

テキストマイニング

— Part 3 —

2022年度 春C
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 — 自然言語処理の最新動向
 - 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 — データ分析の実説明
- Part 5
 - 説明 — ラップアップ

(再掲) 実習で使用するデータ

楽天
トラベル

- ホテルのクチコミ数: 1,237万件 ※年間約60~70万



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

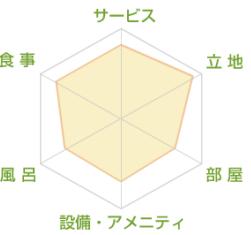
アンケート件数：886件

評価内訳

- 5点 ■■■■■ 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ます。

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

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。

テキストデータ

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。

数値評価

(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

(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) データ理解 — 集計例

件数 (エリア別)

行ラベル	個数 / コメント
■ 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

投稿者の傾向 (年代別, 性別)

行ラベル	個数 / コメン	列ラベル	男性	女性	.	総計
10代			0.03%	0.03%	0.00%	0.06%
20代			1.04%	1.21%	0.00%	2.25%
30代			2.05%	2.14%	0.00%	4.19%
40代			5.66%	3.57%	0.00%	9.23%
50代			9.08%	4.31%	0.00%	13.39%
60代			4.58%	1.53%	0.00%	6.11%
70代			0.92%	0.21%	0.00%	1.13%
80代			0.04%	0.00%	0.00%	0.04%
90代			0.01%	0.00%	0.00%	0.01%
110代			0.00%	0.02%	0.00%	0.02%
120代			0.01%	0.00%	0.00%	0.01%
.			0.00%	0.00%	63.56%	63.56%
総計			23.42%	13.02%	63.56%	100.

投稿者の傾向 (性別, カテゴリ別)

行ラベル	A_レジャー	B_ビジネス	総計
男性	23.36%	23.48%	23.42%
女性	15.62%	10.42%	13.02%
.	61.02%	66.10%	63.56%
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(.)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

(復習) データ理解 — 集計例

投稿者の傾向 (性別, カテゴリ-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	
行ラベル	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	B_ビジネス 集計	総計	
男性	24.70%	24.60%	17.30%	27.20%	23.00%	23.36%	23.30%	25.60%	19.80%	23.60%	25.10%	23.42%	
女性	13.30%	15.00%	16.30%	12.00%	21.50%	15.62%	11.50%	9.20%	10.90%	11.00%	9.50%	10.42%	13.02%
.	62.00%	60.40%	66.40%	60.80%	55.50%	61.02%	65.20%	65.20%	69.30%	65.40%	65.40%	66.10%	63.56%
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)

投稿者の傾向 (年代別, カテゴリ-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					B_ビジネス 集計	総計
行ラベル	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	B_ビジネス 集計	総計	
10代	0.00%	0.00%	0.40%	0.10%	0.00%	0.10%	0.00%	0.00%	0.00%	0.10%	0.00%	0.02%	0.06%
20代	0.80%	3.10%	4.50%	2.10%	3.50%	2.80%	1.90%	1.30%	2.50%	1.70%	1.10%	1.70%	2.25%
30代	4.80%	5.20%	4.20%	4.30%	6.50%	5.00%	3.50%	3.90%	3.10%	3.40%	3.00%	3.38%	4.19%
40代	10.30%	9.40%	6.90%	7.80%	9.10%	8.70%	9.60%	11.40%	8.50%	9.00%	10.30%	9.76%	9.23%
50代	13.20%	13.50%	9.00%	16.40%	14.10%	13.24%	14.20%	12.80%	12.00%	14.00%	14.70%	13.54%	13.39%
60代	7.60%	6.40%	6.80%	7.50%	9.40%	7.54%	4.80%	3.90%	4.30%	5.70%	4.70%	4.68%	6.11%
70代	1.30%	1.90%	1.70%	0.90%	1.70%	1.50%	0.80%	1.40%	0.20%	0.70%	0.70%	0.76%	1.13%
80代	0.00%	0.10%	0.10%	0.10%	0.00%	0.06%	0.00%	0.00%	0.10%	0.00%	0.00%	0.02%	0.04%
90代	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.01%
110代	0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.02%
120代	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
.	62.00%	60.40%	66.40%	60.80%	55.50%	55.50%	55.50%	55.50%	55.50%	55.50%	55.50%	55.50%	55.50%
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

(復習) 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)

投稿者の傾向 (同行者別, カテゴリ)

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
行ラベル		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		27.90%	16.40%	14.60%	45.40%	18.90%	24.64%	61.00%	65.90%	66.40%	58.30%	62.60%	62.84%	43.74%	
家族		57.30%	58.60%	61.20%	41.70%	64.60%	56.68%	25.10%	20.20%	19.40%	25.50%	25.30%	23.10%	39.89%	
恋人		6.60%	16.10%	15.30%	5.70%	9.20%	10.59%	6.50%	5.90%	7.80%	6.90%	4.70%	6.36%	8.47%	
友達		5.20%	7.50%	7.80%	4.40%	5.90%	6.16%	4.70%	3.80%	4.00%	6.90%	4.70%	4.82%	5.49%	
仕事仲間		1.90%	0.80%	0.40%	2.30%	0.50%	1.18%	1.60%	3.20%	1.10%	1.30%	2.20%	1.88%	1.53%	
その他		1.10%	0.60%	0.70%	0.50%	0.90%	0.76%	1.10%	1.00%	1.30%	1.10%	0.50%	1.00%	0.88%	
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

数値評価の構成 (総合)

- 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

- 高評価は、レジャーがビジネスよりも多い

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
行ラベル		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		44.70%	53.80%	48.80%	53.30%	69.20%	53.96%	50.60%	45.30%	53.80%	52.40%	52.10%	50.84%	52.40%	
4		37.99%	32.90%	37.30%	35.20%	23.10%	33.28%	38.50%	39.90%	31.50%	36.60%	34.50%	36.20%	34.74%	
3		10.20%	7.70%	8.30%	7.30%	4.90%	7.68%	7.40%	9.80%	9.60%	7.50%	8.50%	8.56%	8.12%	
2		4.50%	3.90%	3.40%	2.40%	1.90%	3.22%	2.10%	3.30%	2.80%	2.20%	3.30%	2.74%	2.98%	
1		2.70%	1.70%	2.20%	1.80%	0.90%	1.86%	1.40%	1.70%	2.30%	1.30%	1.60%	1.66%	1.76%	
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、登別や箱根が少ない

- ビジネスの高評価は、東京と大阪が多い、名古屋がやや少ない

(復習) データ理解 — 集計例

数値評価の平均 (エリア別, 数値評価別)

- レジャーは、風呂や食事が、設備や部屋に比べて高評価

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	4.10	4.17	4.29	4.36
05_湯布院	4.58	4.39	4.40	4.30	4.52	4.60	4.58
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.22	4.22
06_札幌	4.17	4.42	4.26	4.07	3.96	4.22	4.22
07_名古屋	4.07	4.29	4.17	3.99	3.91	4.17	4.22
08_東京	4.13	4.43	4.20	4.04	3.88	4.21	4.32
09_大阪	4.16	4.42	4.24	4.06	3.97	4.17	4.37
10_福岡	4.17	4.43	4.22	4.05	3.94	4.25	4.32

- 湯布院は、レジャーの中で、軒並み高評価が多い

数値評価の平均 (カテゴリ別, 数値評価別)

- 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

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

(復習) データ理解 — 集計例

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	平均 / 風呂	平均 / 食事	平均 / 総合
■ A_レジャー	4.56	4.42	4.40	4.38			
男性	4.54	4.40	4.41	4.40	4.60	4.53	4.61
女性	4.58	4.44	4.39	4.36	4.58	4.48	4.56
■ B_ビジネス	4.31	4.46	4.28	4.20	4.02	4.31	4.43
男性	4.18	4.45	4.24	4.11	3.91	4.25	4.39
女性	4.48	4.47	4.34	4.31			

- 20~30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

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	4.32	4.33	4.19	4.08			
男性	4.27	4.31	4.15	4.05			
女性	4.39	4.35	4.26				
■ B_ビジネス	4.15	4.39	4.23	4.06			
男性	4.06	4.34	4.18	3.99			
女性	4.36	4.52	4.37	4.24	4.10	4.31	4.41

- 年齢が高くなるに連れてレジャーに対するサービスや風呂,食事の評価に厳しくなる
- 女性は年齢が高くなるに連れてビジネスに対する評価が高くなる

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	平均 / 総合
■ A_レジャー	4.24	4.21	4.14	4.01	4.37	4.26	4.32
男性	4.19	4.17	4.10	3.97	4.33	4.19	4.28
女性	4.38	4.35	4.25	4.14	4.47	4.47	4.42
■ B_ビジネス	4.07	4.48	4.11				
男性	4.05	4.45	4.11				
女性	4.13	4.57	4.11				

- 60~90代男性はアメニティに対する関心が低い
- 60~90代はビジネスに立地に対する期待が高い

(復習) データ理解 — まとめ

	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（“.”）中が、ある年代や性別に偏っている可能性
目的別	<ul style="list-style-type: none"> レジャーは家族が多い、ビジネスは一人が多い（出張は単独） レジャーの中でも、道後は男性の一人客が多い（道後はもはや仕事で行く場所） 	<ul style="list-style-type: none"> レビューの観点が性別によって偏っている可能性 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、少ないのは登別 ビジネスの中で高評価が多いのは大阪と札幌、少ないのは東京都と名古屋だが僅差 	<ul style="list-style-type: none"> 好評価しか投稿しない→コメントが好評価に偏っている可能性 旅行目的によって投稿の動機が異なっている可能性
数値評価 (項目ごと)	<ul style="list-style-type: none"> レジャーの評価は、風呂や食事 > 設備や部屋 ビジネスの評価は、立地 > その他 レジャーの中で湯布院は軒並み高評価 レジャーもビジネスも立地は高評価 	<ul style="list-style-type: none"> 旅行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

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

数値評価の平均 (レジャー, ビジネス別)

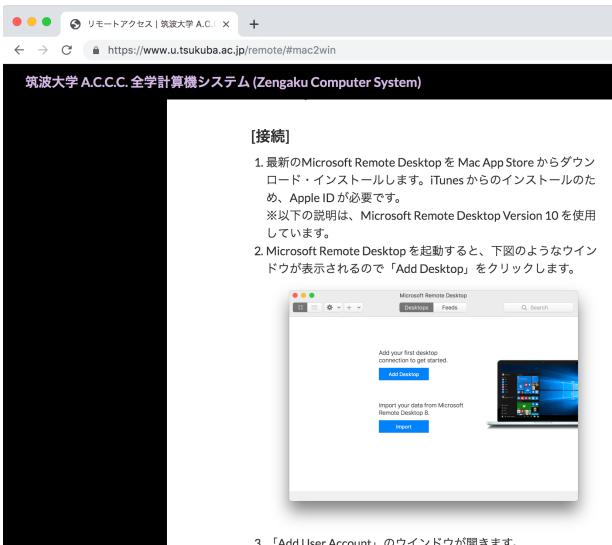
行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.29	4.29	4.18	4.07	4.34	4.29	4.34
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32

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

- ・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 official website for KH Coder. At the top right are language options for Japanese and English. Below the header is a large blue banner with the text "KH Coder". The main content area includes:

- Index**: A section about an on-demand seminar.
- 概要**: A brief introduction to KH Coder.
- 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.

On the right side, there are several social media posts and tweets from users like "khcoderさん" and "JSS 日本社会学会 (非公式)".

参考書 **New**



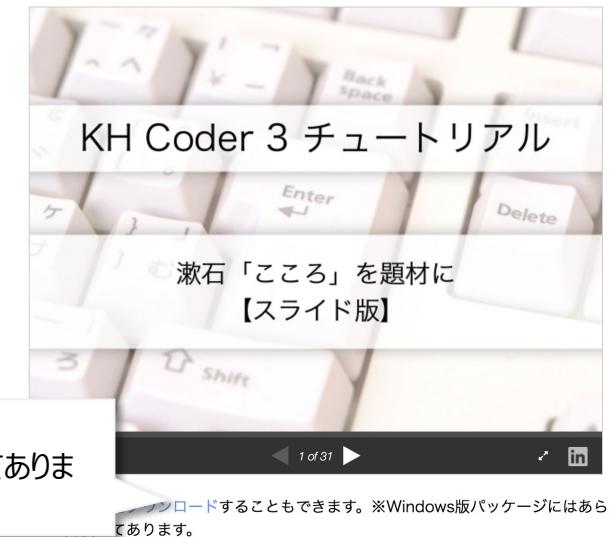
PDFファイルをダウンロードすることもできます。
※Windows版パッケージにはあらかじめ同梱してあります。

チュートリアル

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

チュートリアル & ヒント

[KH Coder]



