テキストマイニングの実践
― 5日目 ―

2020/8/7

ビジネス科学研究科
経営システム科学専攻

スケジュール

- 1日目: 7/1(水)
  - 説明 — テキストマイニングの手順
  - 実習 — データをよく知る (Excel)
- 2日目: 7/10(金)
  - 説明 — テキストマイニング ツールの使い方 (KHCoder)
- 3日目: 7/17(金)
  - 説明 — データ分析の実践 (KHCoder)
  - 実習 — データ分析の実践 (KHCoder)
- 体育の日: 7/24(金)
- 4日目: 7/31(金)
  - Text Mining Studio 利用体験
  - 実習 — データ分析の実践 (KHCoder)
- 5日目: 8/7(金)
  - 発表 — データ分析の実践 (KHCoder)

本日の内容

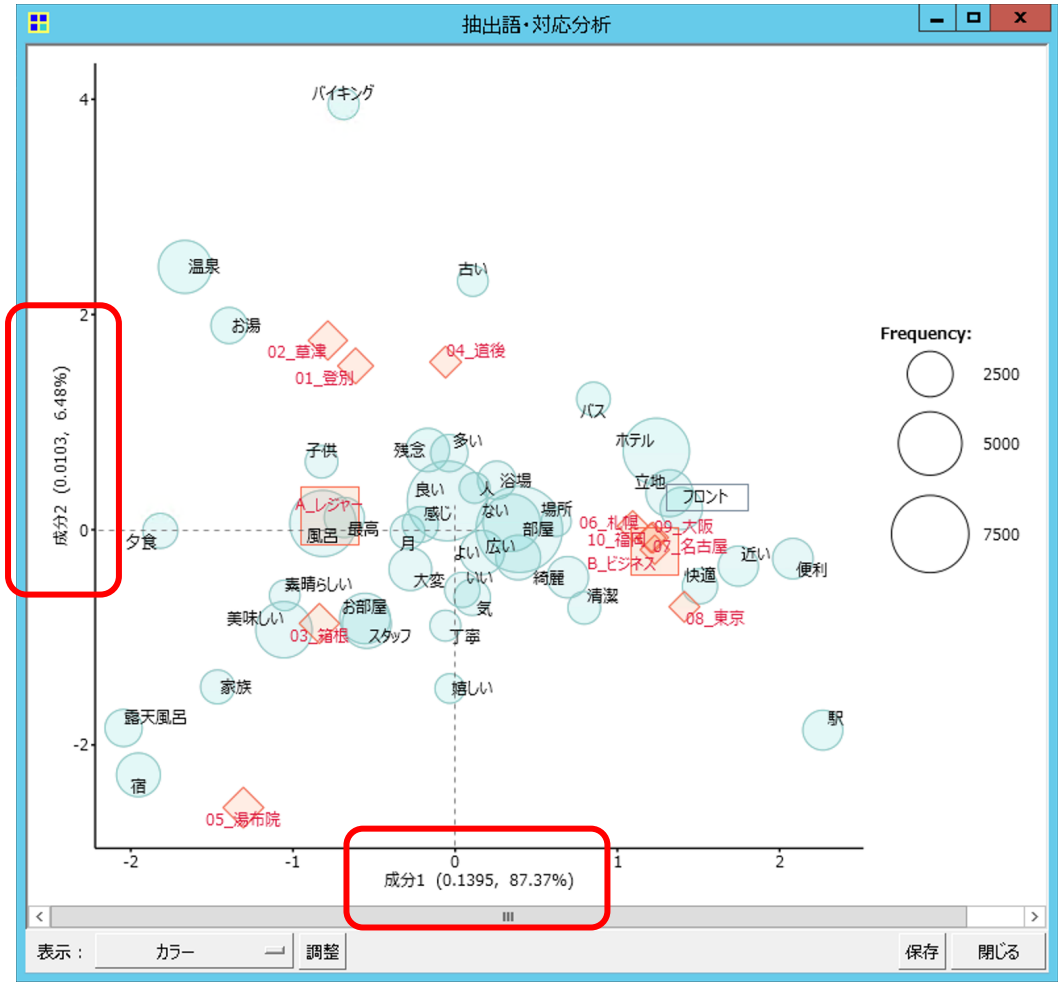
- ~~Q&A~~ (Max ~18:35) → グループワーク (事前確認)

- 発表会 (各グループ 説明10分以内, 質疑5分)
  - 前半 18:35 ~ 19:35
    - グループ1 ~ グループ4
  - 後半 19:45 ~ 21:00
    - グループ5 ~ グループ9

