

テキストマイニング入門

- ワークを行うため、PC環境（Windows）を推奨します。講師はWindowsユーザーのためMacの仕様に関する対応はできかねます。予めご了承ください。
- 本日の配付資料をお配りしています。USBの中に本日使用するデータセットを入れています。各自、ご自身のPCにコピーをお願いします。

参加方法

OK



リラックス
しましょう



徳島大学高等教育研究センター
教育の質保証支援室
塩川 奈々美

配付資料

- 01 KH Coderの基本操作手順2023.pdf
- 02 テキストマイニング入門2023.pdf
- 03 配付資料(ワーク:コーディング).xlsx
- 04 配付資料(ワーク:テキストマイニング).csv
- 05 communication card.pdf

本日お配りしている資料はSPODフォーラムのために作成・公開している資料であり、再配付は不可とさせていただきます。
お取扱いにご注意ください。

講師紹介



徳島大学
TOKUSHIMA UNIVERSITY



塩川 奈々美(しおかわ ななみ)

2019.3 大学院総合科学教育部博士後期課程満期退学
(専門:日本語学・方言学)

*2015-2018 「近畿方言における配慮表現の研究」(15H03211)

2018.5～2020.3

高等教育研究センター教育改革推進部門・特任助教

*文部科学省大学教育再生加速プログラム(AP)事業に携わる

→受講生の自由記述から各学部を取組状況を評価

2020年4月～現在に至る

高等教育研究センター教育の質保証支援室・助教(現職)

高等教育研究センター教育改革推進部門・助教(兼務)

テキストマイニングを行う目的は？

A

既存の（＝別の目的で作られた）文章に隠れた特徴を明らかにしたい

B

アンケートの自由記述回答の傾向を見たい

C

文章の内容の優劣を測りたい

D

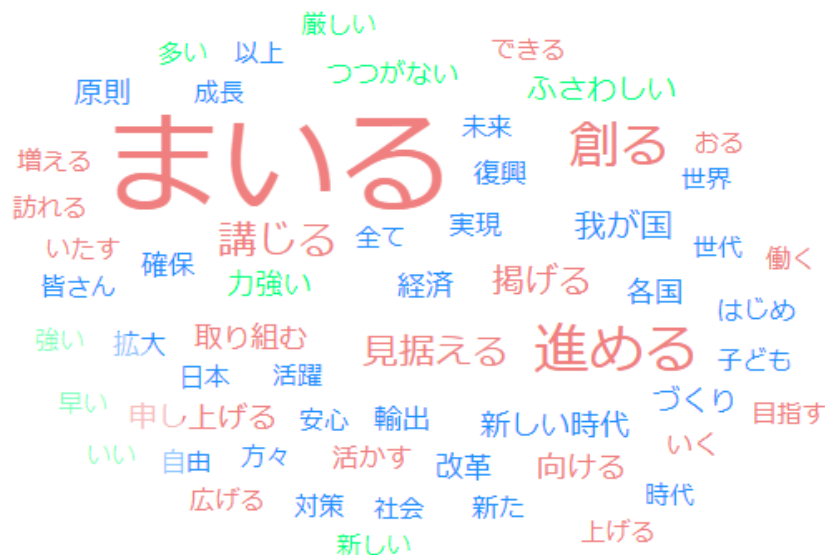
具体的な目的はないけど、興味があって参加した

本日の到達目標

- ① 質的データや質的研究の特徴を把握する
- ② 質的データの分析の流れを理解する
- ③ KH Coderの使い方を学び、今後自身の教育
/研究活動において活用するヒントを得る

本日の流れ

- 質的データと質的調査
- 質的データの分析方法
- ワーク：コーディング
- ワーク：テキストマイニング
- 質的データ分析の注意点



質的データと質的調査

質的データとは

- 数値データまたは数を量的データと見なす。それとは対照的に、質的なデータは、より多様であり、**文字テキストならびに画像、映像、音声記録、人工物など**の他にも多くのものが含まれる。

(Udo Kuckartz著、佐藤郁哉訳2018)

- **「ドキュメント」あるいは「テキスト」と呼ばれ**、古文書、日記、手紙、新聞記事、雑誌記事などをさし、観察やインタビュー調査をもとに得られるものが多い。

Cf. 量的データ

身長、体重、満足度(5件法など)などの数値データのことを表し、0を基点とする比例尺度と、各々のデータの間隔を調べる間隔尺度などがある。

(中嶋2015)

質的データと質的調査方法

・調査方法

(例)

- アンケート調査
- インフォーマル・インタビュー
- 現場観察
- 参与観察
- 自然観察
- 文書・資料の収集
- 歴史的資料・古文書の発掘

● 日常会話に近い言葉による記述であることが多い。

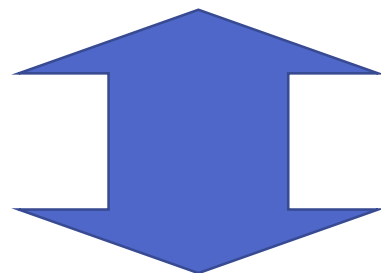
● 言葉のニュアンス、論理の展開(間隔ではなく順序に意味がある)、対象者の思考・特徴などが表出されやすい

質的データにみられる語り手や記録者の思想・考え方などを汲み取ることが重要

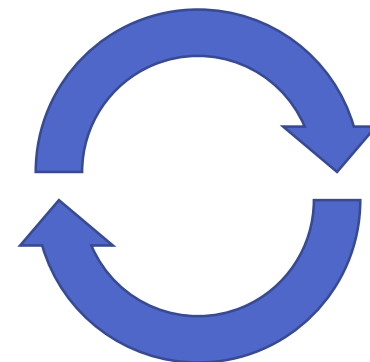
量的研究と質的研究

質的

被調査者の内面を詳細に知りたい
(例: 動機付け、ニーズ把握、ストラテジー等)