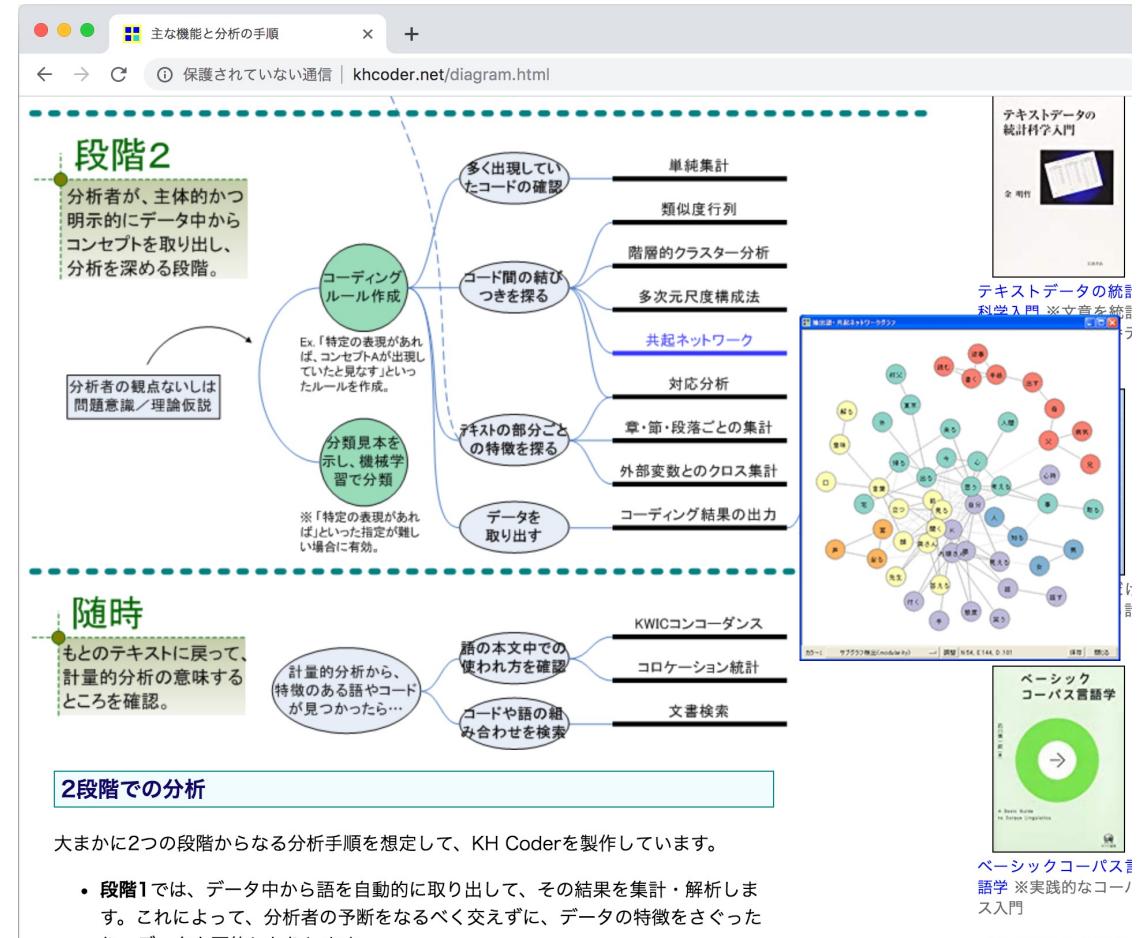
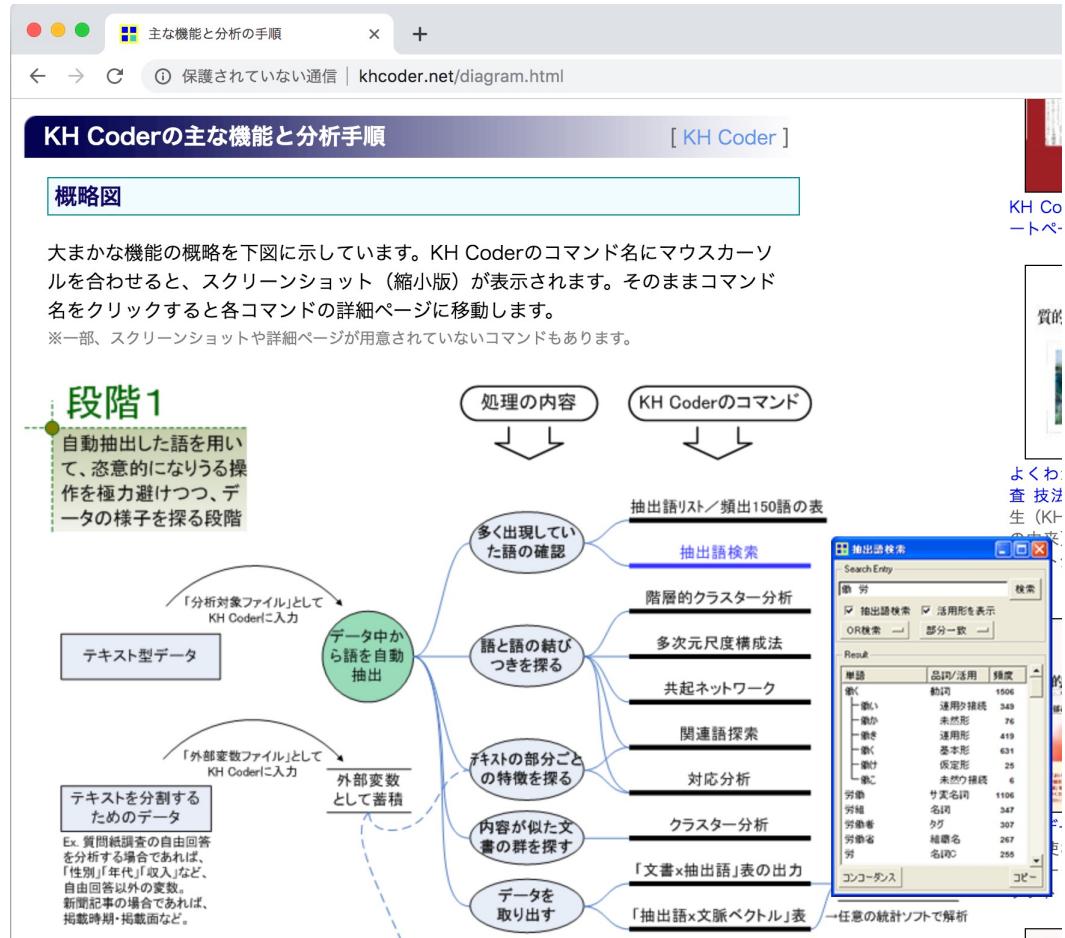
チュートリアル用データ

(2018.01.25)

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。

(参考) KH Coder の分析手順

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



ちょっとその前に…

- 自然言語処理
 - 人間の言語において, 単語や文が持つ意味をどう扱うか (=常識)
 - 分布仮説や分散表現によるアプローチ
- テキストマイニング
 - 対象の文書において, 単語や文が持つ特徴をどう扱うか (=非常識?)
 - 数量的な分析によるアプローチ

具体的には **特徴的な 語 や 関連** を見つける

ちょっとその前に…

他と差がある

- 特徴的な**語**を見つける

- 特定の文書に特有の語を見つける → TF・IDF
 - その文書に特に頻出するが、他の文書ではそれほどではない

他と差がある

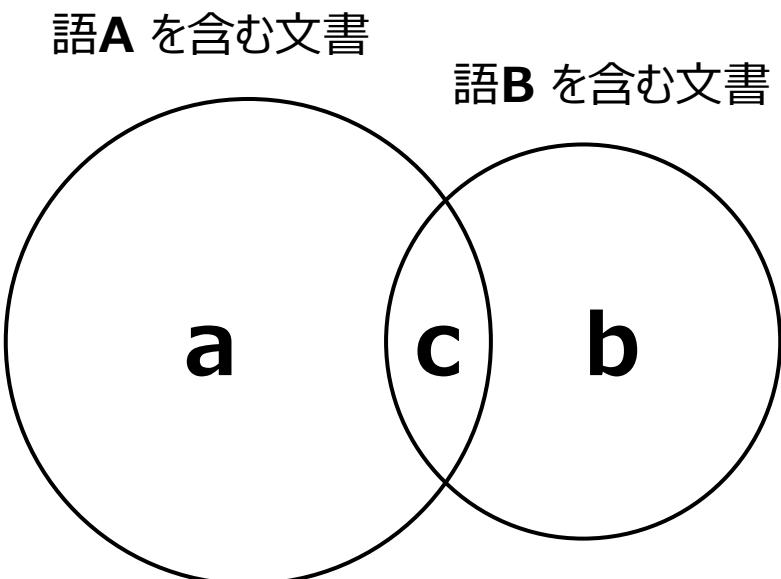
- 特徴的な**関連**(=同時に出現している)を見つける

- 語と語(またはカテゴリー)が共起する
 - 例) 「風呂」と「広い」が共起している
- 語と語(またはカテゴリー)の出現パターンが似てる
 - 例) 「レジャー」と「風呂」の出現パターンが似てる

KHCoder の原理:

関連の強さを測る – 共起尺度

- **共起の強さ**を測る → **Jaccard 係数** (KHCoderで標準的に用いられる)
- 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**



$$\text{語Aと語BのJaccard 係数} = \frac{c}{a+b+c}$$

- 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

KHCoder の原理:

関連の強さを測る — 距離尺度

- ・出現パターンが似てるを測る = ユークリッド距離, コサイン距離
- ・1つひとつの文が長く, 各文中での語の出現回数の大小が重要なケースに向いている (語が1回出現したか, 10回出現したかを区別したい)

ユークリッド距離	コサイン距離
<ul style="list-style-type: none">・サイズ(出現回数の大小)の差までも見る場合向き $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$	<ul style="list-style-type: none">・傾きが似ているかどうかだけを見る場合向き $d(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$

※ \mathbf{x}, \mathbf{y} はそれぞれの単語ベクトル (単語の出現パターン→後述)

※ 文中 (1,000語あたり) における語の出現回数を計算

KHCoder の原理:

単語の出現パターンと文の出現パターン

【列方向】全文中に出現する単語の数を要素とする (単語ベクトル)

【行方向】ある文中に出現する単語の数を要素とする (文ベクトル)

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

KHCoder の原理:

関連の強さを測る – 距離尺度(2)

- 出現パターンが**無関係でない**を測る = カイ²乗距離
- KHCoder の**対応分析**で用いる距離尺度

$$\text{カイ}^2\text{乗距離} = \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

観測度数と期待度数の差が大きく異なるとカイ²乗距離も大きくなり、変数間の関係が期待より強い(=無関係でない)ことを示す

クロス集計表 (観測度数)

	A	B	C	D	E	合計
地質学	3	19	39	14	10	85
生物化学	1	2	13	1	12	29
科学	6	25	49	21	29	130
動物学	3	15	41	35	26	120
物理学	10	22	47	9	26	114
工学	3	11	25	15	34	88
微生物学	1	6	14	5	11	37
植物学	0	12	34	17	23	86
統計学	2	5	11	4	7	29
数学	2	11	37	8	20	78
合計	31	128	310	129	198	796

期待度数

	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410	5.996	9.204	37.000
植物学	3.349	13.829	33.492	13.937	21.392	86.000
統計学	1.129	4.663	11.294	4.700	7.214	29.000
数学	3.038	12.543	30.377	12.641	19.402	78.000
合計	31.000	128.000	310.000	129.000	198.000	796.000

観測度数-期待度数

	A	B	C	D	E	合計
地質学	-0.310	5.332	5.897	0.225	-11.143	0.000
生物化学	-0.129	-2.663	1.706	-3.700	4.786	0.000
科学	0.937	4.095	-1.628	-0.068	-3.337	0.000
動物学	-1.673	-4.296	-5.734	15.553	-3.849	0.000
物理学	5.560	3.668	2.603	-9.475	-2.357	0.000
工学	-0.427	-3.151	-9.271	0.739	12.111	0.000
微生物学	-0.441	0.050	-0.410	-0.996	1.796	0.000
植物学	-3.349	-1.829	0.508	3.063	1.608	0.000
統計学	0.871	0.337	-0.294	-0.700	-0.214	0.000
数学	-1.038	-1.543	6.623	-4.641	0.598	0.000
合計	0.000	0.000	0.000	0.000	0.000	0.000

カイ²乗距離

	A	B	C	D	E	合計
地質学	0.029	2.080	1.050	0.004	5.873	9.036
生物化学	0.015	1.521	0.258	2.913	3.176	7.882
科学	0.173	0.802	0.052	0.000	0.344	1.373
動物学	0.599	0.957	0.703	12.438	0.496	15.194
物理学	6.964	0.734	0.153	4.859	0.196	12.906
工学	0.053	0.702	2.508	0.038	6.700	10.001
微生物学	0.135	0.000	0.012	0.166	0.351	0.663
植物学	3.349	0.242	0.008	0.673	0.121	4.393
統計学	0.671	0.024	0.008	0.104	0.006	0.814
数学	0.354	0.190	1.444	1.704	0.018	3.710
合計	12.343	7.252	6.196	22.899	17.282	65.972

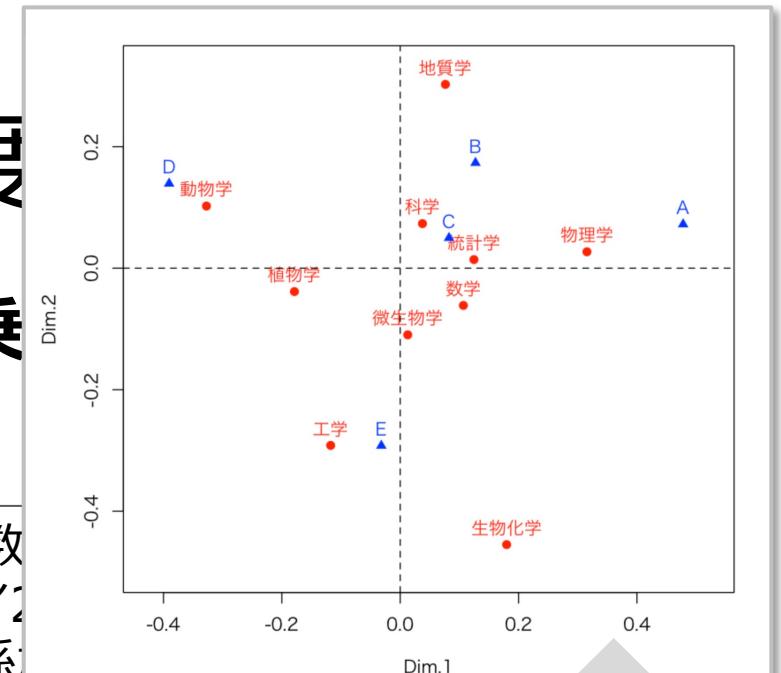
KHCoder の原理:

