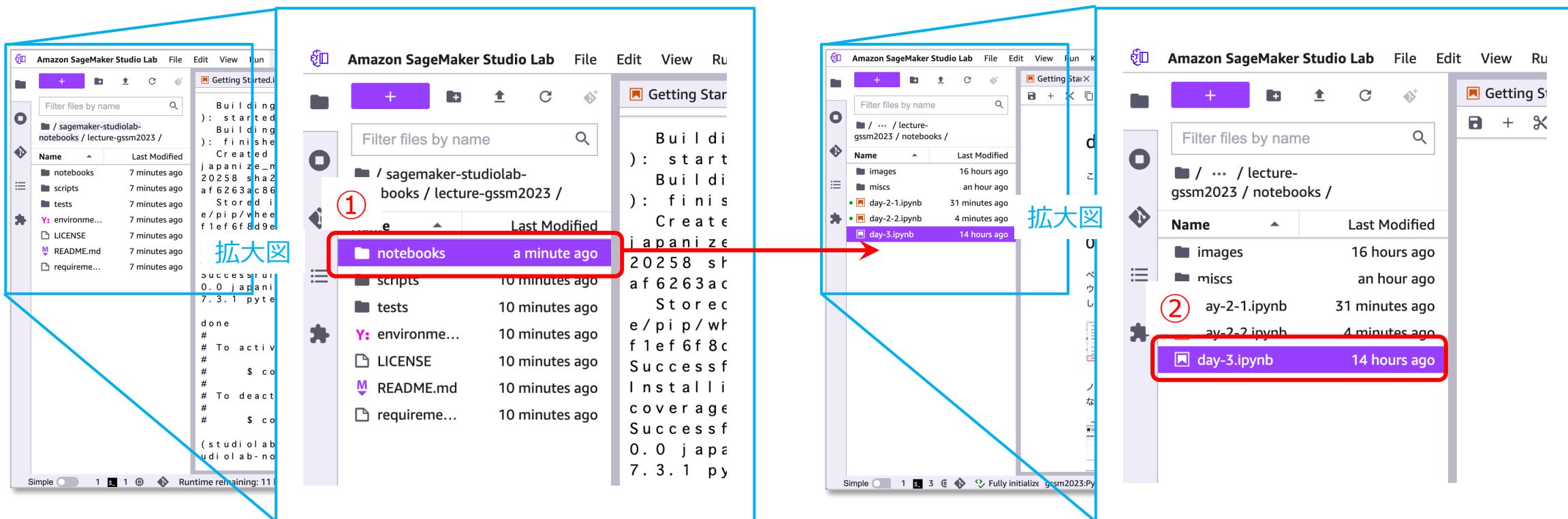
● 教材を最新化する

- 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)

● day-3.ipynb を開いてください

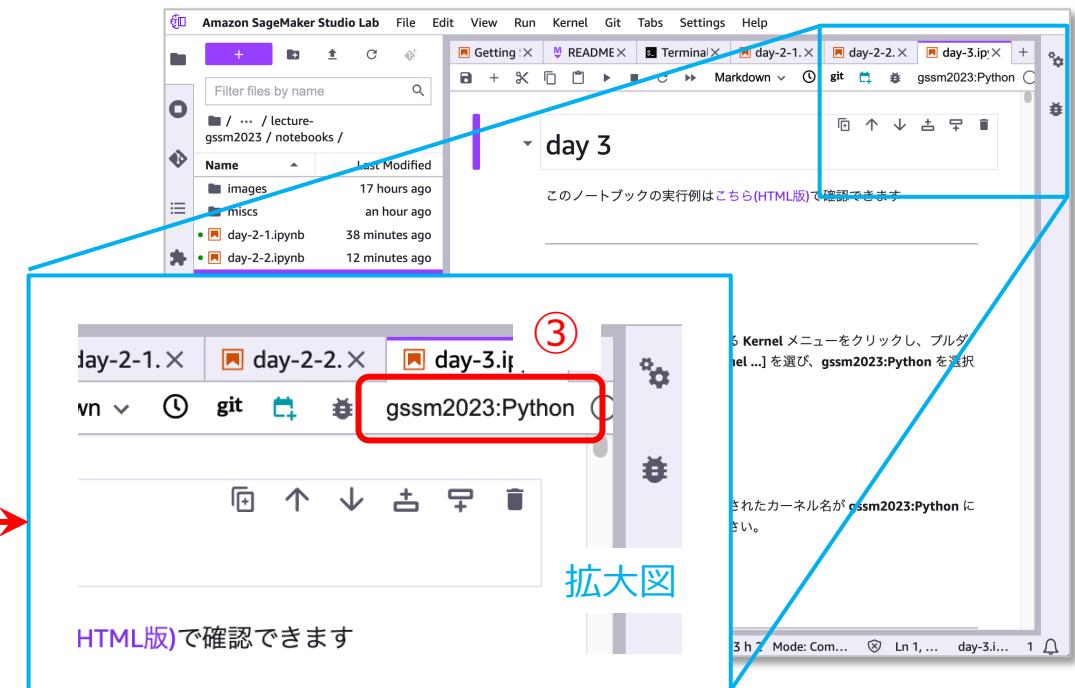
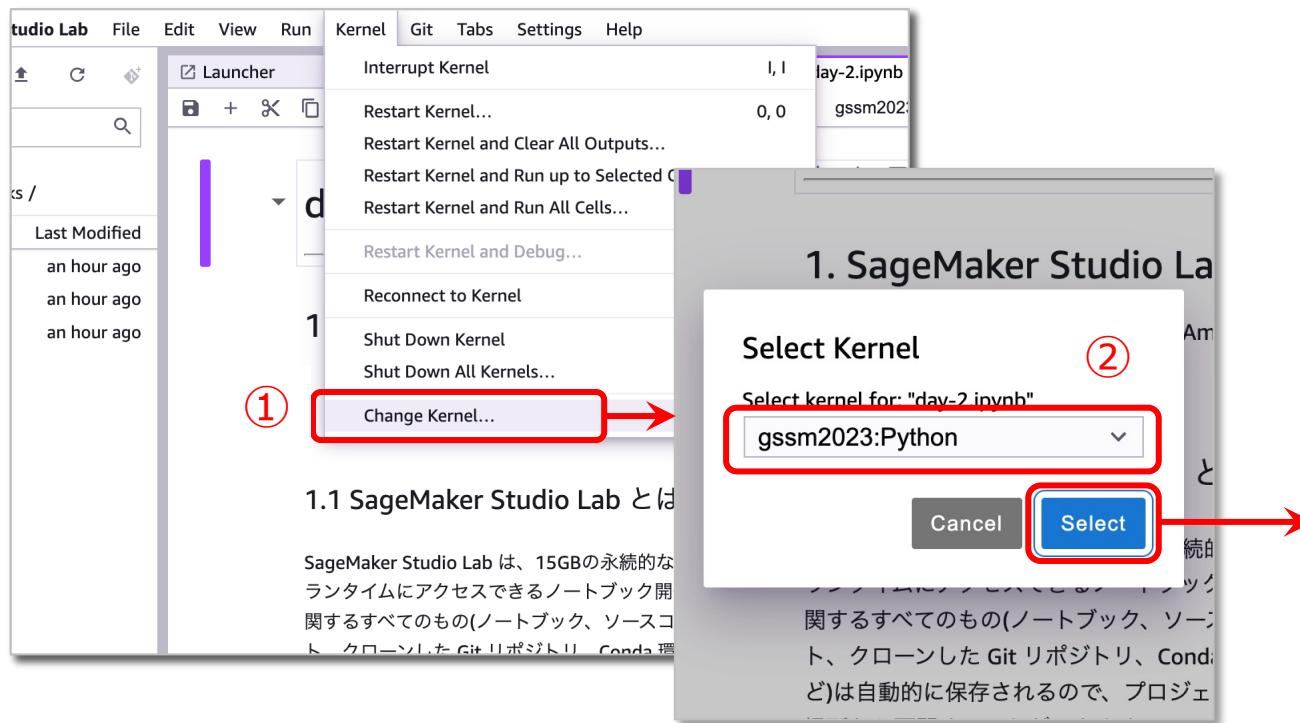
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-3.ipynb** ノートブックを開く



演習 — テキスト解析 (2)

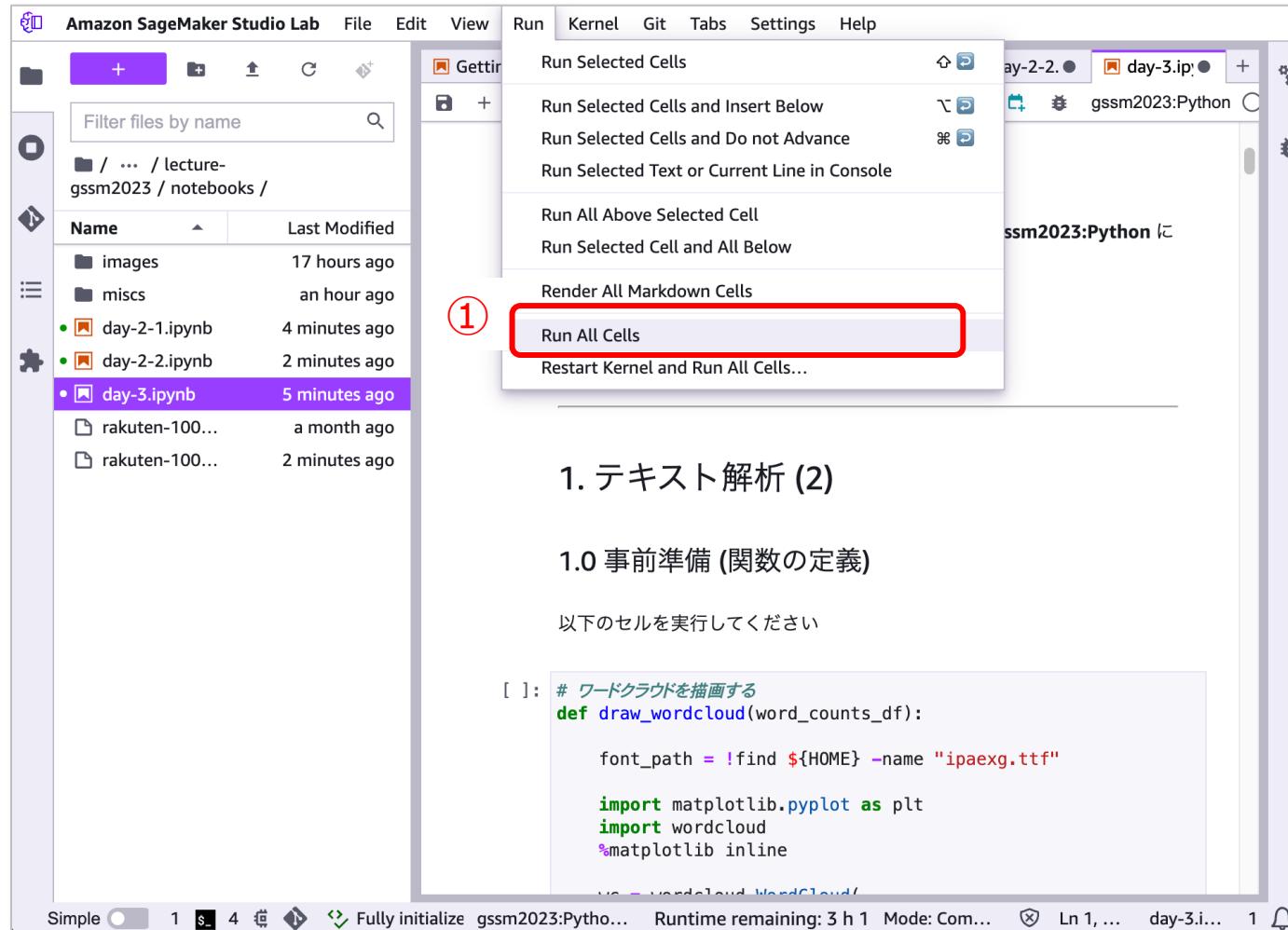
● カーネル gssm2023:Python を選択してください !重要!