混交研究
or
中間的
アプローチ



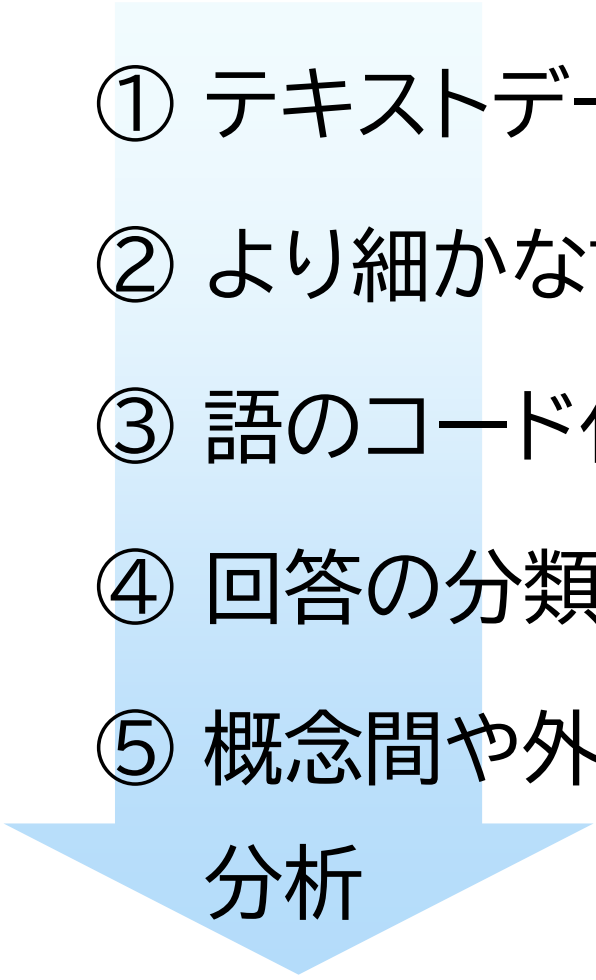
量的

サンプルを多く獲得して結果を
一般化したい、信頼性を高めたい

(大木2017)

質的データの分析方法

分析の基本的な流れ

- 
- ① テキストデータの収集・クリーニング
 - ② より細かな言語単位に分割
 - ③ 語のコード化&カテゴリー化
 - ④ 回答の分類(概念化)
 - ⑤ 概念間や外部変数(属性情報)との関連性を分析