関連の強さを測る — 距離尺度

- 出現パターンが無関係でないを測る = カイ2乗
- KHCoder の対応分析で用いる距離尺度

$$\text{カイ2乗距離} = \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

観測度数
なるとカイ2乗距離
間の関係性
ない)ことを示す



クロス集計表 (観測度数)

	A	B	C	D	E	合計
地質学	3	19	39	14	10	85
生物化学	1	2	13	1	12	29
科学	6	25	49	21	29	130
動物学	3	15	41	35	26	120
物理学	10	22	47	9	26	114
工学	3	11	25	15	34	88
微生物学	1	6	14	5	11	37
植物学	0	12	34	17	23	86
統計学	2	5	11	4	7	29
数学	2	11	37	8	20	78
合計	31	128	310	129	198	796

期待度数

	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410	5.996	9.204	37.000
植物学	3.349	13.829	33.492	13.937	21.392	86.000
統計学	1.129	4.663	11.294	4.700	7.214	29.000
数学	3.038	12.543	30.377	12.641	19.402	78.000
合計	31.000	128.000	310.000	129.000	198.000	796.000

観測度数-期待度数

	A	B	C	D	E	合計
地質学	-0.310	5.332	5.897	0.225	-11.143	0.000
生物化学	-0.129	-2.663	1.706	-3.700	4.786	0.000
科学	0.937	4.095	-1.628	-0.068	-3.337	0.000
動物学	-1.673	-4.296	-5.734	15.553	-3.849	0.000
物理学	5.560	3.668	2.603	-9.475	-2.357	0.000
工学	-0.427	-3.151	-9.271	0.739	12.111	0.000
微生物学	-0.441	0.050	-0.410	-0.996	1.796	0.000
植物学	-3.349	-1.829	0.508	3.063	1.608	0.000
統計学	0.871	0.337	-0.294	-0.700	-0.214	0.000
数学	-1.038	-1.543	6.623	-4.641	0.598	0.000
合計	0.000	0.000	0.000	0.000	0.000	0.000

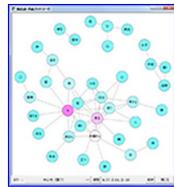
カイ2乗距離

	A	B	C	D	E	合計
地質学	0.029	2.080	1.050	0.004	5.873	9.036
生物化学	0.015	1.521	0.258	2.913	3.176	7.882
科学	0.173	0.802	0.052	0.000	0.344	1.373
動物学	0.599	0.957	0.703	12.438	0.496	15.194
物理学	6.964	0.734	0.153	4.859	0.196	12.906
工学	0.053	0.702	2.508	0.038	6.700	10.001
微生物学	0.135	0.000	0.012	0.166	0.351	0.663
植物学	3.349	0.242	0.008	0.673	0.121	4.393
統計学	0.671	0.024	0.008	0.104	0.006	0.814
数学	0.354	0.190	1.444	1.704	0.018	3.710
合計	12.343	7.252	6.196	22.899	17.282	65.972

KH Coder — 分析手法

共起ネットワーク

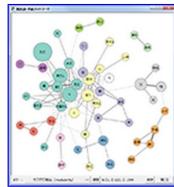
抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



共起の程度が非常に強いものだけを線で結んだ図



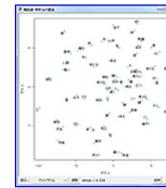
やや弱い共起関係も描画に含め、自動的にグループ分け（色分け）



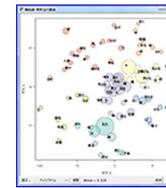
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

多次元尺度構成法 (MDS)

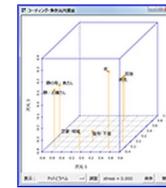
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



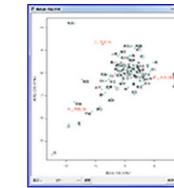
New! クラスタリングと
色分け



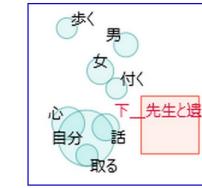
3次元の解

対応分析

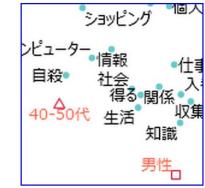
同じく抽出語またはコードを用いての、対応分析です。



同時布置図



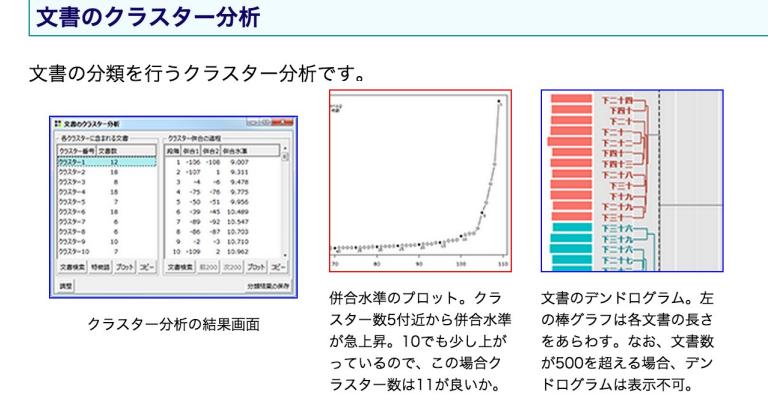
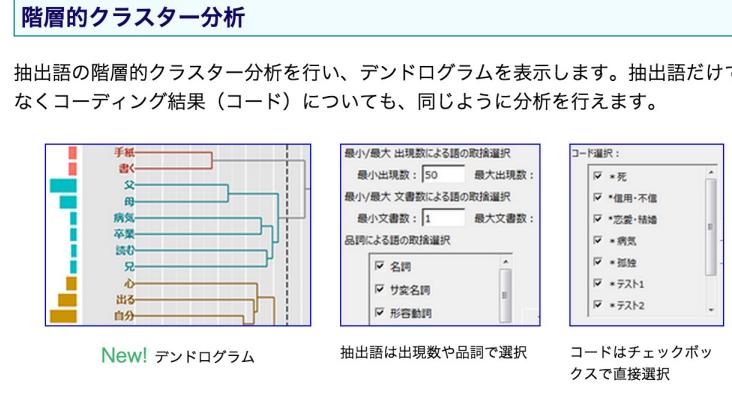
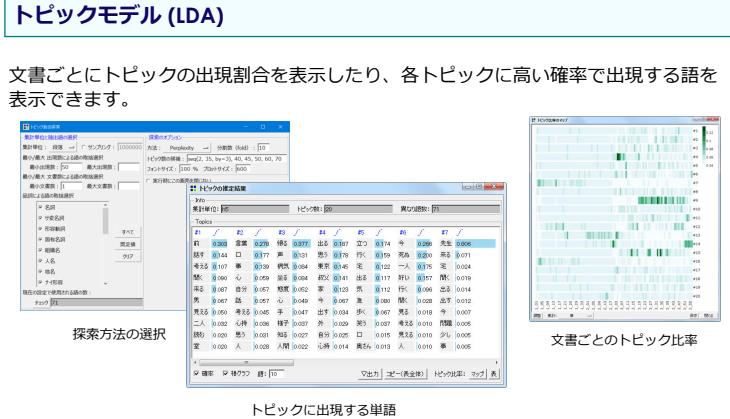
New! バブルプロット



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	<ul style="list-style-type: none">同時に出現した単語同士をネットワークで結んで図示したもの同時に出現したかといった共起の有無を集計し、ネットワークを作成関係の強さ Jaccard 係数で評価、サブグラフは媒介性、クラスタリング精度(エッジ内の密度の高さ)を使って検出
多次元尺度構成法 (MDS)	<ul style="list-style-type: none">出現パターンの似た単語同士を近くに置くよう図示したもの出現パターンは、ある単語がどの文書に出現したかといった単語ベクトルで表現類似度計算には Jaccard, ユークリッド, コサイン距離を用い、クラシカル, Kruskal, Sammon 法のいずれかで2次元にプロット
対応分析 (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、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, 最遠隣接法で樹形図を作成
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, 最遠隣接法で階層クラスターを作成

(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	

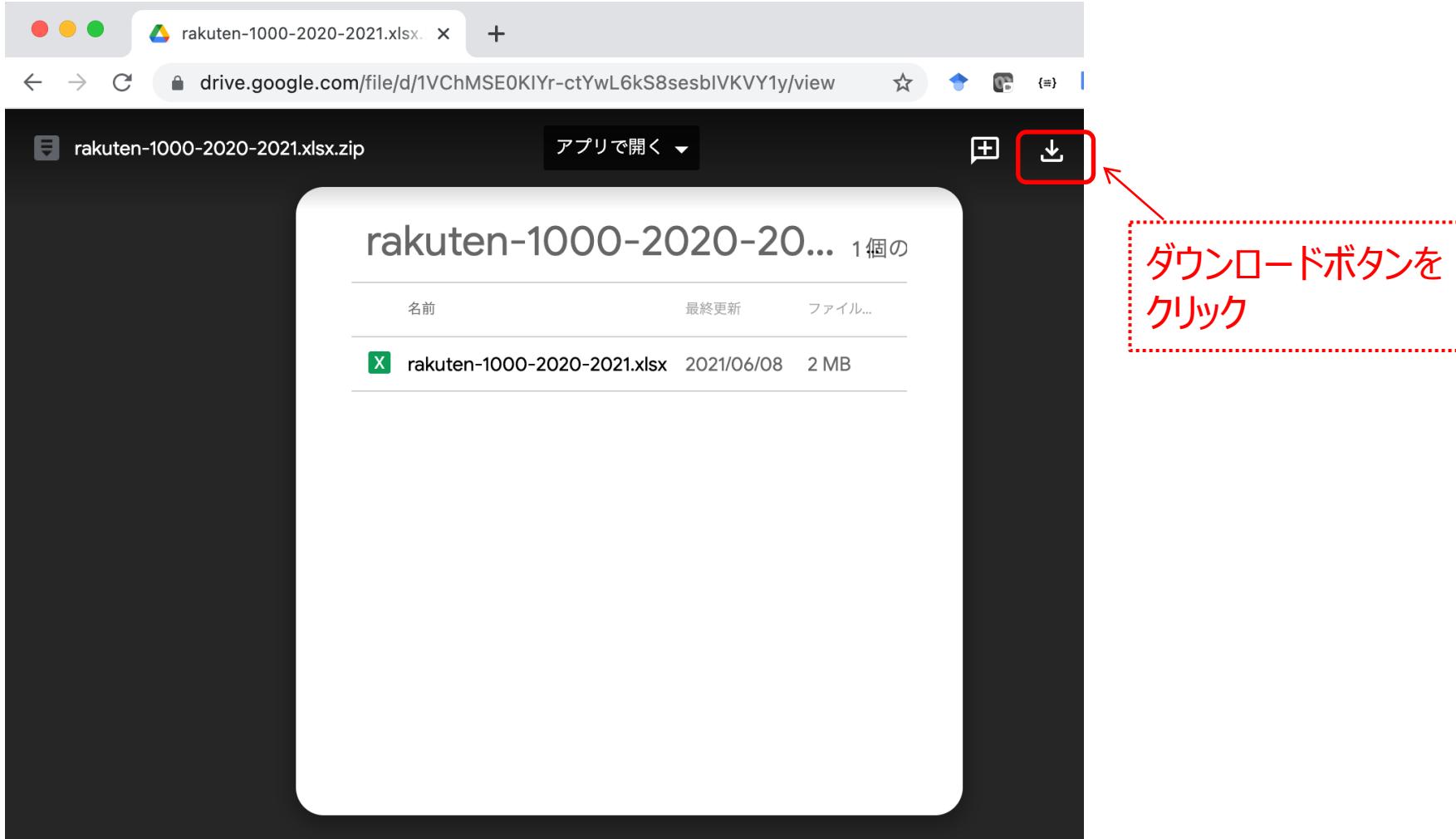
データの取得方法 – 必ず再ダウンロードする

- <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 'lecture / gssm-202207 /'. A red box highlights the '01-data' folder. An arrow points from this folder to the right screenshot, which shows the contents of the '01-data' folder. The right screenshot also has a red box highlighting 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, there is a table with the following data:

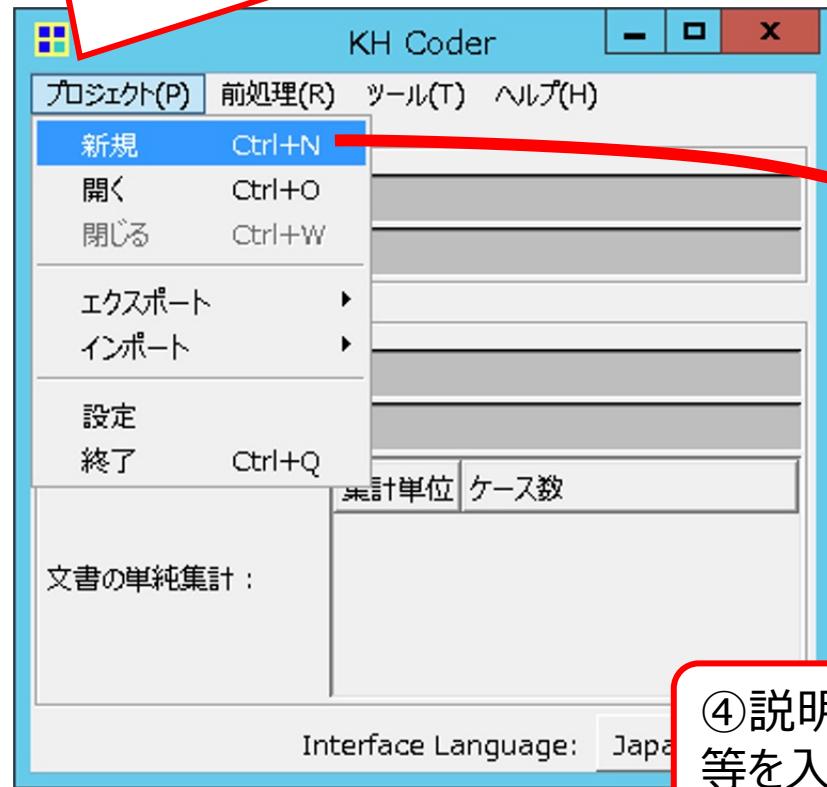
file name	# records	size (zipped)	period
rakuten-1000-2021-2022.xlsx.zip	10,000	2.4 MB	2021/1/1~2022/12/31
rakuten-1000-2019-2020.xlsx.zip	10,000	2.4 MB	2019/1/1~2020/12/31

