心理データ解析演習：第5回
テキストマイニング入門

教育認知心理学講座M1
岡 隆之介

今日みなさんに持ちかえってほしいもの

- テキストマイニングは心理学でも有効
- テキストデータは数字で処理されている
- テキストデータの分析は（ぱっと見）簡単である

発表アウトライン

1. テキストマイニングとは何か
2. テキストマイニングの基礎理論
3. テキストマイニングの分析―クラスター分析編―
4. テキストマイニング演習―KH-Coderを使ってやってみる―

1. テキストマイニングとは何か

1.1 テキストマイニングとは

- テキストマイニングとは、テキストを単語やフレーズなどの単位に分割し、それらの出現頻度や共起関係(同時出現)などを集計し、データ解析やデータマイニングの手法で定量的に解析すること。
- なんらかの意味のある文章(テキスト)を用いて、それを計量的にあつかう心があれば、テキストマイニングになりうる！

1.1 テキストマイニングとは

- テキストマイニングの強み
  - 質的データを数値にコーディングすることで計量的な分析を行うことができるようになる。ゆえに、客観性が高い
  - 社会調査やインターネット上のデータなど、人間の手作業で分類・カウントする作業に強い。したがって、アルゴリズムさえ与えてしまえば計算機が処理してくれる。大規模データも自由自在。

1.2 テキストマイニングの歴史

- 皮切りは計量文体学だと言われている
  - Thomas Corwin Mendenhall(1887)が光学におけるスペクトル分析方法を単語に適用し、単語のスペクトル(単語の長さの分布)によって著者の文体を予測した論文を『サイエンス』に発表したという。
    - 書き手による文体の好みを定量的に分析した
    - もちろん、この時代に個人用計算機などなく、すべて手作業であったという

1.2 テキストマイニングの歴史

- 1950年代には日本にも計量文体学が取り入れられた。
  - 安本(1958,1974)は、「源氏物語」と「宇治十帖」の著者の文体について、心理描写の数、文の長さ、直喩、色彩語、助詞、助動詞などを含む12項目による、心理文章学の視点による作品の比較検討を行ったという。

1.3 心理学における価値・研究の実例

- Kusumi, Matsuda, & Sugimori(2010)
  - 大学生451名を対象に、ノスタルジアを感じる光景・出来事・曲を自由記述してもらった。その後、記述された特徴語を用いて階層的クラスター分析を実施。それぞれに対して共通する概念を探索した。
- 玉利・竹村(2012)
  - マクドナルドとモスバーガーのブランドの好きなところと嫌いなところを自由記述してもらい、自由記述データをもとに潜在的意味解析(LSA)を使用。消費者の背後にある決定フレームを探索した。
- 記述されたデータであれば、なんでもできそう

2. テキストマイニングの基礎理論

2.1 テキストマイニングの基礎概念

1. テキストの電子化
   - まずはなんらかのテキスト(電子掲示板の書き込み、質問紙の自由記述etc)を用意して、txtファイルにする必要がある(PCに読み込んでもらうため)。
2. テキストデータのクリーニング
   - txtファイルをソフトウェアが読み込める形にコード・ソートしてあげる必要がある。

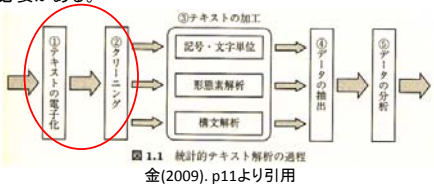


図 1.1 統計的テキスト解析の過程

金(2009). p11より引用

2.1 テキストマイニングの基礎概念

3. テキストの加工
   - ソフトウェアが活躍するところ。データを何らかの基準で「意味の最小単位」にしてあげる必要がある(「質的データ」→「量的データ」の過程)。
   - 形態素解析が特に重要