- 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 drawing 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

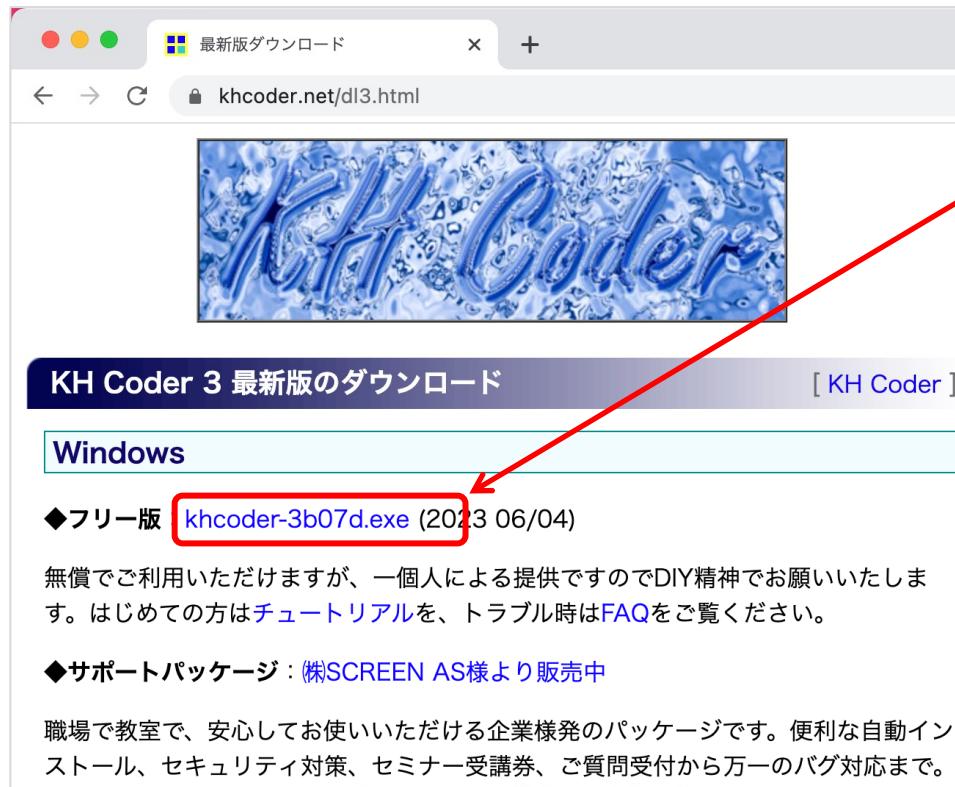
(再掲) 実習環境について

- 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)

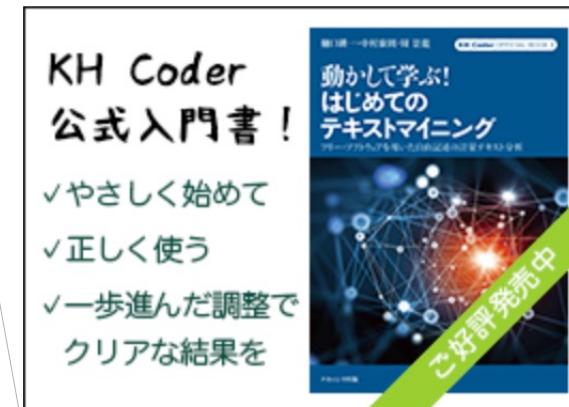
KH Coder の情報

ホームページ <http://khcoder.net/>

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

- Index:** A large blue banner with the text "KH Coder".
- 概要:** Information about the seminar.
- 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 [KH Coder]

KH Coder 3 チュートリアル

漱石「こころ」を題材に
【スライド版】

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。



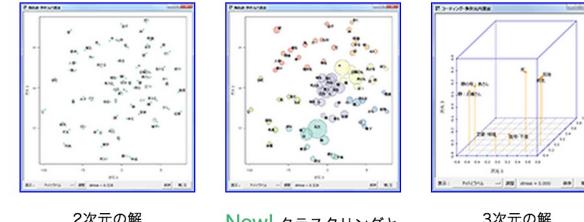
共起の程度が非常に強いものだけを線で結んだ図

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。



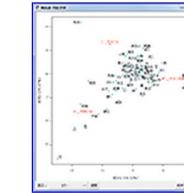
2次元の解

New! クラスタリングと色分け

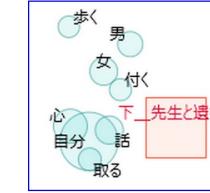
3次元の解

対応分析

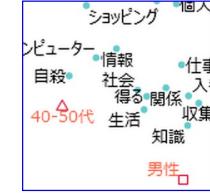
同じく抽出語またはコードを用いての、対応分析です。



同時布置図



New! バブルプロット



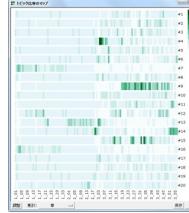
複数の外部変数を用いた多重対応分析

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)

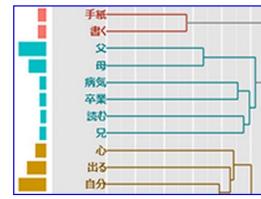
文書ごとにトピックの出現割合を表示したり、各トピックに高い確率で出現する語を表示できます。



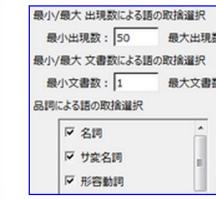
文書ごとのトピック比率

階層的クラスター分析

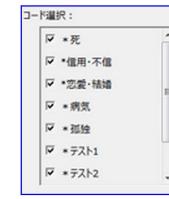
抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけではなくコーディング結果（コード）についても、同じように分析を行えます。



