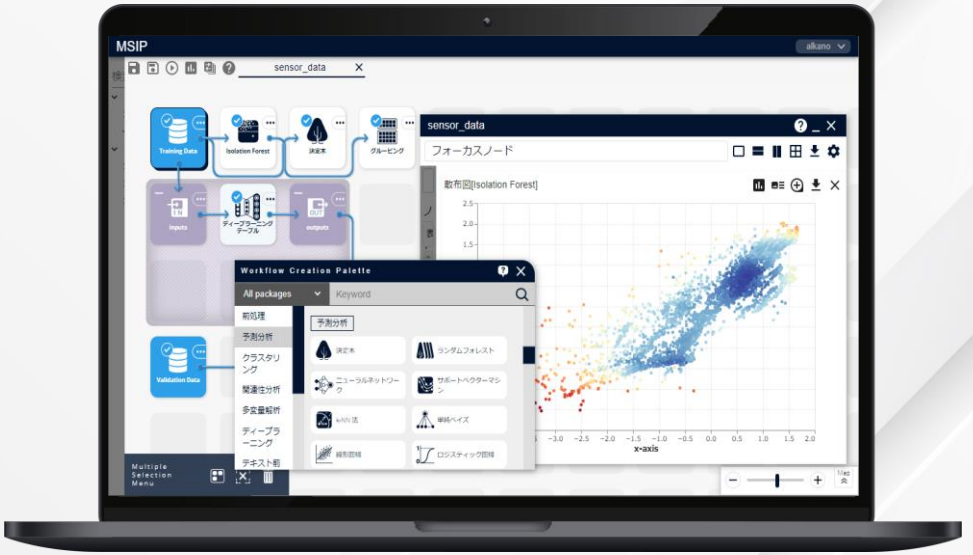
TextExtension

テクニカルサンプルプロジェクト

テキストのクラスタリング

k-means
二項ソフトクラスタリング



株式会社 NTTデータ数理システム

このプロジェクト について

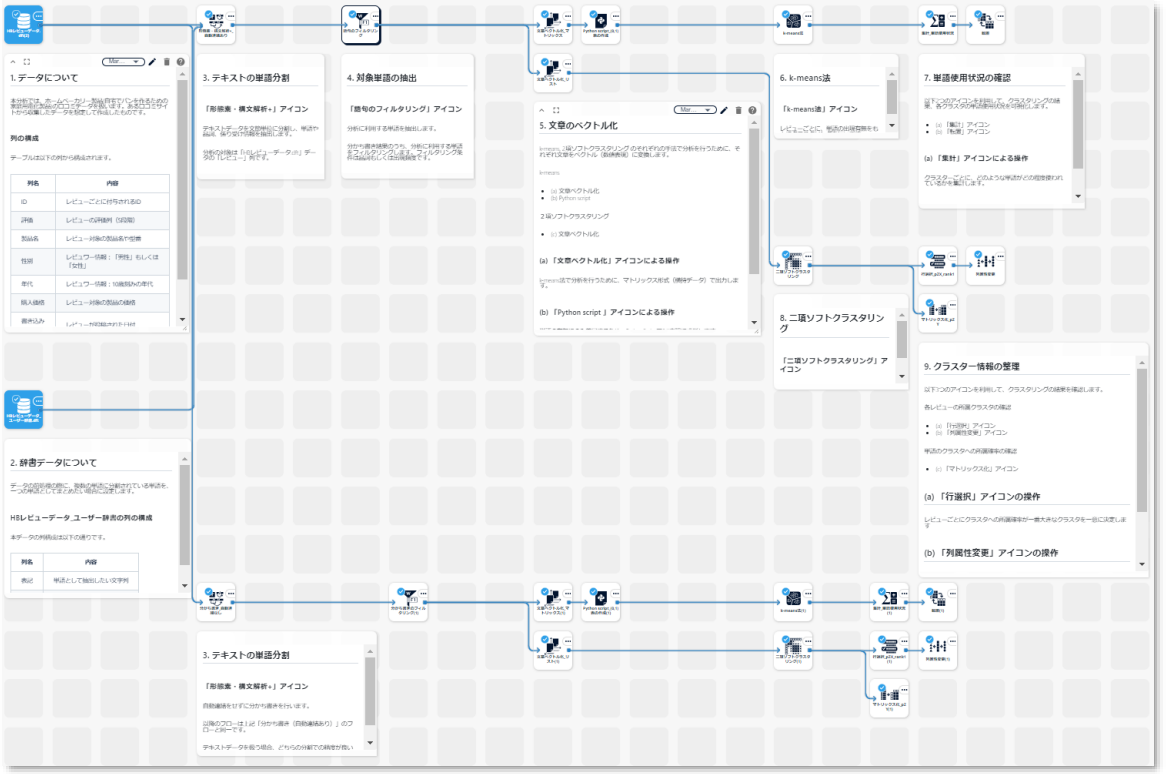
こんな方におすすめします

- テキストデータや属性データを利用してテキストをグループ分け、話題を抽出したい方
- テキストデータを利用した機械学習を行いたい方

何をするプロジェクト？

このプロジェクトでは、いわゆる「教師なし学習」であるクラスタリングという手法を用いてテキストデータをクラスタリング（=グループ分け）する一連の流れを紹介します。
ここでは、クラスタリングとして有名な、k-means法と二項ソフトクラスタリングの2手法をご説明します。