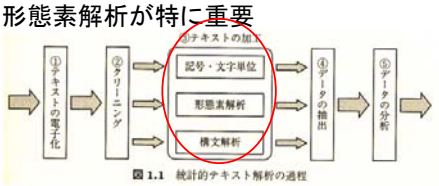


図 1.1 統計的テキスト解析の過程

金(2009). p11より引用

2.1 テキストマイニングの基礎概念

c.f. 形態素解析とは

- 対象言語の文法の知識（文法のルールの集まり）や辞書（品詞等の情報付きの単語リスト）を情報源として用い、自然言語で書かれた文を形態素（*Morpheme*, おおまかにいえば、言語で意味を持つ最小単位）の列に分割し、それぞれの品詞を判別する作業を指す(from wikipedia)。
- テキストマイニングが、さも人間が文章を分類したみたいにふるまうのは形態素解析の恩恵が大きい。

2.1 テキストマイニングの基礎概念

c.f. 形態素解析とは

- 具体例：「白砂がボールを蹴った」

表1.「白砂がボールを蹴った」の形態素解析結果

表層語	基本形	品詞	活用	
白砂	白砂	名詞−一般		
が	が	助詞−格助詞−一般		
ボール	ボール	名詞−一般		
を	を	助詞−格助詞−一般		
蹴っ	蹴る	動詞−自立	五段・ラ行	連用タ接続
た	た	助動詞	特殊・タ	基本形

2.1 テキストマイニングの基礎概念

- c.f. 構文解析とは
  - 統語論で使われる単語。文章を構文木(Syntactic Tree)にして、その文章がどのような文構造を持っているかを明らかにする。
  - 今回は特に使いません。より詳細に知りたい人は後述の参考書などで勉強してください。

2.1 テキストマイニングの基礎概念

4. データの抽出
   - テキストデータの何に関心があるのかによって変わる。一般的なテキストマイニングでは語の共起頻度や、特定の単語の後にどのような単語が出てくるかなどのデータが有益。

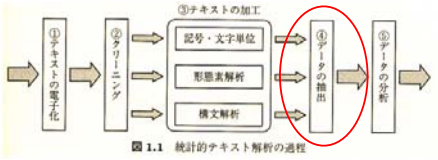


図 1.1 統計的テキスト解析の過程

金(2009). p11より引用

2.1 テキストマイニングの基礎概念

5. データの分析
   - 抽出したデータから何が言いたいのかに関わる。SPSSなどを用いた記述統計量の確認から、後述のような少し高度な分析まで。

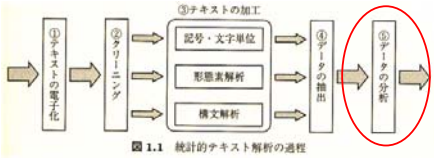


図 1.1 統計的テキスト解析の過程

金(2009). p11より引用

2.2 テキストマイニングの分析の種類

- ざっくりとした説明としては、テキストマイニングの種類は「どのデータを抽出し」「それをどう分析するか」によって分けられる。
  - クラスター分析
  - ネットワーク分析
  - 主成分分析
  - 対応分析
  - 潜在的意味解析 etc…
- 個人的な感想としては、テキストを扱った分析さえしてればテキストマイニングになりうるので、上記の例はあくまでよく使われる方法という認識。

3.テキストマイニングの分析
—クラスター分析編ー

3.1 クラスター分析とは

- 対象となるデータ群のどれとどれが類似しているかを見つけ出すために用いられるさまざまな数学的方法の総称。

クラスター分析の種類

- 大別すると2種類
  - 非階層的クラスター分析
  - 階層的クラスター分析
- 今回は、よりメジャーな階層的クラスター分析を扱います。

3.2 階層的クラスター分析の方法

- 階層的クラスター分析の手順
  1. すべてのクラスターの組(初めは要素)に対して、クラスター間の距離(非類似性)を求める
  2. クラスター間の距離(非類似性)を参照してクラスター間距離が最小のクラスターの組を結合し、新たなクラスターを作成する
  3. 新たなクラスターとその他のクラスター間の距離(更新距離)を求める
  4. クラスター数があらかじめ決められた数(通常は1)になるまで、2・3を繰り返す