New! デンドログラム



抽出語は出現数や品詞で選択



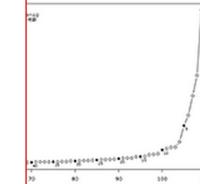
コードはチェックボックスで直接選択

文書のクラスター分析

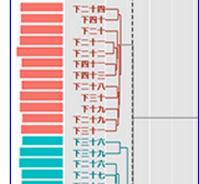
文書の分類を行うクラスター分析です。



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

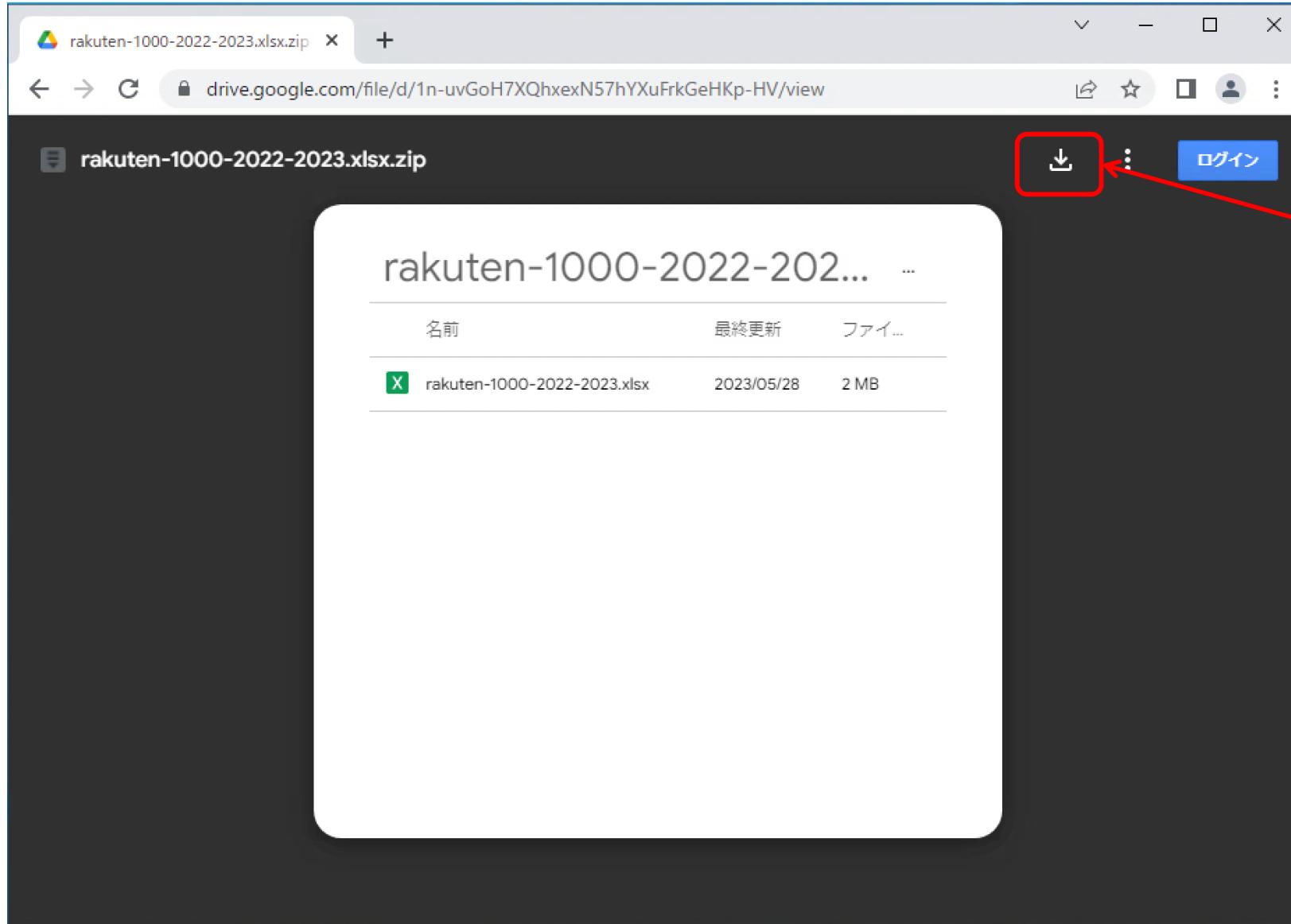
テキスト分析（使い方編）

(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

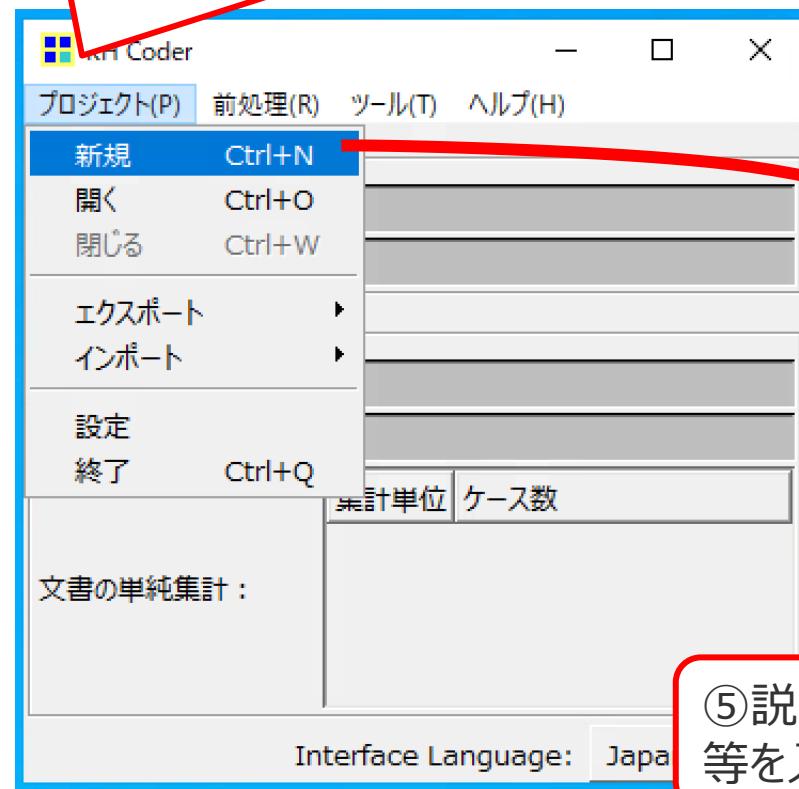
補足:

Jupyter Lab 上からダウ
ンロードした「**rakuten-
1000-2022-2023.
xlsx**」を使用しても構い
ません

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

③「コメント」
を選択（注2）

④「MeCab」
を選択

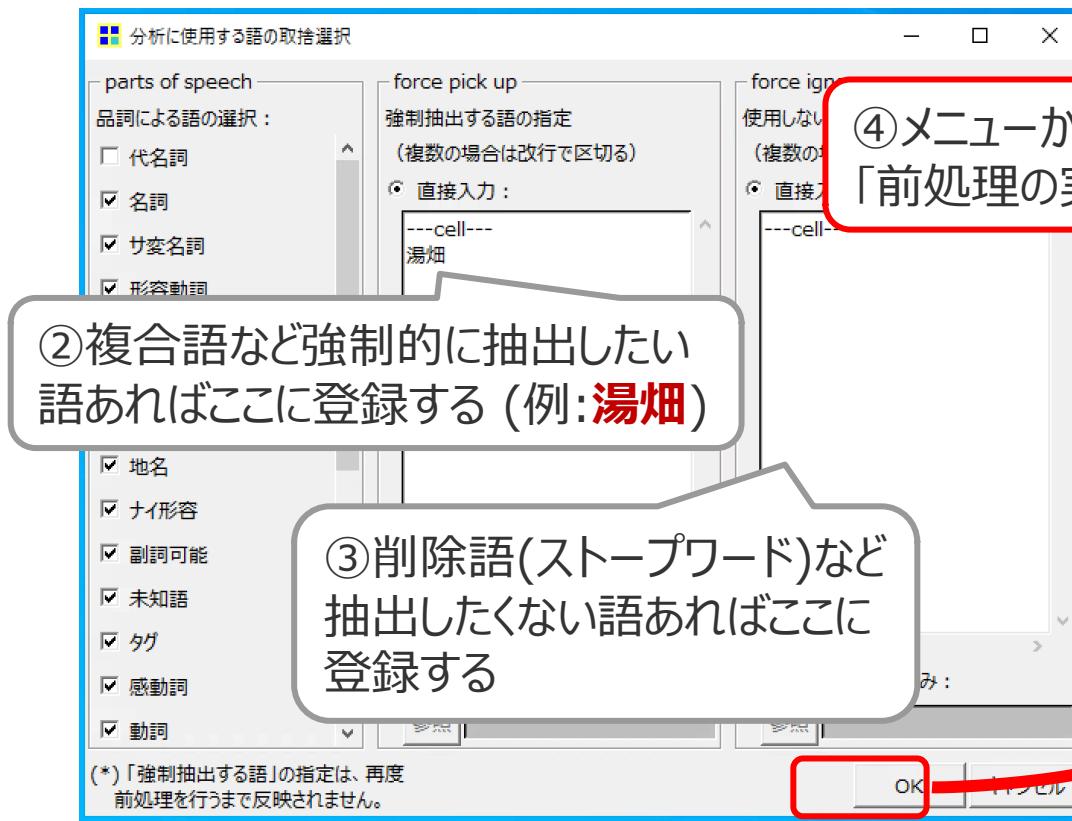
⑤説明「楽天トラベル」
等を入力

⑥「OK」をクリック

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



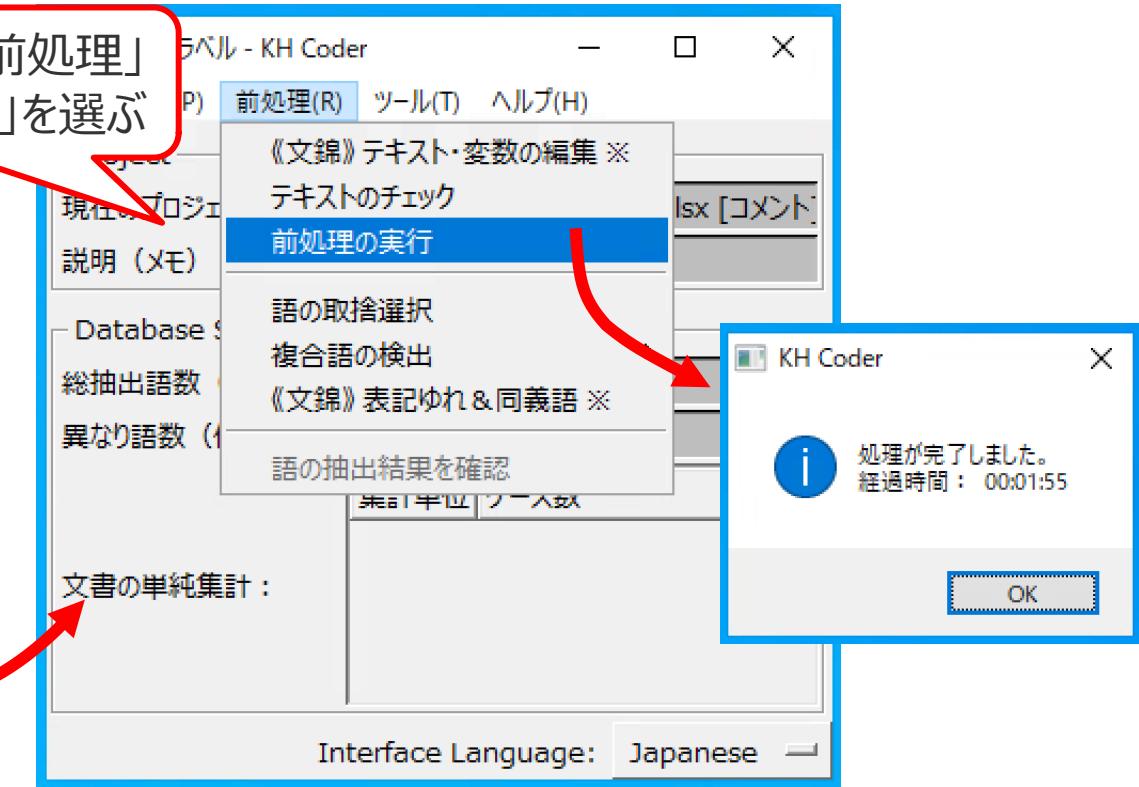
②複合語など強制的に抽出したい語あればここに登録する (例:湯畑)

