

テキストマイニング

— Part 5 —

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
 - 説明 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)
- Part 5
 - 説明 — ラップアップ

(再掲) 実践的な分析 ― ゴール

- カテゴリーやエリアごとに、**注目ポイント**を押さえる
- カテゴリーやエリアごとに、**注目ポイントの評価の違いを見つける**
- 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**
 - ただし、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) 結果の整理

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: ... • ...	• ... • ...

(再掲) 演習 ー 改善案を提案する

- 特徴語と**ポジティブ意見**の共起ネットワーク図を作成し,エリアによって**ポジティブ意見**(とその背景)がどう異なるかを比較することで,何がどう評価されているかを確認する (→P.35)
- 特徴語と**ネガティブ意見**の共起ネットワーク図を作成し,エリアによって**ネガティブ意見**(とその背景)がどう異なるかを比較することで,何がどう評価されているかを確認する (→P.36)
- 高評価エリアに倣って,低評価エリアを**改善プランを提案する** (→P.38)

注: プロットによる可視化 と 宿泊客の生の声(原文) を使って解釈すること

(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

前回の課題 — 改善案を提案する

- 以下をスライドにまとめ **PDF ファイルで提出** してください
 - 演習で作成した**共起ネットワーク図**(P.35,36)と**結果の整理**(P.38)を用いて,選択した低評価エリアを改善するプランを提案する

形式: PDF, 提出先: manaba, 期限: 次週開始時刻(～18:20)

よくある質問

Q1. 単語登録したいときは？

Q2. 表記ゆれ統一(or 同義語登録)したいときは？

Q3. 対応分析の軸って何？ (対応分析)

Q4. Jaccard係数って何？ (関連語検索)

Q5. Web からデータ収集するには？

Q6. その他

Q1. 単語登録したいときは？

- 目的

- 複数の単語に分かれる → 1単語として抽出できるようにする
例) 「湯」「畑」の 2単語 → 「湯畑」として 1単語

- 方法

- 「前処理の実行」前に「強制出力する語の指定」に追加する

- 手順

1. メニューから「前処理」「語の取捨選択」を選ぶ
 - 「強制出力する語の指定」欄に抽出したい単語を登録する
 - 「OK」ボタンで画面を閉じる
2. メニューから「前処理」「前処理の実行」を選ぶ

Q2. 表記ゆれを統一したいときは? (1/2)

出所: <https://github.com/ko-ichi-h/khcoder/issues/101>

- 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

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**

(次ページにつづく)

Q2. 表記ゆれを統一したいときは? (2/2)

- 手順

- 2. プラグインファイル

z1_edit_words3.pm を編集する

```
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    ], # の
18    '偶然' =>
19    [
20        '偶然', # 形容
21    ],
22 };
23
```

編集前

→

```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '部屋' =>
6     [
7         'お部屋',
8     ],
9 };
10
```