分析の基本的な流れ

藤井・小杉・李（2005）

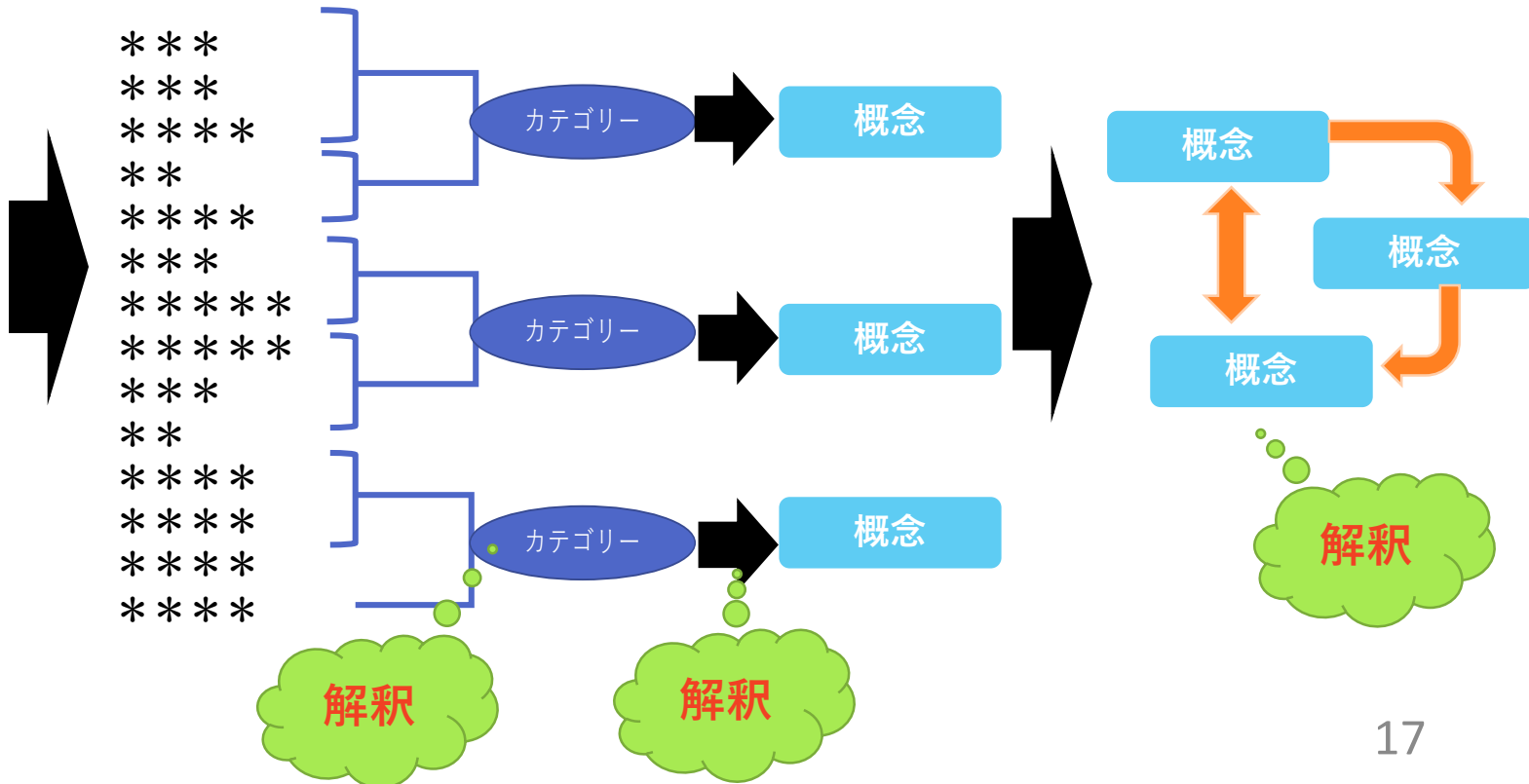
データ
収集・整理

より細かな言
語単位に分割

単語群から考えられるカテゴリー
からより一般的な概念を考える

概念の関係を
検討する

*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*
*	*	*	*	*	*



分析時の重要なポイント

- テキストデータの内容の意味の中心語を抽出すること。

→ **観点(関心)**によってデータ(テキスト)の意味の受け取り方が変わる。

(例) 「あ！あんな所に汚い犬がいる！」

発話者の**行動**に着目→ (発見)

発話者の**感情**に着目→ (嫌悪感)



分析の観点を常に明らかにした状態で
テキストの処理に臨むことが重要！

分析時の重要なポイント

- どういう目的(方針)でテキストを読み込むのか
明確化する
 - ✓ 目的や分類規則の記録を残しつつ作業
- 分析者の語彙力を高める(努力をする)
 - ✓ 先行研究のコーディングを参照
 - ✓ 類語辞典の活用
- 結果を多角的に考察することを心掛ける
 - ✓ 第三者にも納得してもらえるような整理
 - ✓ 見直しや立ち返りの姿勢を心掛ける

ワーク コーディング

コーディング

- 観察記録やインタビュー記録のような1次言語コードから、さらに縮約した2次言語コードを作成すること。（大谷2019）
- 文全体あるいは部分を構成する概念を抽出し、**簡潔に表すための見出しを付けること。**

発表がうまくいった

プレゼンテーション力が身についた

緊張感のあるプレゼンだった

発表の準備を頑張った

プレゼン力

文章の抽象度を上げ、一般化に近づけるための作業

ワーク:コーディングしてみよう

・コーディングの例

竹中（2019）より改変

NO.	回答（自由記述）	見出し	コード
1	配付資料を事前にLMSにアップしてくれたのがよかった。	配付資料のアップ	
2	板書が丁寧だった	丁寧な板書	
3	配付資料がわかりやすかった	わかりやすい配付資料	
4	自分の発表に対する他の学生の意見が聞けてよかった。	発表による意見交換	
5	身近な話題を取り上げてくれて興味が持てた	身近な話題による興味喚起	
6	文字が大きくて板書が見やすかった	板書が見やすい	
7	いろんな学生に発表の機会を与えているのが良かった	発表機会の振り分け	

①自由記述の内容を要約する
＝見出しの作成

ワーク:コーディングしてみよう

・コーディングの例

竹中（2019）より改変

NO.	回答（自由記述）	見出し	コード
1	配付資料を事前にLMSにアップしてくれたのがよかった。	配付資料のアップ	
2	板書が丁寧だった	丁寧な板書	
3	配付資料がわかりやすかった	わかりやすい配付資料	
4	自分の発表に対する他の学生の意見が聞けてよかった。	発表による意見交換	
5	身近な話題を取り上げてくれて興味が持てた	身近な話題による興味喚起	
6	文字が大きくて板書が見やすかった	板書が見やすい	
7	いろんな学生に発表の機会を与えているのが良かった	発表機会の振り分け	

②見出しの中に共通する観点を「コード」として抽出する

ワーク:コーディングしてみよう

• コーディングの例

竹中（2019）より改変

NO.	回答（自由記述）	見出し	コード
1	配付資料を事前にLMSにアップしてくれたのがよかった。	配付資料のアップ	配付資料
2	板書が丁寧だった	丁寧な板書	板書
3	配付資料がわかりやすかった	わかりやすい配付資料	配付資料
4	自分の発表に対する他の学生の意見が聞けてよかった。	発表による意見交換	発表方法
5	身近な話題を取り上げてくれて興味が持てた	身近な話題による興味喚起	話題
6	文字が大きくて板書が見やすかった	板書が見やすい	板書
7	いろんな学生に発表の機会を与えているのが良かった	発表機会の振り分け	発表方法

ワーク:コーディングしてみよう

どのような観点で授業が評価されているでしょうか。

「見出し」「コード」を作成し、指摘されたポイントをまとめてみましょう。

◆配付したエクセルの「ワーク2_コーディング(作業用)」を利用してください。

NO.	回答（自由記述）	見出し	コード
1	先生が優しく、発表しやすい雰囲気づくりをしてくれていたと思う。		
2	授業外でも授業で使用された授業スライドを見られたので、復習しやすかった。		
3	発表ポイントが貰えるのがやる気でした。		
4	課題の量が比較的になく、マイペースに勉強できた。		
5	フィードバックが丁寧だった		

テキストマイニング



テキストマイニング

- 自由記述回答結果(文章)を**単語で区切り(=形態素解析)**、出現頻度や各単語間の相関などを**統計的に解析**することで、有用な情報を取り出す手法。
- **テキストマイニングでできること**
 - ✓ 大量の文章を統計的に解析
 - ✓ 出現頻度が高い語句を抽出できる
 - ✓ 語と語の結びつきを探れる(係り受け、共起関係)
 - ✓ 文中の特定の語彙について分析できる