コラム:階層的クラスター分析と非階層的クラスター分析の違い

階層的クラスター分析と非階層的クラスター分析の特徴

階層的クラスター分析	全対象の類似度(又は非類似度)を計算し、最も類似度の高いものから順次グルーピングする方法。最終的に1つのクラスターになるまで繰り返す。デンドログラムと呼ばれる樹形図で表現され、結びつきの階層構造を明確にできる。
非階層的クラスター分析	分割の数などを与えて全体を一気に分割する方法。最適な分割数を決めるための仮定などを予め決める必要がある。結びつきの階層構造は分からない。

3.3 テキストマイニングにおけるクラスター分析の方法(重要)

- 通常のクラスター分析では、クラスター間の距離(非類似性)を算出して、クラスターを形成していく。
  - e.g. Aさん(快4, 幸福感5, 不満感1)
- しかし、テキストマイニングで用いるデータは文字データ
  - e.g. 白砂がボールを蹴った
- 文字データをどのように数値データとしてコードすればよいのか?

3.4文字データのコーディング:2値データで考える

- じゃあ、それぞれの文章が持っている情報を、形態素解析の結果をもとに、2値データ、つまり、「ある文章に特定の単語が含まれているか(1)、いないか(0)」をデータにしたらどうか、と考える。
- 次表で説明する。

3.4文字データのコーディング：2値データで考える

表3. 文字データのコードの具体例

文章	単語 白砂(n1)	岡(n2)	市村(n3)	ボール(n4)	バット(n5)	蹴る(v1)	打つ(v2)	ベクトル表記
白砂がボールを蹴った(s1)	1	0	0	1	0	1	0	s1=(1,0,0,1,0,1,0)
岡がボールを蹴った(s2)	0	1	0	1	0	1	0	s2=(0,1,0,1,0,1,0)
市村がバットでボールを打った(s3)	0	0	1	1	1	0	1	s3=(0,0,1,1,1,0,1)
白砂が岡を蹴った(s4)	1	1	0	0	0	1	0	s4=(1,1,0,0,0,1,0)
岡が白砂を蹴った(s5)	1	1	0	0	0	1	0	s5=(1,1,0,0,0,1,0)

※sn=(n1,n2,n3,n4,n5,v1,v2)

- s4とs5に注目。この2文のベクトル表記は同じ。つまり、計算機上ではこの2文は区別されていない

3.5テキストマイニングにおける数値データ

- つまり、テキストマイニングにおいてある文章(sn)は、全文中に含まれるすべての単語を要素(次元)とするベクトルとして表現できる。

表4. 計算機上の文章の表現

文章(sn)	単語(n,v,a,pv etc) n1	n2	n3	n4	n5	v1	v2
s1	1	0	0	1	0	1	0
s2	0	1	0	1	0	1	0
s3	0	0	1	1	1	0	1
s4	1	1	0	0	0	1	0
s5	1	1	0	0	0	1	0

3.6 テキストにおけるクラスタリング

- 話を戻して、クラスター分析では要素間の「非類似性」をもとにクラスタリングを行うことが分かっている。
- 文章を数値データにする方法はわかった。じゃあ、これをもとにどうやって「非類似性」を比較し、クラスタリングを行うのか。

→非類似性を知る必要がある。

3.7テキストにおける類似性 2値データの場合：Jaccard係数

- 集計したデータが2値データの場合や、間隔尺度のデータである場合は、それにあった非類似性の指標を用いる必要がある。
- Jaccard係数は上記のようなデータを扱う際の「類似性」の指標。

3.8 Jaccard係数の定義

- 集合XとYの共通要素数を少なくとも1方にある要素の総数で割ったもの
- 今，X∪Yの要素をz1,z2,..,znとして，ベクトルx=(x1,x2,..,xn)を，xi=1 (if zi∈X), xi=0 (otherwise)として定める． ベクトルyも同様に定めると，Jaccard係数は下の式で定められる。

$$\text{Jaccard 係数} = \frac{|x \cap y|}{|x \cup y|} = \frac{x \cdot y}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - x \cdot y}$$