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

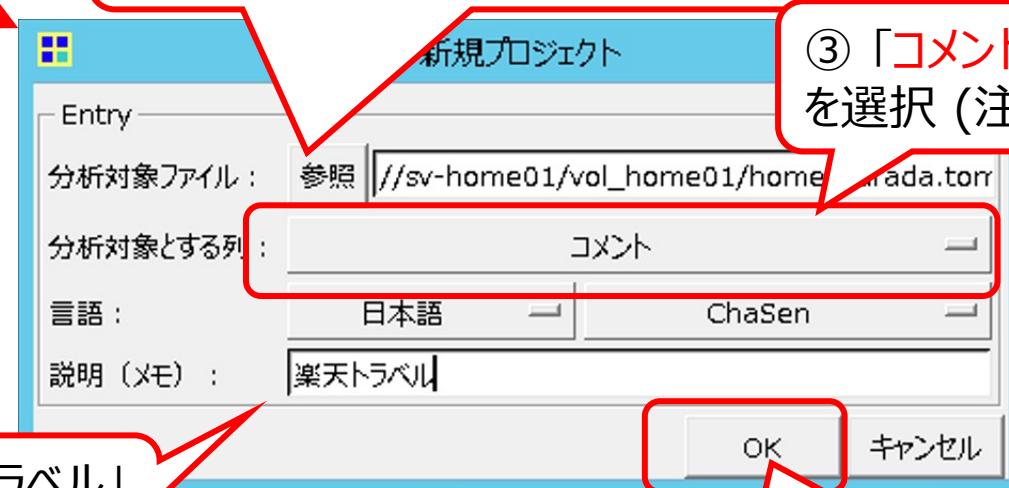
①メニューから「プロジェクト」「新規」を選択 (注1)



