

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 4 (後半)

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 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.45)
- ※ 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

緊急追加

～ ChatGPT登場後のテキストマイニングのカタチ(3)～

LLaMA 2 – 商用利用可能な LLaMA の新モデル

- 7/18: Meta が OSS の大規模言語モデル「Llama 2」を発表

- 2Tトークンで学習、コンテキスト長を4Kに拡張した 7B、13B、70B のモデル
- 研究利用および商用利用向けに無償で提供開始※

ChatGPT 登場以降 – 2つのトレンド		
トレンド	説明	代表例
① OSSモデル	<ul style="list-style-type: none">OpenAI の API を利用する場合、外部(OpenAI)にデータ送信することになるため、手元に LLM を構築するニーズがあるChatGPT(175B)レベルの大規模モデルを載せるにはコンピューティングコストがかかりすぎ、パラメタを減らすと精度が下がる課題がある → AWS等のパブクラで動作可能な軽量で高精度のOSSモデルや学習手法が登場	<ul style="list-style-type: none">LLaMA(22/1, Meta)Alpaca(23/3, Stanford)Vicuna(23/3, UCBerkeley)など
② 自律駆動型AI	<ul style="list-style-type: none">言語モデルは様々なタスクに応用されているか学習時点の(古い)知識しか利用できない自身の内部表現を用いて推論の道筋を生成するため、反応的に探索・推論したり、知識を更新する能力が制限され、また一時的な記憶を持つこともできない → 行動と行動結果に対する推論を繰り返してタスクを達成させる仕組みが登場	<ul style="list-style-type: none">ReAct(22/11, Google)AutoGPT(23/3)BabyAGI (23/4/3, OpenAI)など

R05年度 01KA438, 0ADM126

テキストマイニング

代表的な OSSモデル – 2023/5月末時点								
分類	モデル名	アーキ	提供元	リリース	サイズ	ライセンス	日本語	例:「ベンギンはなぜ空を飛べないですか?」
クラウド	GPT-4	クローズド	OpenAI	2023/3/14	175B	有償API	○	ベンギンが飛べない理由はいくつかあります。主な理由は、彼らの身体…
	Claude	クローズド	Anthropic	2023/3/14	50B	有償API	○	ベンギンは空を飛べない主な理由は次のとおりです。1. 翼がないため…
研究利用のみ	LLaMa	LLaMa	Meta	2023/2/24	7-65B	× 制限あり	△	ベンギンは空の中の風に乗り飛ぶことができます。
	Alpaca (LLaMa)	LLaMa	Stanford	2023/3/13	2B	× 制限あり	△	ベンギンは、空の下で彼は翼を叩くことができます。しかし、彼は午前…
	Alpaca-LoRA (LLaMa)	LLaMa	Er					
	Guanaco (LLaMa)	LLaMa	Jose					
	GPT4All (LLaMa)	LLaMa	N					
	Vicuna (LLaMa)	LLaMa						
	Koala (LLaMa)	LLaMa						
	Stable-Vicuna (LLaMa)	LLaMa	St					
	Cerebras-GPT	GPT-2	C					
商用利用可能	Dolly2.0 (GPT-J-Alpaca)	GPT-J	Databricks	2023/4/12	1B	Apache 2.0	△	よくわかりませんが、ベンギンは空を飛べないのではなく、空を飛ぶことを…
	StableLM-Alpha	GPTNeoX	Stability AI	2023/4/19	2B	CC BY-SA-4.0	△	空を飛べるためには、空気を回し、目的地を知りません。ベンギンは、…
	GPT4All (GPT-J)	GPT-J	Nomic AI	2023/4/24	6B	Apache 2.0	×	It seems like you are having trouble with your writing. …
	RedPajama-INCITE	GPTNeoX	Together	2023/5/5	3-7B	Apache 2.0	△	ベンギンは空を飛べないのですか。
	MPT-7B	MPT	MosaicML	2023/5/5	2B	Apache 2.0	△	もちろん、飛んでくれます！私は飛んでくれます！でも、飛んでくれない…
	open-calm	GPTNeoX	CyberAgent	2023/5/16	1-7B	CC BY-SA-4.0	△	ベンギンが空を飛べない理由は、水中で息ができないから、というのを…
	japanese-gpt-neox-3.6b	GPTNeoX	rinna	2023/5/17	3.6B	MIT	△+	ベンギンは、翼のような推進力を生み出す筋肉がなく、空気力学的に…

R05年度 01KA438, 0ADM126

テキストマイニング

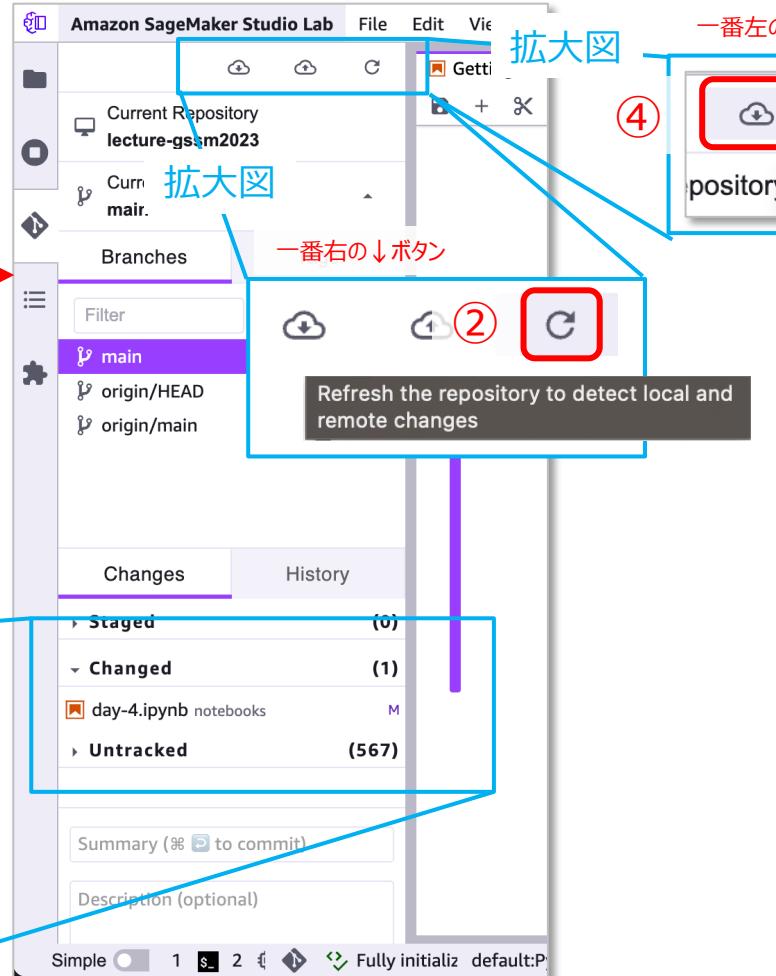
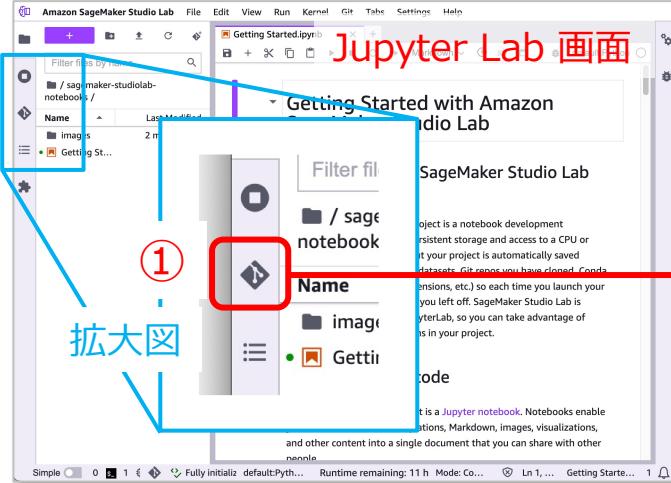
LLaMA が商用利用不可のため、商用利用できるOSSモデルの開発競争が激しくなっていた

※ 7億人超の月間アクティブユーザーがいる場合はMetaから許可を得る必要あり)