形態素解析

- 形態素解析とは、自然言語処理(NLP)の一部で、自然言語で書かれた文を言語上で意味を持つ最小単位(=形態素)に分け、それぞれの品詞や変化などを判別すること。

https://ledge.ai/morpho_analysis_japan/

「庭には二羽ニワトリがいる」

庭(名詞)/に(助詞)/は(助詞)/二(数詞)/羽(助数詞)/ニワトリ(名詞)/が(助詞)/いる(動詞)

テキストマイニングの長所・短所



長所

コンピューターや統計解析を使用することで、分析の客観性と実用性が確保しやすくなる。



短所

コンピューターや統計解析を使用することで、少数派の重要な意見を見逃したり、テキストの内容を誤って解釈する可能性がある。

テキストの内容の誤った解釈の例

Q.本ワークショップに参加して「役に立った」と思う点について自由に記述してください。

(A)「テキストマイニングを体験したことで一通りできるようになったので役に立った。」

(B)「自分の業務には使えないことがわかった。独学で着手する前に気づけて、役に立った。」

◆原文に立ち戻り、文脈を確認することが重要。

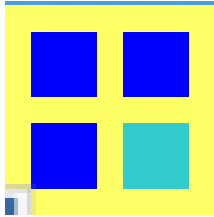
◆可能な限り文脈を指定した設問にする。

(例)

×： この授業に関する感想を自由に記述してください。

○： この授業に関する良かった点を自由に記述してください。
この授業に関する改善点を自由に記述してください。

テキストマイニングソフト



KH Coder

- 立命館大学の樋口耕一先生が開発したテキスト型（文章型）データを統計的に分析するための無料のソフトウェア。

- Windows環境推奨（Win7以上）

- 公式HP

<https://khcoder.net/>



ワーク:KH Coderを使ってみよう

- 出現頻度表
- 階層的クラスタ分析
- 共起ネットワーク図
- (参考資料)コーディング→クロス集計

<使用するデータ>

(質問文) SIH道場のプログラムについて、大学での学修を始める上で役に立った点があればお書き下さい。

→総合科学部・医学部・理工学部
の自由記述回答

(それぞれ93件・162件・395件、計650件)

調査期間:平成30年4月～10月

出現頻度表

抽出語	出現回数
レポート	145
書き方	123
分かる	66
プレゼン	43
協力	43
研究	42
学べる	38
学ぶ	35
人	35
身	34
自分	32
大学	32
良い	31
プレゼンテーション	30
仕方	30
役に立つ	28
理解	28
交流	26
グループ	24
コミュニケーション	24
知る	24
他者	23
学部	22
発表	22
機会	21

抽出語	出現回数
能力	21
知識	20
学科	19
友達	19
思う	16
出来る	16
大切	16
他	15
意見	14
協調	14
考える	14
力	14
やり方	13
学習	13
医療	12
興味	12
書く	12
生活	12
体験	12
専門	11
勉強	11
方法	11
履修	11
基本	10
重要	10

抽出語	出現回数
将来	10
積極	10
知れる	10
内容	10
役立つ	10
考え	9
実際	9
分野	9
訪問	9
話す	9
チーム	8
ディベート	8
ワーク	8
基礎	8
作成	8
授業	8
説明	8
先生	8
調べる	8
普段	8
引用	7
学修	7
活動	7
関係	7
見れる	7

抽出語	出現回数
行う	7
深める	7
前	7
仲間	7
様々	7
パワー	6
ポイント	6
学生	6
向上	6
増える	6
注意	6
登録	6
難しい	6
友好	6
友人	6
話し合い	6
KJ	5
SIH	5
課題	5
感じる	5
教える	5
経験	5
見る	5
見学	5
言う	5

抽出語	出現回数
講座	5
参考	5
使い方	5
姿勢	5
資料	5
持つ	5
取り組む	5
少し	5
色々	5
触れる	5
深まる	5
多い	5
大事	5
非常	5
明確	5
練習	5
話	5
チームワーク	4
現場	4
交換	4
広がる	4
行く	4
高まる	4
今	4
今後	4