この流れは、テキストデータを利用した機械学習の一般的なフローであり、これを応用することで様々な機械学習手法をテキストデータでも扱うことができます。



プロジェクトの解説

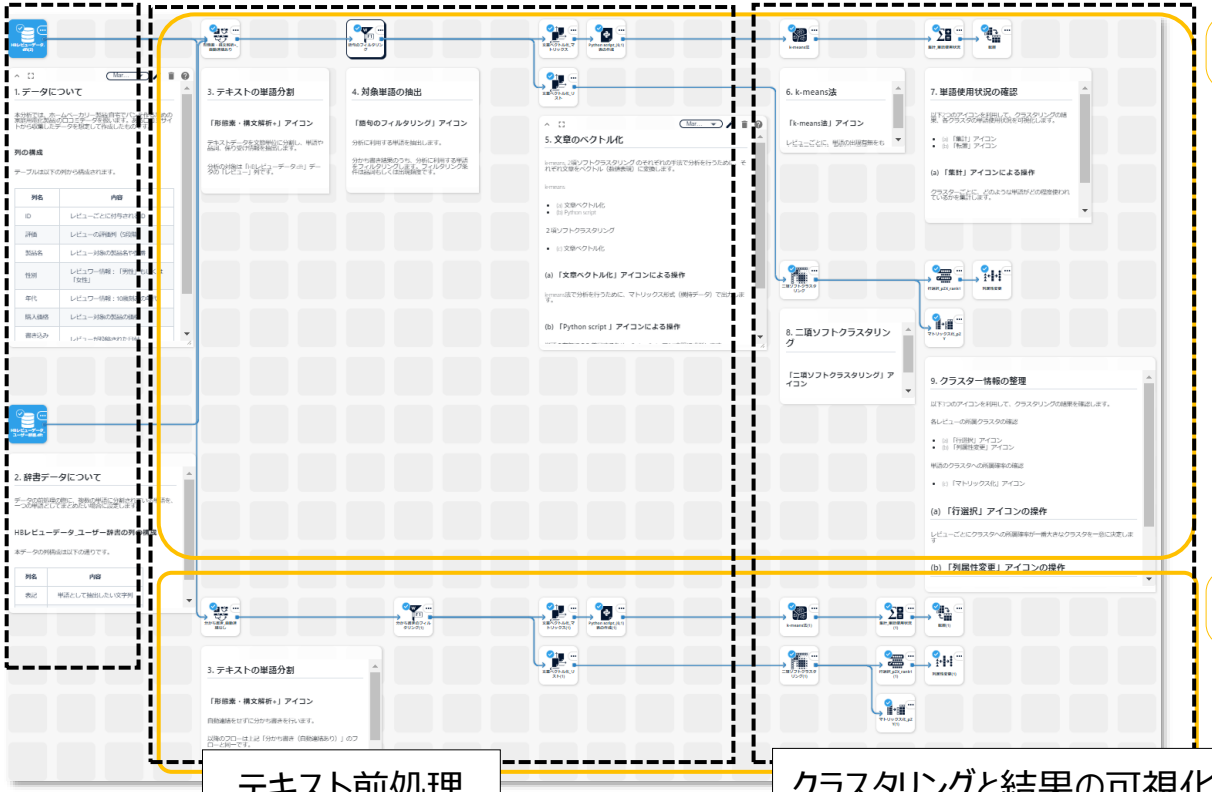
プロジェクト概観

プロジェクトを構成する要素

本プロジェクトは大きく分けて以下の3つの要素に分けられます。

本サンプルプロジェクトでは、文章の分割の粒度を変えてクラスタリングを行っています。「自動連結あり」の結果は、「形態素・構文解析＋」アイコンにて分割の粒度を大きく設定したフロー、「自動連結なし」の結果は、分割の粒度を小さく設定したフローとなっています。

次ページからは「自動連結あり」の結果のフローについて、各要素を構成するアイコンの中身について説明します。



プロジェクト解説 — 対象データ

1. HBレビューデータ.dft

ECサイトで様々なホームベーカリーに対してのレビューをまとめた、仮想の口コミデータです。MSIPの上では、csv形式のデータをdft形式に変換し、シナリオ編集エリア上に配置して使用します。1行が1レビューに対応します。

今回は口コミテキストの入ったレビュー列を利用します。データに含まれる列の詳細については、右の表をご覧ください。

列名	内容
ID	レビューごとに付与されるID
評価	レビューの評価列（5段階）
製品名	レビュー対象の製品名や型番
性別	レビュワー情報：「男性」もしくは「女性」
年代	レビュワー情報：10歳刻みの年代
購入価格	レビュー対象の製品の価格
書き込み日	レビューが投稿された日付
レビュー	レビュー内容 分析対象のテキスト列

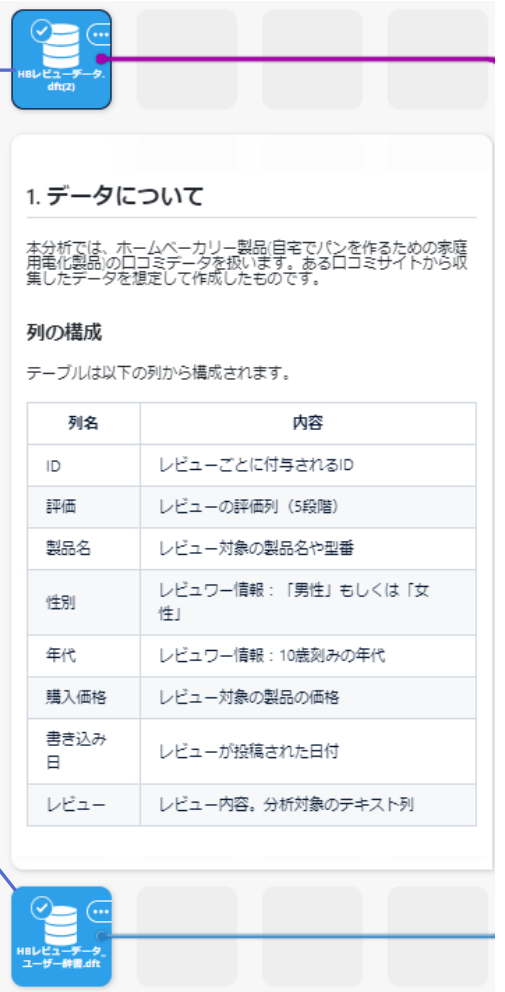
2. HBレビューデータ_ユーザー辞書.dft

既存の辞書にはないような、ユーザー独自の単語を追加するためのデータです。テキストの分割処理を行った結果、つながって出てきてほしい複合語が、いくつかの単語として分かれて出てきてしまう場合などに利用します。

HBレビューデータ_ユーザー辞書.dft-data　列数: 2 行数: 2

No.	表記 Category	品詞 Category
1	パン焼き機	名詞 一般
2	a lot of	形容詞 一般

1.



2.

プロジェクト解説 ― テキスト前処理

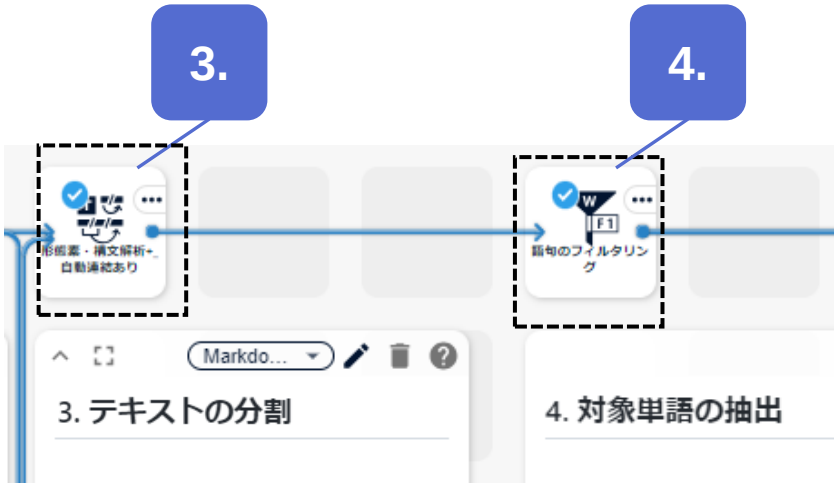
3. テキストの分割

テキストデータを分析する際、記載されている文章の長さや内容が統一されていないため、テキストデータそのままでは分析を行うことができません。そこで、「形態素・構文解析＋」アイコンを利用して、テキストデータを単語単位に分割します。さらに、単語の品詞や係り受け関係などの情報も抽出します。

分析の対象は「HBレビューデータ.dft」データの「レビュー」列です。

4. 対象単語の抽出

クラスタリングの対象とする単語を品詞と頻度の観点から絞り込みます。ここでは意味のある単語でベクトルを作成するために、品詞が**「名詞」「動詞系」「形容詞・形容動詞系」「副詞」**の単語を取り出しています。更に、頻度の大きい単語はクラスタリングに影響するため、**上位5単語を除外**し、ベクトルの次元数を調整するため、**頻度上位100単語**のみを取り出しています。