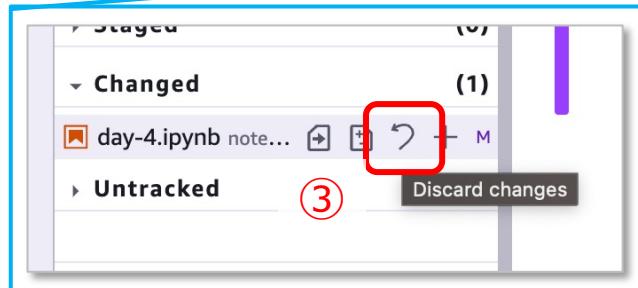
KHCoder の解析・分析手法

(再掲) Jupyter Lab の教材を最新化するには

- 下記の手順で、教材を最新化(pull)することができます



(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」 ボタンを押す
- ⑤ 画面の右下に「Successfully published」が表示されること確認する

- 「単語と単語」、「カテゴリ(外部変数)と単語」の関係に注目した分析が得意

- 特徴的な単語を見つける

- 特定の文書に特徴的な単語を見つける → 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

- 特徴的な関係を見つける

- 関係性のある単語と単語と見つける → **共起ネットワーク(Jaccard係数)**
例) 「風呂」と「広い」に関係がありそう
 - 関係性の強い単語と外部変数を見つける → **対応分析(カイ2乗値)**
例) 「レジャー」と「風呂」に関係がありそう

KHCoder で使われるデータ表

- 「文書-抽出語」表 [【行】ある文中に出現する単語の数を要素とする文ベクトル
【列】全文中に出現する単語の数を要素とする単語ベクトル]

「文書-抽出語」頻度表

「抽出語-抽出語」共起頻度表（共起ネットワーク）

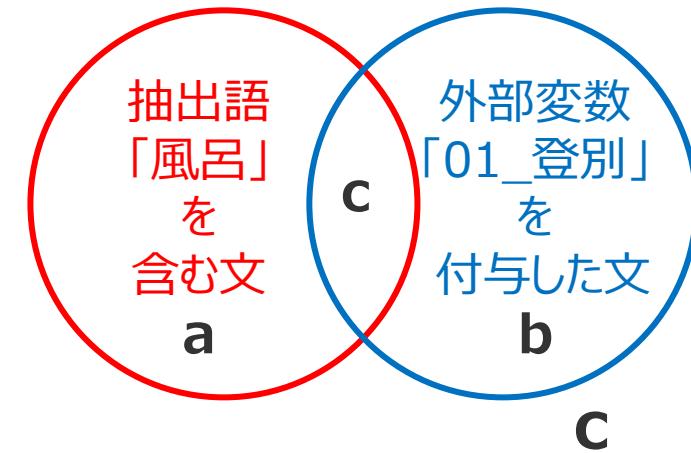
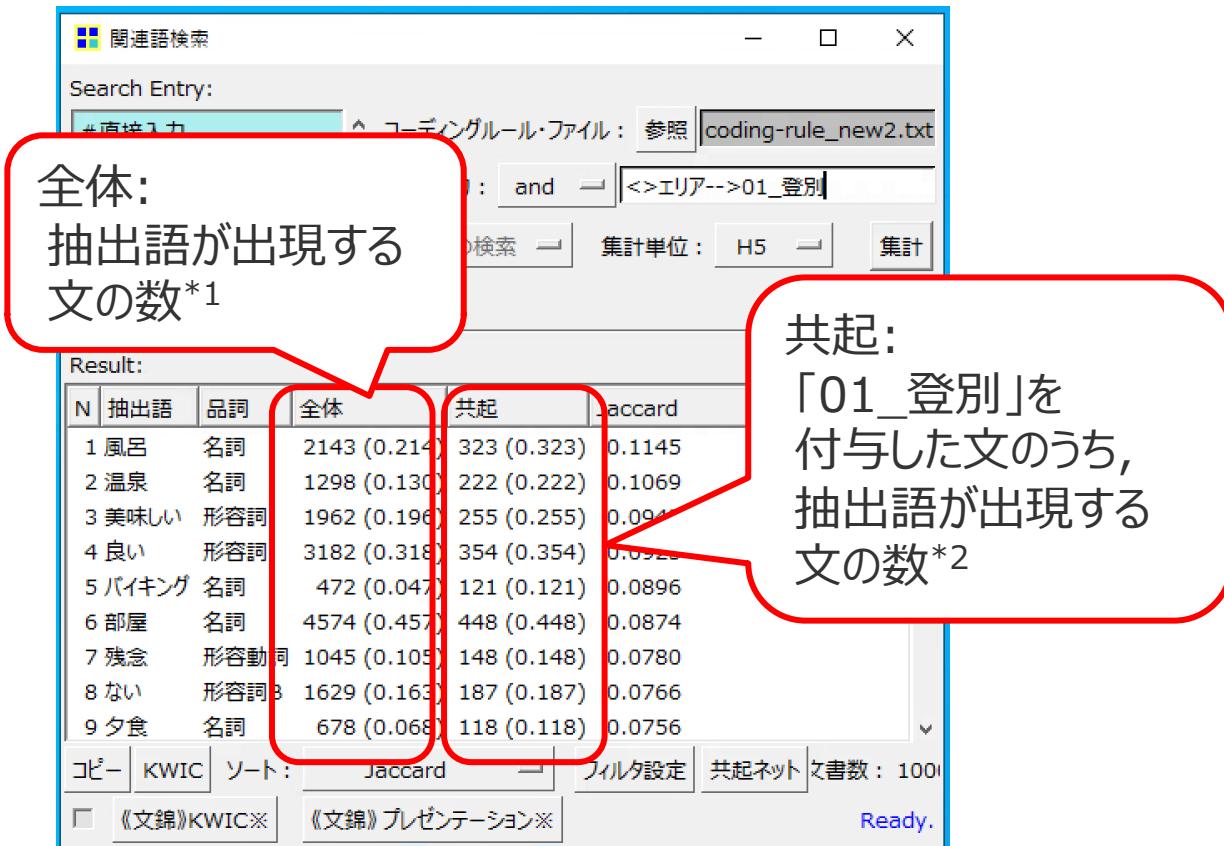
	部屋	良い	ホテル	風呂	美味しい	ない	スタッフ	温泉	よい	立地	広い	綺麗だ	宿	大変だ	残念だ	最高
部屋	4568	1961	1091	1273	1068	1012	783	706	676	691	955	817	461	512	635	526
良い	1961	3701	892	1026	896	692	712	638	339	701	558	494	406	439	468	356
ホテル	1091	892	2041	391	354	506	411	235	267	356	291	339	72	207	283	198
風呂	1273	1026	391	2143	598	503	376	359	333	317	414	280	334	277	325	294
美味しい	1068	896	354	598	1962	379	432	429	229	215	299	255	300	308	249	270
ない	1012	692	506	503	379	1629	337	296	269	289	260	185	202	193	341	159
スタッフ	783	712	411	376	432	337	1411	275	216	201	182	184	206	214	203	175
温泉	706	638	235	359	429	296	275	1298	210	176	221	120	293	197	182	246
よい	676	339	267	333	229	269	216	210	1205	238	167	135	124	124	158	113
立地	691	701	356	317	215	289	201	176	238	1328	167	186	120	139	157	243
広い	955	558	291	414	299	260	182	221	167	167	1186	209	122	153	152	128
綺麗だ	817	494	339	280	255	185	184	120	135	186	209	1132	76	120	119	117

「外部変数-抽出語」クロス集計表（対応分析）

	部屋	良い	ホテル	風呂	美味しい	ない	スタッフ	温泉	よい	立地	広い	綺麗だ	宿	大変だ	残念だ	最高
A_レジャー	2398	2046	757	1535	1430	880	888	1188	631	518	631	459	769	652	609	697
B_ビジネス	2170	1655	1284	608	532	749	523	110	574	810	555	673	57	355	436	294
01_登別	447	409	194	323	255	187	148	222	114	38	125	88	76	120	148	124
02_草津	488	434	181	352	274	180	154	275	117	155	114	109	195	136	122	165
03_箱根	548	436	134	326	355	202	212	212	133	57	140	93	161	137	150	82
04_道後	416	349	191	181	174	130	135	225	137	176	129	87	59	108	102	123
05_湯布院	499	418	57	353	372	181	239	254	130	92	123	82	278	151	87	203
06_札幌	452	346	255	121	129	151	114	38	103	166	129	142	10	87	92	74
07_名古屋	434	310	241	116	97	144	102	18	133	141	84	138	11	58	89	42
08_東京	441	338	240	106	99	131	99	12	104	166	98	128	14	69	82	56
09_大阪	431	317	297	135	88	162	93	20	121	161	132	144	9	62	82	62
10_福岡	412	344	251	130	119	161	115	22	113	176	112	121	13	79	91	60