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

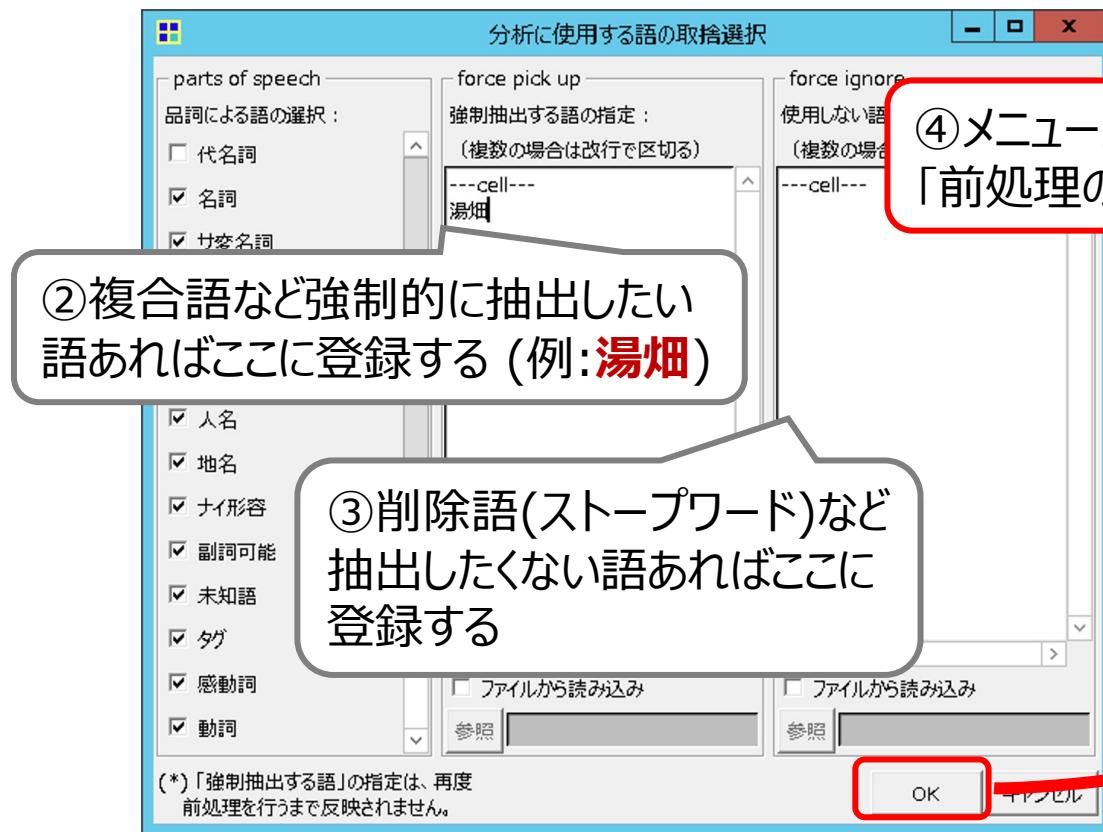


④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

使い方 – 前処理（形態素解析）

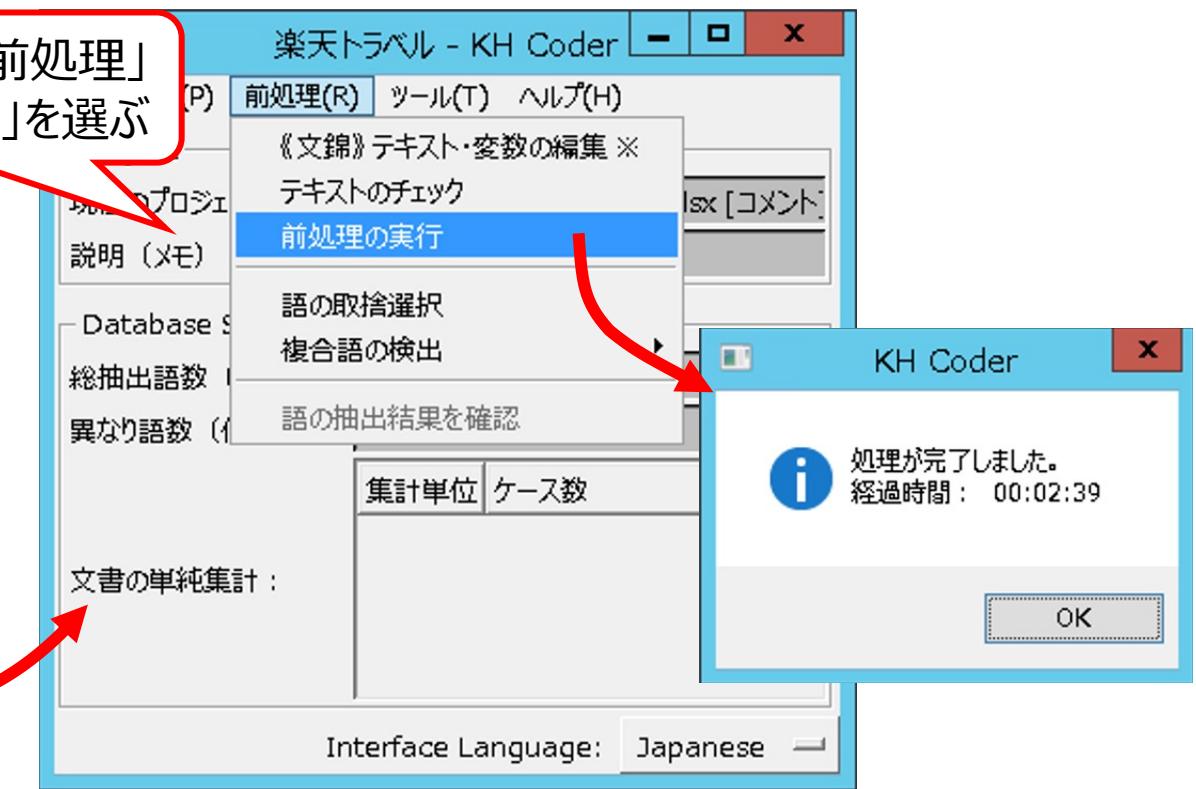
①メニューから「前処理」「語の取捨選択」を選ぶ



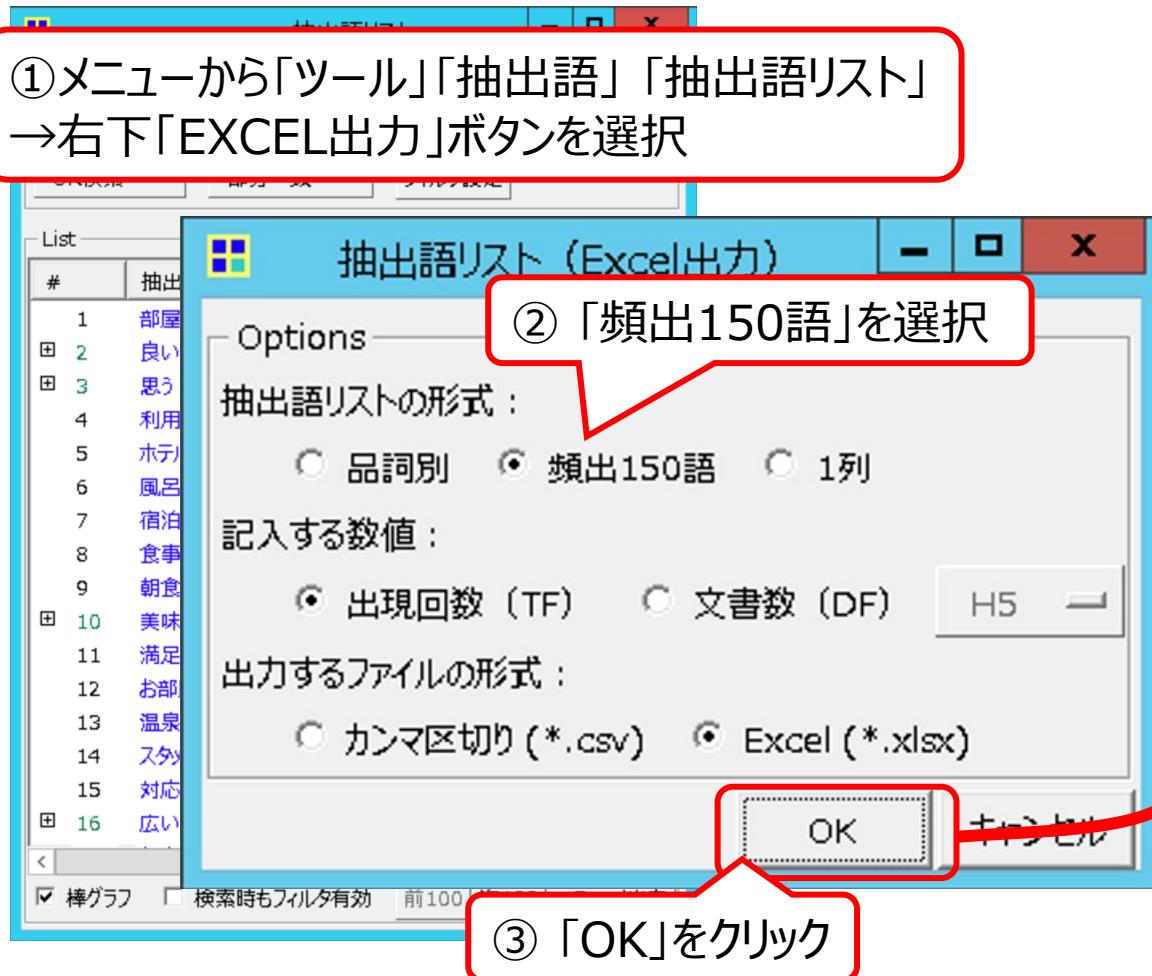
②複合語など強制的に抽出したい
語あればここに登録する（例：湯畠）

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

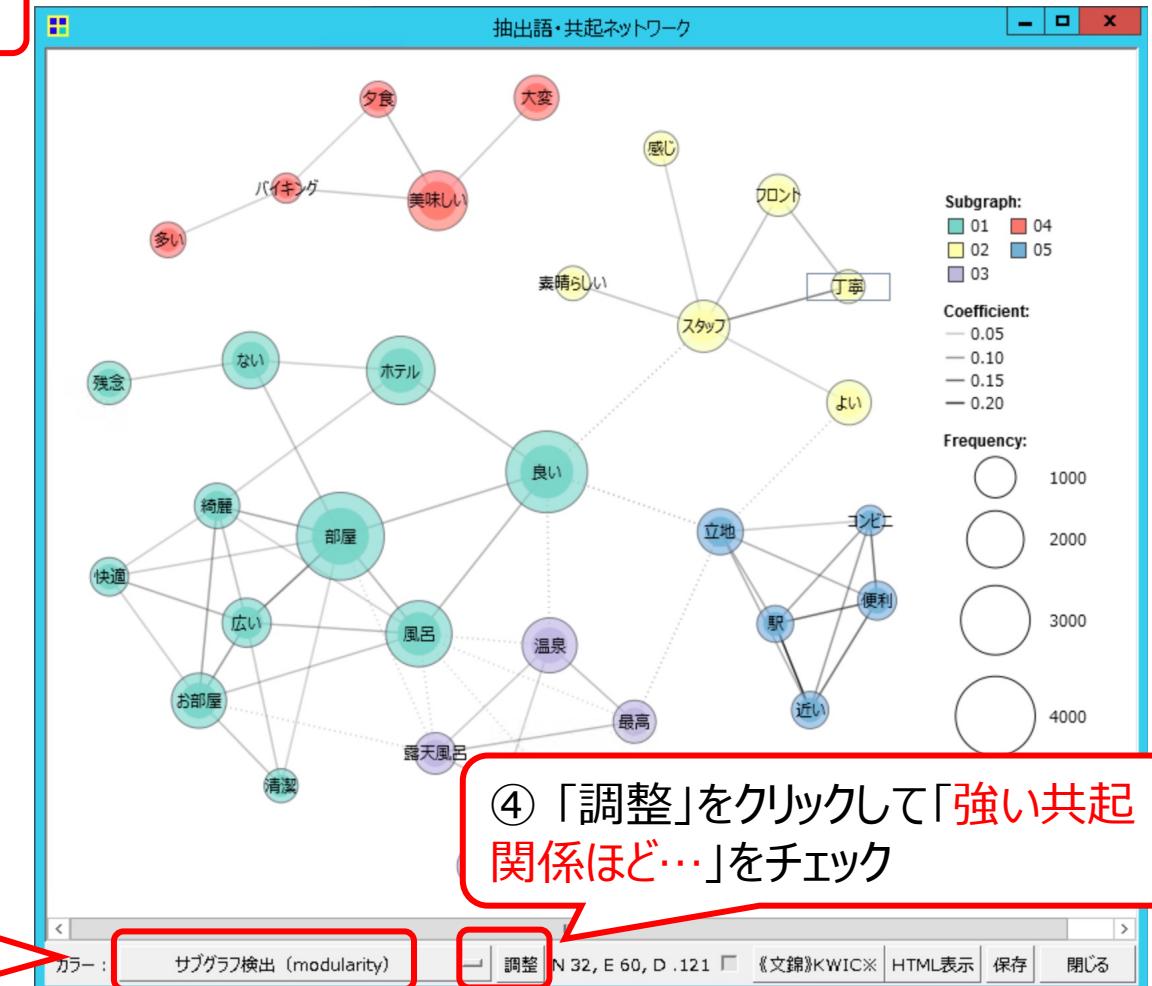
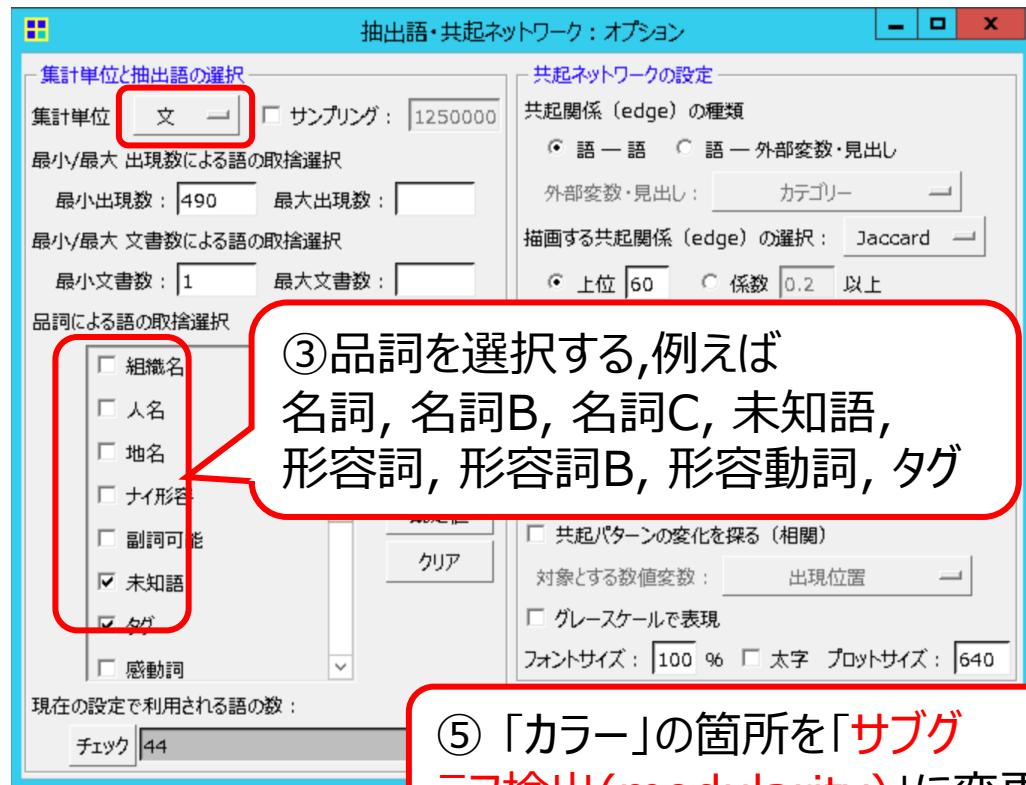


A	B	C	D	E	F	G	H
1 抽出語	出現回数	抽出語	出現回数	抽出語	出現回数		
2 部屋	4876	素晴らしい	696	ご飯	405		
3 良い	4213	過ごす	694	清掃	403		
4 思う	4126	感じ	686	高い	397		
5 利用	3554	清潔	673	無料	396		
6 ホテル	2915	過ごせる	669	悪い	395		
7 風呂	2724	丁寧	656	新しい	387		
8 宿泊	2723	バス	640	設備	383		
9 食事	2385	家族	635	安心	380		
10 朝食	2221	月	619	旅館	372		
11 美味しい	2184	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

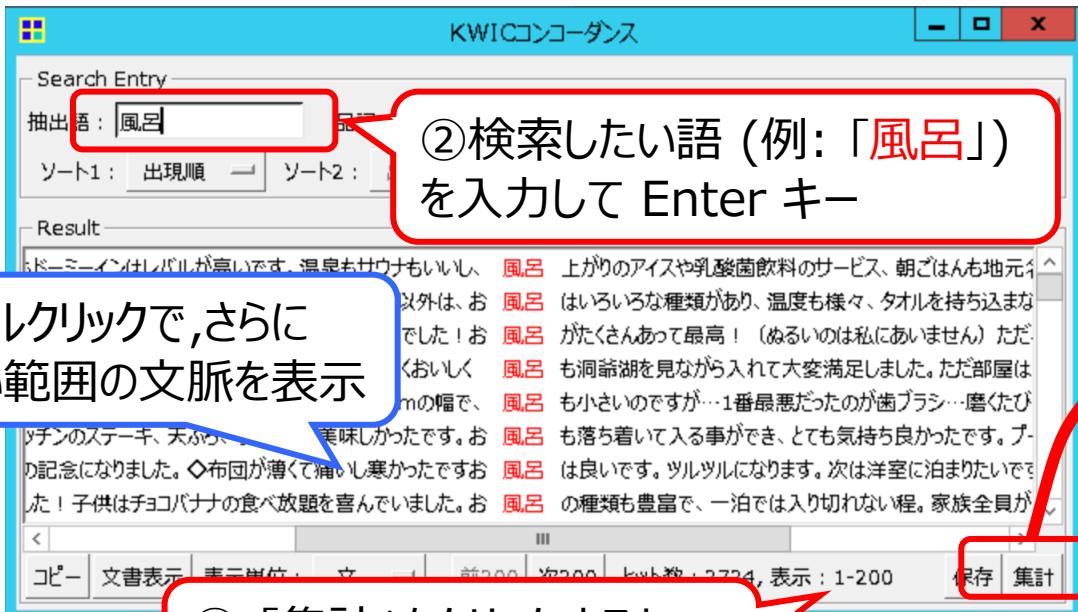
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般（漢字を含む 2 文字以上の語）
名詞 B	名詞一般（平仮名のみの語）
名詞 C	名詞一般（漢字 1 文字の語）
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立（漢字を含む語）
動詞 B	動詞-自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平仮名のみの語）
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 titled 'Result' with data for the word '風呂'. The table includes columns for 'N' (rank), '抽出語' (target word), '品詞' (part of speech), '合計' (total), and various '左' (left) and '右' (right) count columns. 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 highlights the '右合計' button in the toolbar at the bottom.

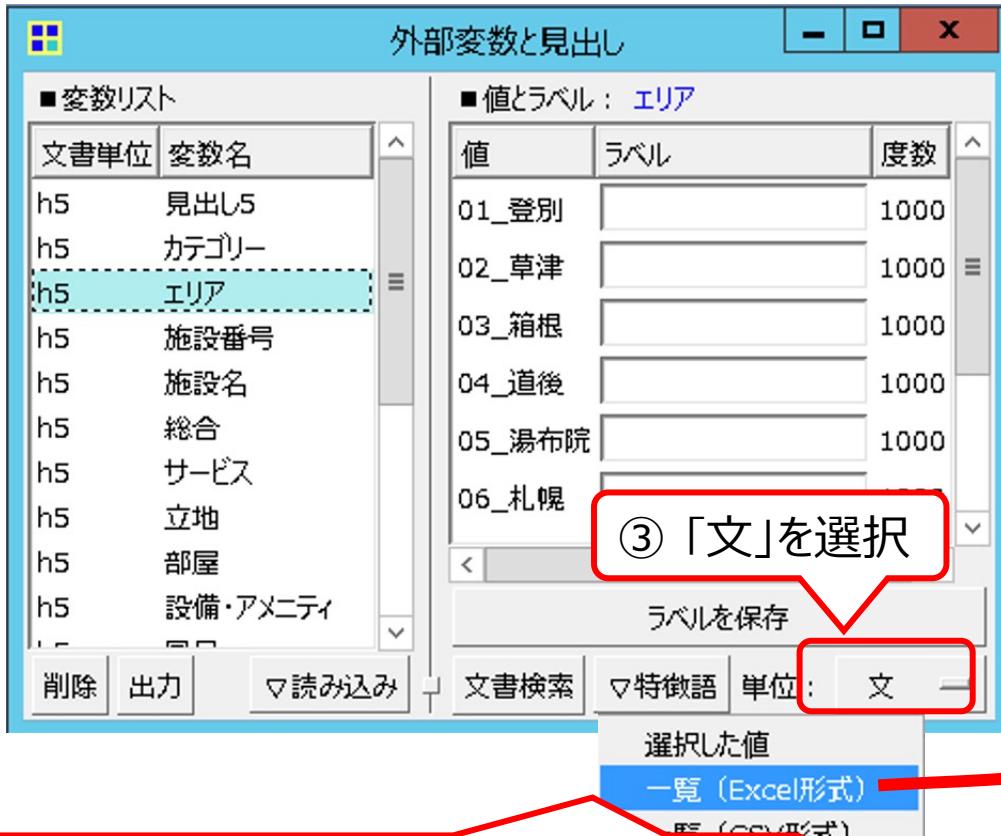
N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	スコア
1	良い	形容詞	223	74	149	42	15	13	4	0	1	51	39	26	32	70.88
2	広い	形容詞	189	48	141	10	8	20	8	2	1	91	24	17	8	77.01
3	綺麗	形容動詞	100	41	59	9	12	17	3	0	34	8	11	6	35.58	
4	よい	形容詞B	76									15	13	10	9	23.50
5	ない	形容詞B	56									9	5	11	12	18.20
6	清潔	形容動詞	46						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

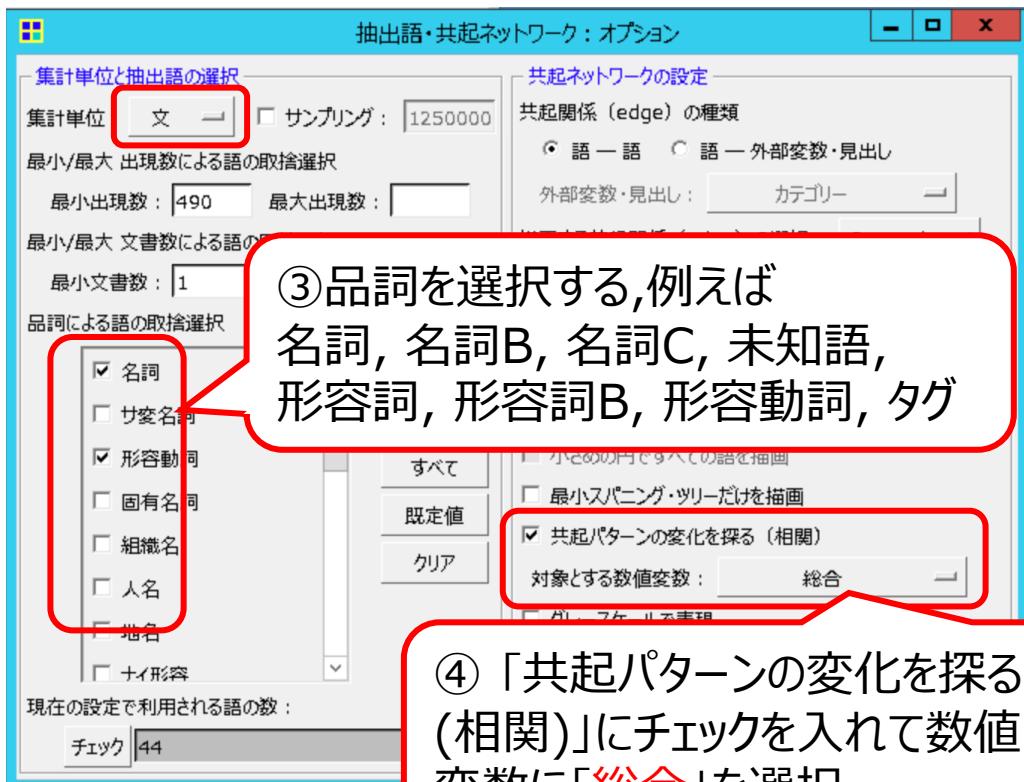
A	B	C	D	E	F	G	H	I	J	K
1										
2	01_登別		02_草津		03_箱根		04_道後			
3	食事	.059	温泉	.068	思う	.066	温泉	.054		
4	良い	.058	湯畑	.064	食事	.064	良い	.051		
5	風呂	.057	風呂	.062	良い	.060	朝食	.045		
6	思う	.054	良い	.061	風呂	.053	ホテル	.042		
7	温泉	.049	食事	.056	美味しい	.049	美味しい	.042		
8	美味しい	.044	草津	.055	露天風呂	.048	道後	.041		
9	宿泊	.043	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