プロジェクト解説 ― テキスト前処理

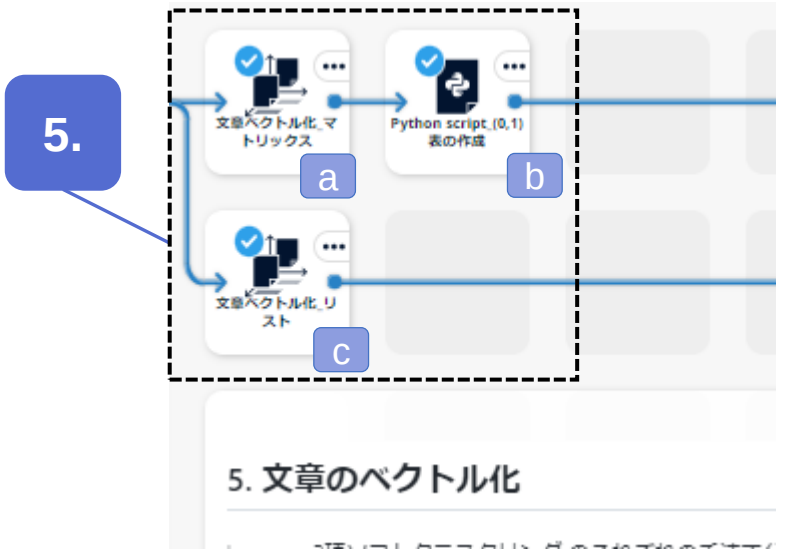
5. 文章のベクトル化

文章に現れる単語をもとに、文章を数値で表したベクトル表現を獲得します。テキストデータを機械学習で用いるためによく使われる手段です。

ここでは、BoW（Bag of Words）を用いて、「どの単語がどのくらい出現しているか」という数値のベクトル表現を獲得します。出現しているかどうかのみに着目する場合には、単語の頻度ではなく、有無を表す「0/1」の数値表現を利用します。

k-means法の入力として、(a)横持データであるマトリックス形式のベクトル表現を獲得したのち、単語が出現するかのみに着目するため、(b)「Python Script」アイコンを用いて 0/1 のベクトル表現を作成しています。

また二項ソフトクラスタリングの入力として、(c)縦持データであるリスト形式のベクトル表現を作成しています。



プロジェクト解説 ― クラスタリングと結果の可視化

6. k-means法

どのような単語が出現しているかをもとに、文章をクラスタリングします。

ここではクラスター数を3に設定し、1行のテキストデータを1件としてクラスタリングしているため、レビューの1件ずつがクラスター1～3のいずれか一つのみに所属します。

7. 単語使用状況の確認

各クラスターがどのような性質をもつか、どのような話題が多いかを確認します。

(a)クラスターごとに、どのような単語がどの程度使われているかを集計し、(b)集計結果を転置し折れ線グラフで可視化します。

プロジェクト解説 ― クラスタリングと結果の可視化

8. 二項ソフトクラスタリング

レビューデータとレビュー全体で利用されている単語のクラスタリングを同時に行います。各レビューが所属するクラスターを把握するとともに、単語のクラスタリング結果をもとに各クラスターがどのような話題を持つかを見ることができます。

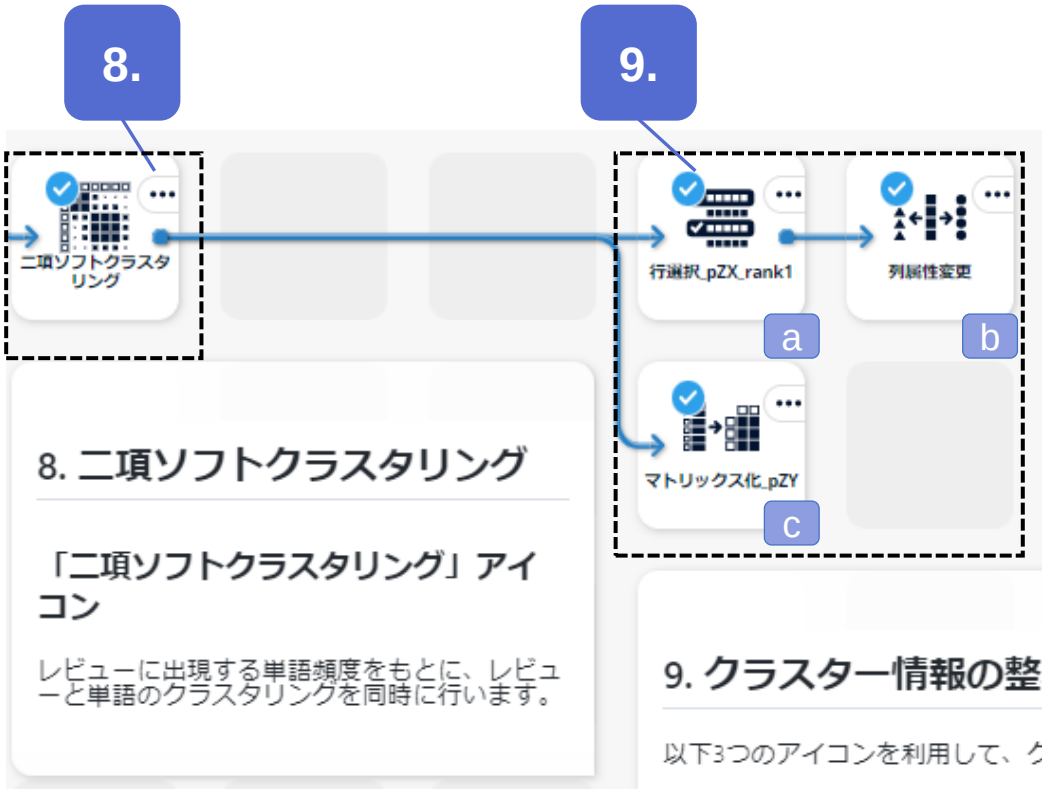
二項ソフトクラスタリングは、レビューデータや単語がそれぞれのクラスターに所属する確率を算出します。レビューデータが複数の話題を持っている、ひとつの単語が複数の話題で語られるということを見ることができます。

9. クラスター情報の整理

クラスタリングの結果を確認します。

極端に偏りがないか、などを確認するために、(a)各レビューの所属確率が一番高いクラスターを抽出し (b)各クラスターに所属するデータ件数を可視化します。

単語のクラスターへの所属確率を確認します。二項ソフトクラスタリングの結果はリスト形式で出力されるため、確認しやすくするために、(c)テーブル形式に整形します。



アウトプットの説明

アウトプット – k-means法

「k-means法」アイコンの結果を確認します。

resultテーブルには、cluster_id、元データの順で列が並びます。「cluster_id」列が、各行が所属するクラスターの値です。

cluster_infoテーブルでは、id, size, 単語, residual列があります。id が各クラスターを表し、クラスターごとにsize(データ件数)とクラスター中心となる単語（各次元）の値を確認できます。residualは中心値との残差の絶対値の総和を表します。