Jaccard 系数

- Jaccard 系数は、共起の強さを測る尺度 (KHCoderで標準的に使用)
 - どちらの語も含まない文書を無視 → 言語のようなスパースデータ分析に向いている



$$\text{Jaccard 系数} = \frac{c}{a+b+c}$$

抽出語「部屋」の場合:
 $c = 323$ ("共起"列の値)
 $a = 2143$ ("全体"列の値) - 323 = 1820
 $b = (323 / 0.323) - 323 = 677$

*1 括弧内はデータ全体に対する割合(前提確率) *2 括弧内は「01_登別」を付与したデータに対する割合(条件付き確率)

カイ2乗値

- カイ2乗値は「無関係でない」度合いを測る尺度 → カテゴリと変数間の関連性を測定

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

「観測度数」: カテゴリと変数に従ってクロス集計された度数

「期待度数」: 変数が互いに独立している場合に期待される度数

「観測度数 - 期待度数」: 実際の度数と独立と期待される度数の差

- カイ2乗値も大きい → カテゴリと変数間の関係が**期待より強い**を示す

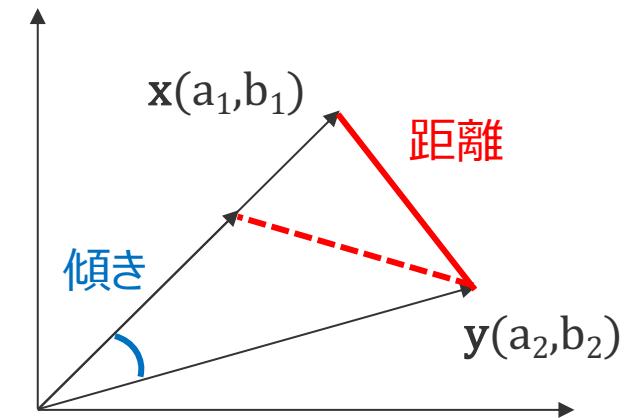
クロス集計表 (観測度数)						期待度数					観測度数-期待度数					カイ2乗値				
	A	B	C	D	E	合計		A	B	C	D	E	合計		A	B	C	D	E	合計
地質学	3	19	39	14	10	85	地質学	3.310	13.668	33.103	13.775	21.143	85.000	地質学	-0.310	5.332	5.897	0.225	-11.143	0.000
生物化学	1	2	13	1	12	29	生物化学	1.129	4.663	11.294	4.700	7.214	29.000	生物化学	-0.129	-2.663	1.706	-3.700	4.786	0.000
科学	6	25	49	21	29	130	科学	5.063	20.905	50.628	21.068	32.337	130.000	科学	0.937	4.095	-1.628	-0.068	-3.337	0.000
動物学	3	15	41	35	26	120	動物学	4.673	19.296	46.734	19.447	29.849	120.000	動物学	-1.673	-4.296	-5.734	15.553	-3.849	0.000
物理学	10	22	47	9	26	114	物理学	4.440	18.332	44.397	18.475	28.357	114.000	物理学	5.560	3.668	2.603	-9.475	-2.357	0.000
工学	3	11	25	15	34	88	工学	3.427	14.151	34.271	14.261	21.889	88.000	工学	-0.427	-3.151	-9.271	0.739	12.111	0.000
微生物学	1	6	14	5	11	37	微生物学	1.441	5.950	14.410	5.996	9.204	37.000	微生物学	-0.441	0.050	-0.410	-0.996	1.796	0.000
植物学	0	12	34	17	23	86	植物学	3.349	13.829	33.492	13.937	21.392	86.000	植物学	-3.349	-1.829	0.508	3.063	1.608	0.000
統計学	2	5	11	4	7	29	統計学	1.129	4.663	11.294	4.700	7.214	29.000	統計学	0.871	0.337	-0.294	-0.700	-0.214	0.000
数学	2	11	37	8	20	78	数学	3.038	12.543	30.377	12.641	19.402	78.000	数学	-1.038	-1.543	6.623	-4.641	0.598	0.000
合計	31	128	310	129	198	796	合計	31.000	128.000	310.000	129.000	198.000	796.000	合計	0.000	0.000	0.000	0.000	0.000	0.000

その他の尺度

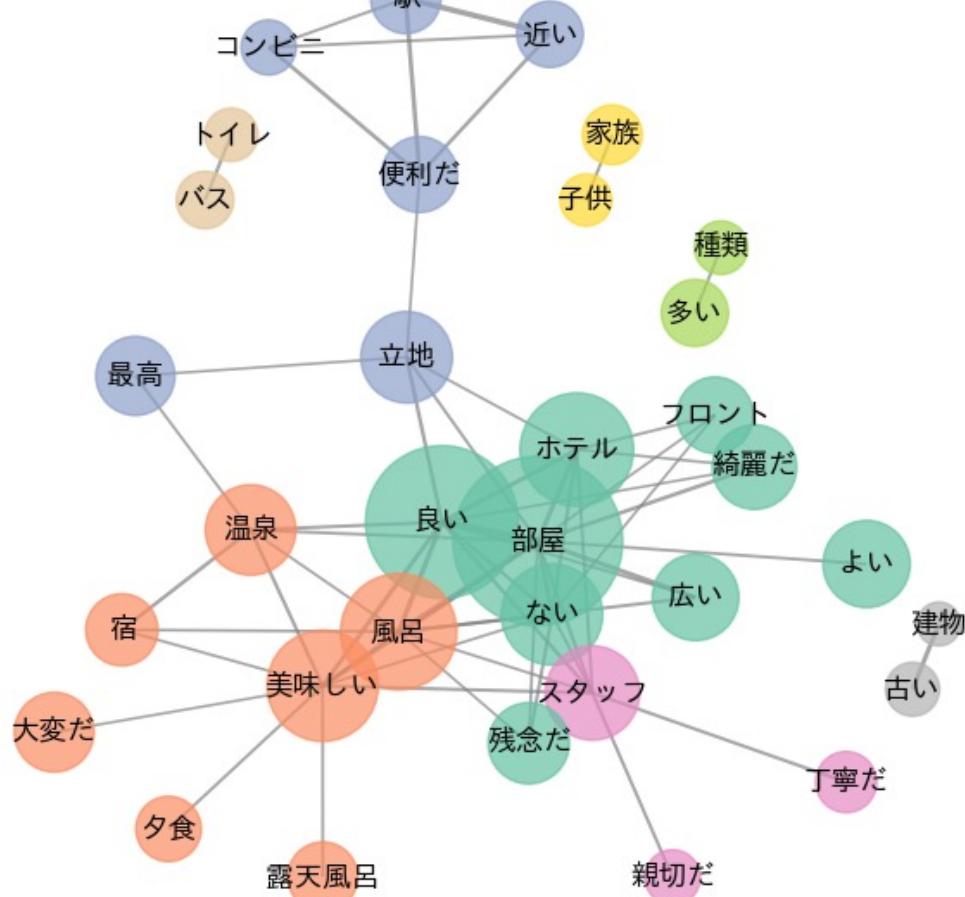
- 出現パターンが似てる を測る = ユークリッド距離、コサイン距離
- 1つひとつの文が長く、各文中での語の出現回数の大小が重要なケースに向く
(語が1回出現したか、10回出現したかを区別したい)

ユークリッド距離	コサイン距離
サイズ(出現回数の大小)の差まで見る場合向き	傾きが似ているかどうかだけを見る場合向き
$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$	$d(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$