編集後

↓

- 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
 - 6. 分析を続ける

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

抽出語リスト

Filter Entry

部屋 検索 クリア

OR検索 部分一致 フィルタ設定

List

#	抽出語	品詞/活用	頻度
1	部屋	名詞	6737
	部屋		4876
	お部屋		1861
2	大部屋	名詞	3

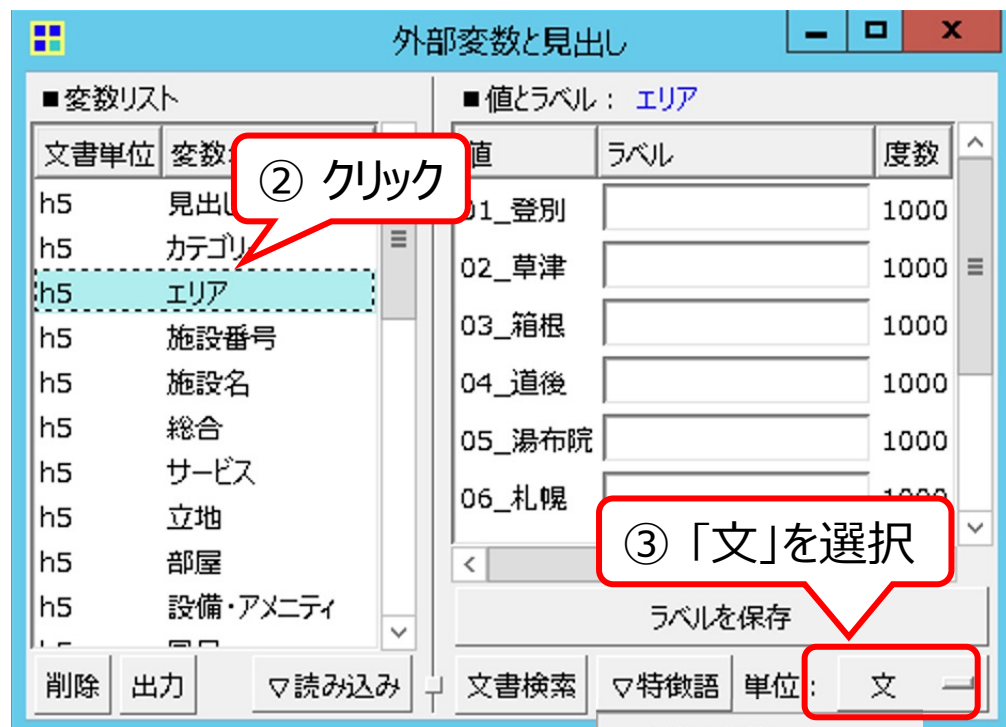
Q3. 対応分析の軸って何？



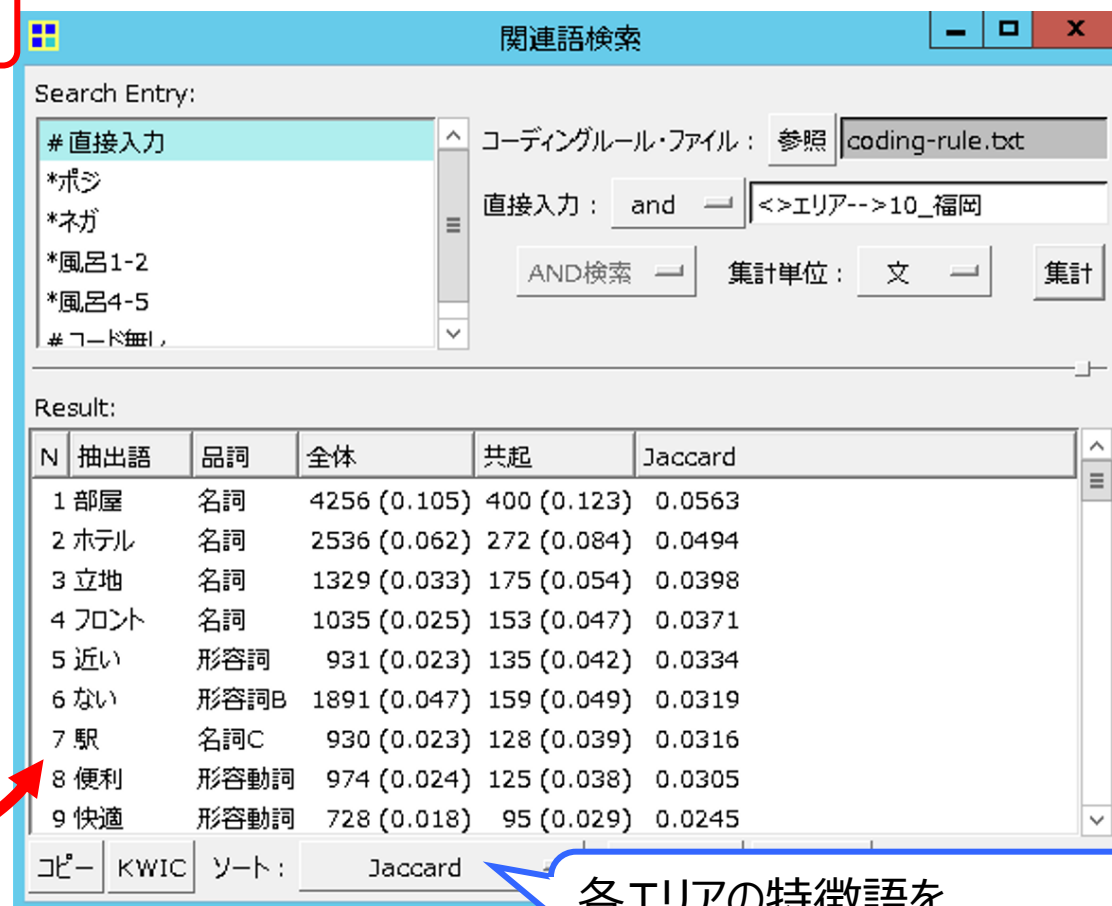
- KHCoder の対応分析は R の **MASS** パッケージにある **corresp** 関数を使用
- 軸ラベルの数値は、固有値および寄与率を示す
- 左図の場合、第2固有値までの累積寄与率は 93.85% で非常に高い
→第1,2固有値に対応する軸のみを分析すればよい
- 寄与率が高い固有値に対応する行や列の得点の大小とその相対関係について分析する