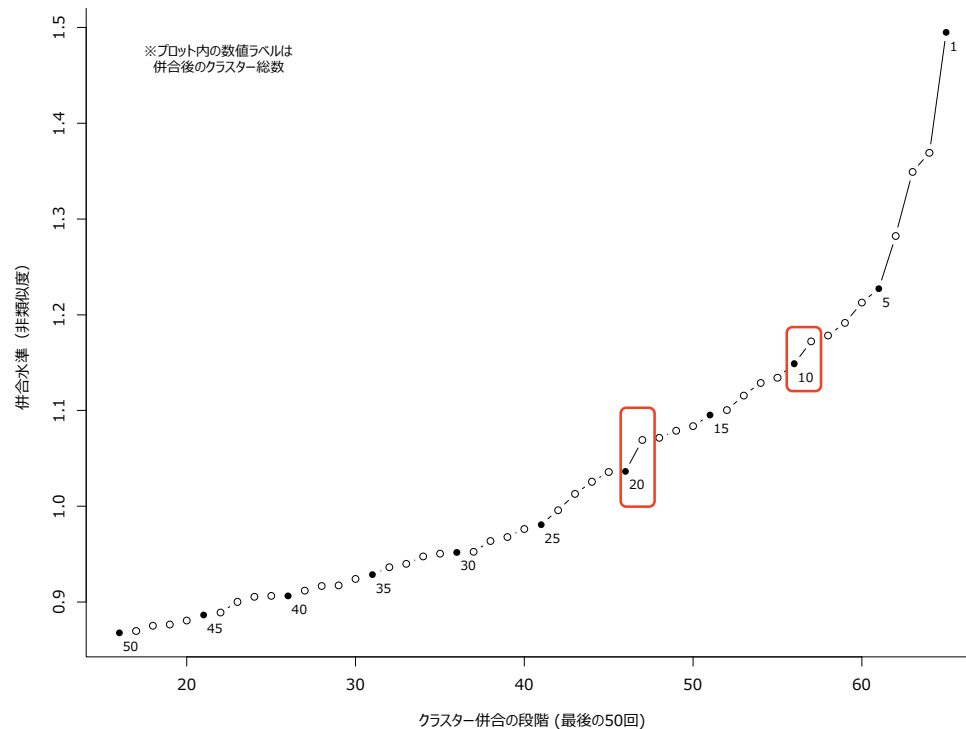
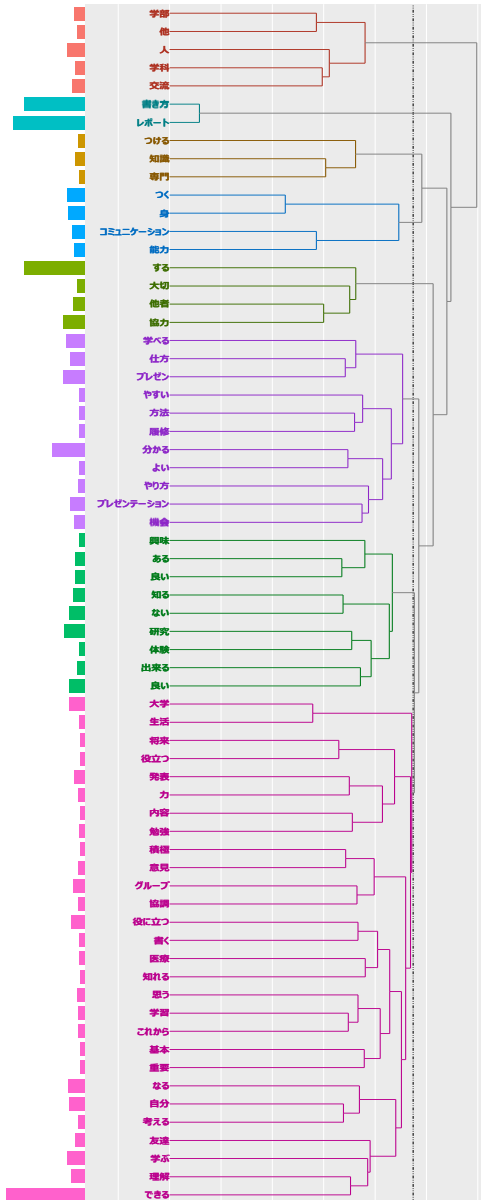
抽出語	出現回数
作る	4
始める	4
視点	4
視野	4
持てる	4
質問	4
取る	4
初めて	4
設ける	4
相手	4
他人	4
大学生	4
働	4
必要	4
文章	4
ある程度	3
お話	3
ダム	3
ブレインストーミング	3
モチベーション	3
ワークショップ	3
意識	3
意欲	3
違う	3
改善	3

- 読み込んだテキストデータにみられた単語の出現回数一覧。品詞別・上位150語・一覧から選択。

→テキストマイニングを行う際の基礎資料となる。データの概観ができる上、表記の揺れ修正作業などのデータクリーニングやコーディングを目的とした単語の振り分けを行う際にも利用。