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

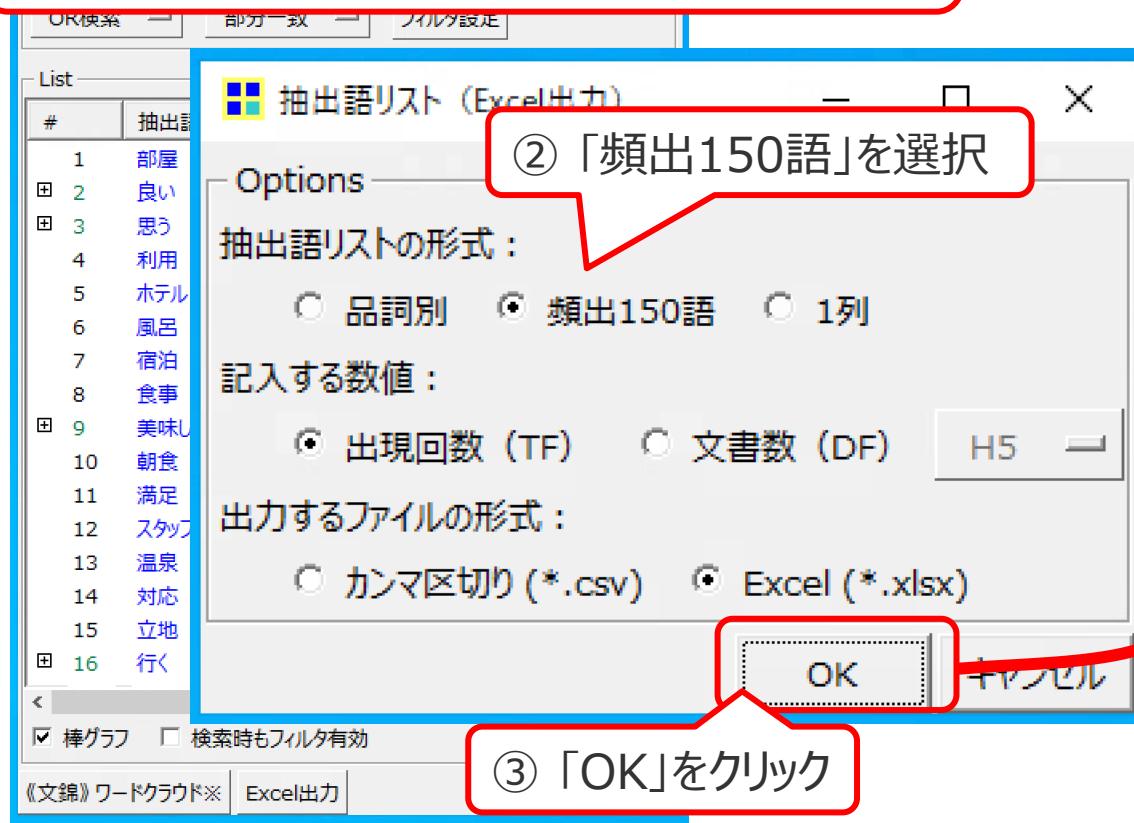
注2: メニューから「前処理」「複合語の検出」を選ぶと,複合語候補の一覧を出力できます



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

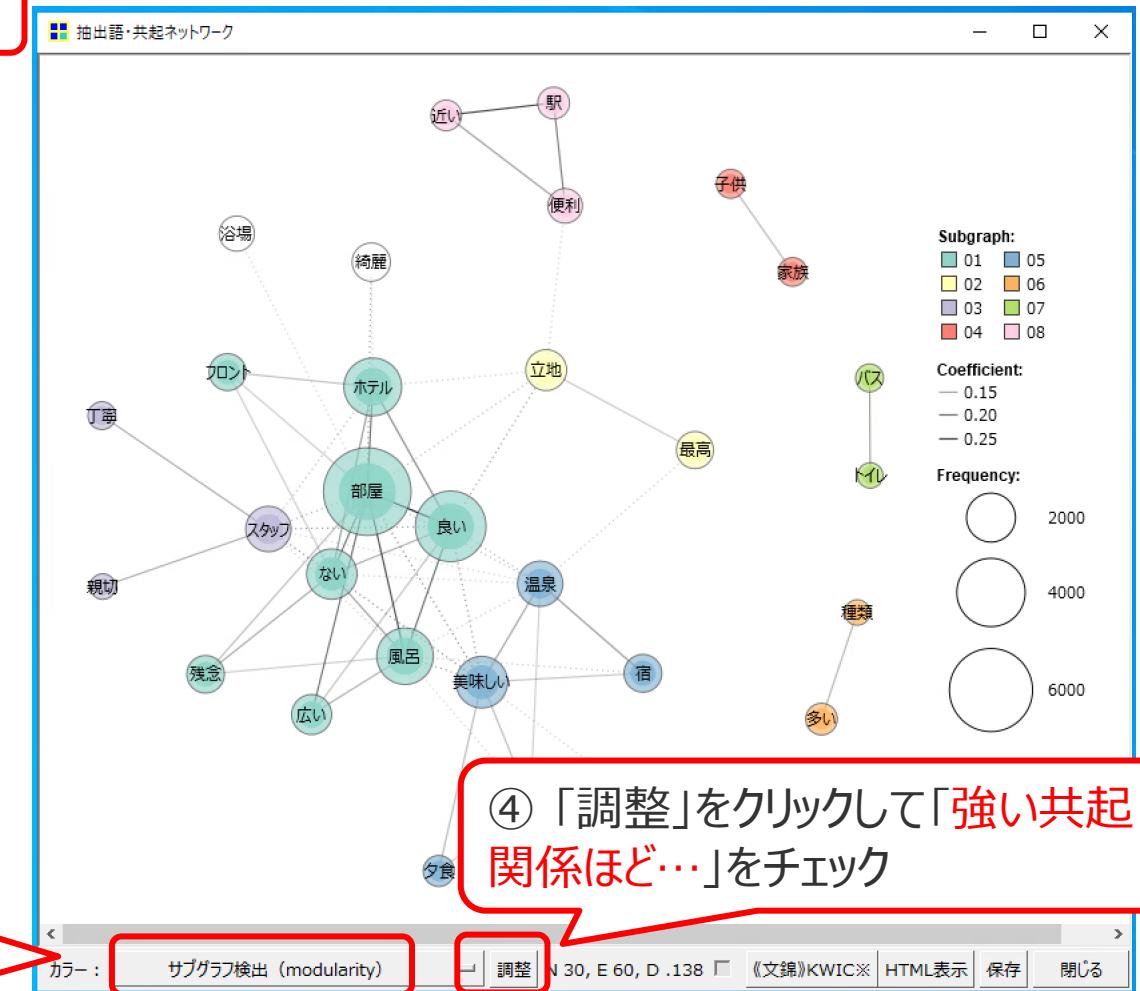
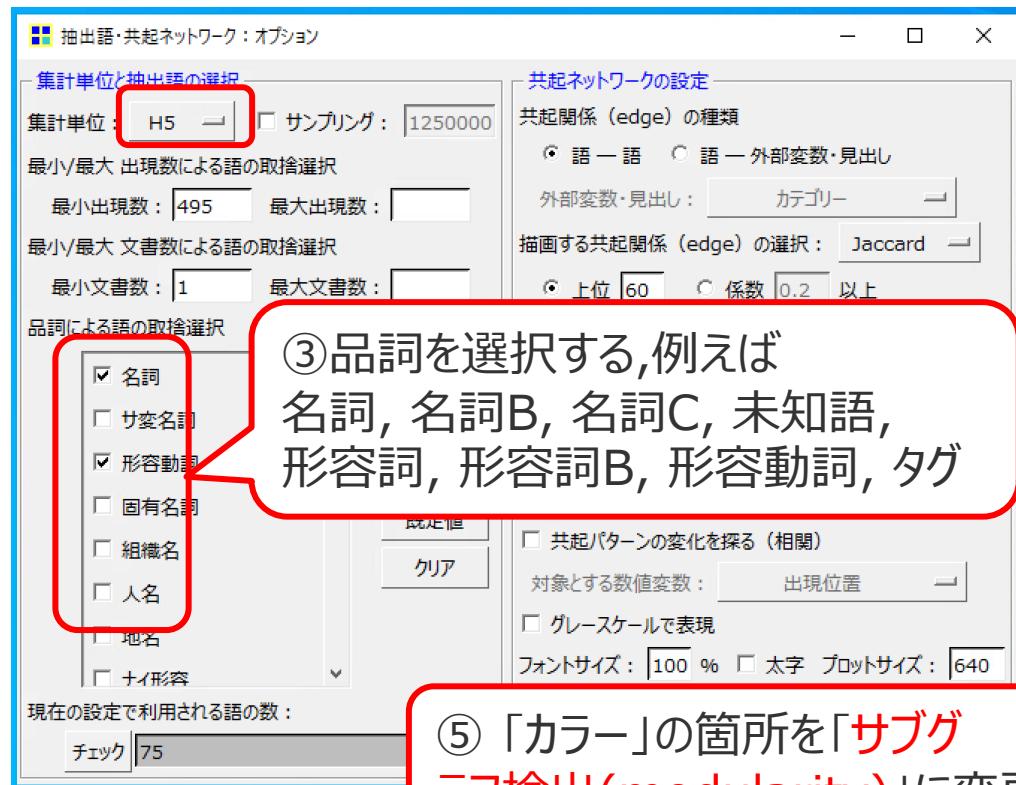


A	B	C	D	E	F	G	H
1	抽出語	出現回数	抽出語	出現回数	抽出語	出現回数	
2	部屋	6689	子供	661	プラン	389	
3	良い	4302	過ごす	657	見える	388	
4	思う	3976	家族	648	機会	387	
5	利用	3481	予約	636	設備	387	
6	ホテル	2831	過ごせる	626	旅館	386	
7	風呂	2702	駐車	613	置く	384	
8	宿泊	2649	素晴らしい	612	きれい	377	
9	食事	2447	月	611	歩く	368	
10	美味しい	2249	バス	610	湯	359	
11	朝食	2172	丁寧	610	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