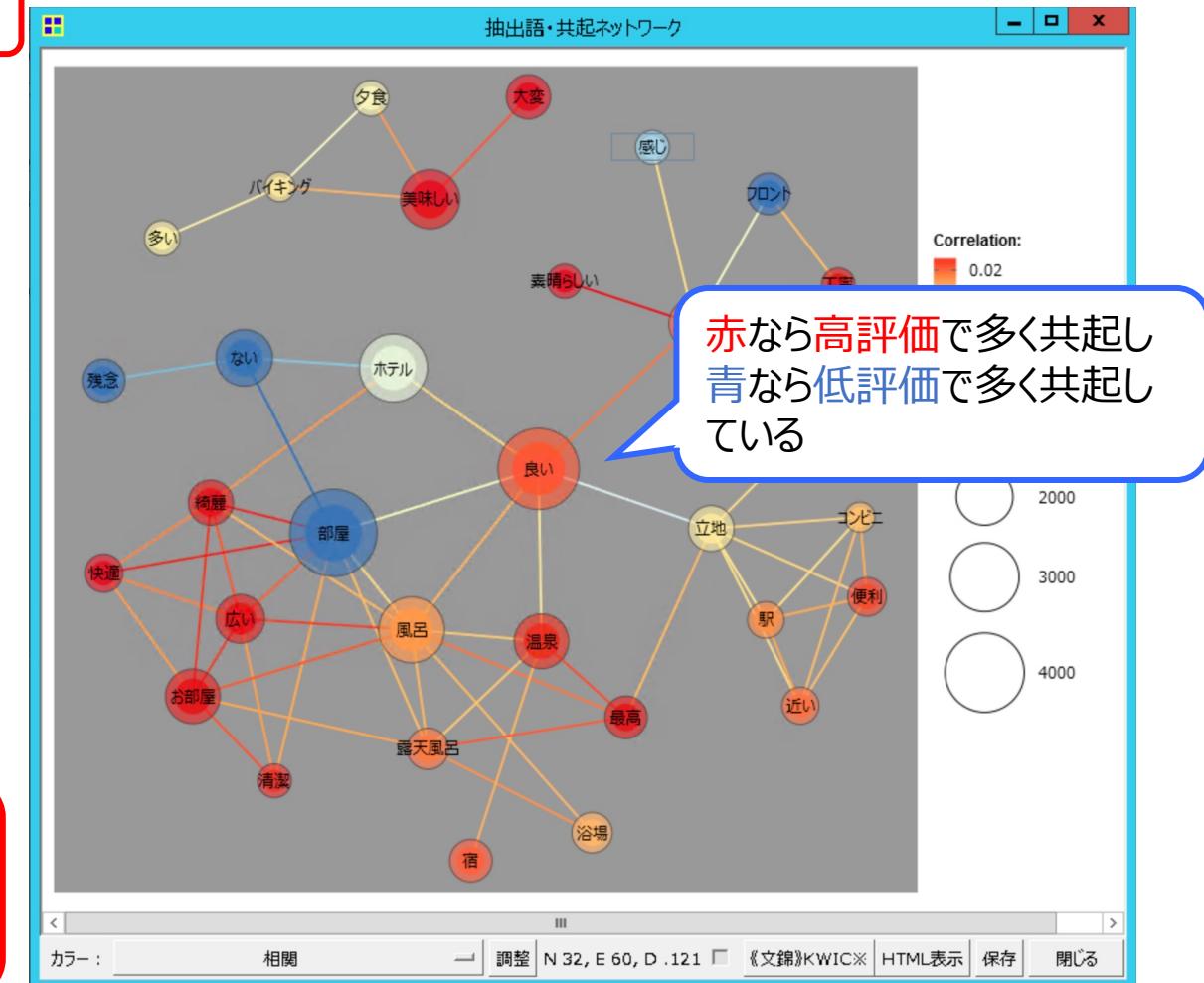
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「文」を選んで「OK」をクリック



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

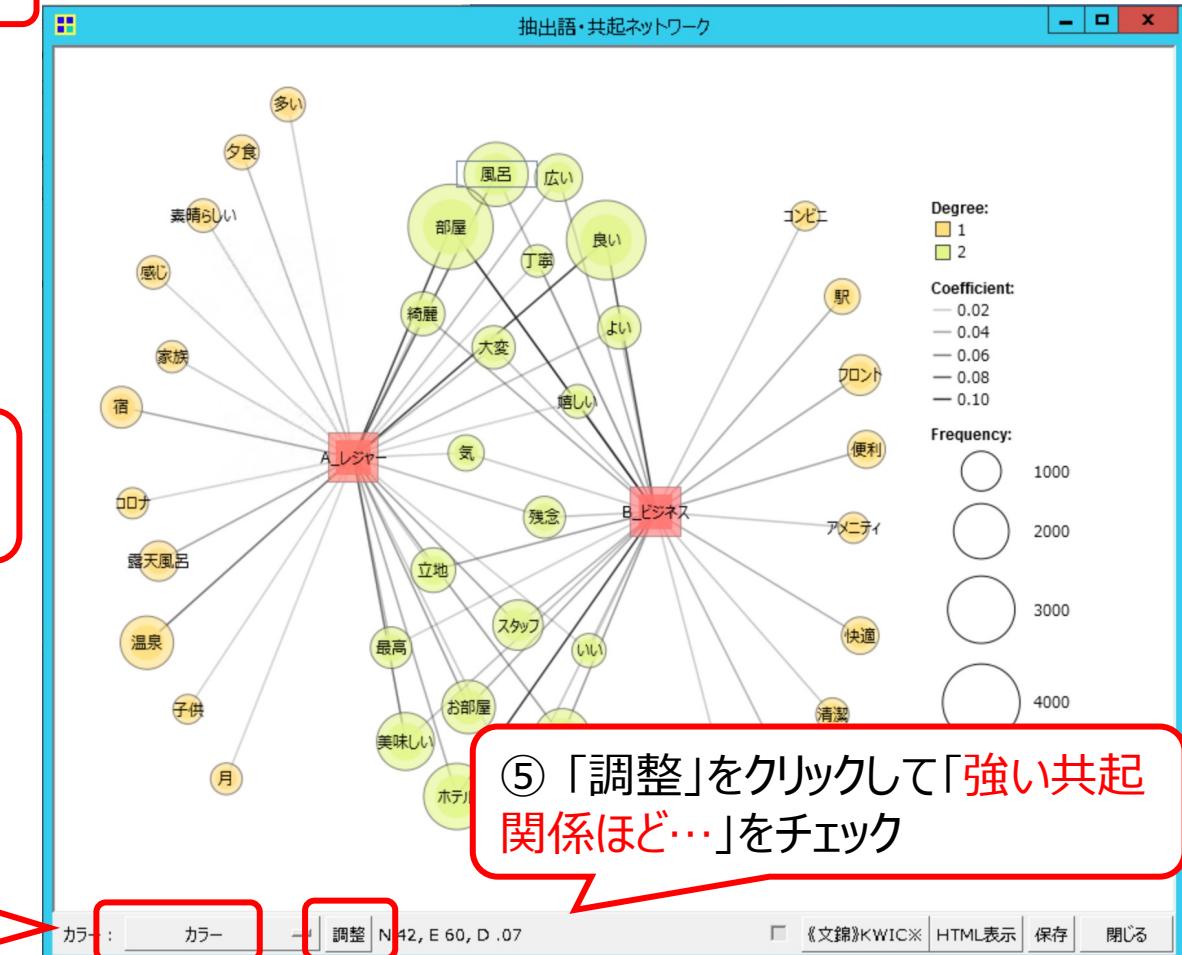
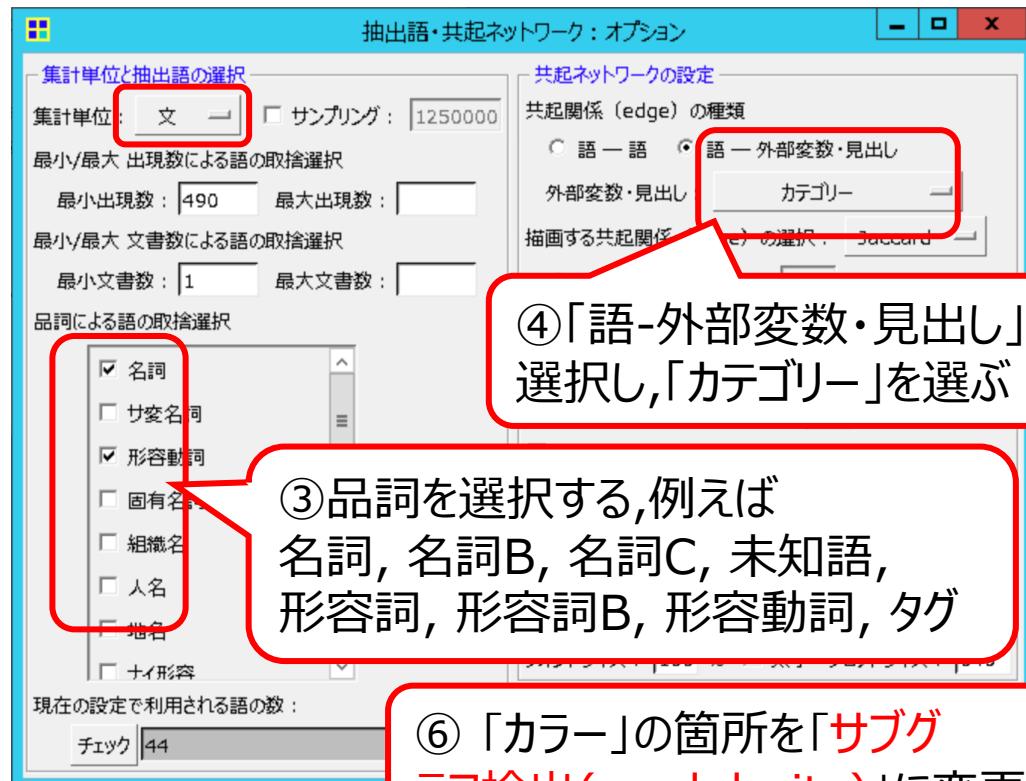
④「共起パターンの変化を探る
(相関)」にチェックを入れて数値
変数に「総合」を選択



使い方 – 共起ネットワークの作成3

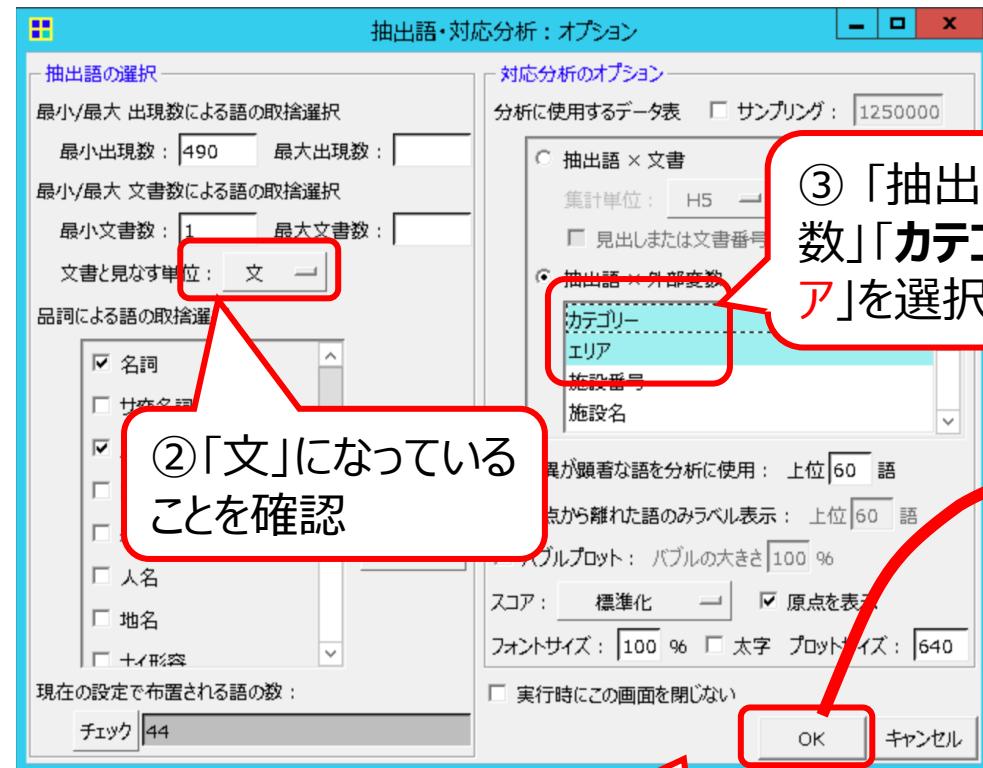
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「文」を選んで「OK」をクリック