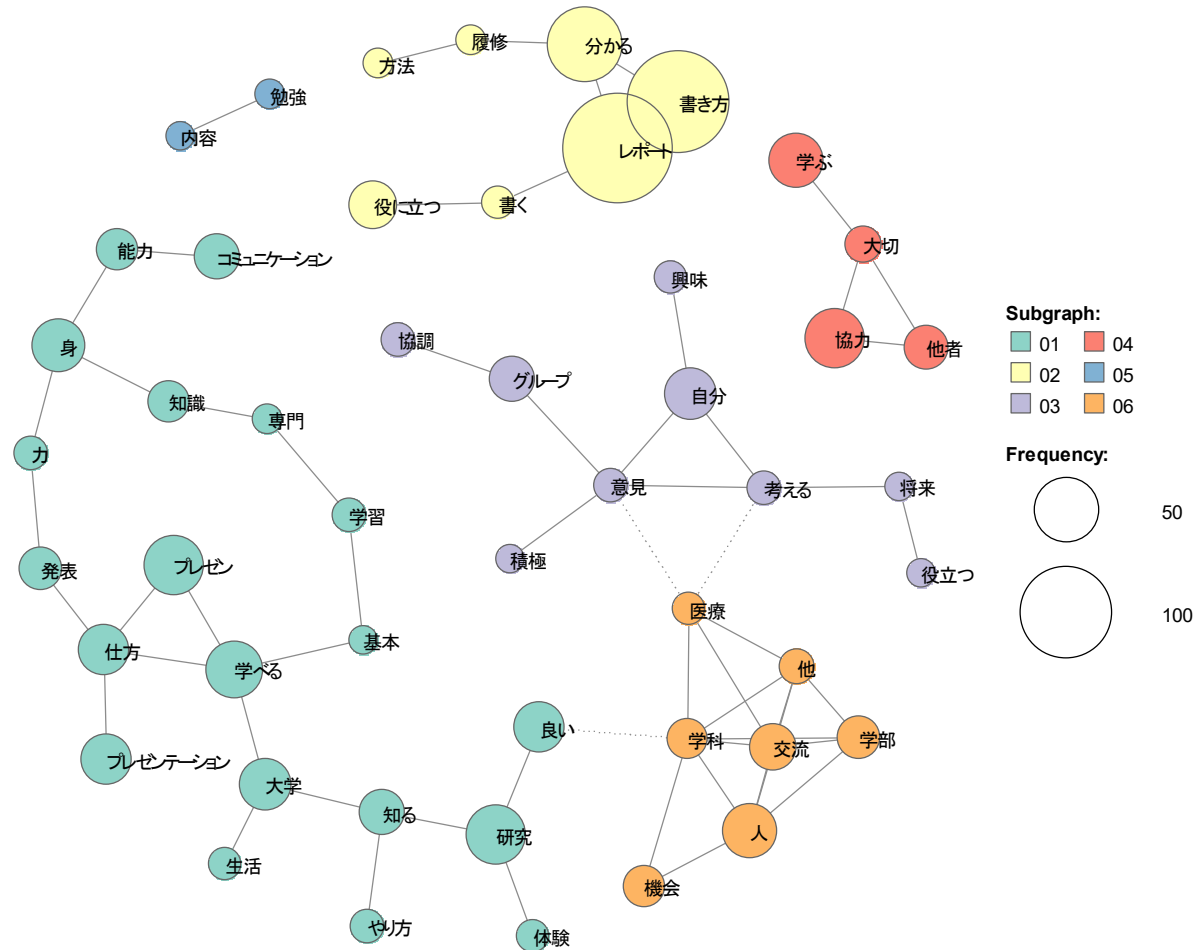
- 外部変数別に集計すれば、それぞれの変数に紐づくデータの特徴も明らかになる。

階層的クラスタ分析



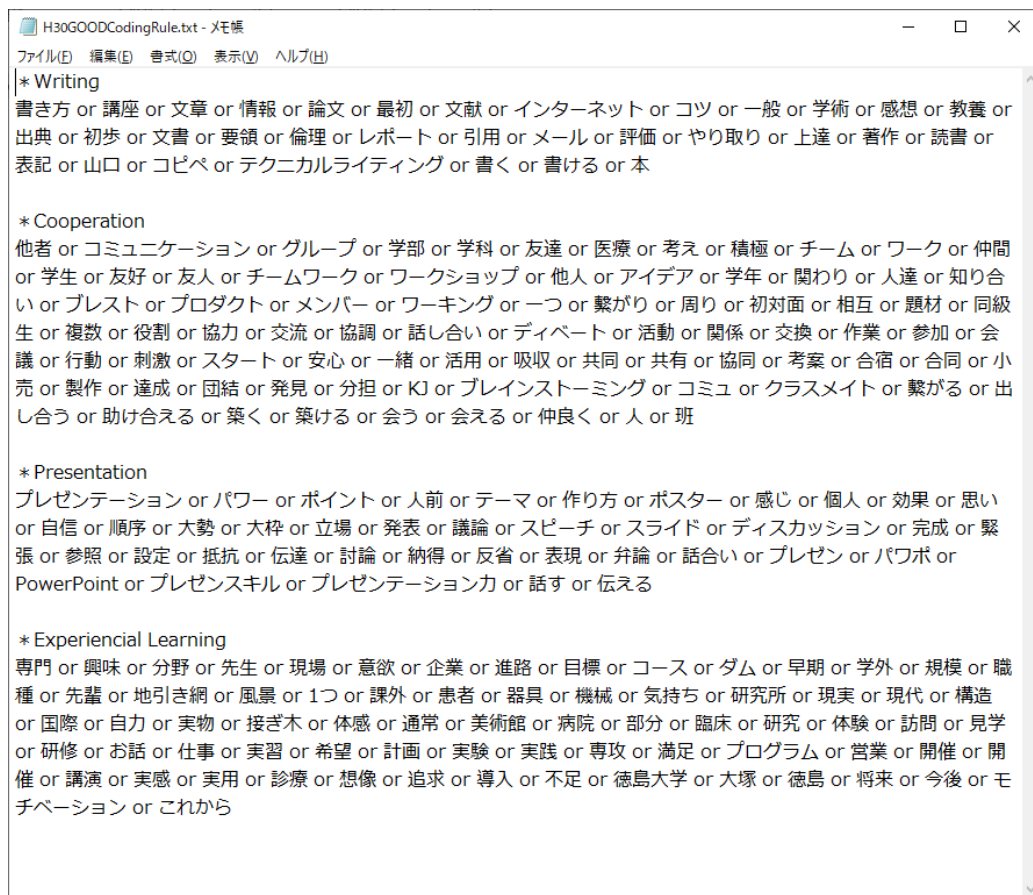
- 数量的裏付けに基づき類似性の高い語同士をグループ化 (クラスタ化) することができる。
- 図の左側で結合する語ほど似ている関係にあると言える。
- 点線の位置でクラスターの数設定され、クラスター数は分析者が自由に調整することができる (併合水準のグラフを参考に決めると良い)。

共起ネットワーク図



- 円が大きいほど出現回数が多い。語と語が線で結ばれているかどうか共起性や関連性の有無を表し、線の太さで関連の強さを表現。
- 語の出現パターンを機械的に分類することで、テキストデータ全体の中に隠された主要なメッセージを抽出することができる。