(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

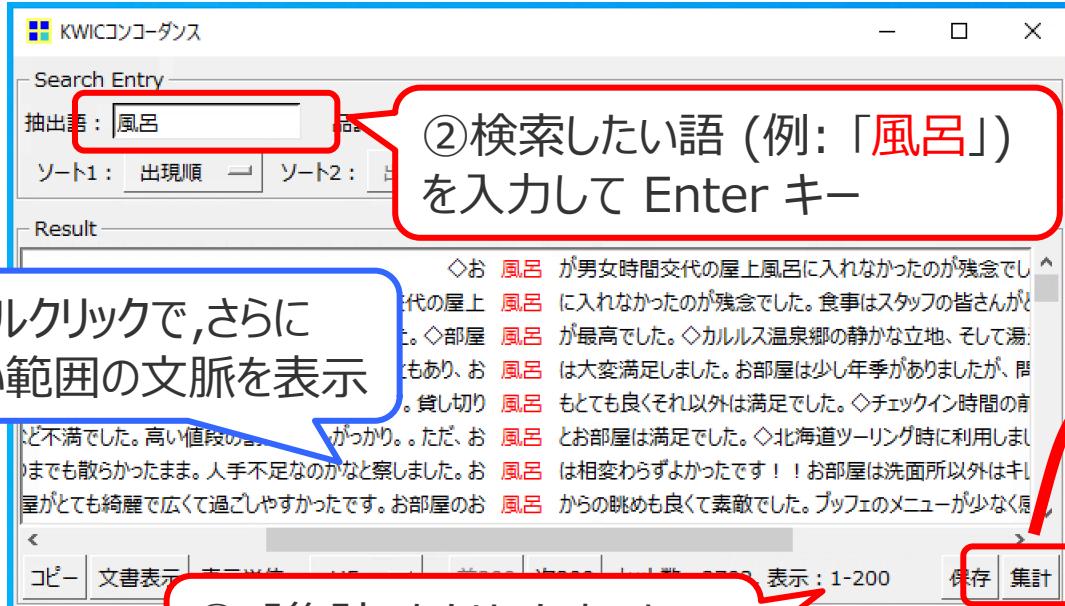
注: どの品詞を選択すべきかは、分析対象のデータや分析目的により異なります。

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



- ③「集計」をクリックするとコロケーション統計(右)を開く

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	

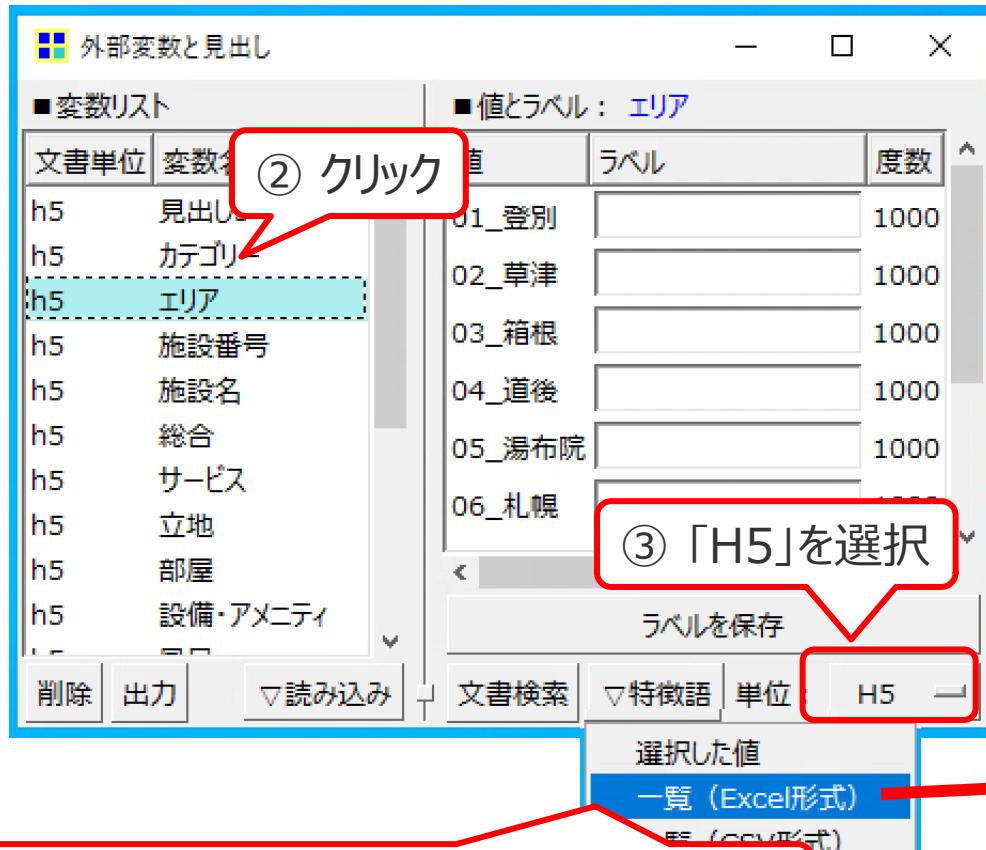
- ④表示する語の品詞を選択
(例: 形容詞, 形容詞B, 形容動詞)

- ⑤「右合計」でソート

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



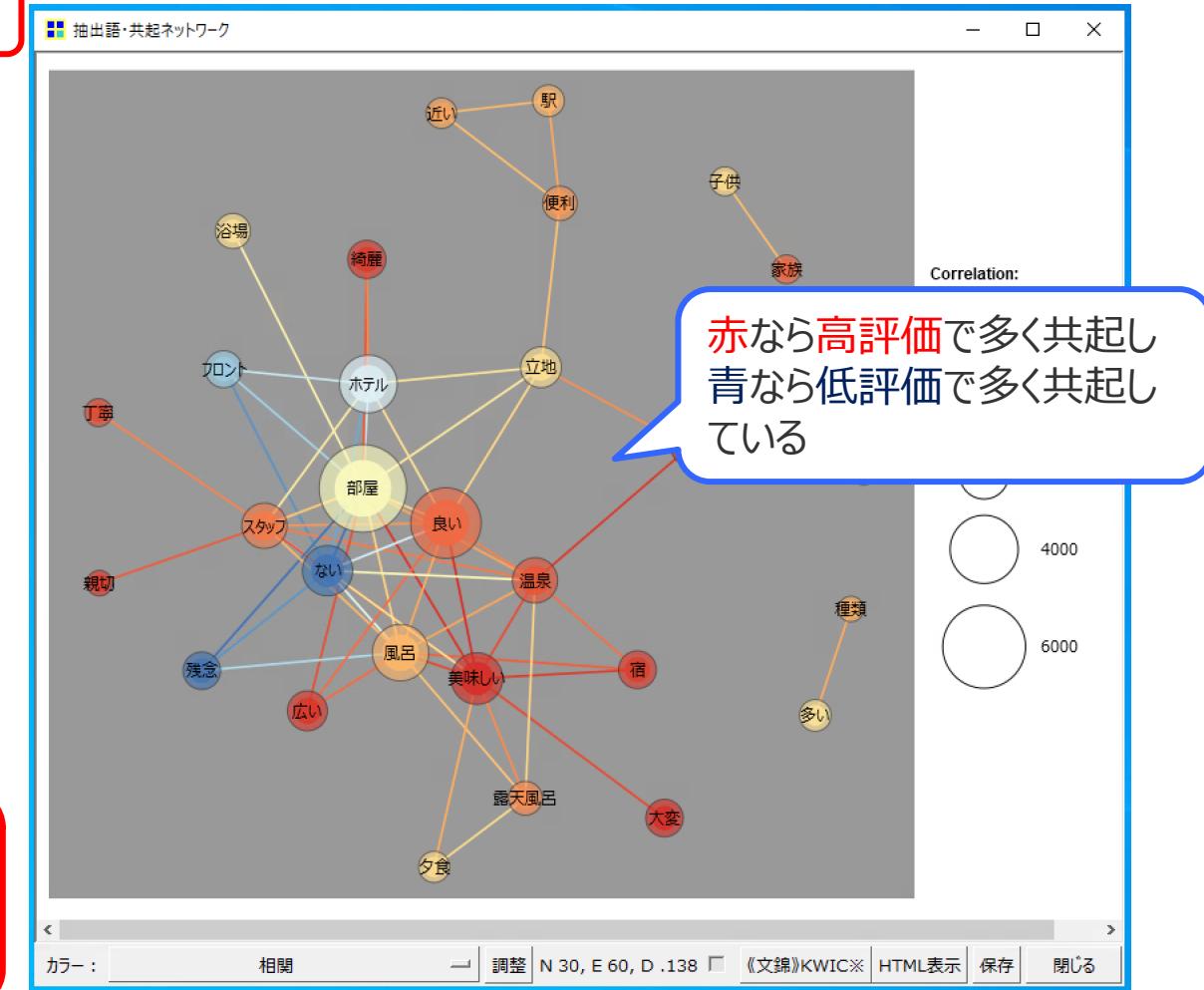
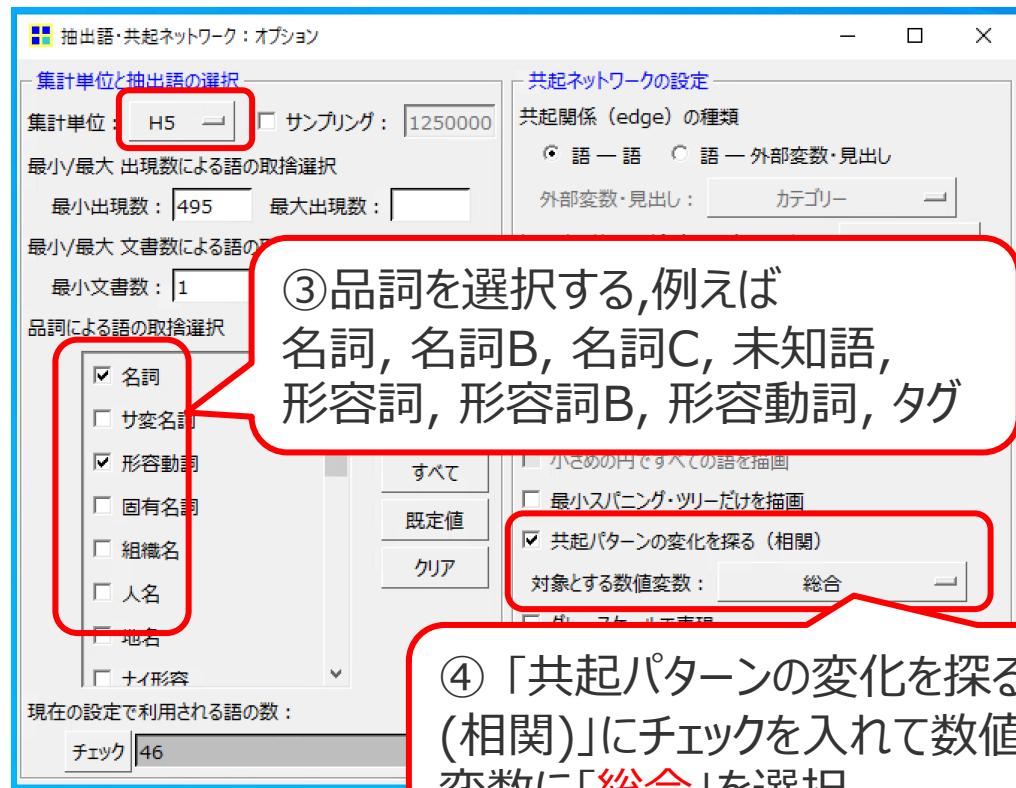
	A	B	C	D	E	F	G	H	I	J	K
1											
2	01_登別		02_草津		03_箱根		04_道後				
3	風呂	.115	湯畑	.327	美味しい	.136	温泉	.109			
4	温泉	.107	温泉	.136	露天風呂	.134	立地	.082			
5	美味しい	.094	風呂	.126	風呂	.116	最高	.066			
6	良い	.093	宿	.120	部屋	.109	広い	.063			
7	ピング	.090	美味しい	.102	良い	.106	浴場	.059			
8	残念	.078	良い	.100	温泉	.102	よい	.058			
9	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