3.8 Jaccard係数の定義

- さっきの「文章×単語」行列を「単語×文章」行列に転置すると、以下のような行列になる。
- この行列をもとに、Jaccard係数を算出

表5. 各単語の1文章あたりの出現(したか・しないか)

単語	文章(sn) 白砂がボールを蹴った(s1)	岡がボールを蹴った(s2)	市村がバットでボールを打った(s3)	白砂が岡を蹴った(s4)	岡が白砂を蹴った(s5)
白砂(n1)	1	0	0	1	1
岡(n2)	0	1	0	1	1
市村(n3)	0	0	1	0	0
ボール(n4)	1	1	1	0	0
バット(n5)	0	0	1	0	0
蹴る(v1)	1	1	0	1	1
打つ(v2)	0	0	1	0	0

3.8 Jaccard係数の定義

- 下がその単語行列のJaccard係数

表6. 各単語のJaccard係数(類似度)

	n1	n2	n3	n4	n5	v1	v2
白砂(n1)	–						
岡(n2)	0.50	–					
市村(n3)	0.00	0.00	–				
ボール(n4)	0.20	0.20	0.33	–			
バット(n5)	0.00	0.00	1.00	0.33	–		
蹴る(v1)	0.75	0.75	0.00	0.40	0.00	–	
打つ(v2)	0.00	0.00	1.00	0.33	1.00	0.00	–

- 値が1に近いほど、それぞれの単語の類似度が高い

3.9 Jaccard距離の定義

- Jaccard係数はあくまで2つの単語の類似度を測るもの
- 分析で用いる非類似性はJaccard距離で定まる

$$\text{Jaccard 距離} = 1 - (\text{Jaccard 係数})$$

3.9 Jaccard距離の定義

- さっきのデータのJaccard距離は下のようになる

表7. 各単語のJaccard距離(非類似性)

	n1	n2	n3	n4	n5	v1	v2
白砂(n1)	–						
岡(n2)	0.50	–					
市村(n3)	1.00	1.00	–				
ボール(n4)	0.80	0.80	0.67	–			
バット(n5)	1.00	1.00	0.00	0.67	–		
蹴る(v1)	0.25	0.25	1.00	0.60	1.00	–	
打つ(v2)	1.00	1.00	0.00	0.67	0.00	1.00	–

コラム:文章のクラスター分析と特徴語のクラスター分析

- 特徴語のクラスター分析(今回)
  - 「ある単語をもとに、その単語が他の単語とどのように共起しているかをそれぞれの文章を参照して調べ、文章内での共起の頻度が高い順に単語のクラスターを形成する」
- 文章のクラスター分析
  - 「ある文章をもとに、その文章が含む単語がどのように共起しているかを調べ、他の似たような単語の共起を示す文章とクラスターを形成する」

3.10 クラスターの形成

- それぞれのデータ間の距離はJaccard距離を利用して表現できた。いよいよクラスタリングがしたいが、どうするか。
- 比較的よく使われているクラスタリング法としてWard法がある。

3.11 Ward法

- 分散の情報を用いる。データをグループ分けしたとき、全体の分散は、グループ内の分散とグループ間の分散の合計に等しい。偏差の2乗の和を用いても同じことがいえる。
- 全体の偏差の2乗和をT、グループ内の偏差の2乗和をW、グループ間の偏差の2乗和をBで示すと次の式が成り立つ。

$$T = W + B$$

- ウォード法では、グループ内の分散が小さく、かつグループ間の分散が大きい組み合わせでグループ分けする。

3.11 Ward法

- 「なんのこっちゃ」という感じでしょうから、もう少し説明します(細かい部分が聞きたい人は僕と議論しましょう)。
- 下がさっきのJaccard距離。距離が小さいものをグルーピングする

	n1	n2	n3	n4	n5	v1	v2
白砂(n1)	–						
岡(n2)	0.50	–					
市村(n3)	1.00	1.00	–				
ボール(n4)	0.80	0.80	0.67	–			
バット(n5)	1.00	1.00	0.00	0.67	–		
蹴る(v1)	0.25	0.25	1.00	0.60	1.00	–	
打つ(v2)	1.00	1.00	0.00	0.67	0.00	1.00	–