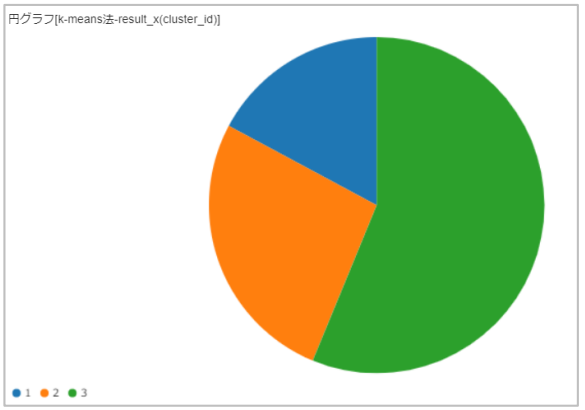
クラスターごとのデータ件数を円グラフで可視化することも有効です。

k-means法-result　列数: 102 行数: 361

No.	cluster_id Category	RowID Integer	焼き立て Integer	食べる Integer
1	1	1	1	1
2	2	2	0	1
3	3	3	0	0
4	2	4	0	1
5	2	5	0	0
6	1	6	0	0
7	2	7	0	0
8	3	8	0	1
9	3	9	1	1
10	2	10	0	1

k-means法-cluster_info　列数: 103 行数: 3

No.	id Category	size Integer	焼き立て Float	食べる Float
1	1	62	0.080645	0.161290
2	2	96	0.042036	0.354742
3	3	203	0.078654	0.231190

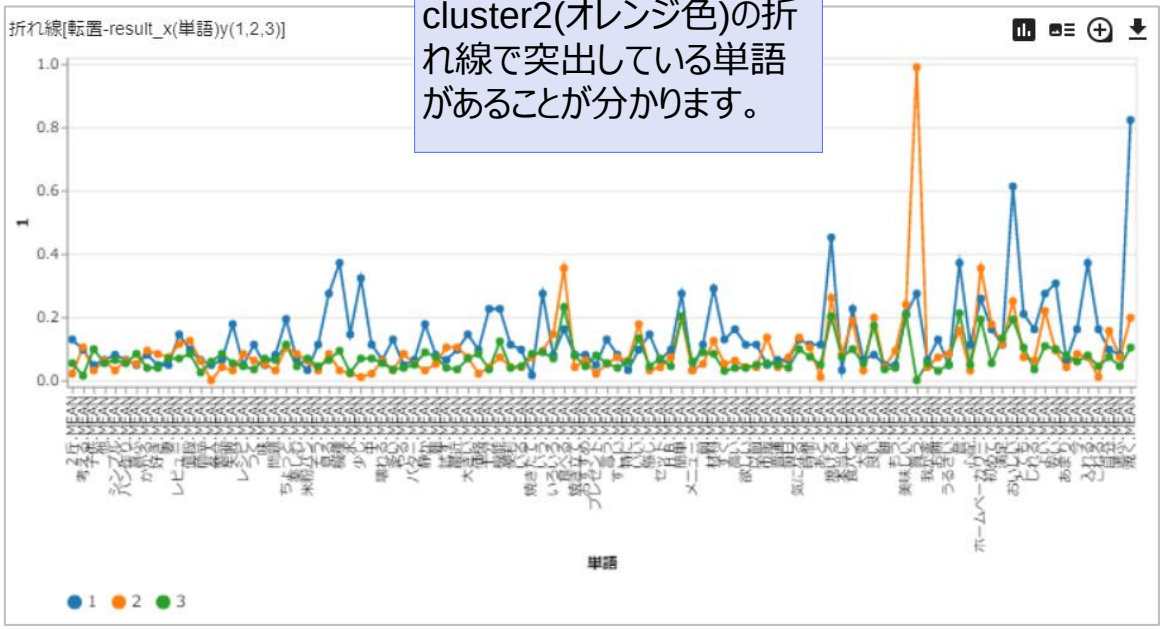


アウトプット – k-means法（単語使用状況）

各クラスターに割り当てられた文章ごとに、どのような単語がどの程度使われているかを集計し、折れ線グラフで可視化します。
それぞれのクラスターにおける単語の影響度合いを見ることができます。値の大きいものほど、どのクラスターに対する影響が大きい単語とみなします。
cluster3（緑色）は特に突出した値の単語はなさそうです。複数の話題を含むために各単語の影響度合いがならされていることが考えられます。cluster3の結果から、クラスター数を増やすことでより話題を分割できそうなことが考えられます。cluster2（橙色）は、「買う」という単語が最も影響が強いことが分かります。

転置-result　列数: 4 行数: 100

No.	単語 String	1 Float	2 Float	3 Float
1	2斤.MEAN	0.129032	0.020833	0.054187
2	考える.MEAN	0.096774	0.104167	0.014778
3	子供.MEAN	0.048387	0.031250	0.098522
4	他.MEAN	0.064516	0.062500	0.054187
5	シンプル.MEAN	0.080645	0.031250	0.064039
6	パン作り.MEAN	0.064516	0.062500	0.054187
7	喜ぶ.MEAN	0.048387	0.052083	0.083744
8	かかる.MEAN	0.080645	0.093750	0.039409
9	好き.MEAN	0.048387	0.083333	0.039409
10	妻.MEAN	0.048387	0.072917	0.068966



cluster2(オレンジ色)の折れ線で突出している単語があることが分かります。

アウトプット – 二項ソフトクラスタリング

二項ソフトクラスタリングの結果には複数のテーブルが表示されます。pZXテーブルでは、X（レビュー）のZ（クラスター）への所属確率を確認します。RowID 1 のレビューはz=2クラスターに43%、z=1クラスターに33%、z=3クラスターに24%の割合で所属していることが分かります。pYZテーブルでは、Z（クラスター）ごとのY（単語）の所属確率を見るため、クラスターの特色を表す単語を確認することができます。z=1クラスターでは「焼ける」「美味しい」などの一般的な単語の他、「音」や「静か」が上位に現れるため、音に関する話題を持つクラスターと考えられます。またz=3クラスターは、「お餅」や「米粉パン」などパン以外メニューの話題を持つと考えられます。

【pZXテーブル】

No.	X Category	Z Integer	pZX Float	rank Integer
1	1	2	0.429060	
2	1	1	0.332568	
3	1	3	0.238372	
4	10	3	0.757824	1
5	10	1	0.242176	2
6	10	2	0.000000	3

各クラスタ(Z)への所属確率を確認します。

【pYZテーブル　z=1】

No.	Y Category	Z Integer	pYZ Float	rank Integer
1	音	1	0.071793	1
2	焼ける	1	0.057133	2
3	美味しい	1	0.041141	3
4	食べる	1	0.037988	4
5	簡単	1	0.035084	5
6	おいしい	1	0.032401	6
7	気になる	1	0.029272	7
8	静か	1	0.027919	8
9	大きい	1	0.024729	9
10	ちょっと	1	0.024038	10