Q&A

Q1. 対応分析の軸は何？ (対応分析)
Q2. Jaccard係数って何？ (関連語検索)
Q3. カイ2乗値で分かることは何？ (クロス集計)

Q1. 対応分析の軸は何？



- KHCoder の対応分析は R の MASS パッケージにある corresp 関数を使用
- 軸ラベルの数値は,固有値および寄与率を示す
- 左図の場合,第2固有値までの累積寄与率は 93.85% で非常に高い
  →第1,2固有値に対応する軸のみを分析すればよい
- 寄与率が高い固有値に対応する行や列の得点の大小とその相対関係について分析する

文の出現パターンと単語の出現パターン

【行】ある文中に出現する単語の数を要素とする (文ベクトル)

【列】全文中に出現する単語の数を要素とする (単語ベクトル)

h5	bun	部屋	ホテル	風呂	温泉	お部屋	スタッ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイキ	家族	場所	トイレ	子供	ベット	コンヒ	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	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	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	0
1	4	0	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	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	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	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	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	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	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	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	0
4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

KH Coder で使われるデータ表

「文書-抽出語」頻度表 (文書のクラスター分析)

h5	bun	部屋	ホテル	風呂	温泉	お部屋	スタッ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイキ	家族	場所	トイレ	子供	ベット	コンヒ	良い	美味し	広い	近い	多い	素晴ら	古い	嬉しい	ない	よい	いい	おいし	宿	駅	気	月	人
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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	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	0	1	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	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	0	0	0	0	0	0	0	0	1	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	0	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

「抽出語-文書」頻度表 (対応分析以外)

h5	1	1	1	1	1	1	1	2	3	3	3	3	4	4	4
bun	1	2	3	4	5	6	7	1	1	2	3	4	1	2	3
id	2	3	4	5	6	7	8	10	12	13	14	15	17	18	19
部屋	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ホテル	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
風呂	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
温泉	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
お部屋	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
スタッフ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
立地	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
フロント	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
最高	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
浴場	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

「外部変数-抽出語」クロス集計表 (対応分析)

	部屋	ホテル	風呂	温泉	お部屋	スタッ	立地	フロン	最高	浴場	お湯	露天風	感じ	夕食	バス	バイキ
A_レジャー	2723	1157	2113	1657	1095	1014	531	436	691	518	756	788	504	730	326	501
B_ビジネス	2340	1839	668	85	419	455	812	806	222	383	113	19	280	47	438	135
01_登別	541	251	429	280	168	198	49	123	128	119	77	122	81	141	47	162
02_草津	532	290	493	469	236	173	160	81	157	95	308	102	111	186	129	164
03_箱根	621	250	476	301	283	267	65	89	130	136	133	254	132	172	76	79
04_道後	464	284	216	319	120	118	170	104	79	100	73	56	80	78	58	81
05_湯布院	565	82	499	288	288	258	87	39	197	68	165	254	100	153	16	15
06_札幌	503	351	131	24	77	95	168	161	49	95	20	4	56	4	70	38
07_名古屋	454	377	141	14	80	70	135	164	39	71	31	3	47	13	77	29
08_東京	431	350	106	2	91	98	157	151	41	83	10	3	57	9	81	13
09_大阪	472	350	150	24	91	116	176	183	45	83	25	5	56	9	84	29
10_福岡	480	411	140	21	80	76	176	147	48	51	27	4	64	12	126	26