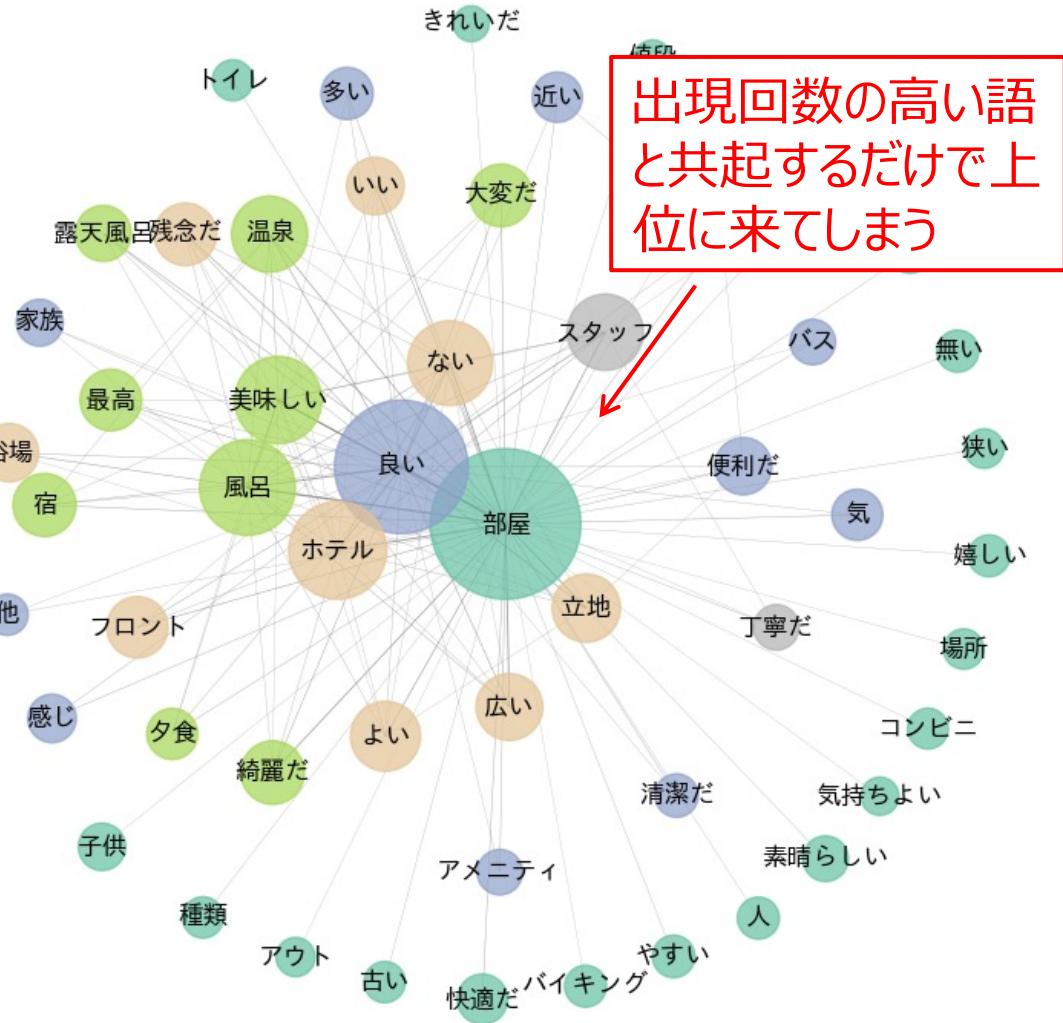
※ x, y はそれぞれの単語ベクトル (単語の出現パターン)



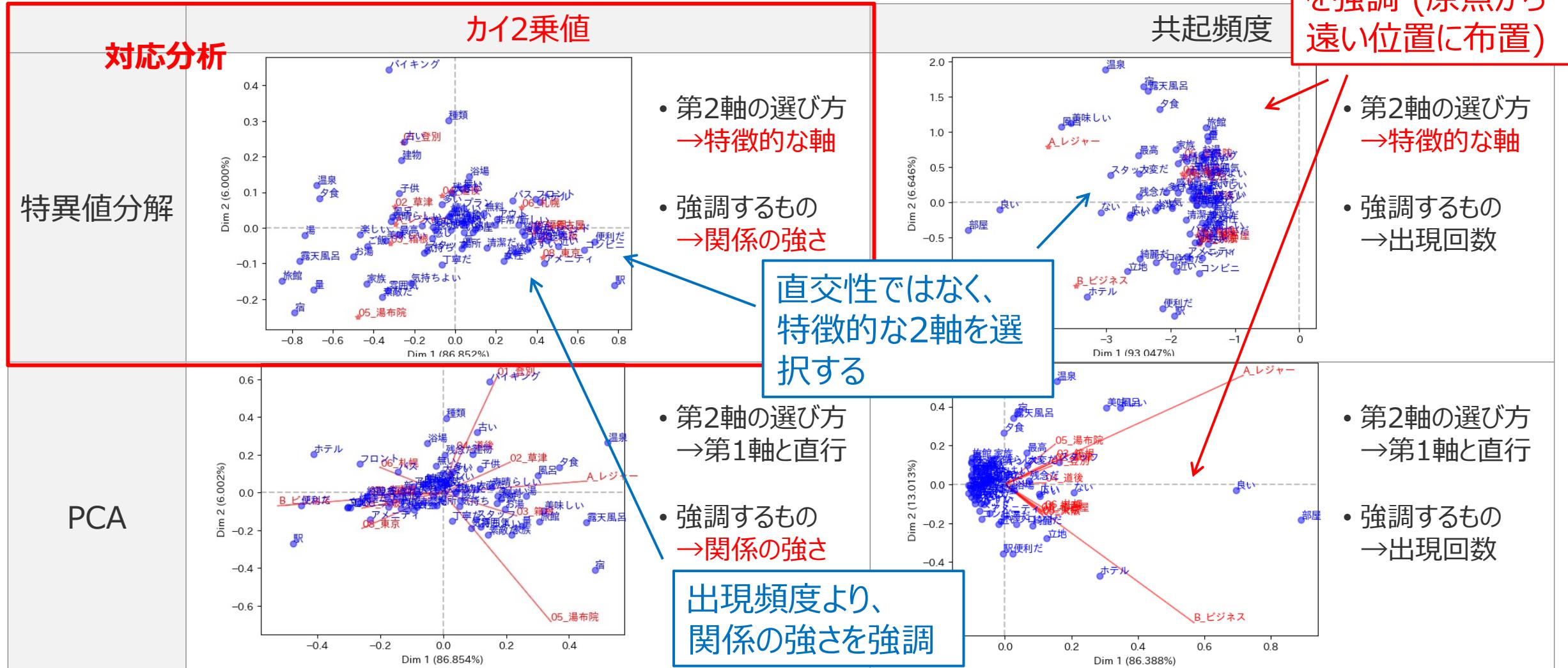
- Jaccard 係数が上位のエッジを残す



- 共起頻度の高いエッジを残す



- 対応分析は、カイ2乗値を利用して、関係の強さを強調して表現できる

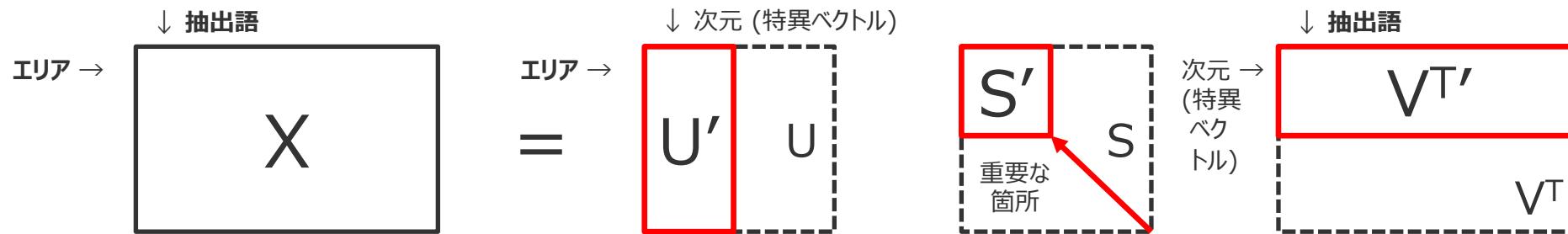


特異値分解

- 特異値分解 $X = USV^T$



- S の特異値が小さいものを削る



テキスト分析 (実践編)

(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

実践的な分析

- 実践1: カテゴリーやエリアごとのユーザーの注目ポイントを押さえる
- 実践2: カテゴリーやエリアごとのユーザーの注目ポイントの評価の違いを見つける
- 実践3: 高評価のエリアに倣って、低評価のエリアを改善するプランを提案する
→ 注意: プロットによる可視化と宿泊客の生の声(原文)を使って解釈する

例) 実践3のまとめ方

対象エリア	エリアX の評価ポイント	エリアY の課題	エリアYの改善案
エリアX: XXX エリアY: XXX	• 風呂が広い 根拠原文: ... • ...	• エアコンが臭い 根拠原文: ... • ...	• ... • ...