【pYZテーブル　z=3】

No.	Y Category	Z Integer	pYZ Float	rank Integer
239	満足	3	0.009200	39
240	好き	3	0.008890	40
241	商品	3	0.008650	41
242	餅	3	0.008399	42
243	あまり	3	0.008114	43
244	バター	3	0.007640	44
245	言う	3	0.007546	45
246	普通	3	0.007310	46
247	もう	3	0.007261	47
248	すごい	3	0.006601	48
249	つく	3	0.006392	49
250	安い	3	0.006324	50
251	お餅	3	0.005971	51
252	考える	3	0.005840	52
253	今	3	0.005733	53

アウトプット – 二項ソフトクラスタリング（単語のクラスタリング結果）

各単語がどのクラスターにどのくらいの確率で所属しているかを把握するには pZY テーブルで確認します。

ただし、二項ソフトクラスタリングの結果では、リスト形式で表示されているため、各クラスターでまとめて確認したい場合には不便です。そのため、マトリックス形式に整形し、結果を確認することをお勧めいたします。

マトリックス化_pZY-result　列数: 4 行数: 100

No.	Y Category	pZY.2 Float	pZY.1 Float	pZY.3 Float
1	あと	1.000000	0.000000	0.000000
2	あまり	0.130169	0.431374	0.438457
3	いい	0.000000	0.501736	0.498264
4	いう	0.904389	0.000000	0.095611
5	いる	0.000000	1.000000	0.000000
6	いろいろ	0.000000	0.000000	1.000000
7	うるさい	0.000000	1.000000	0.000000
8	おいしい	0.292957	0.312449	0.394595
9	おすすめ	0.000000	1.000000	0.000000
10	お餅	0.381104	0.351876	0.267020

単語ごとに、どのクラスタにどの程度所属しているかが一覧で見やすくなります。

アウトプット – 二項ソフトクラスタリング（データ数の確認）

二項ソフトクラスタリングの特徴として、各要素は複数のクラスタに所属しうるということがあります。一方でどれか一つのクラスターにのみ所属するとみなしたい場面も起こりえます。そのようなときには、所属確率の一番大きいクラスターに所属するとみなして、所属先を一意に決めることがあります。

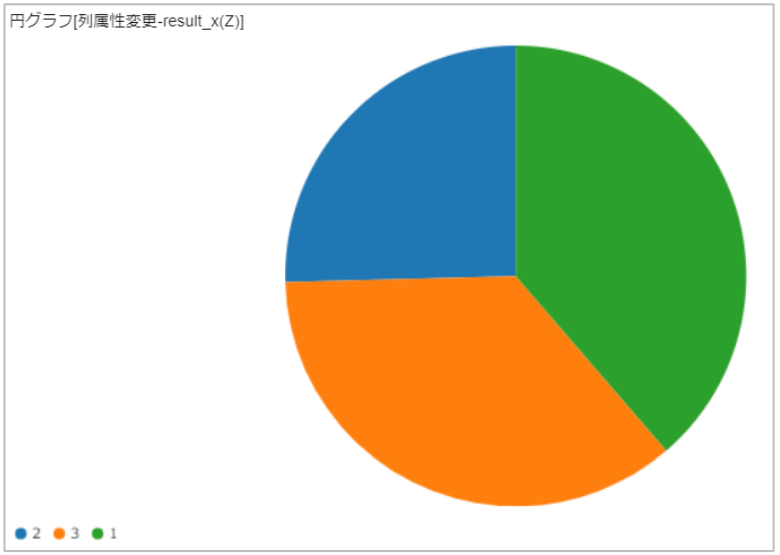
二項ソフトクラスタリング結果のpZXにて、rank=1の行のみを抽出することで、各レビューと所属確率が一番大きいクラスターの行を抜き出すことができます。その結果を円グラフで可視化することで、各クラスターのサイズをおおよそ把握することが可能です。

列属性変更-result　列数: 4 行数: 362

No.	X Category	Z Category	pZX Float	rank Integer
1	1	2	0.429060	1
2	10	3	0.757824	1
3	100	3	0.770846	1
4	101	3	1.000000	1
5	102	2	1.000000	1
6	103	1	0.701432	1
7	104	3	0.535244	1
8	105	2	0.524179	1
9	106	1	1.000000	1
10	107	2	0.429607	1

Zの値はカテゴリとして扱います。

rank=1 のみを抽出します。



アイコンの設定

アイコンの入力設定や処理実行時の設定項目について

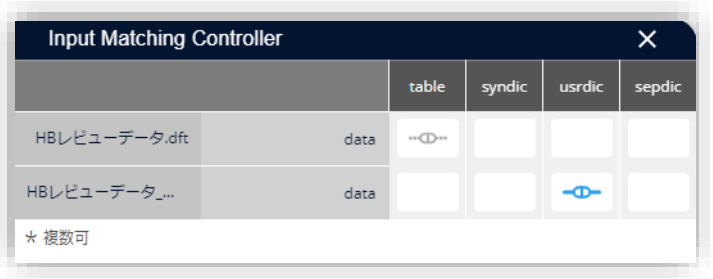
アイコン – 形態素・構文解析+_自動連結あり

インプット設定

テキストデータと辞書ファイルの設定を行います。

ここでは、分割処理の対象のテキスト列を含むデータを「table」、HBレビューデータ_ユーザー辞書を「usrdic」に指定します。

辞書はそれぞれ、類義語辞書を「syndic」、ユーザー辞書を「usrdic」、分割辞書を「sepdic」に設定します。いずれの辞書も必須ではありません。詳細は補足情報の『辞書ファイル』をご参照ください。



対象テキスト列

- **テキスト列**

  分割処理の対象としたい列を指定します。1列のみの指定が可能です。ここでは「レビュー」列を対象とします。

アイコン – 語句のフィルタリング

品詞フィルタ

よく利用される品詞セットは「デフォルト品詞セット」として設定されています。**名詞/動詞系/形容詞・形容動詞系/副詞**の選択が可能です。詳細に設定する場合には「オリジナル設定」を選択し、利用する品詞を個別に指定します。

頻度フィルタ

- **対象列**

  頻度を指定したい単語列を指定します。

- **上位N単語を除外する**

  頻度上位から指定した数の単語を除外します。ここでは初期設定の「5」を指定します。

- **上位N単語を抽出する**

  頻度上位から指定した数の単語を抽出します。ここでは「100」を指定します。

アイコン – 文章ベクトル化_マトリックス・文章ベクトル化_リスト①

変数選択

- **単語列**

  ベクトル化の対象となる単語列を指定します。ここでは置換語列である「replaced」列を選択します。

- **キー列**

  ベクトルを生成するキー列を指定します。「形態素・構文解析+」アイコンの結果を利用する場合、以下の列を選択します。

  - 1行（1セル）単位のベクトル化：RowID
  - 1文単位のベクトル化：RowID, SntID

  ここでは、1行単位でベクトル化を行うため、「RowID」列を選択します。

アイコン – 文章ベクトル化_マトリックス・文章ベクトル化_リスト②

モデルの選択

ベクトル表現のモデルを選択します。モデルの種類は、単語の出現状況から文章データをベクトル化する手法として、

- BoW（Bag of Words）
- tf-idf（Term Frequency-Inverse Document Frequency）

単語の埋め込み表現を利用してベクトル化する手法として、

- SWEM（Simple Word-Embedding-based Methods）

があります。詳細はマニュアルをご参照ください。

ここでは「BoW」を選択します。

出力形式

ベクトル化したデータの出力形式を指定します。マトリックス形式はキー列で指定した単位1行ごとにベクトル表現を出力します。リスト形式は、キー列・単語・値の組を出力します。