(参考資料) コーディング→クロス集計

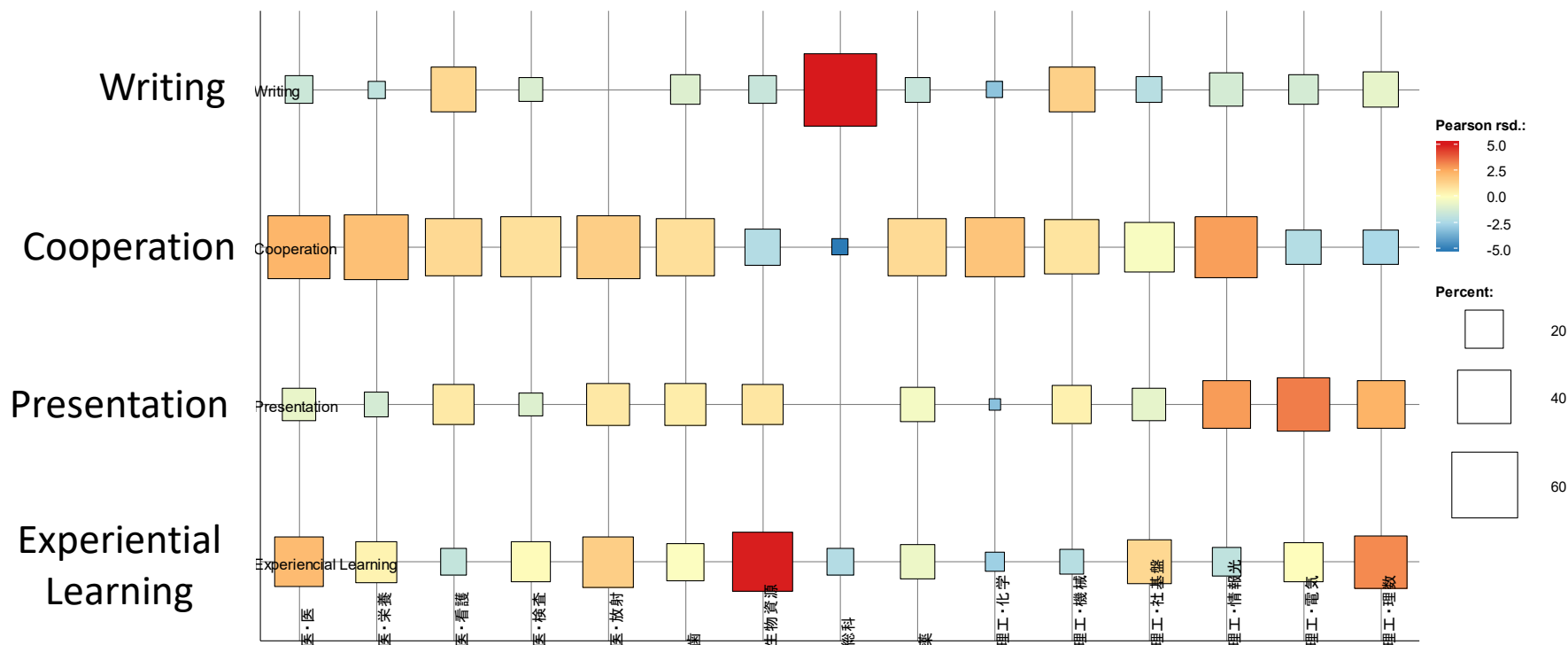


1. 抽出語リストを作成

2. KWICコンコーダンス
を利用して単語が使わ
れる文脈を確認し、各
コードに分類(左図、手
作業)

3. 出力されるコード×外
部変数のクロス集計結
果をもとに、バブルプ
ロットを作成(次スライ
ド)

(参考資料) コーディング→クロス集計



全学共通で実施する初年次教育プログラムがどういう点で役に立ったかを受講生が記述したものをプログラムに盛り込む共通項目×学部学科で分析。プログラムの強みや弱みを読み取る。

テキストマイニングのポイント

分析前～分析時の準備作業が重要

- 固有名詞や人名の辞書登録(強制抽出語の設定)
例:英会話→英/会話、Apple Watch→Apple/Watch
- 分析対象外にする語の登録
- 表記の揺れの統一
例:ユーザー/1-ザー、わかる/分かる、(株)/株式会社
- 共起ネットワークで表示する単語数
- クラスターの数

分析結果の解釈と原文参照を行って、
準備作業への立ち戻りを繰り返す

テキストマイニングを用いた研究例

① 既存の(別の目的で書かれた)文章の分析

手紙、新聞、小説、TwitterやFacebook等のSNS、歌詞など

② アンケートを用いて自ら収集した文章を分析

意識調査、満足度調査などの自由記述

テキストマイニングを用いた研究例

①既存の(別の目的で書かれた)文章の分析

- 大瀧友織(2003)「夫に対する役割期待の変化:妻からの不満をとおして」
- 石井健一(2011)「FacebookとTwitterの発言における特徴語の比較」
- 陣内正敬(2011)「高校野球・選手宣誓の時代性」
- 金城克哉(2013)「槇原敬之の歌詞の数量的分析:『君が笑うとき君の胸が痛まないように』から『Heart to Heart』まで」

テキストマイニングを用いた研究例

②アンケートを用いて自ら収集した文章を分析

- 川端亮・樋口耕一(2003)「インターネットに対する人々の意識:自由回答の分析から」
- 木村清・水田恵三(2008)「学生の大学生活、授業に対する満足・不満点に関する自由記述の分析」
- 原美穂子・藤森和美(2013)「自然災害時に支援者として心のケアに携わる学校教員のニーズに関する研究:養護教諭へのアンケート調査から」
- 熊本市都市政策研究所(2020)「新型コロナウイルス感染症に対する3月上旬時点での熊本市民の声の緊急分析」