実践1 — ユーザーの注目ポイントを押さえる

- カテゴリーやエリアごとの注目する観点の違いを確認する
 - カテゴリー「レジャー」と「ビジネス」を比較する
 - カテゴリー「レジャー」(or「ビジネス」) の 5エリアを比較する
- 手順:
 - カテゴリーやエリアごとの特徴語の違いから、宿泊客が注目する観点を調べる

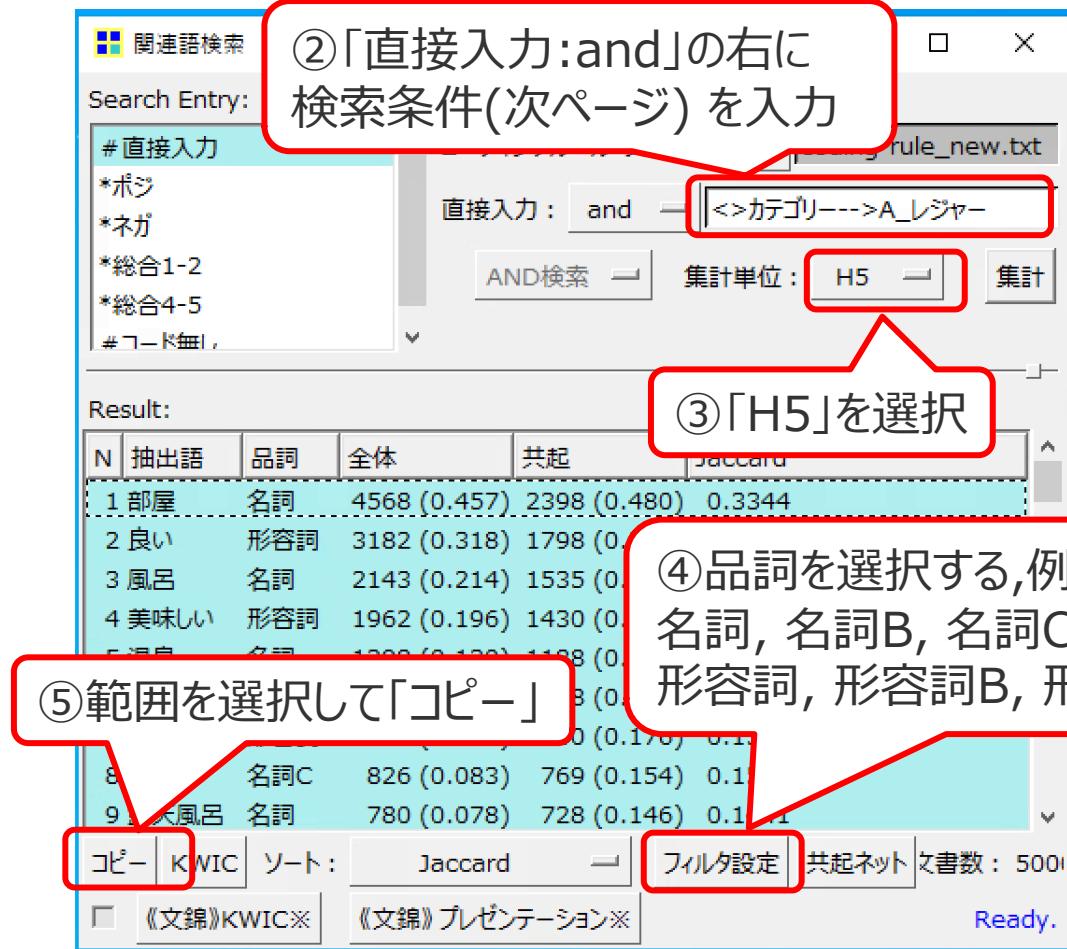
「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]”<>カテゴリ-->A_レジャー”」「集計単位:H5」→「フィルタ設定」→「品詞=名詞, 形容動詞, 未知語, タグ, 形容詞, 名詞B, 形容詞B, 名詞C」を選択→「集計」→結果を選択し「コピー」

「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]”<>エリア-->01_登別”」「集計単位:H5」→「フィルタ設定」→「品詞=名詞, 形容動詞, 未知語, タグ, 形容詞, 名詞B, 形容詞B, 名詞C」を選択→「集計」→結果を選択し「コピー」

実践1 — ユーザーの注目ポイントを押さえる

● カテゴリーやエリアごとの特徴語を抽出する

- ①メニューから「ツール」「抽出後」「関連語検索」を選ぶ



A	B	C	D	E	F
1	1 部屋	名詞	4568 (0.457)	2398 (0.480)	0.3344
2	2 良い	形容詞	3182 (0.318)	1798 (0.360)	0.2816
3	3 風呂	名詞	2143 (0.214)	1535 (0.307)	0.2737
4	4 美味しい	形容詞	1962 (0.196)	1430 (0.286)	0.2585
5	5 温泉	名詞	1298 (0.130)	1188 (0.238)	0.2325
6	6 スタッフ	名詞	1411 (0.141)	888 (0.178)	0.1608
7	7 ない	形容詞B	1629 (0.163)	880 (0.176)	0.1531
8	8 宿	名詞C	826 (0.083)	769 (0.154)	0.1521
9	天風呂	名詞	780 (0.078)	728 (0.146)	0.1441
10	高	名詞	992 (0.099)	698 (0.140)	0.1318
11	変	形容動詞	1007 (0.101)	652 (0.130)	0.1218
12	食	名詞	678 (0.068)	611 (0.122)	0.1206
13	小	形容詞	1186 (0.119)	631 (0.126)	0.1136
14	14 残念	形容動詞	1045 (0.105)	609 (0.122)	0.112
15	15 よい	形容詞B	892 (0.089)	495 (0.099)	0.0917
16	タヨ	名詞	864 (0.080)	467 (0.099)	0.0866

実践1 — ユーザーの注目ポイントを押さえる

- **直接入力: [and]** の右側に入力する条件式

レジヤー:

<>カテゴリ-->A_レジヤー

<>エリア-->01_登別

<>エリア-->02_草津

<>エリア-->03_箱根

<>エリア-->04_道後

<>エリア-->05_湯布院

ビジネス:

<>カテゴリ-->B_ビジネス

<>エリア-->06_札幌

<>エリア-->07_名古屋

<>エリア-->08_東京

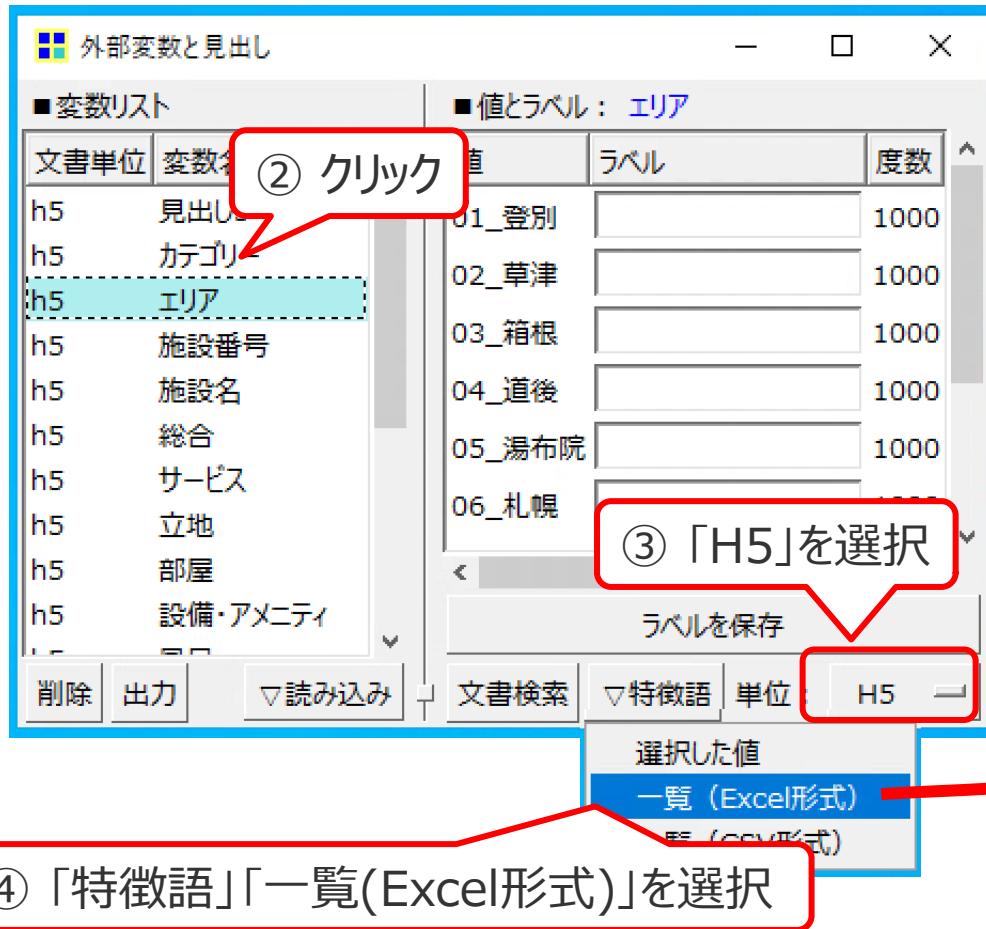
<>エリア-->09_大阪

<>エリア-->10_福岡

実践1 — ユーザーの注目ポイントを押さえる

- カテゴリーやエリアごとの特徴語を抽出する（一括でEXCELに出力）

- 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	食事	.134	湯畑	.327	食事	.159	道後	.128			
4	風呂	.115	草津	.171	美味しい	.136	温泉	.109			
5	温泉	.107	温泉	.136	露天風呂	.134	松山	.098			
6	美味しい	.094	食事	.129	風呂	.116	朝食	.092			
7	良い	.093	風呂	.126	部屋	.109	立地	.082			
8	満足	.090	宿	.120	良い	.106	最高	.066			
9	バイキング	.090	美味しい	.102	温泉	.102	広い	.063			
10	宿泊	.088	良い	.100	思う	.100	浴場	.059			
11	思う	.082	部屋	.096	料理	.100	駐車	.057			
12	料理	.082	思う	.096	宿	.097	フロント	.057			
13	05_湯布院		06_札幌		07_名古屋		08_東京				
14	宿	.180	札幌	.158	名古屋	.150	駅	.102			
15	食事	.159	朝食	.099	朝食	.097	利用	.087			
16	美味しい	.144	ホテル	.092	ホテル	.086	ホテル	.086			
17	露天風呂	.135	利用	.085	利用	.083	便利	.078			
18	風呂	.127	立地	.077	便利	.072	立地	.077			
19	温泉	.124	便利	.077	駅	.070	東京	.072			
20	料理	.123	綺麗	.071	綺麗	.069	近い	.071			
21	最高	.114	浴場	.070	フロント	.066	朝食	.070			
22	スタッフ	.110	対応	.066	立地	.065	綺麗	.064			
23	満足	.107	フロント	.065	近い	.059	快適	.063			
24	09_大阪		10_福岡								
25	大阪	.111	博多	.126							
26	ホテル	.108	ホテル	.090							
27	便利	.097	便利	.087							
28	駅	.096	利用	.085							
29	立地	.080	立地	.074							
30	綺麗	.072	朝食	.064							
31	福岡	.067	駅	.064							
32	フロント	.067	立地	.064							
33	快適	.064	綺麗	.064							
34	広い	.064	駅	.064							

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

実践1 — ユーザーの注目ポイントを押さえる

- 数値評価ではすべての項目に回答
→ レジャーとビジネスでは注目する項目にかなり偏りがありそう

● 特徴語の抽出結果の例

A_レジャー	数値評価指標
部屋	.334
良い	.282
風呂	.274
美味しい	.259
温泉	.233
スタッフ	.161
ない	.153
宿	.152
露天風呂	.144
最高	.132

01_登別	02_草津	03_箱根	04_道後	05_湯布院
風呂	.115	湯畑	.327	美味しい
温泉	.107	温泉	.136	露天風呂
美味しい	.094	風呂	.126	風呂
良い	.093	宿	.120	部屋
バイキング	.090	美味しい	.102	良い
残念	.078	良い	.100	温泉
ない	.077	部屋	.096	宿
夕食	.076	最高	.090	スタッフ
種類	.075	夕食	.085	夕食
露天風呂	.074	ない	.074	便利

B_ビジネス	数値評価指標
ホテル	.223
立地	.147
便利	.134
駅	.124
綺麗	.123
フロント	.107
近い	.091
快適	.090
アメニティ	.072
コンビニ	.069

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡
ホテル	.092	ホテル	.086	駅
立地	.077	便利	.072	ホテル
便利	.077	駅	.070	駅
綺麗	.071	綺麗	.069	便利
浴場	.070	フロント	.066	立地
フロント	.065	立地	.065	近い
広い	.063	近い	.059	綺麗
快適	.056	アメニティ	.056	フロント
駅	.056	快適	.055	綺麗
ベッド	.055	コンビニ	.051	便利

Tips: 「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=カテゴリー」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→「フィルタ設定」→「品詞=名詞、形容動詞、未知語、タグ、形容詞、名詞B、形容詞B、名詞C」を選択→「▽特徴語」→「一覧(EXCEL形式)」で連続実行

(参考) 表記ゆれを吸収する方法 (1/2)

目的：同じ意味の単語を同一視する別の単語として扱わない

例) 「部屋」「お部屋」の 2単語 → どちらも「部屋」としてカウント

方法：「表記揺れを吸収」プラグインを利用する

手順：(出典 <https://github.com/ko-ichi-h/khcoder/issues/101>)

1. プラグインをダウンロードし、解凍して plugin_jp 配下へコピー

ダウンロードURL https://github.com/ko-ichi-h/khcoder/files/4809463/z1_edit_words3.zip

解凍後ファイル名 z1_edit_words3.zip → **z1_edit_words3.pm**

配置後のパス **khcoder3¥plugin_jp¥z1_edit_words3.pm**

(参考) 表記ゆれを吸収する方法 (2/2)

手順 (続き):

2. プラグインファイル「z1_edit_words3.pm」を編集する

編集前

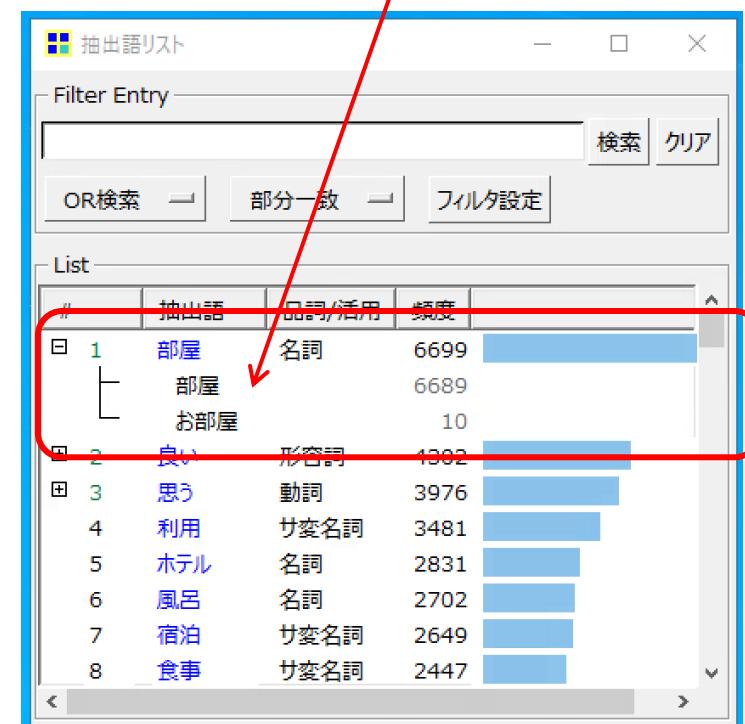
```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '友達' =>
6         [
7             '友人',
8             '旧友',
9             '親友',
10            '盟友',
11            '友',
12        ],
13        '格別' =>
14        [
15            '特別',
16            '格別', # 通常
17        ],
18        '# の',
19        '偶然' =>
20        [
21            '偶然', # 形容
22    ];
23}
```

編集後

```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '部屋' =>
6         [
7             'お部屋',
8         ],
9 };
```