3.11 Ward法

- グルーピング後の2値データは下表のようになる。

表8. c1形成後の各単語の1文章あたりの出現したかしないか)

単語	文章				
	白砂がボールを蹴った(s1)	岡がボールを蹴った(s2)	市村がバットでボールを打った(s3)	白砂が岡を蹴った(s4)	岡が白砂を蹴った(s5)
白砂(n1)	1	0	0	1	1
岡(n2)	0	1	0	1	1
市村・バット・打つ(c1)	0	0	1	0	0
ボール(n4)	1	1	1	0	0
蹴る(v1)	1	1	0	1	1

3.11 Ward法

- さっきと同様に、Jaccard係数を算出し、Jaccard距離を求めると下表のようになる。

表9. c1形成後のJaccard距離

	白砂(n1)	岡(n2)	バット・打つ・ボール(n4)	蹴る(v1)	
白砂(n1)	–				
岡(n2)	0.50	–			
市村・バット・打つ(c1)	1.00	1.00	–		
ボール(n4)	0.80	0.80	0.67	–	
蹴る(v1)	0.25	0.25	1.00	0.60	–

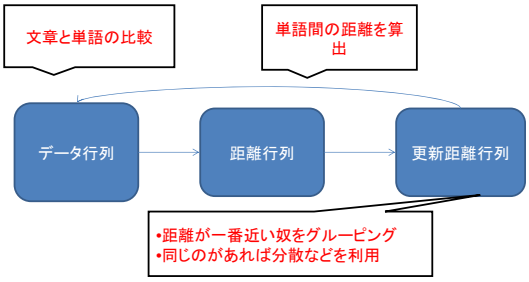
- 次はここらへんでグルーピングができそう。でもどっちだろう？→Ward法の出番

3.11 Ward法

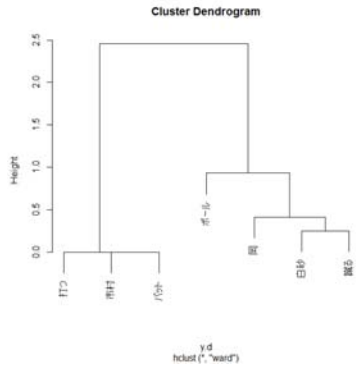
- 「白砂・蹴る」のグループを組んだときと、「岡・蹴る」のグループを組んだときで、グループ内分散が小さく、グループ間分散が大きくなるほうのクラスターを採用する
- （計算すると今回のデータの場合、どちらでも同じ結果になります・・・予想外）
- クラスターの分類の基本は「どのクラスターが近いか」できめる。もし、クラスター間の距離が等しいときにWard法を使うというイメージでOK。

3.12 分析方法の確認と分析結果

- こんな感じのが階層的クラスター分析
- フローで書くとこんな感じ



3.12 分析方法の確認と分析結果



講義のまとめ

- テキストマイニングは心理学でも有効
- テキストデータは数字で処理されている
  - 文字データは数値データとして処理されている
- テキストデータの分析は（ぱっと見）簡単である
  - 今回紹介した方法はほんの一例。単語の抽出の仕方も頻度をとる方法もあるし、クラスターの分類もユークリッド距離を求めたり…いろいろあります。

4.テキストマイニング演習
ーKH-Coderを使ってやってみるー

KH-Coderとは

- 立命館大学の樋口耕一准教授が開発した、計量テキスト分析をグラフィカルユーザーインターフェイスで行えるフリーソフトウェア
- R(統計解析ソフト)・chasen(形態素解析ソフト)・mysql（フリーのデータベース検索ソフト）を用いて、各種計量テキスト分析を可能にしている

KH-Coderの利点

- 計算式で表現することができなくても、簡単に多くのテキストマイニングを行うことができる。
- できる分析の種類
  - 抽出語検索
  - 階層的クラスター分析
  - 共起ネットワーク分析
  - 多次元尺度構成法
  - 関連語分析
  - 対応分析　などなど