https://www.city.kumamoto.jp/hpkiji/pub/detail.aspx?c_id=5&type=top&id=28027

質的データ分析の注意点

テキストデータの管理

- Excelを利用して自由記述の一覧を作成する
- 記述に対する属性情報(回答者の性別、年齢等)を付加

→自由記述の分析の幅が広がる。

* 2種類以上の要素をもつ変数を追加したい場合は、専用の列を追加し、「=A2 & B2」のように&で結合→値で貼り付け。

- LEN関数で文字数をカウント

→ 平均的な記述量の計算や分量の基準を設けることが可能

* SUBSTITUTE関数と組み合わせて特定の語の出現頻度をカウントすることも可能

ID	Q1:この授業全体を通して学んだことや、考えや行動	Q2:この授業を良くするために改善すべきことがあれ	性別	学年	所属	性別*学年
1	授業だけでなく日常生活においても見直すことや立	対面に来ている人が固定化しているので対面からオ	男		1 理工学部	1男
2	以前よりも勉強面やそれ以外の様々な場面で振り返	前期授業のやり方で非常によかったと思います。	男		2 理工学部	2男
3	お金を払ってわざわざ大学に来ているのだから、	受講生同士の交流の機会が欲しい。どんな学生が	男		1 総合科学部	1男
4	誠実さを得るようになった	オンライン受講生と対面受講生の交流がもっとあれ	男		1 医学部	1男
5	最初は予習も復習もなぜこんなにしなければならな	課題をもっと簡潔にしてほしい。振り返りは大事だけ	女		2 医学部	2女
6	自分の意見を発表することにハードルを感じにくくな	発表したときのポイントをもっと増やしてほしい。	男		2 理工学部	2男
7	振り返りを行うことがあったが、客観的な視点を付け	対面で受けている人の声が聞き取りにくい時があっ	女		1 総合科学部	1女
8	一般教養が単位だけのためではないことが分かった	ネットの環境が不安定な時があった。声や映像がき	男		1 総合科学部	1男

質的データに対する認識

- 量的データと質的データに対する印象、重視するポイントが人によって異なる。

質的研究には、「信頼性が『求められていない』のである」…量的研究とは違って質的研究では「真実は、文脈や研究者の立場によって多様にあり、多様な解釈で成立」すると考えるため…(大谷2019)

- 妥当性があることを論理的に示す必要がある。

量的データは統計的手法を用いることそのもので妥当性が示せることがある。

- 1件のデータも無視できないことがある。

質的データ分析の注意点

- 都合の良い記述だけを切り取っていないか
- コードやキーワードが内容とあっているか
- 複数の要因間の関係を適切に示せているか
- 記述間の関係を論理的に説明できるか
- 文を簡潔にまとめられる余地はないか
- 文脈を考慮できているか

本日の到達目標

A

到達した！

B

まあまあかな

C

イマイチかな

- ① 質的データや質的研究の特徴を把握する
- ② 質的データの分析の流れを理解する
- ③ KH Coderの使い方を学び、今後自身の教育/研究活動において活用するヒントを得る

参考文献

- Udo Kuckartz著、佐藤郁哉訳(2018)『質的テキスト分析法:基本原理・分析技法・ソフトウェア』新曜社
- 石井健一(2011)「FacebookとTwitterの発言における特徴語の比較」筑波大学
- 牛澤賢二(2018)『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
- 大木俊英(2017)「テキストマイニングを使った自由記述式アンケートの分析」JLTA 2017 Workshop (JLTA Web Publication)
- 大瀧友織(2003)「夫に対する役割期待の変化:妻からの不満をとおして」家族研究年報(28)pp.51-62、家族問題研究会
- 大谷尚(2019)「質的研究の考え方ー研究方法論からSCATによる分析までー」名古屋大学出版会、参考リンク<http://www.educa.nagoya-u.ac.jp/~otani/scat/>
- 金城克哉(2013)「槇原敬之の歌詞の数量的分析:『君が笑うとき君の胸が痛まないように』から『Heart to Heart』まで」琉球大学欧米文化論集(57)pp.23-42、琉球大学法文学部国際言語文化学科(欧米系)
- 川端亮・樋口耕一(2003)「インターネットに対する人々の意識:自由回答の分析から」『大阪大学大学院人間科学研究科紀要』(29)
- 木村清・水田恵三(2008)「学生の大学生活、授業に対する満足・不満点に関する自由記述の分析」尚絅学院大学紀要(56)pp.227-235、尚絅学院大学
- 熊本市都市政策研究所(2020)「新型コロナウイルス感染症に対する3月上旬時点での熊本市民の声の緊急分析」https://www.city.kumamoto.jp/hpkiji/pub/detail.aspx?c_id=5&type=top&id=280」
27(2020年9月20日アクセス)

参考文献

- ・ 陣内正敬(2011)「高校野球・選手宣誓の時代性」九州大学言語学論集(32)pp.267-280、九州大学大学院人文科学研究院言語学研究室
- ・ 末吉美喜(2019)『テキストマイニング入門:ExcelとKH Coderでわかるデータ分析』オーム社
- ・ 竹内理・水本篤(2012)『外国語教育研究ハンドブック:研究手法のより良い理解のために』松柏社
- ・ 竹中喜一(2019)「質的データを分析する」IRer養成講座in愛媛(愛媛大学教育・学生支援機構教育企画室主催2019)
- ・ 中嶋洋(2015)「初学者のための質的研究26の教え」医学書院
- ・ 中島英博(2018)「インタビューで集めたデータをまとめる」『名古屋大学生のためのアカデミック・スキルズ・ガイド』<http://www.cshe.nagoya-u.ac.jp/asg/summarizedatagatheredinterview.html>(2020年9月20日アクセス)
- ・ 原美穂子・藤森和美(2013)「自然災害時に支援者として心のケアに携わる学校教員のニーズに関する研究:養護教諭へのアンケート調査から」『武蔵野大学心理臨床センター紀要』(13)pp.19-30、武蔵野大学
- ・ https://www.wordminer.org/wp-content/uploads/2013/04/63_20.pdf
- ・ 樋口耕一(2020)『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』(第二版)ナカニシヤ出版
- ・ 藤井美和・小杉考司・李 政元(2005)『福祉・心理・看護のテキストマイニング入門』中央法規出版

ご清聴
ありがとうございました。