対応分析で使われる距離尺度

- χ2 距離でカテゴリー変数間の関連性を測定
  - χ2 は独立性の検定で用いられる指標

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

「観測度数」 カテゴリー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

  - 観測度数と期待度数の差が大きく異なると χ2 値も大きくなり，変数間の関係が期待より強いことを示す

対応分析のプロット

クロス集計表

	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

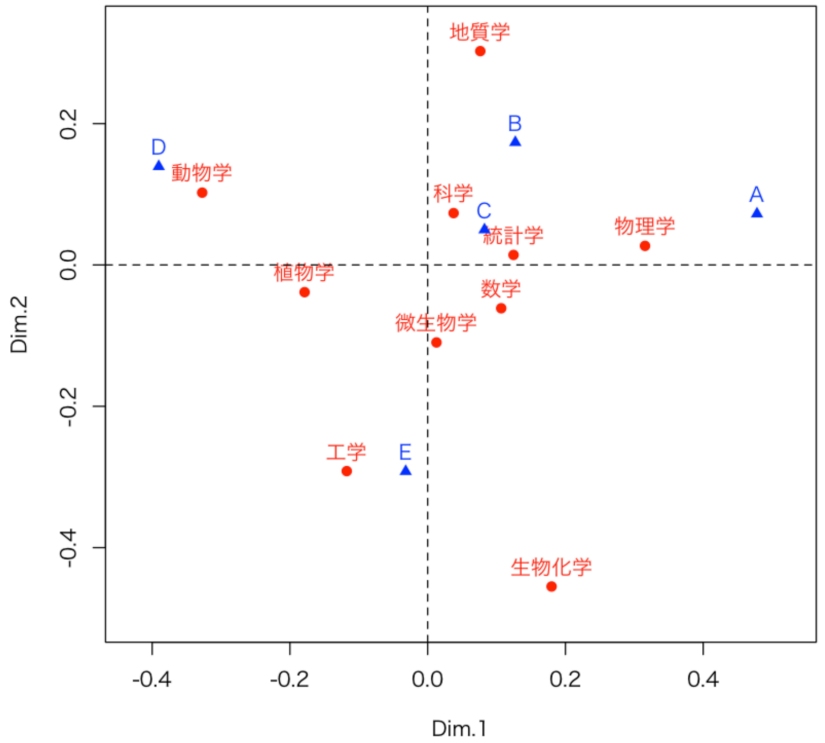
	A	B	C	D	E	合計
地質学	3.310	13.668	33.103	13.775	21.143	85.000
生物化学	1.129	4.663	11.294	4.700	7.214	29.000
科学	5.063	20.905	50.628	21.068	32.337	130.000
動物学	4.673	19.296	46.734	19.447	29.849	120.000
物理学	4.440	18.332	44.397	18.475	28.357	114.000
工学	3.427	14.151	34.271	14.261	21.889	88.000
微生物学	1.441	5.950	14.410	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

カイ二乗距離

	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

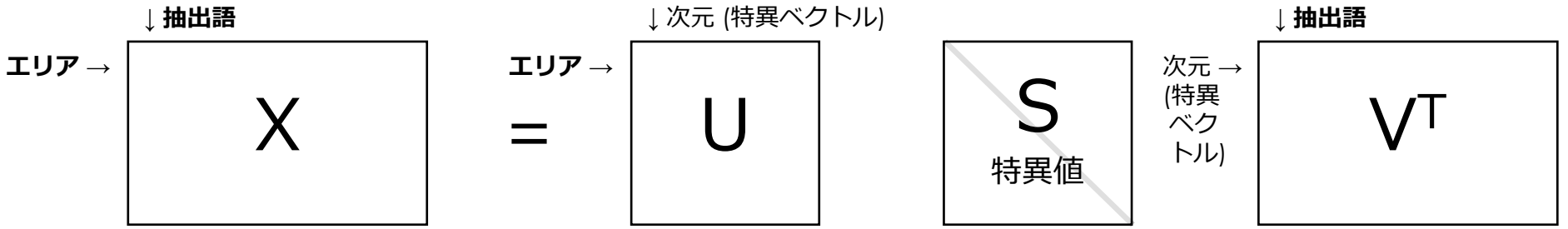


特異値分解してプロット

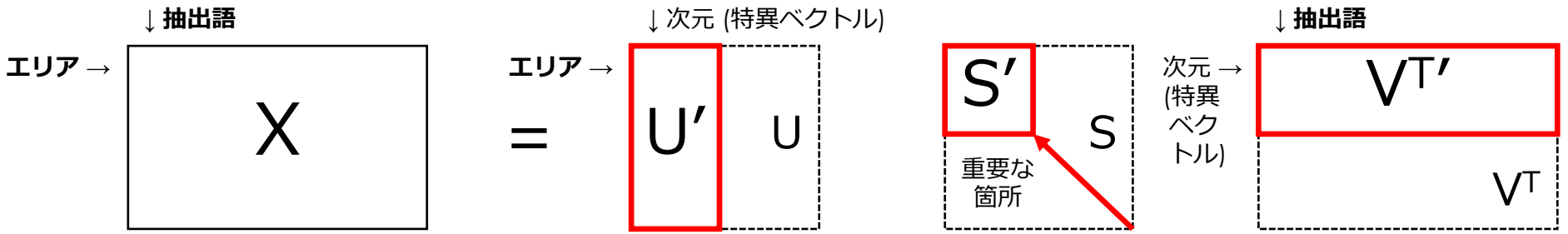
固有値 = {0.0391, 0.0304, 0.0109, 0.0025, 0}
寄与率 = {47.2%, 36.66%, 13.11%, 3.03%, 0%}

特異値分解 (SVD)