k-meansはマトリックス形式、二項ソフトクラスタリングはリスト形式の出力を利用します。

アイコン – Python script(0,1)表の作成

Python script

BoWでベクトル化したデータをもとに、出現有無の情報に置き換えたベクトル表現を獲得します。

【スクリプト内で行っていること】

- （MSIP）DataFrame を pandas.DataFrame に変換する
- ベクトル表現のテーブルにおいて、単語のキー列ごとの出現頻度である各セルの値に対して、以下の置換を行う
  - 値 ≧ 1 … 新しい値：1 (=単語が出現している)
- pandas.DataFrame を（MSIP）DataFrame に変換する

アイコン – k-means法

変数選択

- **説明変数**

  クラスタリングのもととなる変数を指定します。ここではRowIDを除く単語列をすべて指定します。

基本設定

- **距離計算方法**

  クラスターの中心と各要素のベクトルの距離を定義します。単語を扱う場合、ユークリッド距離の他、コサイン距離などもよく利用します。

- **クラスター数**

  いくつのクラスターに分けるかという値を指定します。ここでは「3」に指定します。

初期クラスターの設定方法

初期クラスターの設定方法を指定します。ここでは KMeans++を選択します。初期のクラスターを離れた位置に定める手法で、効率よくクラスタリングを行います。

グラフ – 円グラフ（k-means法）

グラフの種類

作成するグラフの種類を指定します。ここでは「その他」の「円グラフ」を選択します。

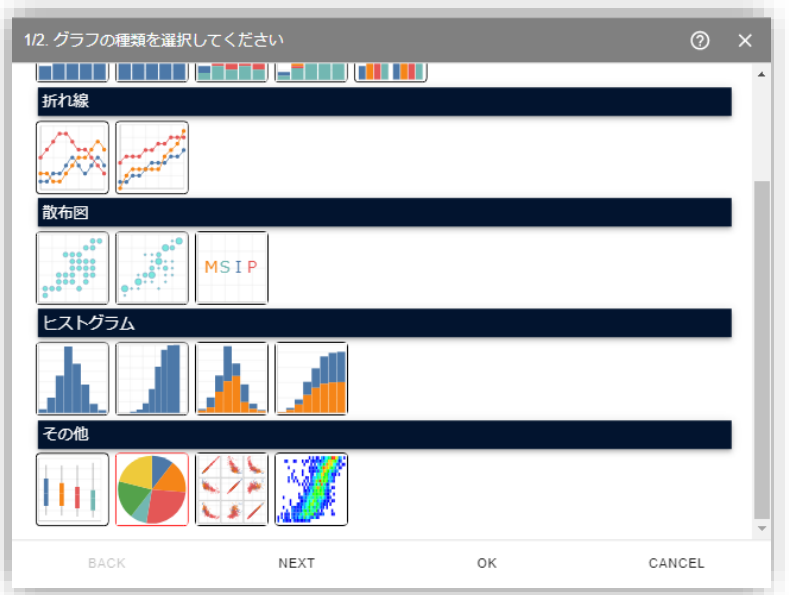
データの列

- **データ選択**

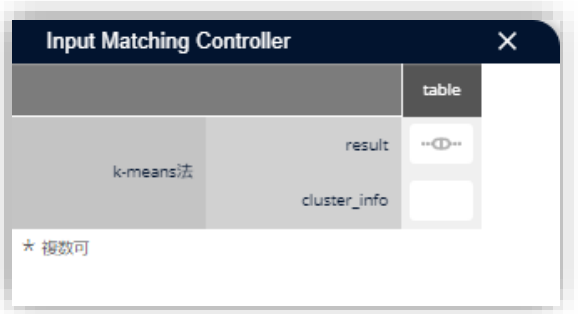
  円グラフを作成するデータを選択します。分析結果テーブルからグラフを作成する場合は自動的に入力されています。