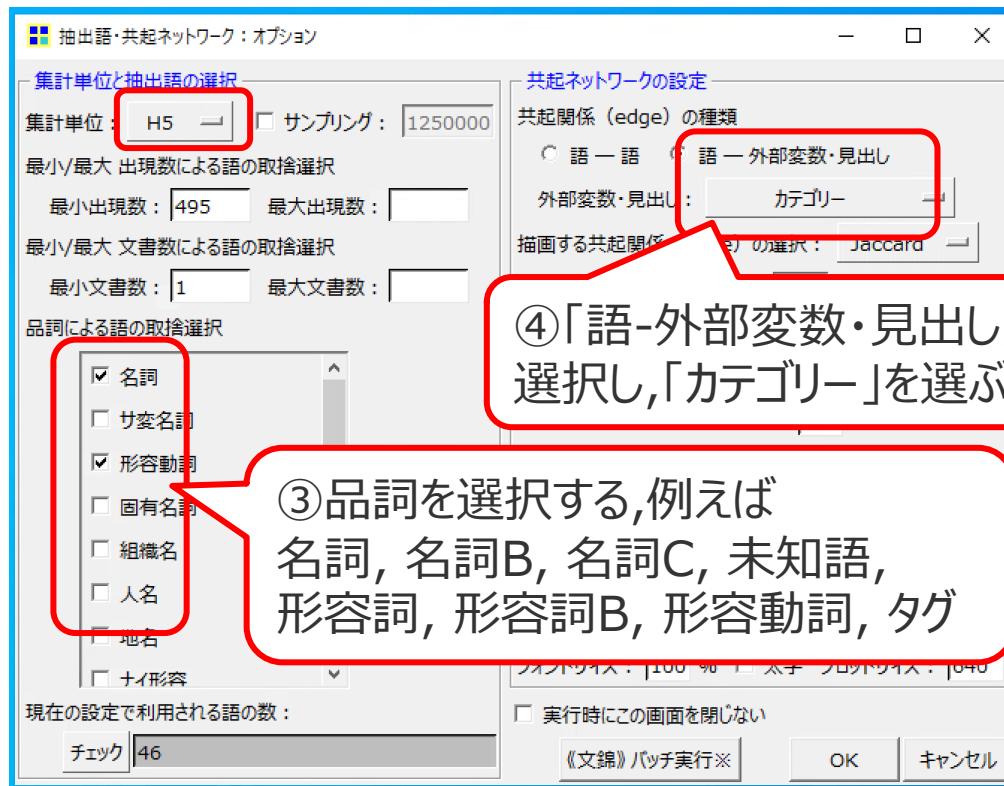


KHCoder の使い方

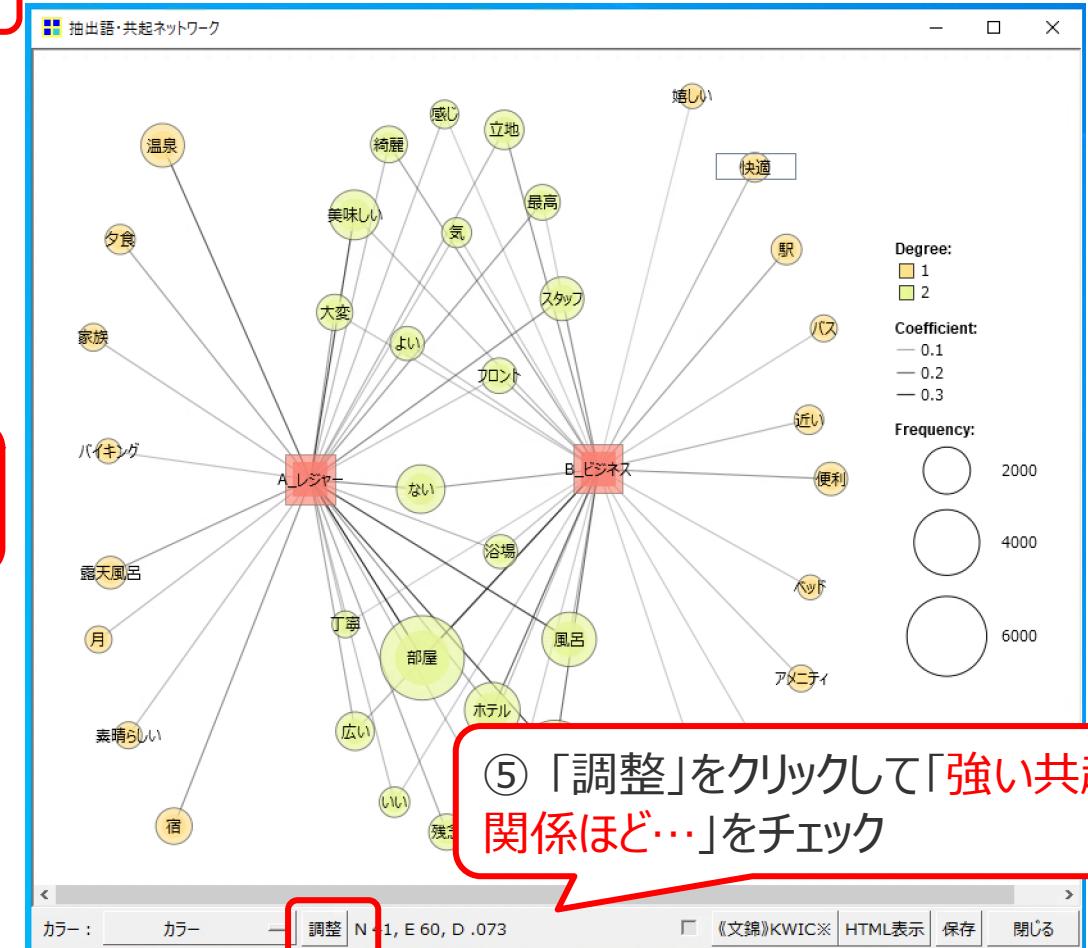
● 共起ネットワークの作成(3)

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



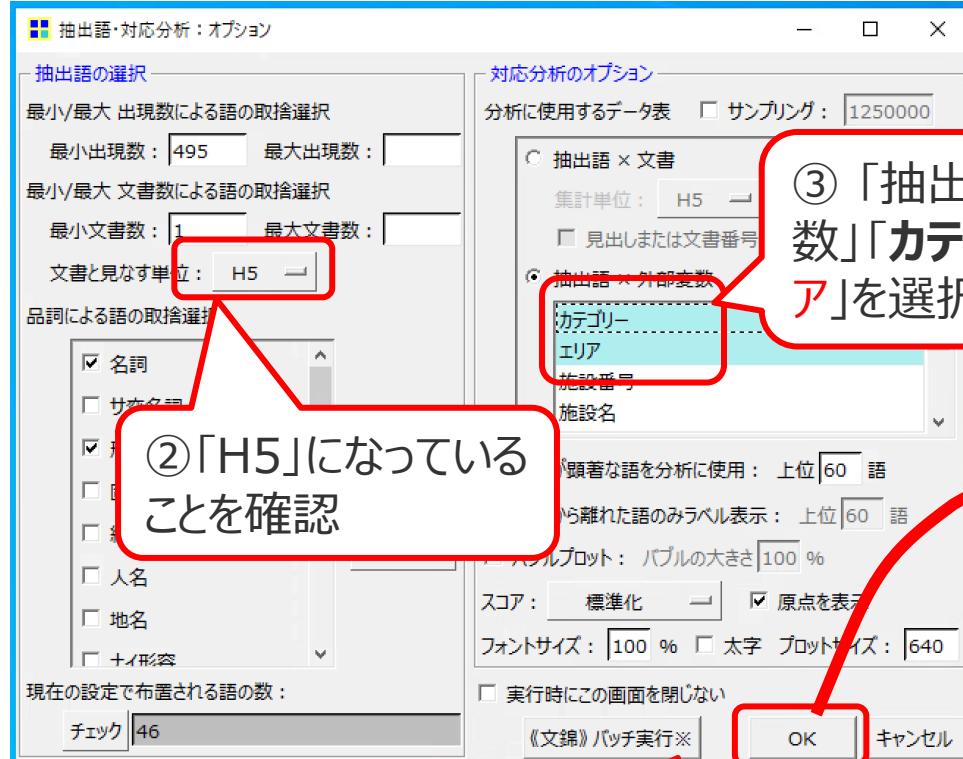
③品詞を選択する,例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B, 形容動詞, タグ



KHCoder の使い方

● 対応分析による探索(1)