- 特異値分解　$X = USV^T$



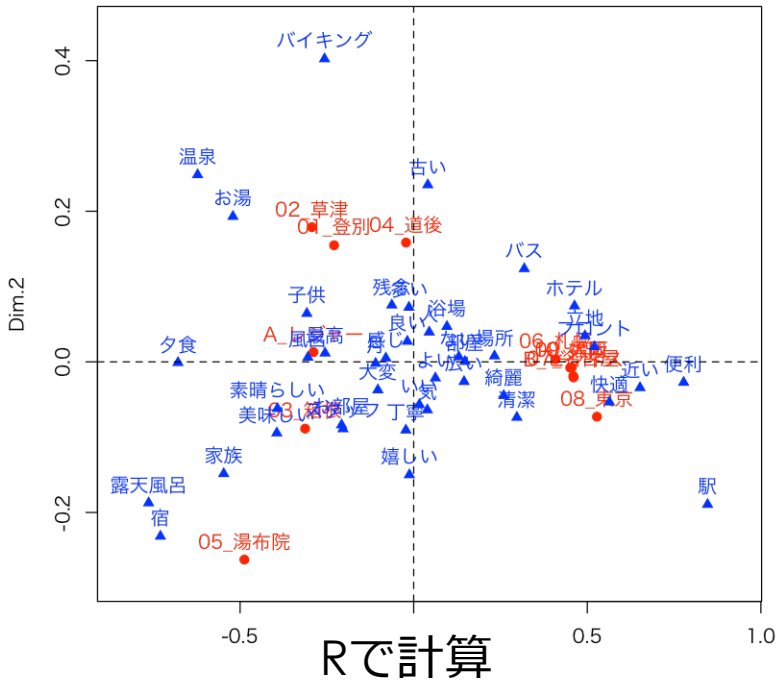
- S の特異値が小さいものを削る



Rによる計算過程の解説

- https://github.com/haradatm/lecture/blob/master/gssm-202007/03-samples/practice-5_sample.ipynb

KHCoder



Rで計算

固有値:
{ 0.1395, 0.0103, … }
寄与率:
{ 87.37%, 6.48%, … }

参考文献

(対応分析)

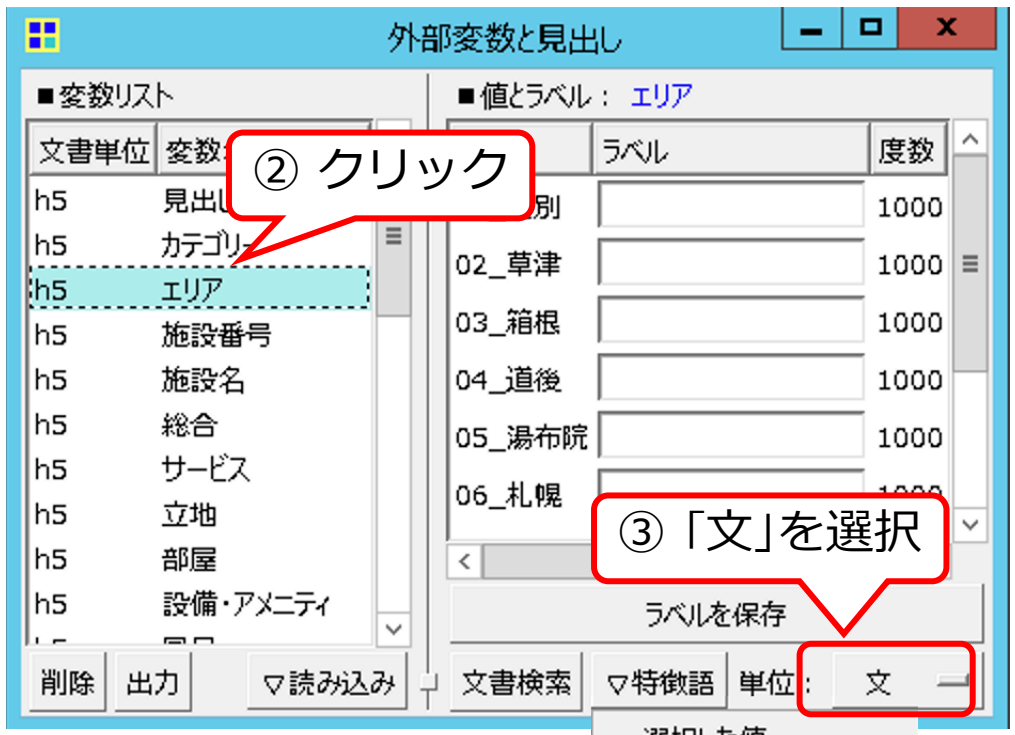
[1] 中山慶一郎. “< 研究ノート> 対応分析によるデータ解析.” 関西学院大学社会学部紀要 108 (2009): 133-145.