- **カテゴリ**

  円グラフで表現したい列を指定します。ここでは「cluster_id」を選択します。



アイコン – 集計_単語使用状況



インプット設定

集計対象のテーブルを指定します。k-means法の「result」テーブルが対象です。

対象テキスト列

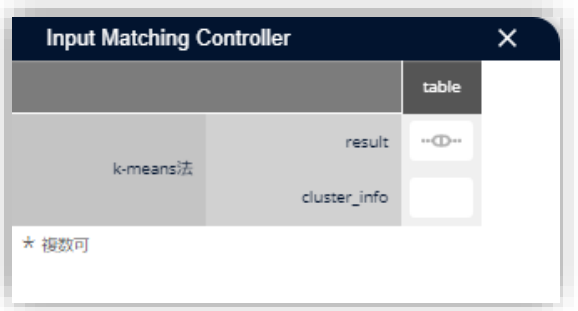
- **集計項目**

  集計対象列と集計方法を設定します。ここでは、単語列を集計対象とし、平均を算出します。

- **キー列**

  指定したキー列ごとに、集計項目設定した集計が行われます。ここではクラスターごとに単語の平均を算出するため、「cluster_id」を指定します。

アイコン – 転置



転置設定

- **対象列**

  行と列の転置を行いたい対象の列を指定します。ここでは可視化（折れ線グラフの作成）のため、各単語列名を値に持つ単語列を作成します。

詳細設定

- **転置前の列名：**

  転置前の列名の扱いを指定します。「元の列名を転置後の一列目とする」と指定することで、元の列名を1列目のデータと指定します。

- **転置後の列名**

  転置後の列名を指定します。「列名となる列を選択する」を指定することで、いずれかの1列の値を列名として扱います。

グラフ – 折れ線グラフ（転置）

グラフの種類

作成するグラフの種類を指定します。ここでは「折れ線」の「折れ線」グラフを選択します。

データの列

- **データ選択**

  折れ線を作成するデータを選択します。分析結果テーブルからグラフを作成する場合は自動的に入力されています。

- **X軸**

  折れ線グラフのX軸（横軸）を指定します。ここでは「単語」を選択します。

- **Y軸**

  折れ線グラフのY軸（縦軸）を指定します。ここでは各クラスターごとの単語の利用状況を描画するため「1」「２」「3」を選択します。

アイコン – 二項ソフトクラスタリング

変数選択

- **X列**

  クラスタリング対象の列を指定します。ここでは「RowID」列を選択します。

- **Y列**

  X列と同時にクラスタリングしたい対象の列を指定します。ここでは「word」列を選択します。

- **スコア列**

  X列とY列の共起度合いを表す列を指定します。ここでは、各列の単語の出現頻度である「value」列を選択します。

パラメータ設定

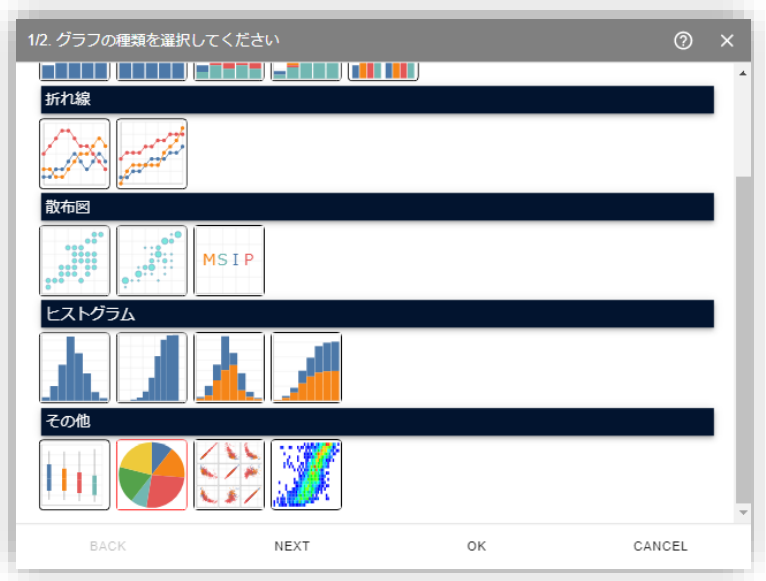
- **隠れ変数(Z)の数/クラスター数**

  いくつのクラスターに分けるかという値を指定します。ここでは「3」に指定します。

グラフ – 円グラフ（二項ソフトクラスタリング）

グラフの種類

作成するグラフの種類を指定します。ここでは「その他」の「円グラフ」を選択します。



データの列

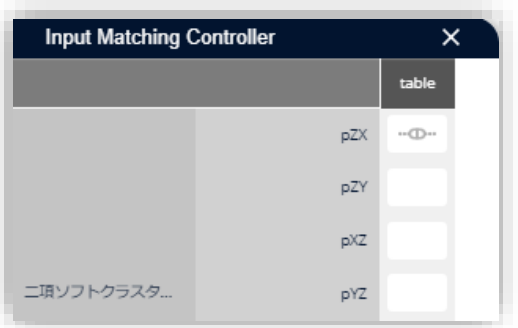
- **データ選択**

  円グラフを作成するデータを選択します。分析結果テーブルからグラフを作成する場合は自動的に入力されています。

- **カテゴリ**

  円グラフで表現したい列を指定します。ここでは「Z」を選択します。

アイコン – 行選択_pZX_rank1



インプット設定

行選択を行いたい対象列を指定します。ここでは、レビューのクラスタリング結果を見るため、「pZX」テーブルを指定します。

対象列

行選択を行う条件の対象列を指定します。ここでは、各レビューの所属確率が一番大きいデータを抽出するため、「rank」列を指定します。

rank

「rank」列に対する条件を指定します。

- **演算子**

  条件の演算子を指定します。「==」を指定し、一致する条件を抽出します。

- **式**

  条件式を指定します。ここでは rank=1 の行のみを抽出するため、「1」とします。

アイコン – 列属性変更

対象列

列属性を変更したい対象列を指定します。円グラフ作成のため「Z」列（クラスターID列）のデータ型を変更します。

Z

「Z」列の変更の設定を行います。

- **新列名**

  新しい列名に変更できます。特に変更がなければそのまま利用します。

- **新列型**

  新しい列型を指定します。円グラフの変数として指定するため「category」とします。

アイコン – マトリックス化_pZY

インプット設定