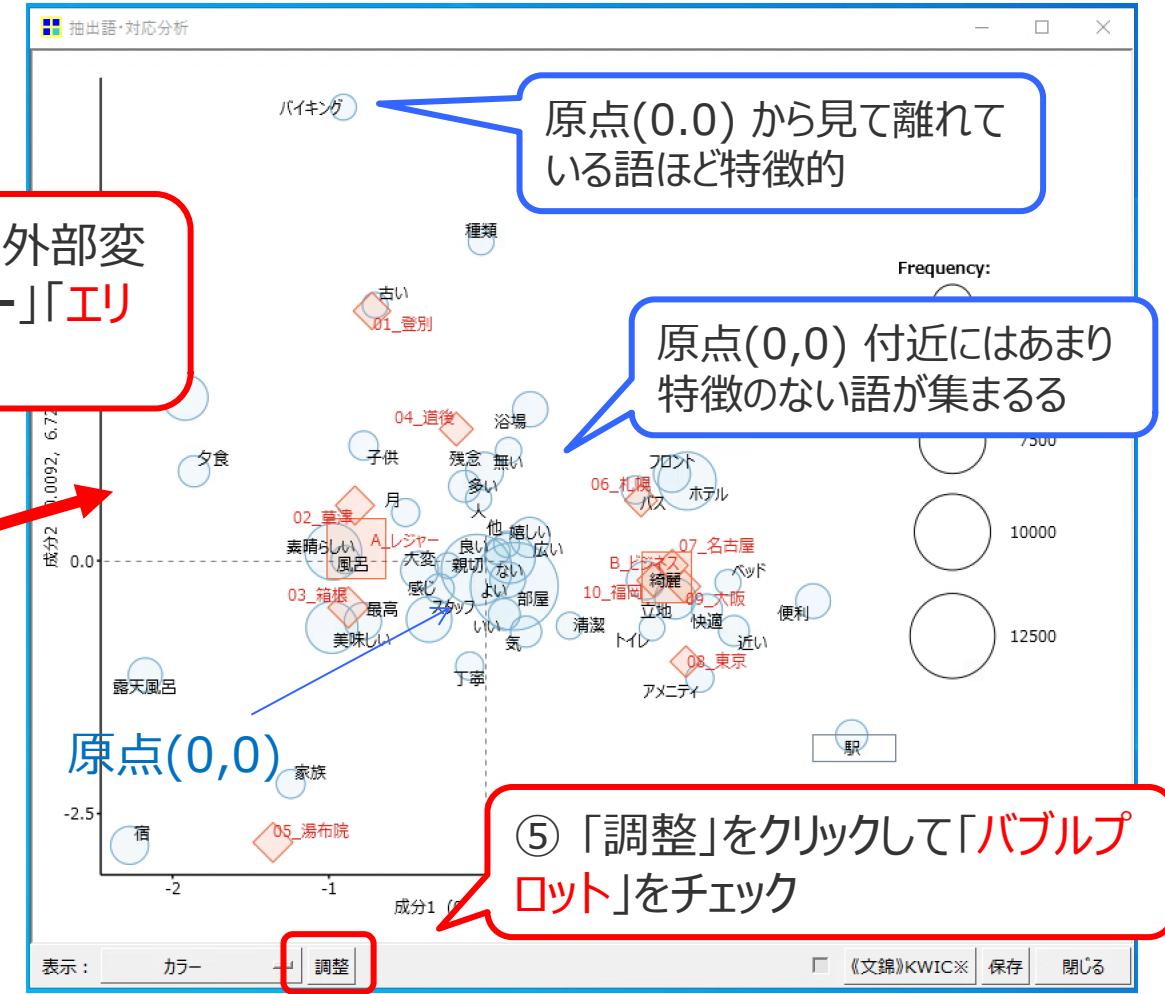
- ① メニューから「ツール」「抽出語」「対応分析」を選ぶ



② 「H5」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

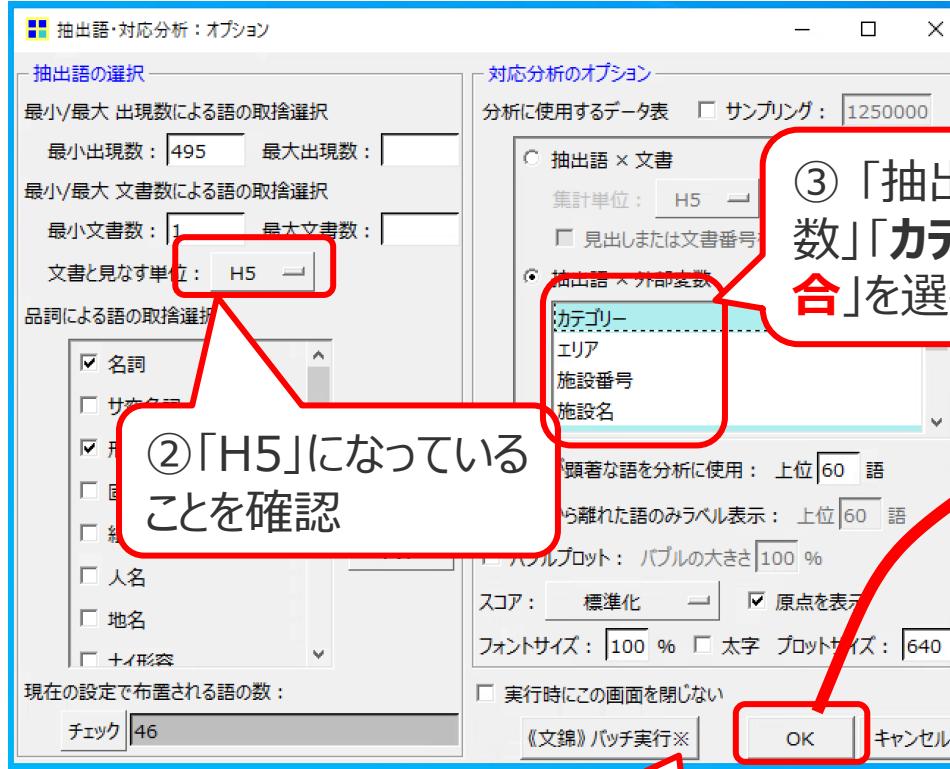
④ 「OK」をクリック



KHCoder の使い方

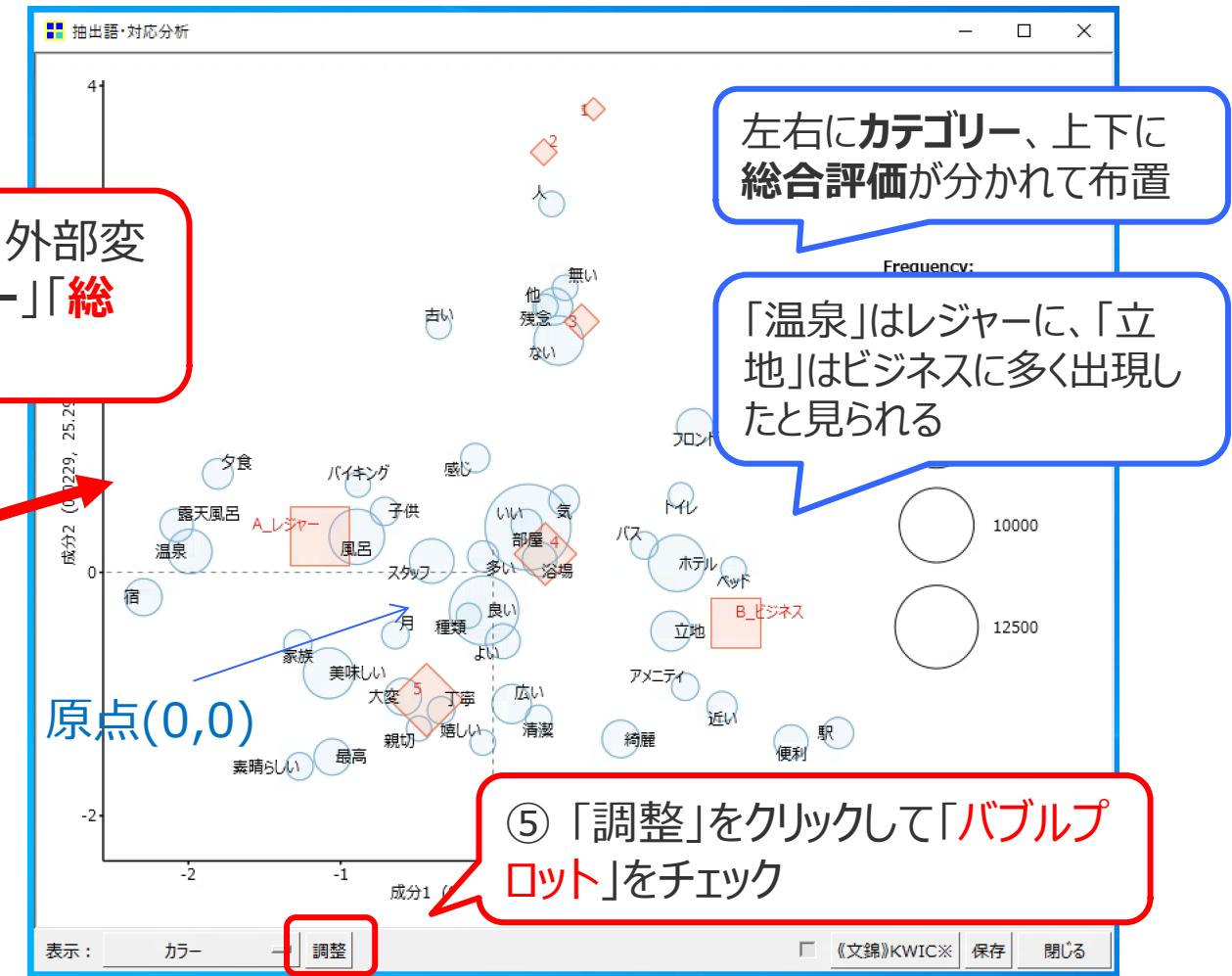
● 対応分析による探索(2)