[2] 金明哲. Rによるデータサイエンス: データ解析の基礎から最新手法まで. 森北出版, 2007. (P.85「7.2 対応分析」)

[3] 使用したRのコード. https://github.com/haradatm/lecture/blob/master/gssm-202007/03-samples/practice-4_sample.ipynb

Q2. Jaccard係数って何？

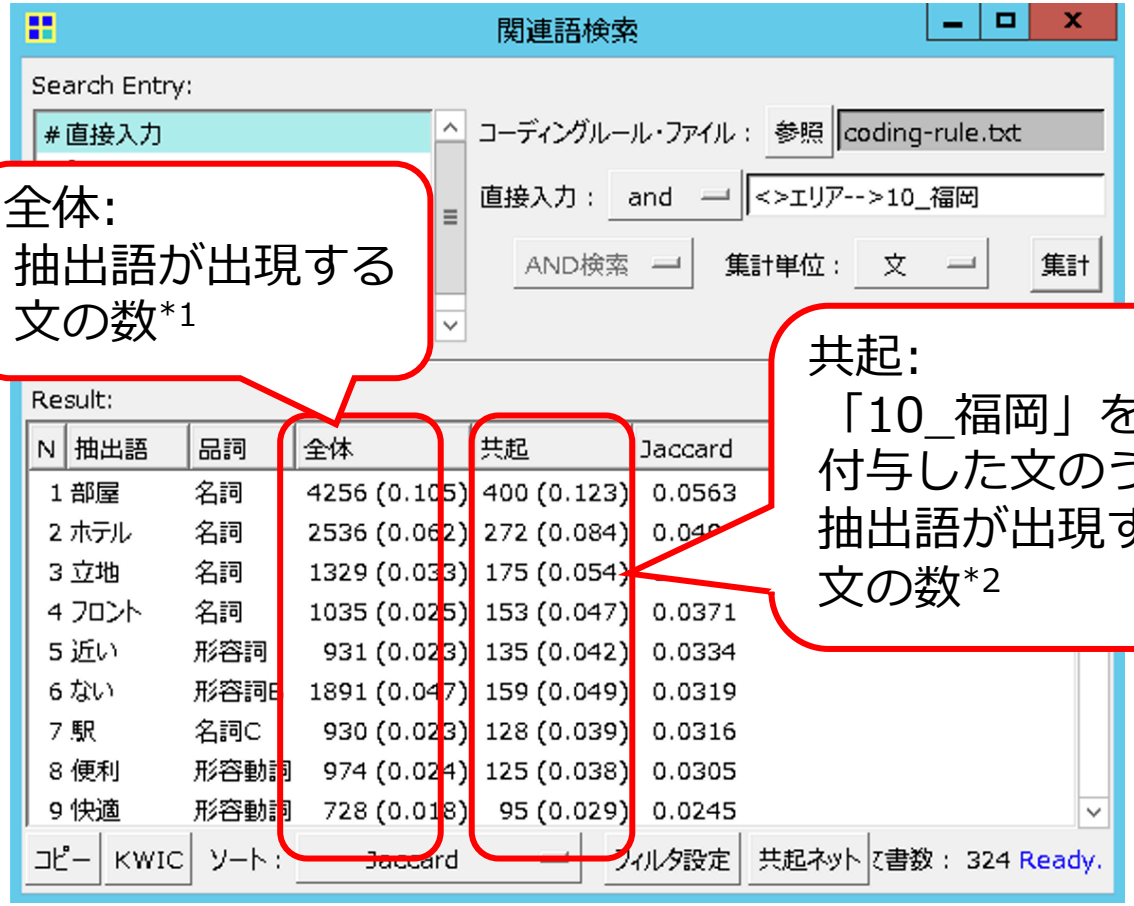
①メニューから「ツール」「外部変数と見出し」「リスト」を開く

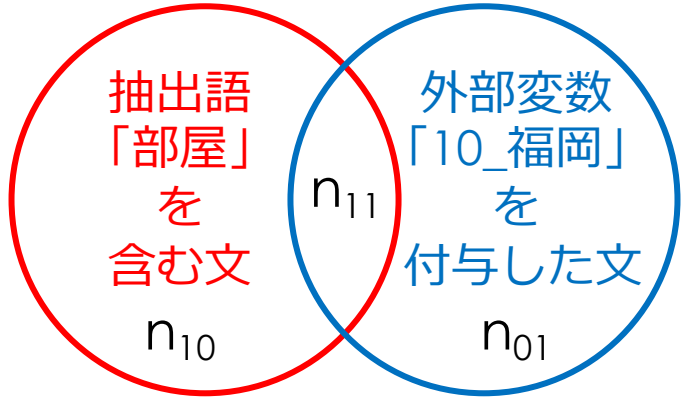


④「特徴語」「一覧(Excel形式)」を選択

各エリアの特徴語を
Jaccard 係数の降順で表示

Jaccard 係数 — 関連の強い語が分かる





Jaccard 係数 $_JS = \dfrac{n_{11}}{n_{11} + n_{10} + n_{01}}$

抽出語「部屋」の場合:

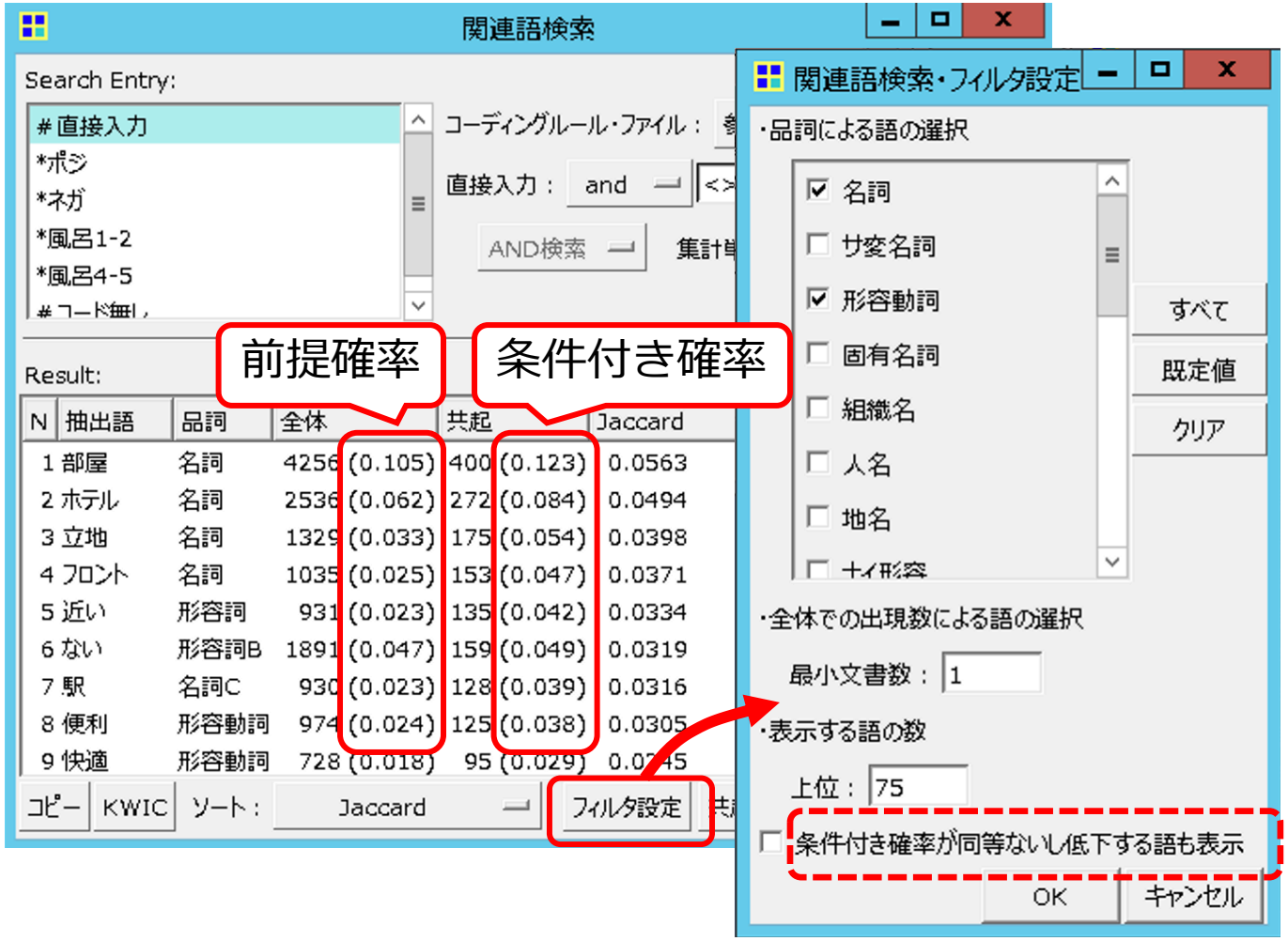
n_{11} = 400 ("共起"列の値)

n_{10} = 4256 ("全体"列の値) – 400 = 3856

n_{01} = (400 / 0.123) – 400 = 2852

*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「10_福岡」を付与したデータに対する割合(条件付き確率)