Q4. 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

関連語検索

Search Entry:
直接入力

コーディングルール・ファイル: 参照 coding-rule.txt

直接入力: and <>エリア-->10_福岡

AND検索 集計単位: 文 集計

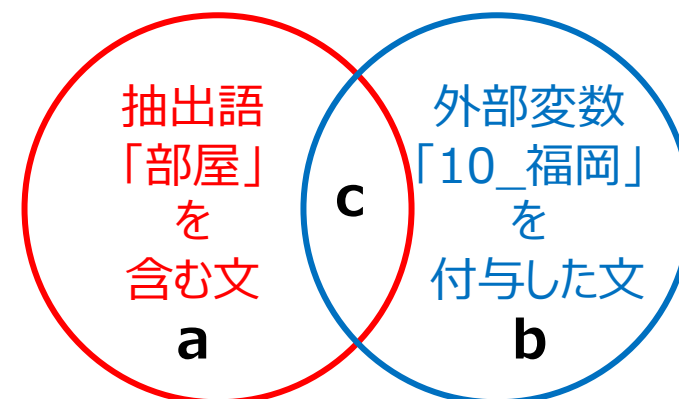
Result:

N	抽出語	品詞	全体	共起	Jaccard
1	部屋	名詞	4256 (0.105)	400 (0.123)	0.0563
2	ホテル	名詞	2536 (0.062)	272 (0.084)	0.049
3	立地	名詞	1329 (0.033)	175 (0.054)	
4	フロント	名詞	1035 (0.025)	153 (0.047)	0.0371
5	近い	形容詞	931 (0.023)	135 (0.042)	0.0334
6	ない	形容詞	1891 (0.047)	159 (0.049)	0.0319
7	駅	名詞C	930 (0.023)	128 (0.039)	0.0316
8	便利	形容動詞	974 (0.024)	125 (0.038)	0.0305
9	快適	形容動詞	728 (0.018)	95 (0.029)	0.0245

全体: 抽出語が出現する文の数*1

共起: 「10_福岡」を付与した文のうち、抽出語が出現する文の数*2

コピー KWIC ソート: Jaccard フィルタ設定 共起ネット 文書数: 324 Ready.



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

抽出語「部屋」の場合:

$c = 400$ (“共起”列の値)

$b = 4256$ (“全体”列の値) $- 400 = 3856$

$a = (400 / 0.123) - 400 = 2852$

*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

関連語検索

Search Entry:

- # 直接入力
- * ポジ
- * ネガ
- * 風呂1-2
- * 風呂4-5
- # コード無し

コーディングルール・ファイル: 参照

直接入力: and

AND検索 集計

Result:

N	抽出語	品詞	全体	共起	Jaccard
1	部屋	名詞	4256 (0.105)	400 (0.123)	0.0563
2	ホテル	名詞	2536 (0.062)	272 (0.084)	0.0494
3	立地	名詞	1329 (0.033)	175 (0.054)	0.0398
4	フロント	名詞	1035 (0.025)	153 (0.047)	0.0371
5	近い	形容詞	931 (0.023)	135 (0.042)	0.0334
6	ない	形容詞B	1891 (0.047)	159 (0.049)	0.0319
7	駅	名詞C	930 (0.023)	128 (0.039)	0.0316
8	便利	形容動詞	974 (0.024)	125 (0.038)	0.0305
9	快適	形容動詞	728 (0.018)	95 (0.029)	0.0245