- ① メニューから「ツール」「抽出語」「対応分析」を選ぶ



③ 「抽出語×外部変数」「カテゴリー」「総合」を選択

④ 「OK」をクリック



左右にカテゴリー、上下に総合評価が分かれて布置

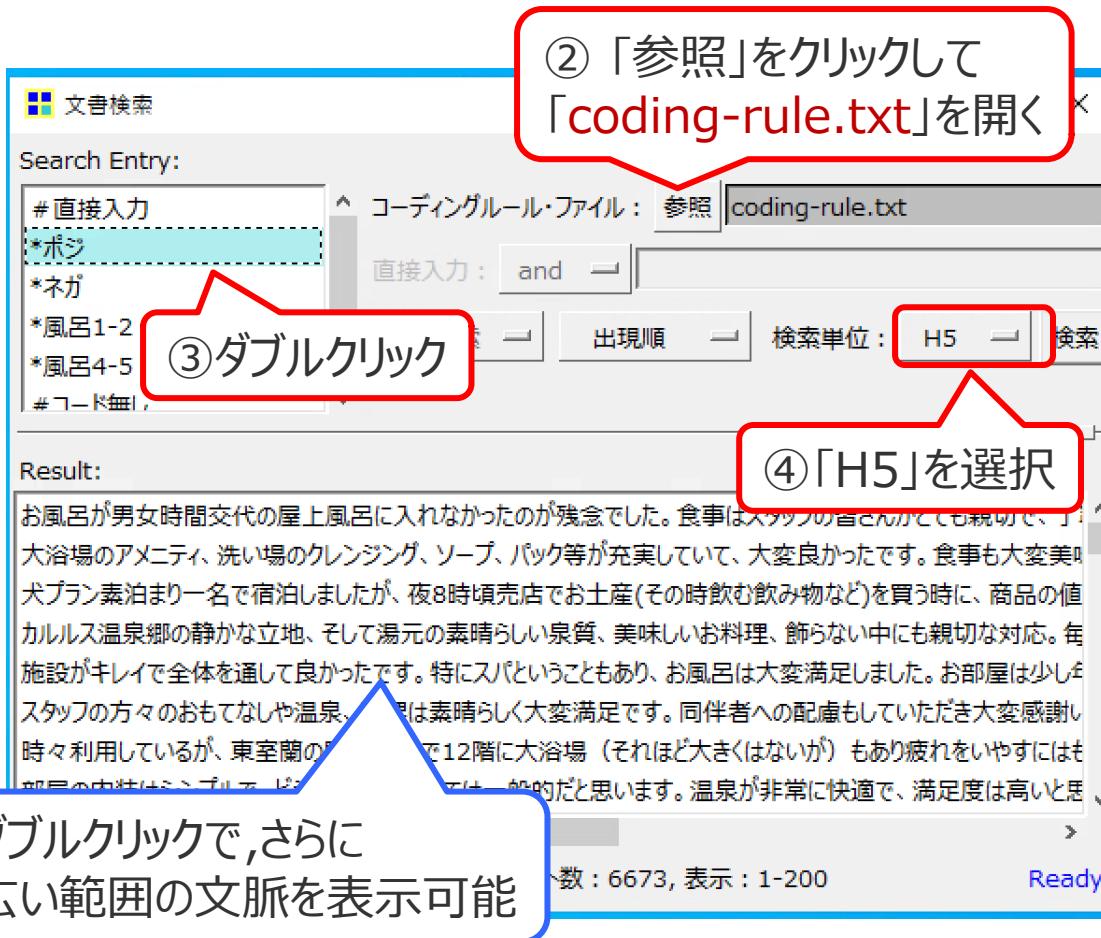
「温泉」はレジャーに、「立地」はビジネスに多く出現したと見られる

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

*風呂4-5

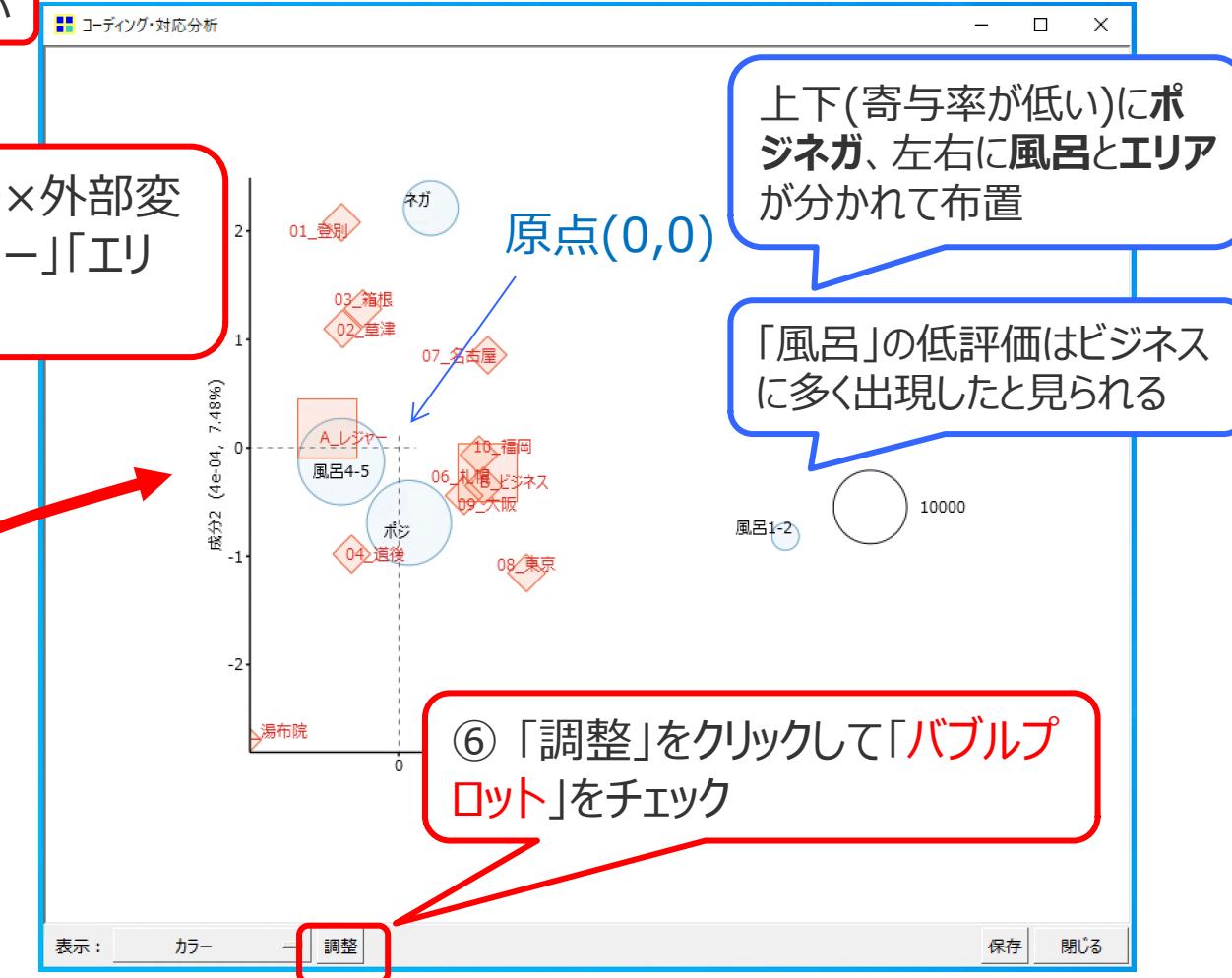
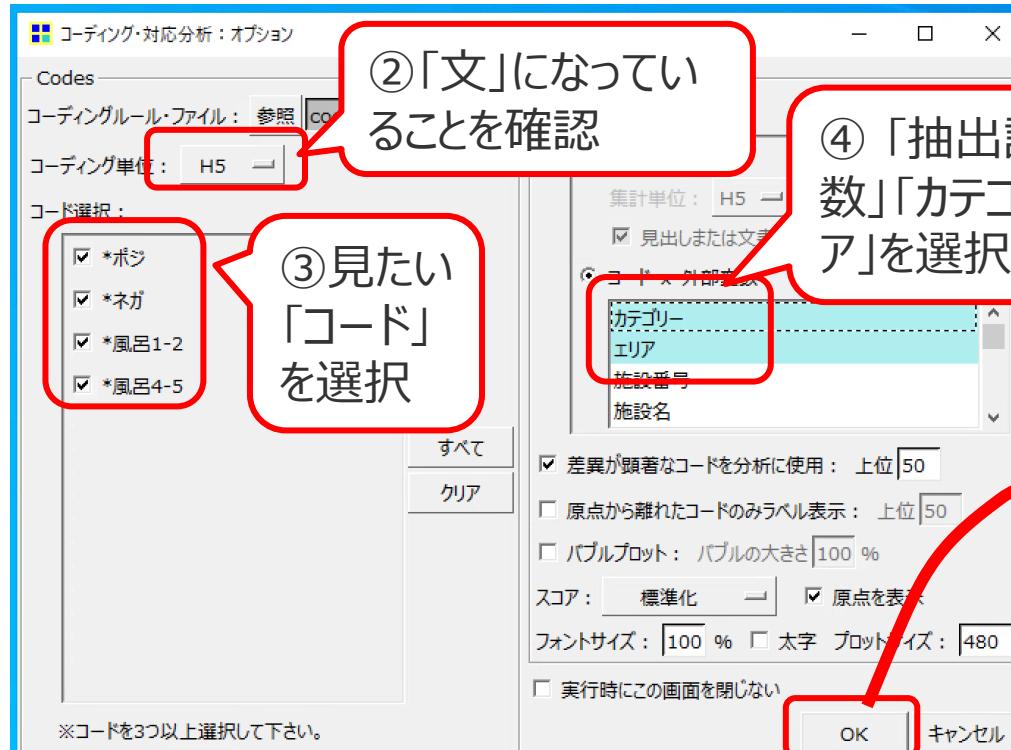
<>風呂-->4 | <>風呂-->5

外部
変
数

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



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

③ H5 (=セル)を選択

④ エリアを選択

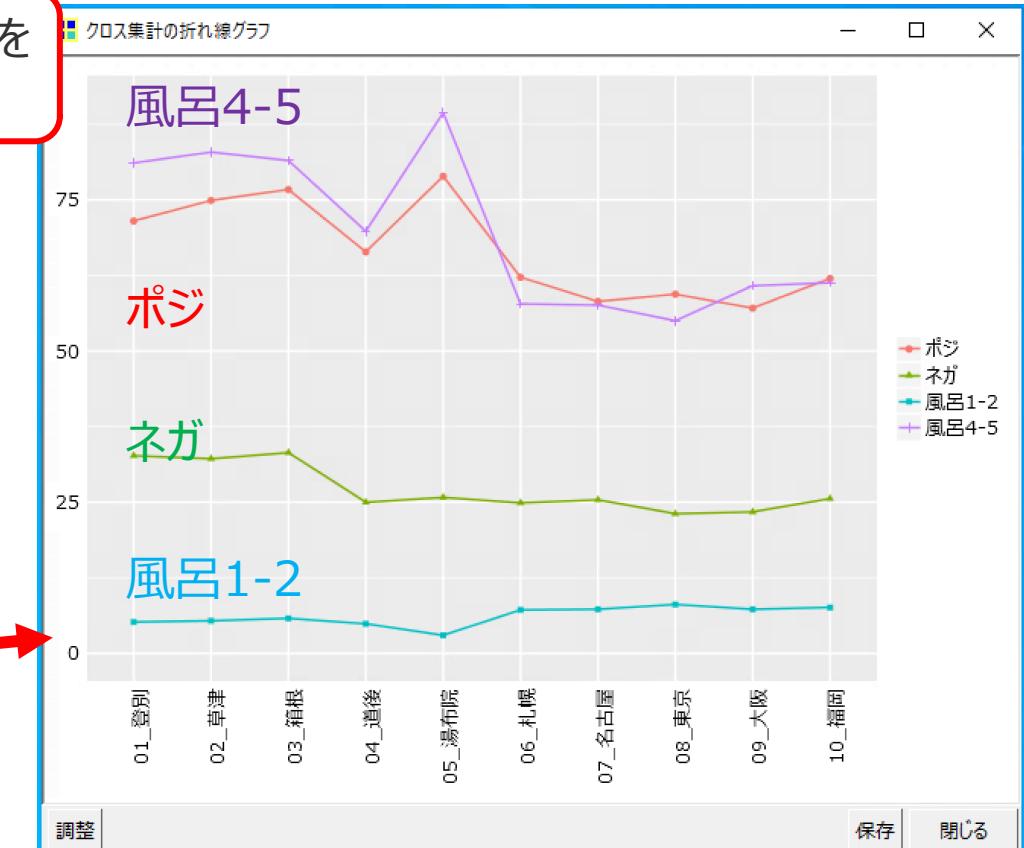
⑥ すべて

② 参照

⑤ 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

注: プロット左側のラベルは表示されません



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.49)

※ 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.