今回の演習

- **今回の演習はKH-Coderダウンロード時についてくるサンプルデータを使って、実際にテキストマイニングを行ってみます。**
- **サンプルデータは夏目漱石の『こころ』**
- **こころの登場人物やその特徴がテキストマイニングによって少しでも読みとれたら嬉しいです**

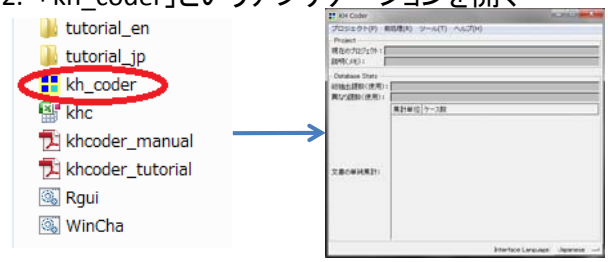
※今回やる程度の分析に関しては、KH-Coderのチュートリアルにも載っている基本的なものです！アドバンスなことがやりたい人は僕と一緒にRを勉強しましょう

演習の流れ

1. KH-Coderの起動
2. データの読み込み・前処理の実行
3. データの概要をつかむ：抽出語の検索
4. 抽出語間の特徴をつかむ：階層的クラスター分析
5. 抽出語の用いられ方を調べる：KWICコンコーダンス

1. KH-Coderの起動

1. デスクトップ上に落としてもらった「khcoder」というフォルダを開く
2. 「kh_coder」というアプリケーションを開く



2. データの読み込み・前処理の実行

データの読み込み

1. メニューから「プロジェクト」→「新規」を選択
2. 分析対象ファイルの横の「参照」ボタンをクリック
3. 「kokoro2」を選択→OKを選択
4. この画面が開く

2. データの読み込み・前処理

データの前処理

1. メニューから「前処理」→「語の取捨選択」を選択
2. 全角で「K」と入力(注意：右下の「入力モード」を選択して「全角英数」で入力してください！)
3. OKを選択

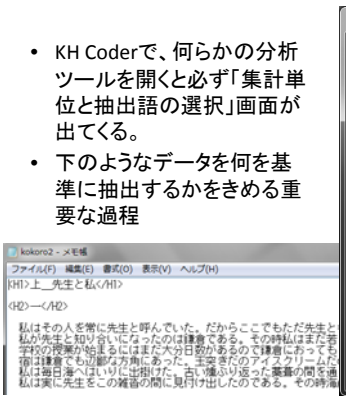
4. メニューから「前処理」→「前処理の実行」を選択→OKをクリック(24秒くらいで終わります)

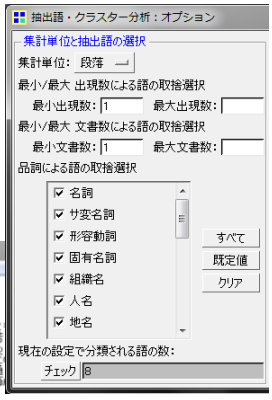
コラム:データの読み込み

- 心理学で使う場合、一番利用可能性が高いのはある質問項目に対する自由記述と考えられる。
- KH Coderでは、与えたテキストデータをどのように分析してほしいかを指定することができる(e.g. 文章/段落/ヘッダーによる指定)

コラム:データの読み込み

- KH Coderで、何らかの分析ツールを開くと必ず「集計単位と抽出語の選択」画面が出てくる。
- 下のようなデータを何を基準に抽出するかをきめる重要な過程



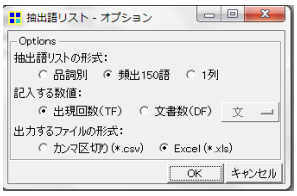


コラム：データの前処理の意味

- KH-Coderは「有意味なデータ」のみを対象として分析している。
- あまりに短すぎる文字(e.g. アルファベット一文字)や記号列(e.g. カッコや句読点)は無視して分析される。
- 夏目漱石の『こころ』の下巻において「K」は重要な役割を持っているが、KH-Coder上では無意味なデータとして無視されてしまう
- 「K」が有意味なデータであることをKH-Coderに教えてあげる必要がある→語の取捨選択はこれを教える作業。