「条件付き確率が同等ないし低下する語も表示」とは



【注意】

- デフォルトでは「前提確率」より「条件付き確率」が高くなっていない語はリストアップされない
- データ全体における出現確率と同等以下の確率でしか出現していない語は，「関連の強い」「特徴的な語」ではないという考え方
- ただし「フィルタ設定」ボタンをクリックして，「**条件付き確率が同等ないし低下する語も表示**」にチェックを入れると条件付き確率の方が低い語も表示できる

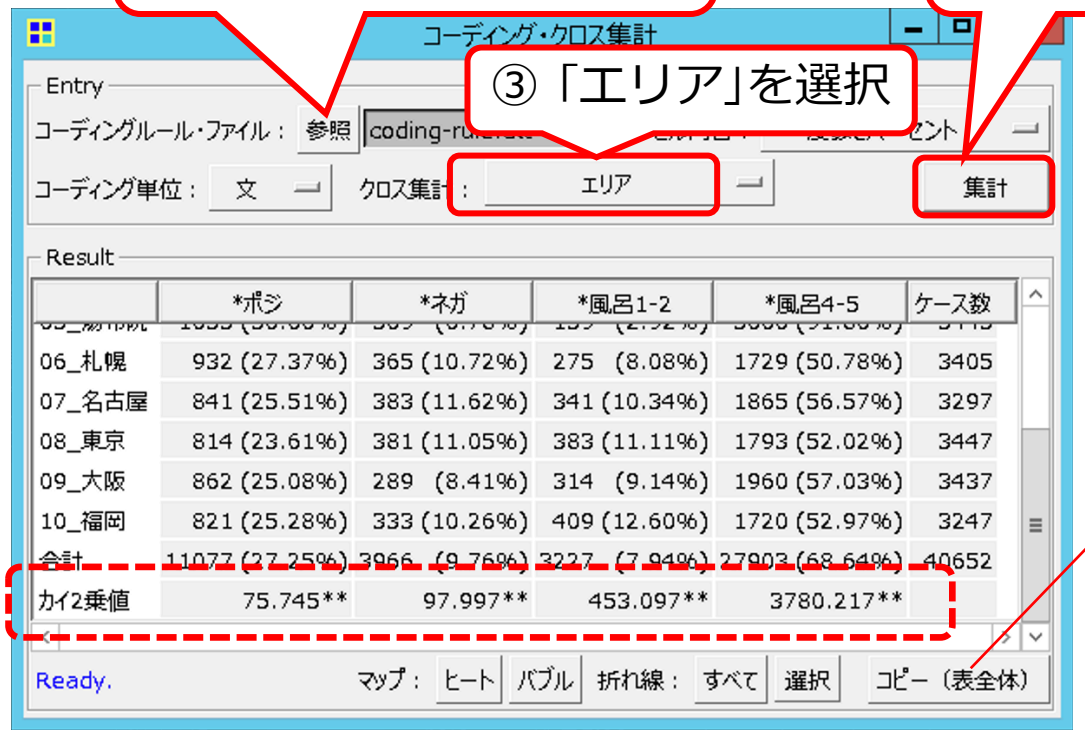
Q3. カイ2乗値の使い方?