→

適用後の例:
「部屋」と「お部屋」がひとつの単語にまとまっている



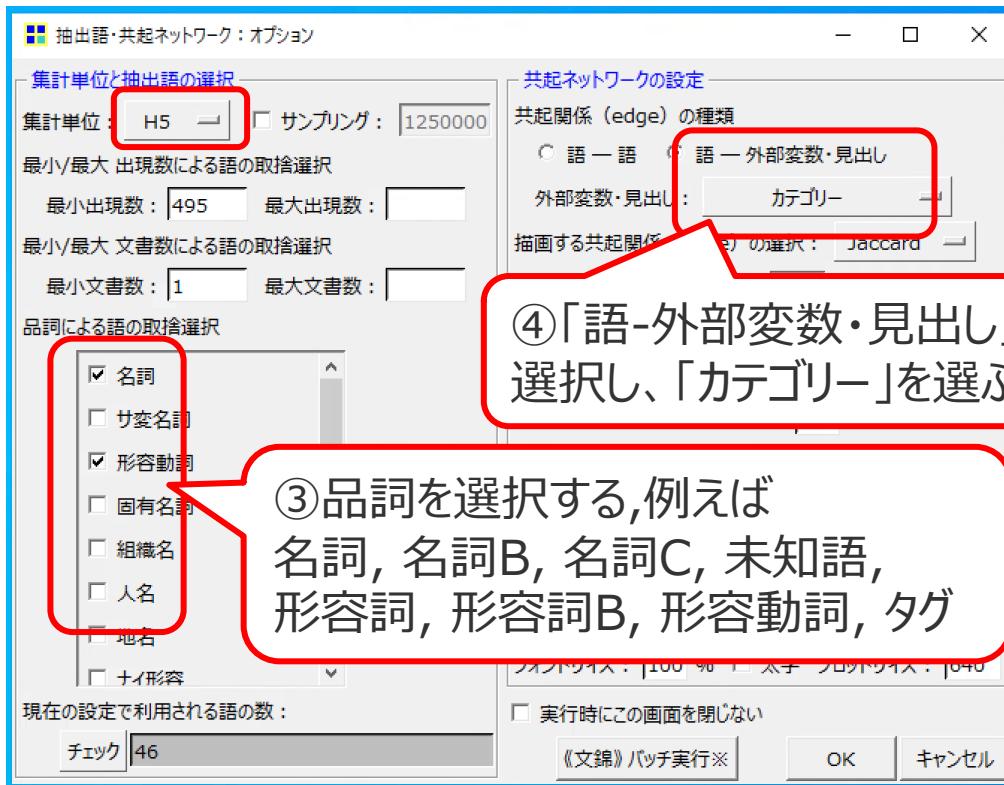
3. KH Coder を再起動する
4. プロジェクトファイルを開く
5. メニューから「ツール」「プラグイン」「表記ゆれの吸収」を選ぶ