単語のクラスターへの所属確率を確認するため、「pZY」テーブルを指定します。

キー列

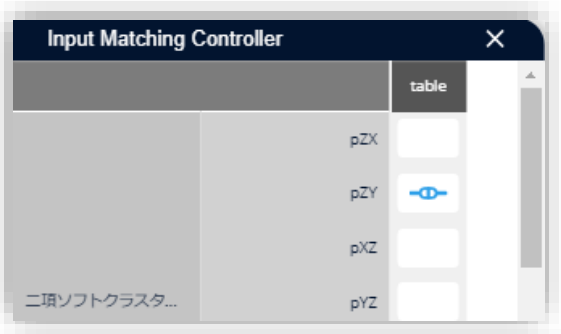
マトリックスのキーとなる列を指定します。単語である「Y」列を選択します。

横展開列

キー列以降の列名になる列を指定します。ここでは「Z」列を選択します。

内容列

所属確率である「pZY」列を指定します。



補足情報

技術的な情報や利用規約について

辞書ファイル

「形態素・構文解析＋」アイコンを利用する際には、ユーザー辞書、類義語辞書、分割辞書を利用することができます。

ユーザー辞書

単語の切れ目を変える辞書です。主に、つながって出てきてほしい複合語が、いくつかの単語として分かれて出てきてしまう場合などに利用します。

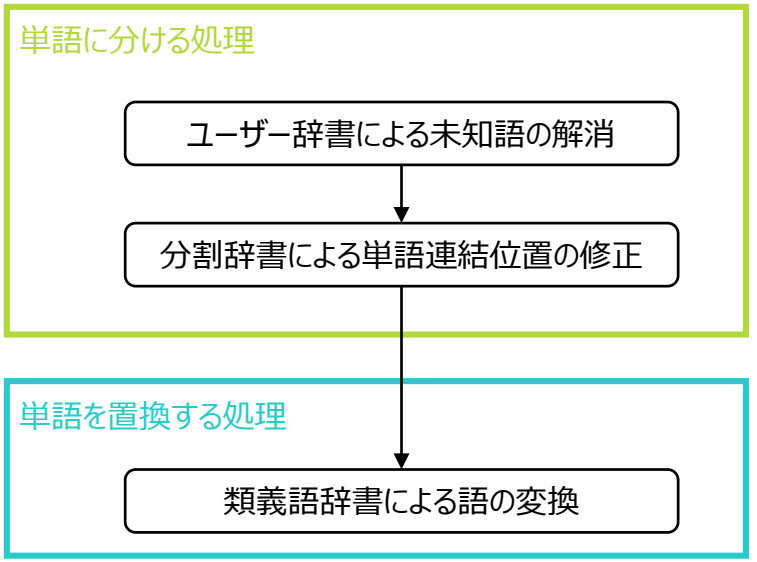
分割辞書

単語の切れ目を変えるために用いる辞書です。「構文解析と自動連結を行う」にチェックを入れて分かち書きする際に、登録した内容に応じて「連結しないように」します。

類義語辞書

キー列以降の列名になる列を指定します。ここでは「Z」列を選択します。類義語をまとめ上げるために用いる辞書です。表記ゆれのまとめ上げに有用です。

これらの辞書はテキストの分割処理の際、右図のような流れで用いられます。

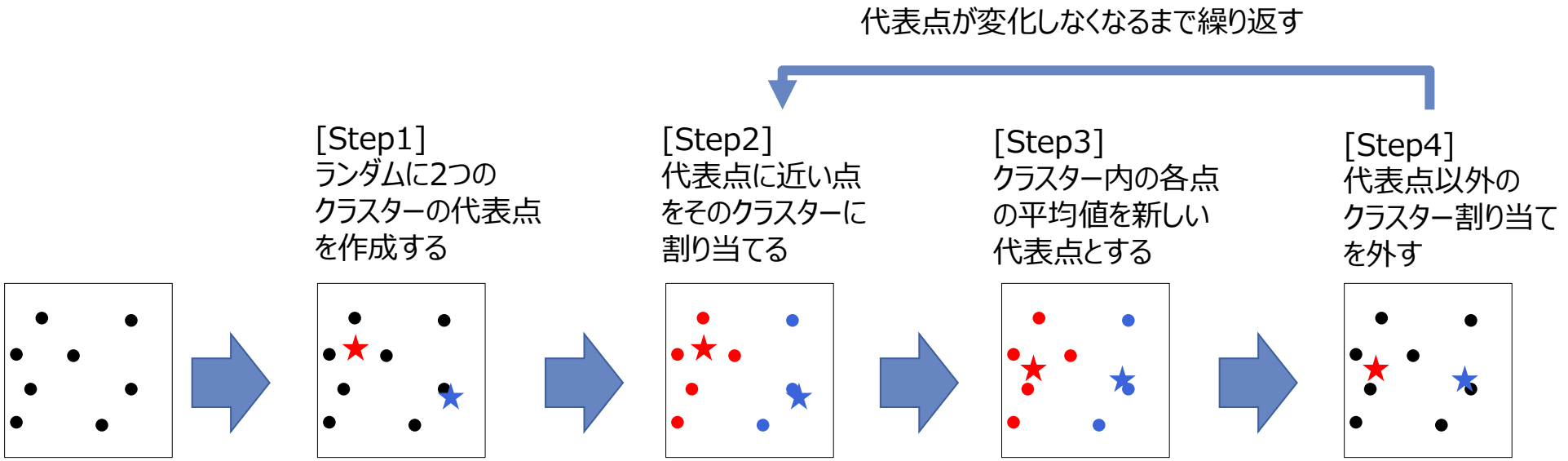


k-means法

クラスターの「中心」の探索、および、クラスター対象の要素がどの重心に一番近いかを計算し所属するクラスターを決めるという操作を繰り返し行うことで、全ての要素をいずれかのクラスターに割り当てます。

一度、分類した後にそのグループ分けが本当に最適か計算し、最適でなければ少しだけ分類を変えてグループを分けなおす作業を繰り返し行います。そのため、Step1 の最初の点が異なると、最終的な分類結果が少し異なることがあります。（初期値依存性）

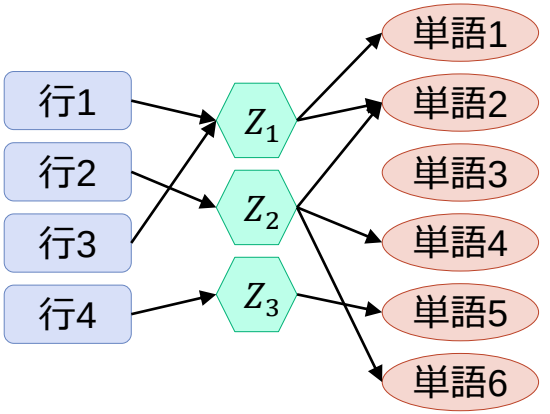
- クラスター数2の場合



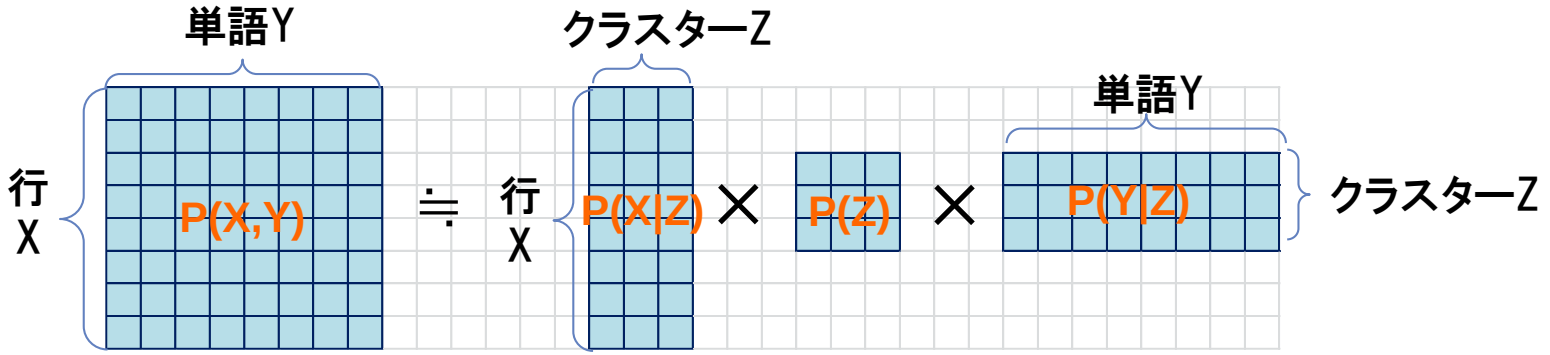
二項ソフトクラスタリング

二項ソフトクラスタリングは、2つの項目 X と Y の組み合わせで表されるデータに対して、同時に発生している組み合わせとクラスターとして抽出する方法です。テキストデータにおいては、1行のテキストデータと単語に対して適用することで、特許テキストと技術用語やレビューと評価単語のクラスタリングを行うことができます。

二項ソフトクラスタリングは、隠れクラスター $Z = \{Z_1, \ldots, Z_k\}$ があると仮定し、行と単語の間に隠れクラスターが介在し、それらを通じて行から単語が発生するというモデルを考えます。（右図）



内部的な計算のロジックとしては、行と単語の二項目の組み合わせの行列データを3つの行列に分解したうえで、右辺の3つの行列について、はじめにランダムな値を初期値として設定し、その積がXYの共起行列であるP(X,Y)と等しくなるように更新していく、という学習をしています。（下図）



本文書・プロジェクトファイルのご利用にあたって

本文書ならびにプロジェクトファイルは、（株）NTT データ数理システム （以下「弊社」）が開発・販売する分析プラットフォーム MSIP および Alkano と TextExtension の機能についての情報提供として弊社が作成を行ったものです。弊社による事前の許可なしに、本文書の再配布や引用の範囲を超える複製といった行為、およびリバースエンジニアリングを禁じます。

本文書ならびにプロジェクトファイルのご利用に際して、ご利用者様および第三者に損害が発生したとしても、弊社は責任を負わないものとします。

プロジェクトファイルは、その中に同梱されているデータを利用し、本文書内で解説している設定可能なパラメータで動作させた場合についてのみ、弊社にて動作の検証を行っております。これを超えるような状況における動作は保証いたしません。

本プロジェクトファイル は、MSIP1.10.0 および Alkano1.4.0、TextExtension1.2.0 にて動作確認を行っております。

TextExtension

お問い合わせ: 株式会社NTTデータ数理システム　営業部
WEB: https://www.msi.co.jp/solution/analytics/index.html

株式会社 NTTデータ数理システム