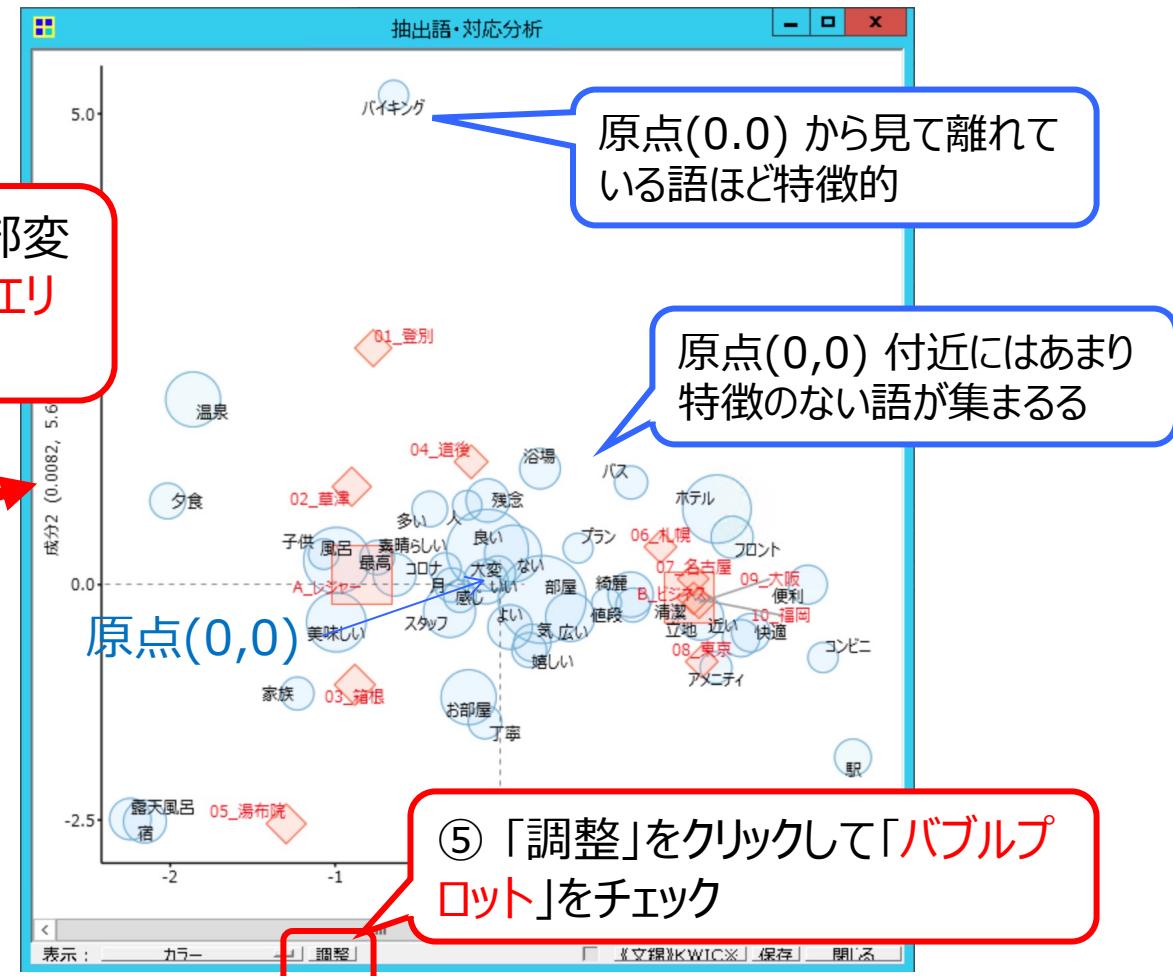


使い方 – 対応分析による探索1

①メニューから「ツール」「抽出語」「対応分析」を選ぶ



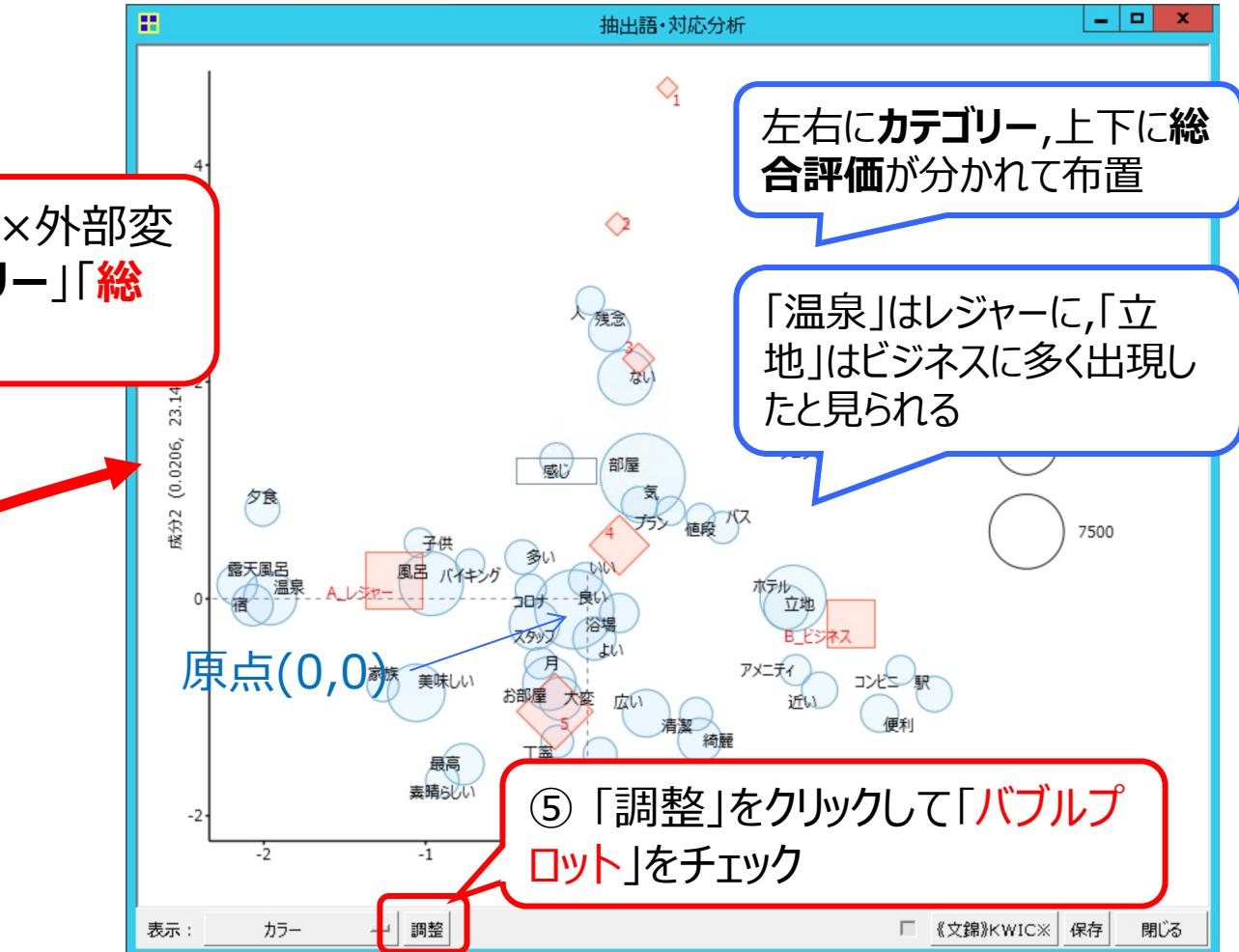
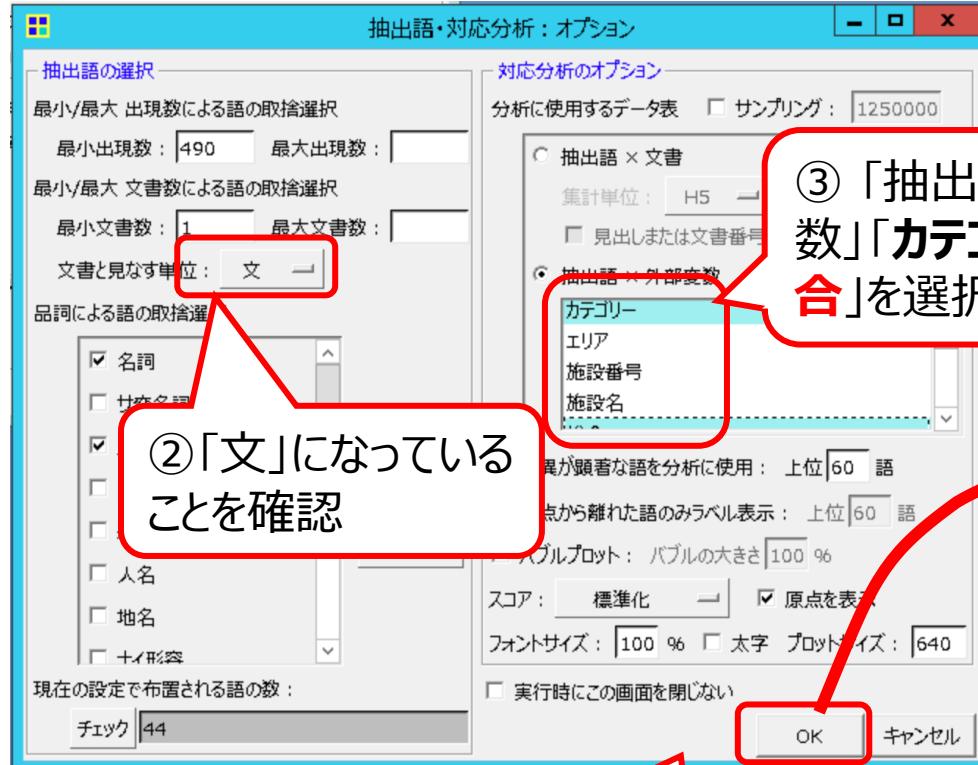
④ 「OK」をクリック