3. データの概要をつかむ：抽出語の検索

1. メニューから「ツール」→「抽出語」→「抽出語リスト」を選択。
2. 抽出語リストの形式を「頻出150語」に変更
3. その他の設定はいじらず、OKを選択



3. データの概要をつかむ：抽出語の検索

- 右のようなエクセルデータが出てくるはず
- 抽出語の横にそれぞれの単語が『こころ』の文章中で何回出てきたかが確認できる
- 質問紙で言うところの、「ローデータを眺める」「ざっくり、どういうデータかわかる」作業

	抽出語	出現回数	抽出語	出現回数	抽出語	出現回数
2	先生	597	兄	62	時々	36
3	K	411	出す	62	入る	36
4	奥さん	388	病気	62	医者	35
5	思う	296	様子	61	驚く	35
6	父	269	声	60	若い	35
7	自分	264	外	59	書物	35
8	見る	225	卒業	58	信じる	35
9	聞く	219	話す	58	与える	35
10	出る	185	取る	57	頼む	35
11	人	182	心持	57	通去	34
12	母	170	急	54	決して	34
13	お嬢さん	168	坐る	54	新しい	34
14	前	163	態度	54	身体	34
15	帰る	155	付く	54	世の中	34
16	今	139	問題	54	他	34
17	顔	135	笑う	53	変る	34
18	来る	131	返事	53	気分	33
19	考える	130	歩く	52	記憶	33
20	言葉	126	読む	51	起る	33
21	眼	123	強い	49	向う	33
22	知る	118	持つ	49	減む	33
23	心	106	分る	49	得る	33
24	妻	104	違う	48	聞える	32

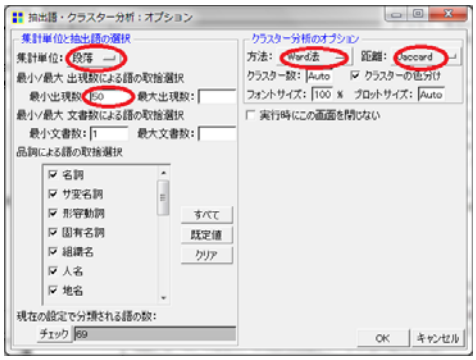
4. 抽出語間の特徴をつかむ：階層的クラスター分析

- 出現パターンが似通っていた単語はどんなのだろう？→クラスター分析が使える！
  - c.f.クラスター分析：対象となるデータ郡のどれとどれが類似しているかを見つけ出すために用いられるさまざまな数学的方法の総称
- 説明はさっきしたので、さっそく試す

4. 抽出語間の特徴をつかむ：階層的クラスター分析

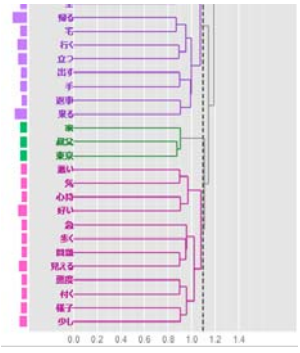
1. メニューのから「ツール」→「抽出語」→「階層的クラスター分析」を選択
2. 「集計単位と抽出語の選択」の集計単位を「段落」に変更。最小出現数を「50」に変更。
3. 「クラスター分析のオプション」の方法と距離がそれぞれ「Ward法」「Jaccard」になっていることを確認。
4. 右下のOKをクリック

4. 抽出語間の特徴をつかむ：階層的クラスター分析



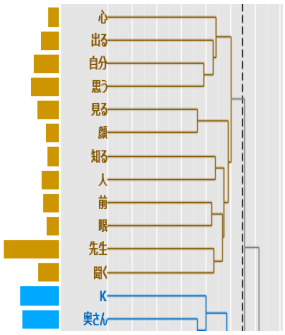
階層的クラスター分析：出力結果

- 下の数値は各クラスター間の結合距離をあらわす
- 左のバーは主成分得点をあらわす
- Jaccard法によって項目間の距離(非類似性)を決め、ここで求められた非類似性をもとにWard法によるクラスリングを行っている
- 色分けはクラスターをあらわす