前提確率

条件付き確率

フィルタ設定

関連語検索・フィルタ設定

品詞による語の選択

- ☒ 名詞
- ☐ サ変名詞
- ☒ 形容動詞
- ☐ 固有名詞
- ☐ 組織名
- ☐ 人名
- ☐ 地名
- ☐ ナイ形容

すべて 既定値 クリア

全体での出現数による語の選択

最小文書数: 1

表示する語の数

上位: 75

☐ 条件付き確率が同等ないし低下する語も表示

OK キャンセル





【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

Q5. Web からデータ収集するには？

- Web からデータ収集する方法 = **Web スクレイピング**
 - **Python** (プログラミング言語) とそのライブラリ “**BeautifulSoup**” を使うと、比較的手軽に HTML からデータを抽出&整形できる
- 楽天トラベルからデータを収集するサンプル
 - <https://github.com/haradatm/lecture/tree/master/gssm-202207/05-colab>

More sample scripts (not used in the course)

file name	memo
scraping_example.ipynb  Open in Colab  Open Studio Lab	A toy example of web scraping (using the data in the course)
rakuten_example.ipynb  Open in Colab  Open Studio Lab	A toy example of rakuten dataset analysis (using the data in the course)



クリックすると Colab にジャンプ

(参考) Colaboratory とは

• 機械学習の教育・研究を目的とした研究用ツール



<https://studiolab.sagemaker.aws/>

AWS から同様に、無料のクラウドサービスが登場しました (ファイルの永続化ができます)



- **設定不要** (最初から **Python** や機械学習に必要なものが入っている)
- **無料**で使える (**Googleアカウント**さえあれば良い)
- **ブラウザ**で動作する (PCのスペックが低くても関係なし)
- **GPUが無料**で使える (計算時間を大幅に短縮できる)
- **ただし, 90分&12時間ルール** あり *1

*1 Colab Pro (1,072円/月)にすることで各種制限を緩和できます

(参考) Colab による Python 入門サイト

- 東大が無料公開している **Python の初心者 にオススメの教材**
 - <https://utokyo-ipp.github.io/>



The screenshot shows the 'Python プログラミング入門' (Python Programming Introduction) page. It features a navigation menu on the left with links to various topics, including '1-0. Colaboratoryによるノートブックの使い方' (How to use notebooks with Colaboratory). The main content area is titled 'Python プログラミング入門' and includes a warning about the content being for hobbyists and a list of topics to be covered, such as 'Colaboratoryの立ち上げ' (Setting up Colaboratory) and 'セルの操作' (Cell operations).

Python プログラミング入門

▲で始まる項目は授業では扱いません。興味にしたがって学習してください。
ノートブック全体に▲が付いているものもありますので注意してください。

- 1-0. Colaboratoryによるノートブックの使い方
 - Colaboratoryの立ち上げ
 - ノートブックのアップロード
 - 教材のオープン
 - ノートブックのダウンロード
 - ノートブックのアップロード (再び)
 - ノートブックの作成
 - ノートブックの操作
 - セル
 - セルの編集
 - 練習
 - セルの挿入
 - セルの実行が止まらないとき
 - セルの操作
 - ノートブックの参照

- 教材と照らし合わせながら Python に関する学習を進められる
- Google Colaboratory による **ノートブックの使い方, 文字列, 条件分岐, 繰り返し, 関数, NumPy ライブラリ, pandas ライブラリ, scikit-learn ライブラリ** など幅広く扱っている

Q.6 その他

Q: 樋口先生のチュートリアル, 夏目漱石の小説を一体どうやって読み込んだの?

A: 青空文庫 (著作権が消滅した作品や著者が許諾した作品のテキストを公開している電子図書館) で公開されています

- <https://www.aozora.gr.jp/>

01_登別	
風呂	.058
温泉	.045
美味しい	.043
残念	.034
お部屋	.032
最高	.030
バイキング	.030
露天風呂	.029
大変	.028
夕食	.027

Q: この最下位の0.027は非常に数値が低いいため特徴とはいえない…という解釈ですか?

A: データは, その収集方法や時期, 掲載元の特徴などの影響を受けるため, 単独で数値の高(頻度や共起率)を議論するのではなく, エリア内/外で比較するなど, 常に**相対的に見る**のがポイントです

お疲れ様でした!