⑤ 「調整」をクリックして「バブルプロット」をチェック

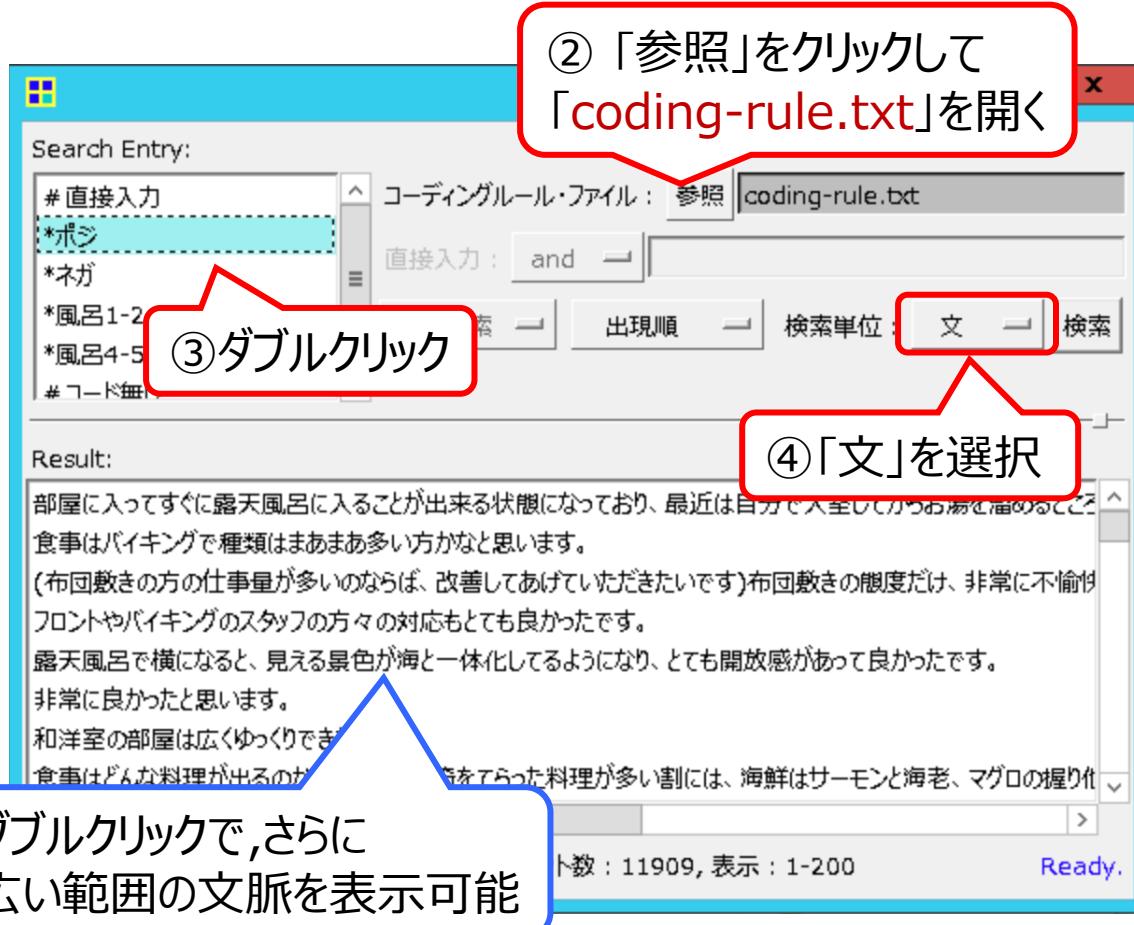
使い方 — 対応分析による探索2

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



※ 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

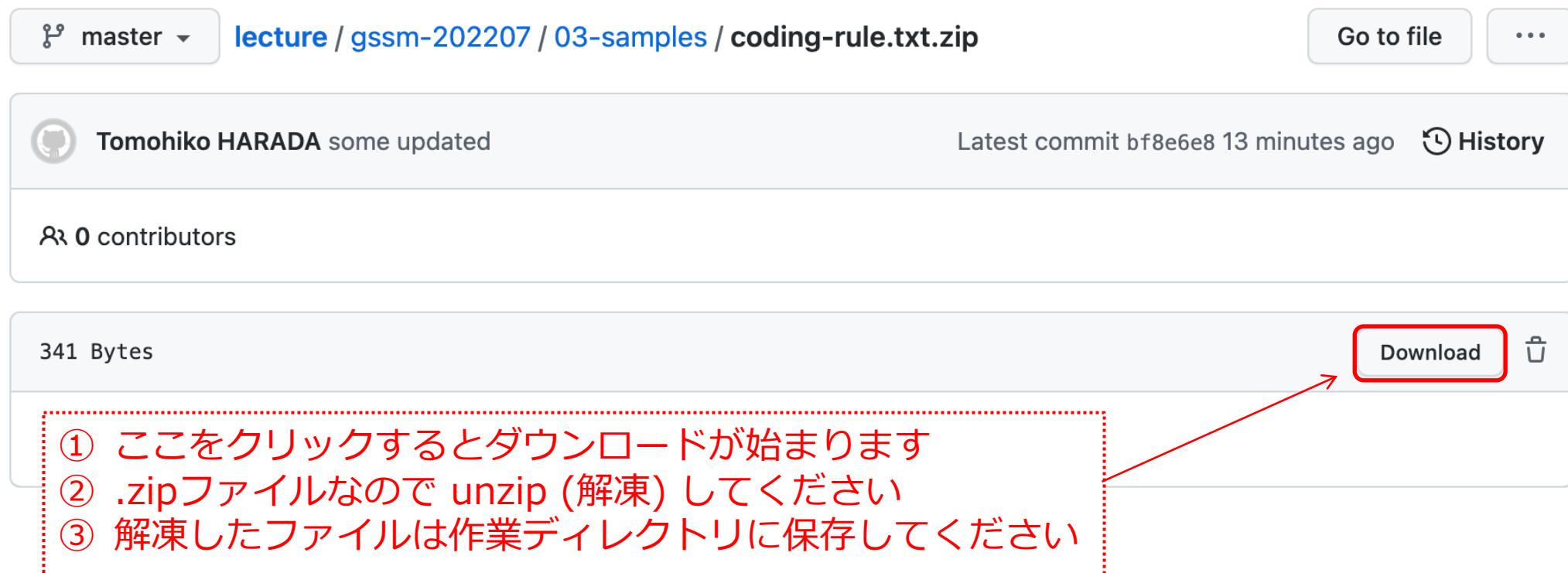
<>風呂-->4 | <>風呂-->5

外部変数

(参考) コーディングルールのサンプル

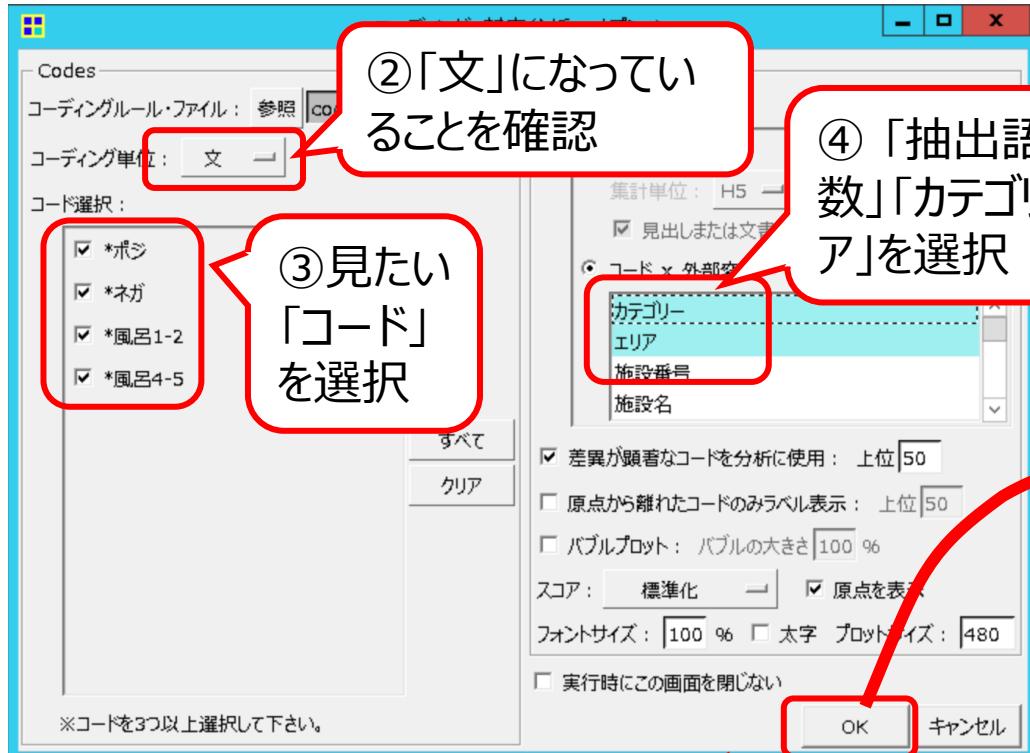
<https://github.com/haradatm/lecture/blob/master/gssm-202207/03-samples/coding-rule.txt.zip>

- 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

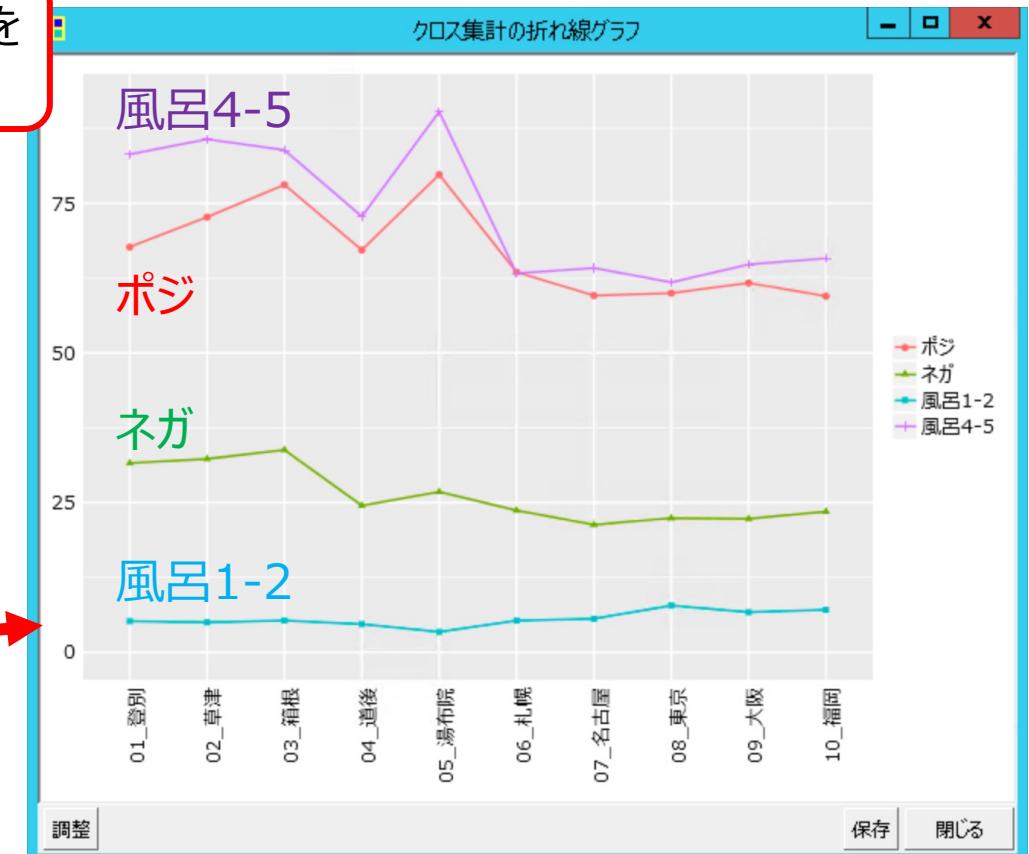
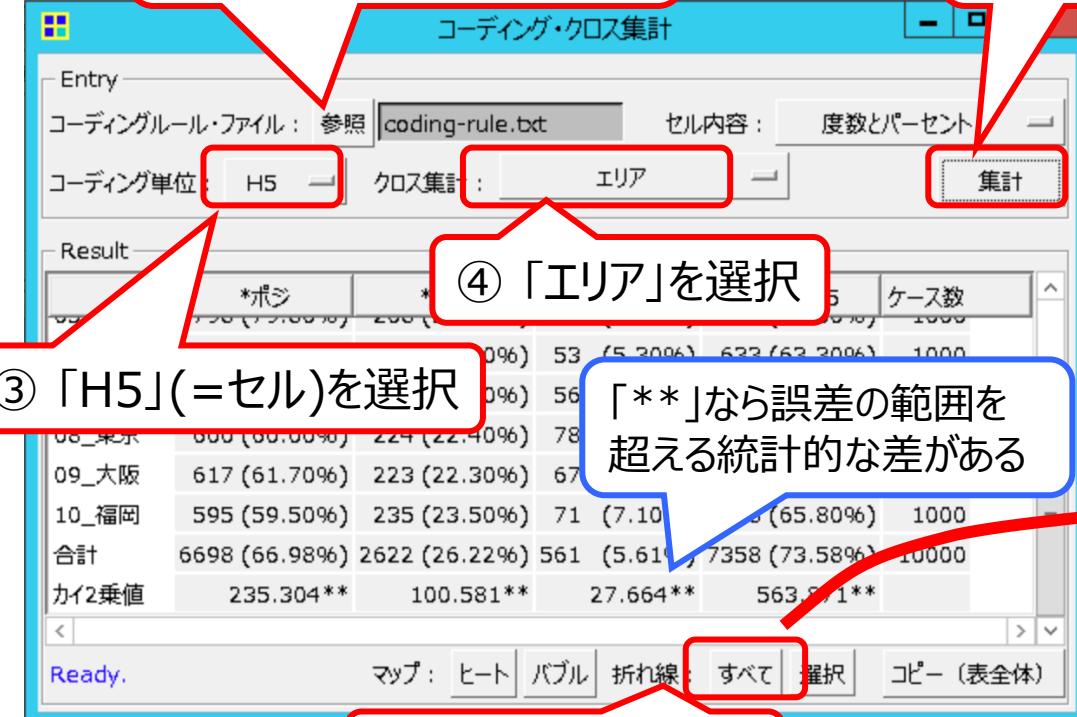


使い方 – クロス集計

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

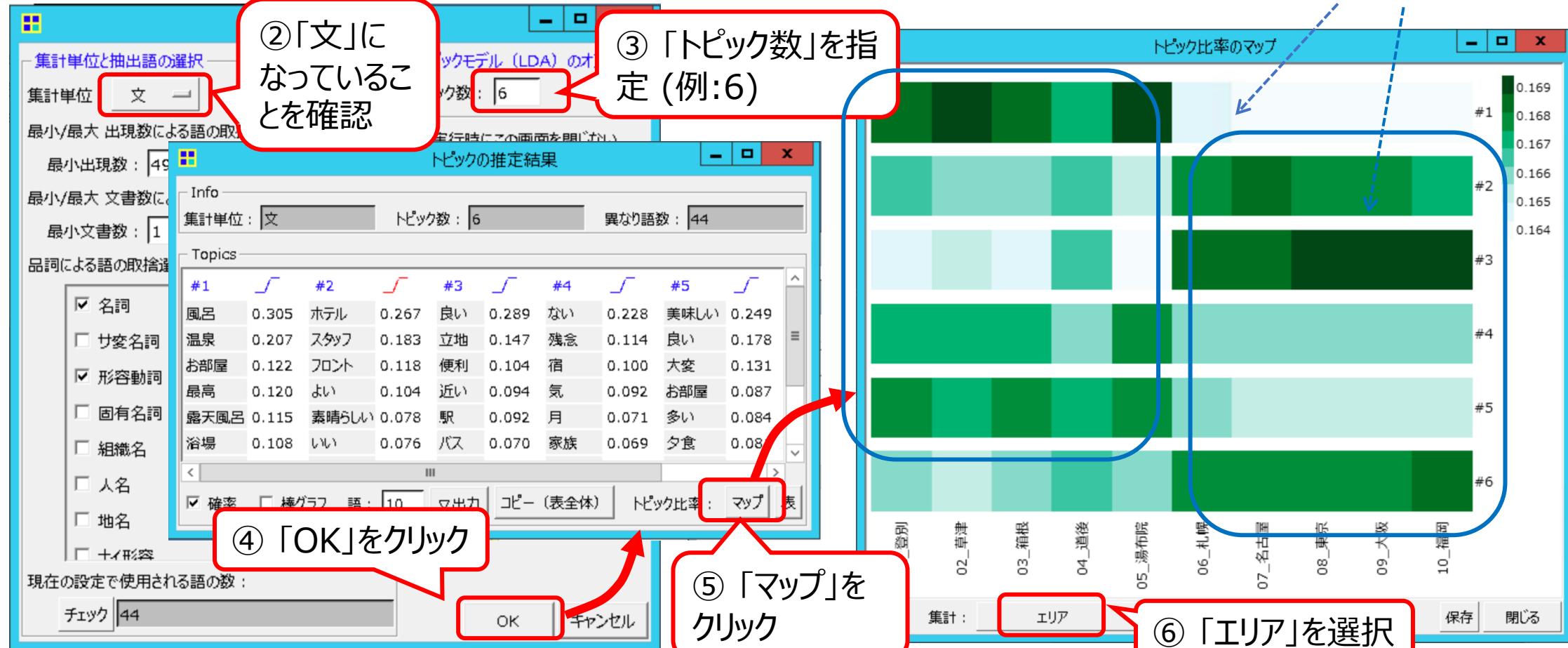
②「参照」をクリックして
「coding-rule.txt」を開く

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
 - コード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
 - P.48 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.