実践1 — ユーザーの注目ポイントを押さえる

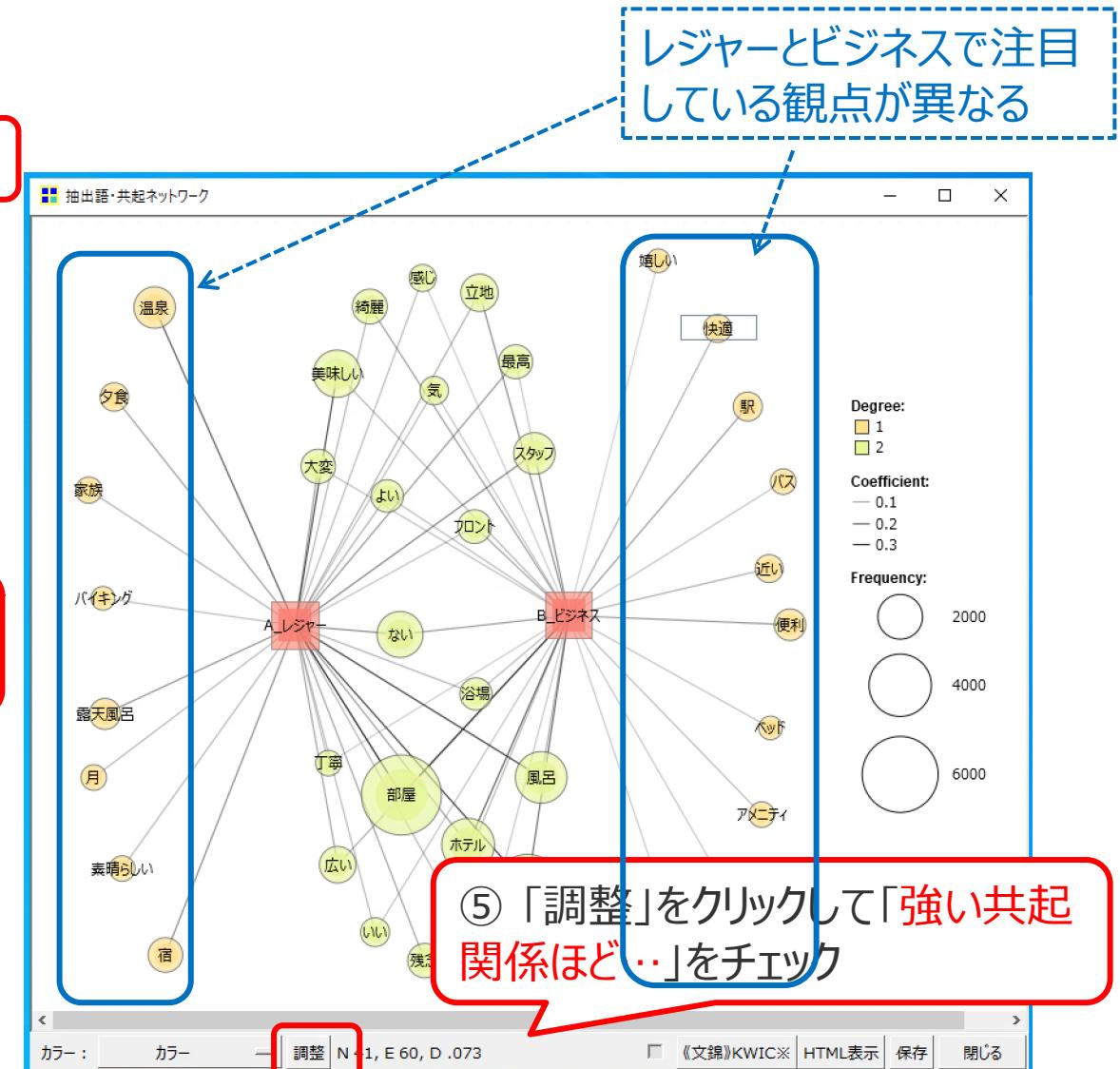
● 共起ネットワークを使う（カテゴリー）

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

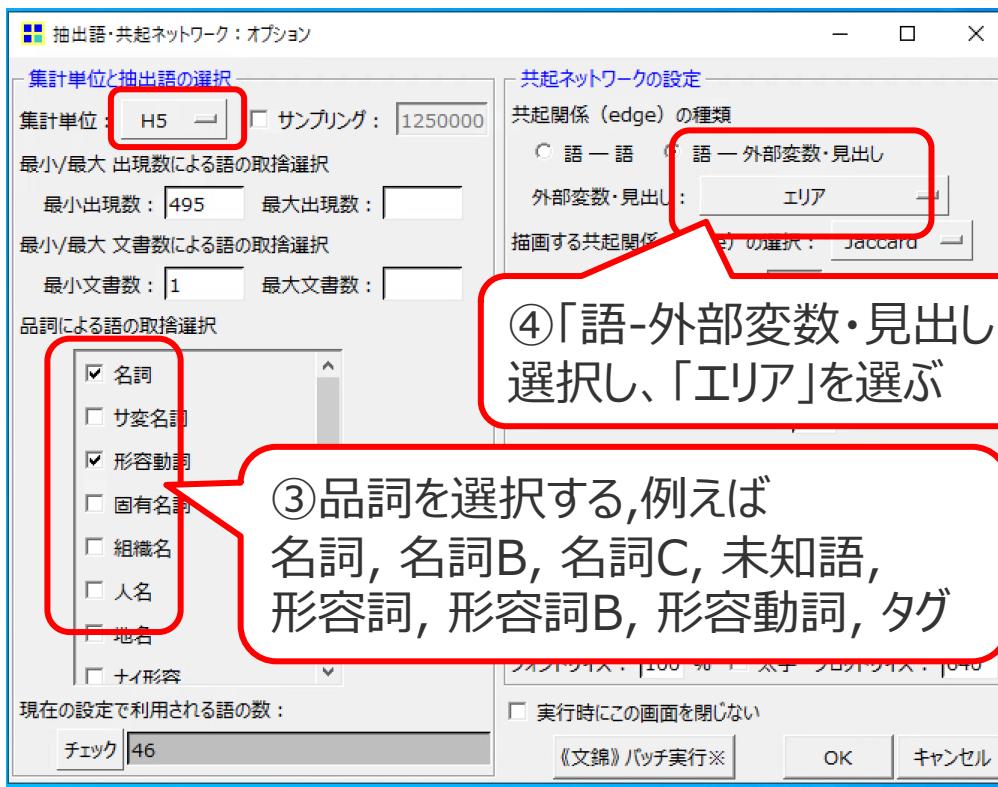


実践1 — ユーザーの注目ポイントを押さえる

● 共起ネットワークを使う（エリア）

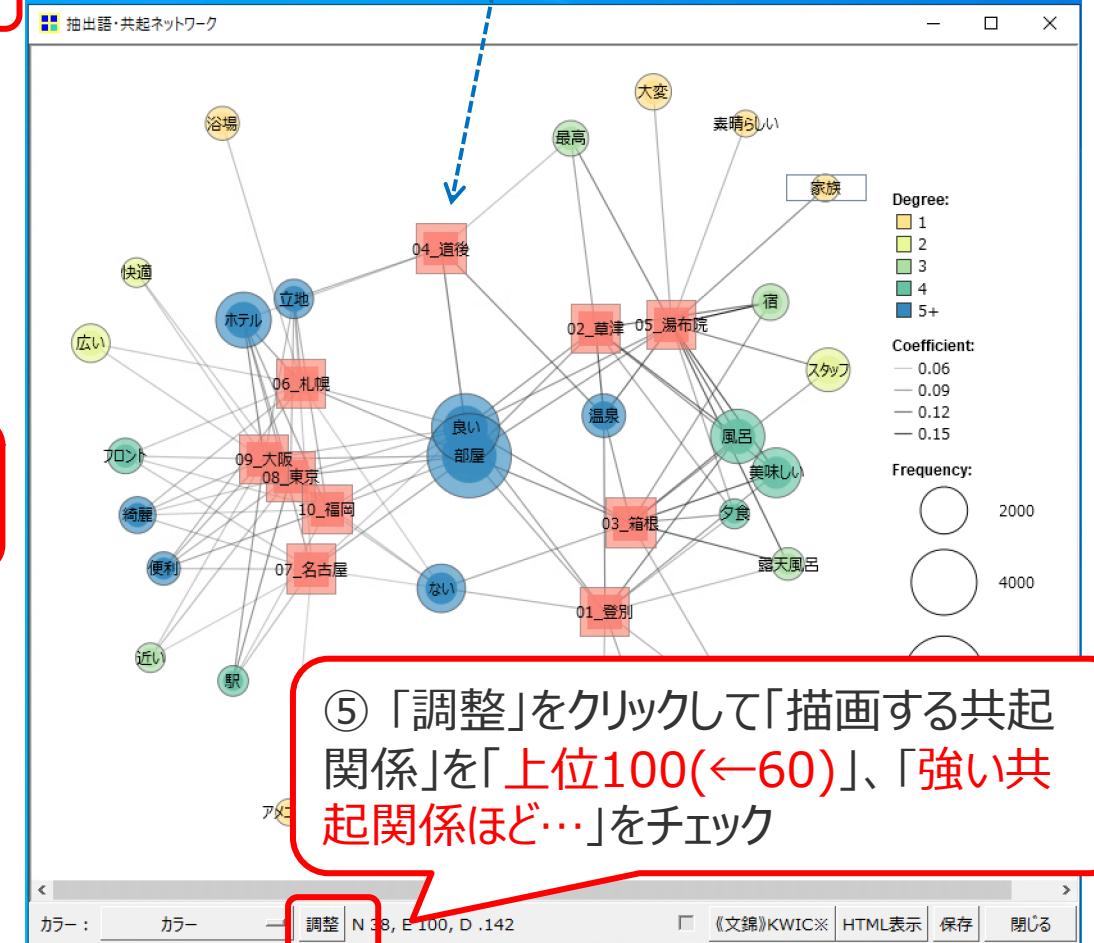
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

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

- ・特徴語抽出と似た傾向が確認できる
- ・道後はレジャーとビジネスの中間的な位置付け

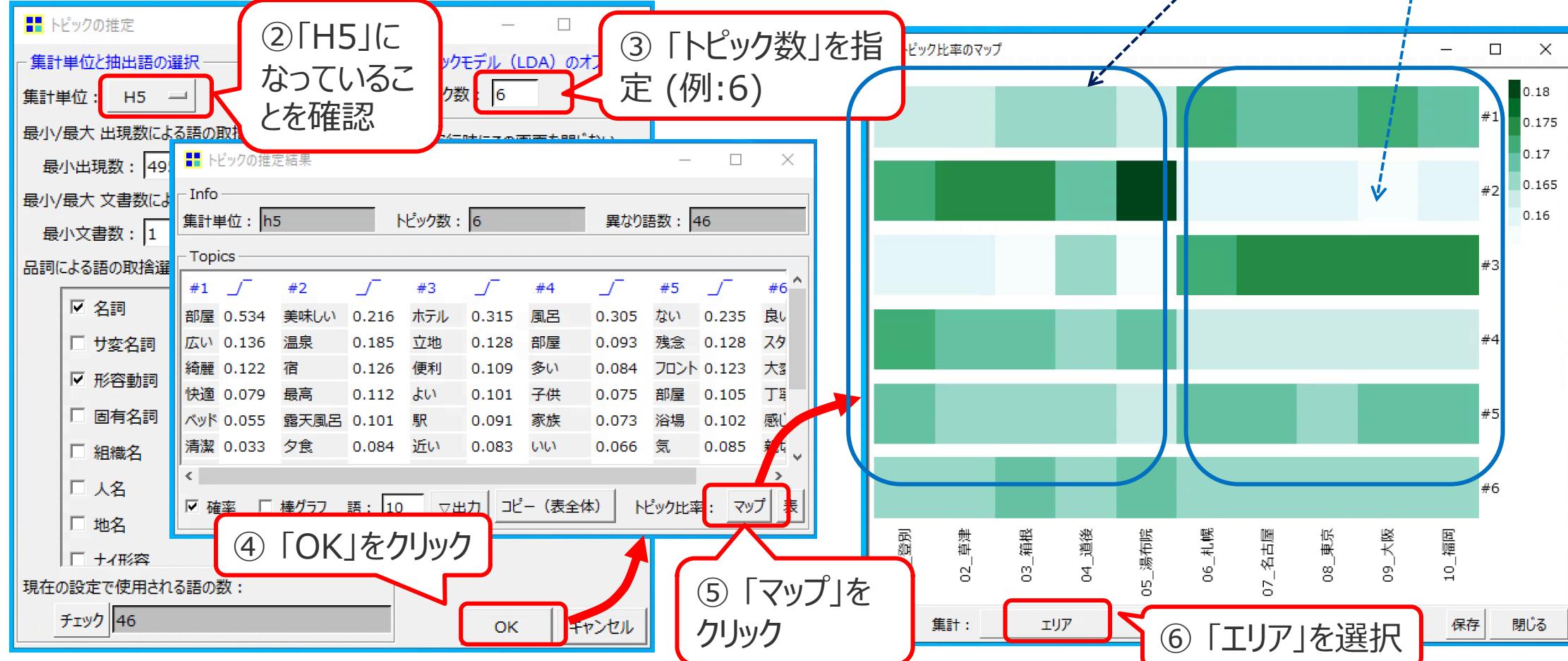


⑤「調整」をクリックして「描画する共起関係」を「上位100(←60)」、「強い共起関係ほど…」をチェック

実践1 — ユーザーの注目ポイントを押さえる

● トピックモデルを使う（エリア）

- 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 4 – レポート課題

- 以下の課題 A と B 両方について PDF ファイルで提出してください
- A. 演習用のデータを用いて、以下の2つのネットワーク図を作成し、キャプチャーを撮るとともに、2つのネットワーク図の違いを観察し、違いについて考察してください
1. KHCoder の共起ネットワーク図
 2. TextMining Studio のことばネットワーク図
- B. 今後、テキストマイニングを使ってご自身で分析したいデータやテーマ(目的)を挙げてください (1件以上)
- ※ 何らかの事情で A. の2つ以上のツールが試せない場合、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.