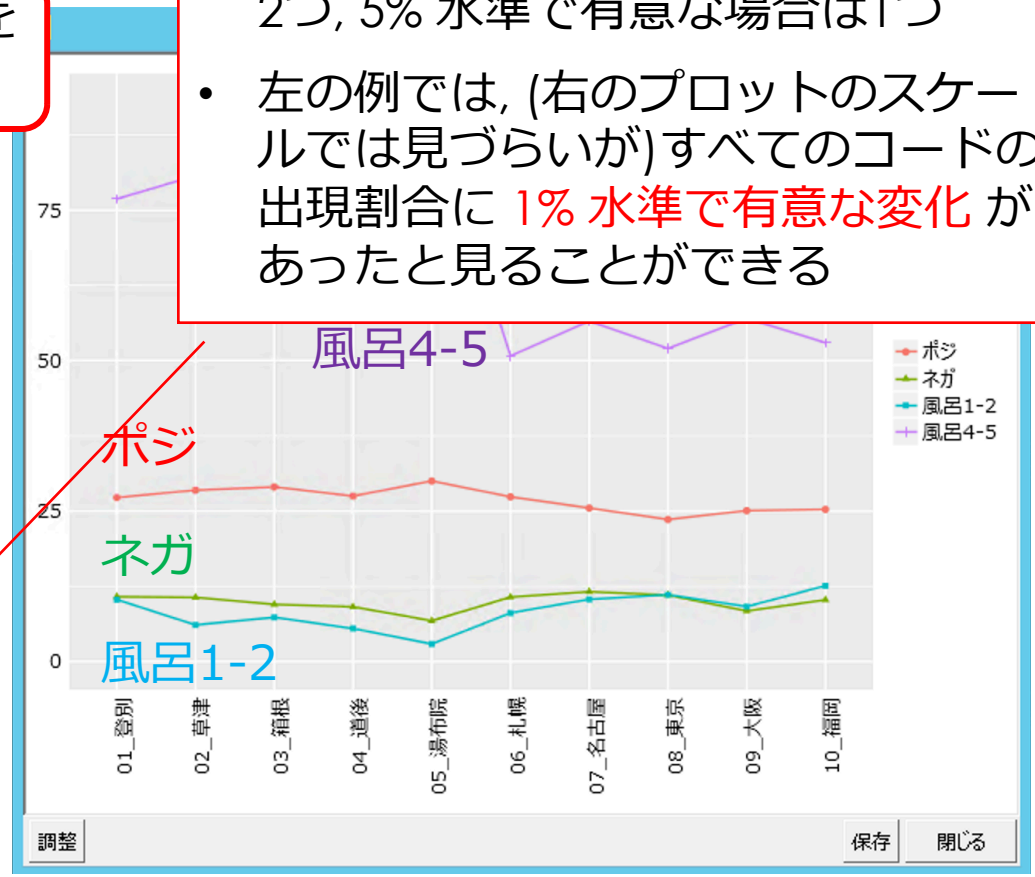
①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

②「参照」をクリックして「coding-rule.txt」を開く

③「エリア」を選択

④「集計」をクリック





- χ2 値の欄に表示されるアスタリスク「*」の数は,1% 水準で有意な場合は2つ, 5% 水準で有意な場合は1つ
- 左の例では, (右のプロットのスケールでは見づらいが)すべてのコードの出現割合に 1% 水準で有意な変化 があったと見ることができる

発表会の進行について

- 各グループの発表時間は，**説明10分以内**，**質疑5分**
  - 前半 18:35~19:35 (グループ1~グループ4)
  - 後半 19:45~21:00 (グループ5~グループ9)

- 次に発表するグループには，司会をお願いします
  - タイムキーパー: 説明10分が過ぎたら，終了を知らせる
  - 発表内容に関する質問を考える: 各人で1件以上 →課題レポート
  - 質疑の司会: 質問が出なければ，司会グループから質問(1件)する

課題 (5日目)

- 司会を担当したグループの発表内容に関する質問(1件以上),
  および 自身グループワークに関する感想 を提出してください
- 発表者は, **必ず**,発表スライドも提出してください
- レポートには, 以下の記載をお願いします
  - 自身が所属する グループ名
  - 発表タイトル
  - メンバ全員の名前 (発表者および司会者名には印)

- 形式: PPT(PDF)，提出先: manaba，期限: 8/14 23:55

発表前の事前確認

- 発表者は, 発表内容を確認してください
- 司会の 進行係(質疑含む), タイムキーパー係 を決めてください

発表会

発表スケージュール

- 各グループ 説明10分以内, 質疑5分
- 司会は, 次に発表するグループが担当
  司会の役割: タイムキーパー,質問(≧1件)

グループ1 18:35~18:50 司会: グループ2	201940108
	202040051
	202040055
	202040074
	202040077
グループ2 18:50~19:05 司会: グループ3	201945014
	202040057
	202040063
	202040080
	MSI
グループ3 19:05~19:20 司会: グループ4	202040052
	202040062
	202040070
	202040072
	202040170

グループ4 19:20~19:35 司会: グループ5	201947529
	202040058
	202040060
	202040068
	202040069
グループ5 19:45~20:00 司会: グループ6	201940015
	201947523
	202040059
	202040066
	202040073
グループ6 20:00~20:15 司会: グループ7	201840109
	201940129
	202040064
	202040078
	202040079

グループ7 20:15~20:30 司会: グループ8	202040071
	202040076
	202040409
	202040413
	201540111
グループ8 20:30~20:45 司会: グループ9	202020027
	202020051
	202040067
	202040075
グループ9 20:45~21:00 司会: グループ1	202040053
	202040054
	202040056
	202040061

皆さん，大変お疲れ様でした