階層的クラスター分析：出力結果

- どういう意味で使われていたのかが気になるところ
- 茶色のクラスターをみると「先生」と「聞く」の結合距離が近そう→なにか読み取れないか？
- 元データに立ち返って確認する必要がある



5. 抽出語の用いられ方を調べる：KWICコンコーダンス

- KWICコンコーダンスは「ある特徴語がどのような文章中に共起しているのか」を確認するツール
- KWICコンコーダンスを使うことで、元データ上でどのように使用されていたのか、吟味することができる。

5. 抽出語の用いられ方を調べる：KWICコンコーダンス

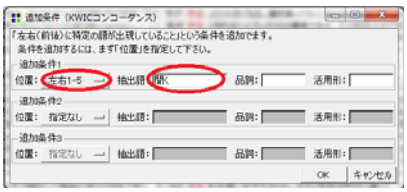
1. メニューから「ツール」→「抽出語」→「KWICコンコーダンス」を選択
2. 抽出語の欄に「先生」と入力
3. OKを選択

5. 抽出語の用いられ方を調べる：KWICコンコーダンス

5. 抽出語の用いられ方を調べる：KWICコンコーダンス

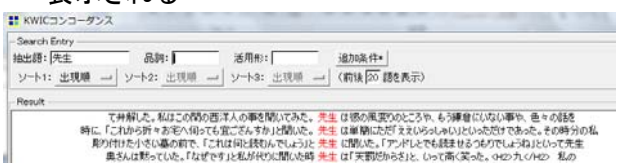
- これだけだと、「先生」と「聞く」の関係がわからないので、聞くと共起している文章に注目したい

1. Search Entryから「追加条件を選択」
2. 追加条件1の位置を「左右1-5」に、抽出語に「聞く」を入力
3. OKをクリック



5. 抽出語の用いられ方を調べる：KWICコンコーダンス

- 先生と聞くが共起している文章が表示される
- 文脈を考慮した検討が可能になる
- さらに、特定の文章をクリックして、左下の「文書表示」をクリックすると、その文章が詳しく表示される



演習のまとめ

- KH-Coderを用いて、データの読み込みとクリーニング(前処理)を行った
- KH-Coderが形態素解析をしてくれた
- データから出現頻度の高い単語をあつめ
- 階層的クラスター分析を行い、それぞれの文脈を確認した

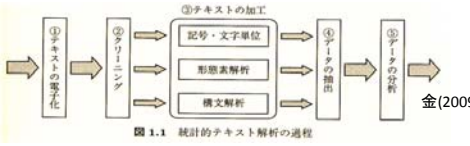


図 1.1 統計的テキスト解析の過程

金(2009). p11より引用

引用文献

- 北海道　市町村間の結びつきの分析方法(市民向けに北海道の市町村のデータのテキスト分析の結果を報告しているため、わかりやすい) http://www.pref.hokkaido.lg.jp/ss/cks/grp/17/set202d04.pdf
- 石田基広 (2008). Rによるテキストマイニング入門. 森北出版株式会社
- 金明哲 (2009). テキストデータの統計科学入門. 岩波書店
- Jin`s HP（金明哲 先生のHP。Rの分析が超わかりやすいです） http://mjin.doshisha.ac.jp/R/
- KH-Coder (樋口耕一准教授の開発したKH-CoderのHP) http://khc.sourceforge.net/
- Kusumi, T., Matsuda, K., & Sugimori, E. (2010). The effects of aging on nostalgia in consumers' advertisement processing. *Japanese Psychological Research*, **53**, 3, 150-162.
- 齊藤堯幸・宿久洋 (2006). 関連性データの解析法　多次元尺度構成法とクラスター分析法. 共立出版
- 玉利祐樹・竹村和久 (2012). 言語プロトコルの潜在意味解析モデルによる消費者の選好分析. 心理学研究, **82**, 6, 497-504.
- 豊田秀樹 (2008). データマイニング入門 Rで学ぶ最新データ解析. 東京書店