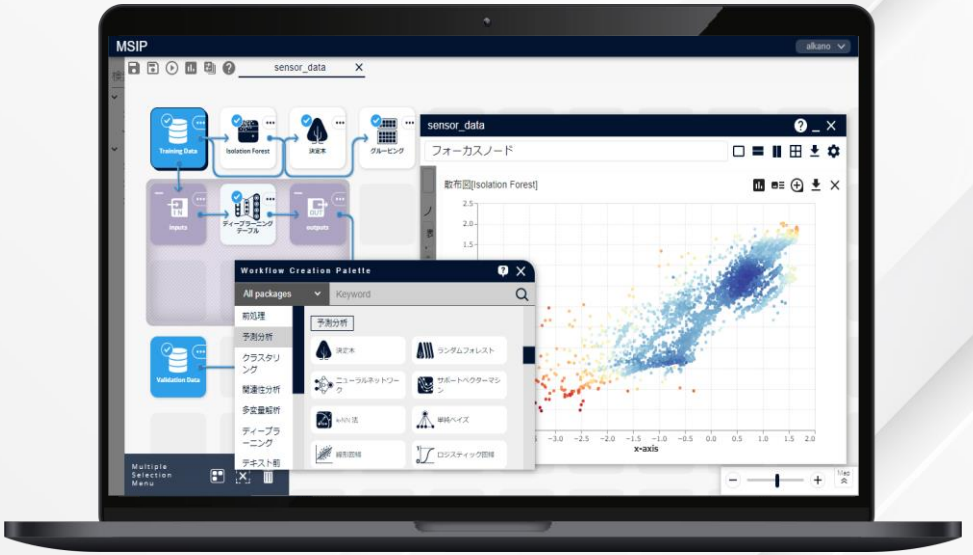
TextExtension

テクニカルサンプルプロジェクト

テキストの話題分析
対応分析



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

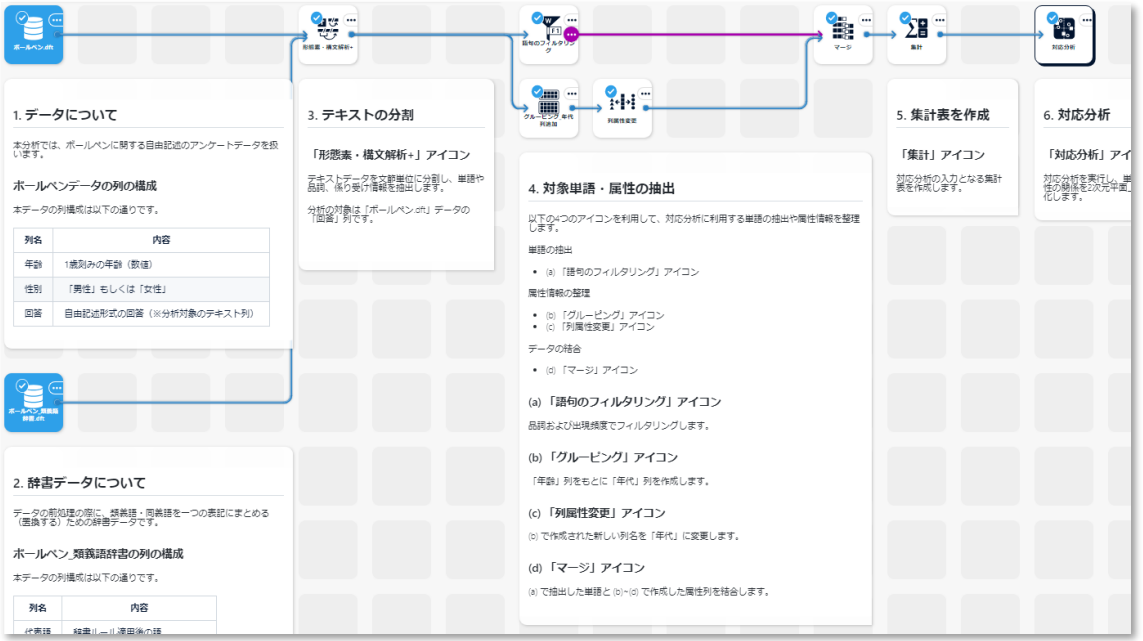
- **テキストデータに含まれる単語を介して、属性情報の傾向や話題を把握したい方**

何をするプロジェクト？

テキストデータの分析にでは、単語そのものだけでなく、テキストデータに付随する属性情報との関係を見ることも重要です。

このプロジェクトでは、対応分析を用いて、単語と属性の情報を合わせて次元圧縮し2次元平面上に可視化することで、話題や属性値の傾向を把握するための分析を行っています。

対応分析は、要素の関係の近さ遠さを2次元平面上の距離で把握することが可能なため、ポジショニングマップとしても有効です。特にテキストデータに利用する場合、テキスト中のことばと属性の関係を2 次元空間上に分布させることにより、ことばを介した属性の分布を見ることができます。



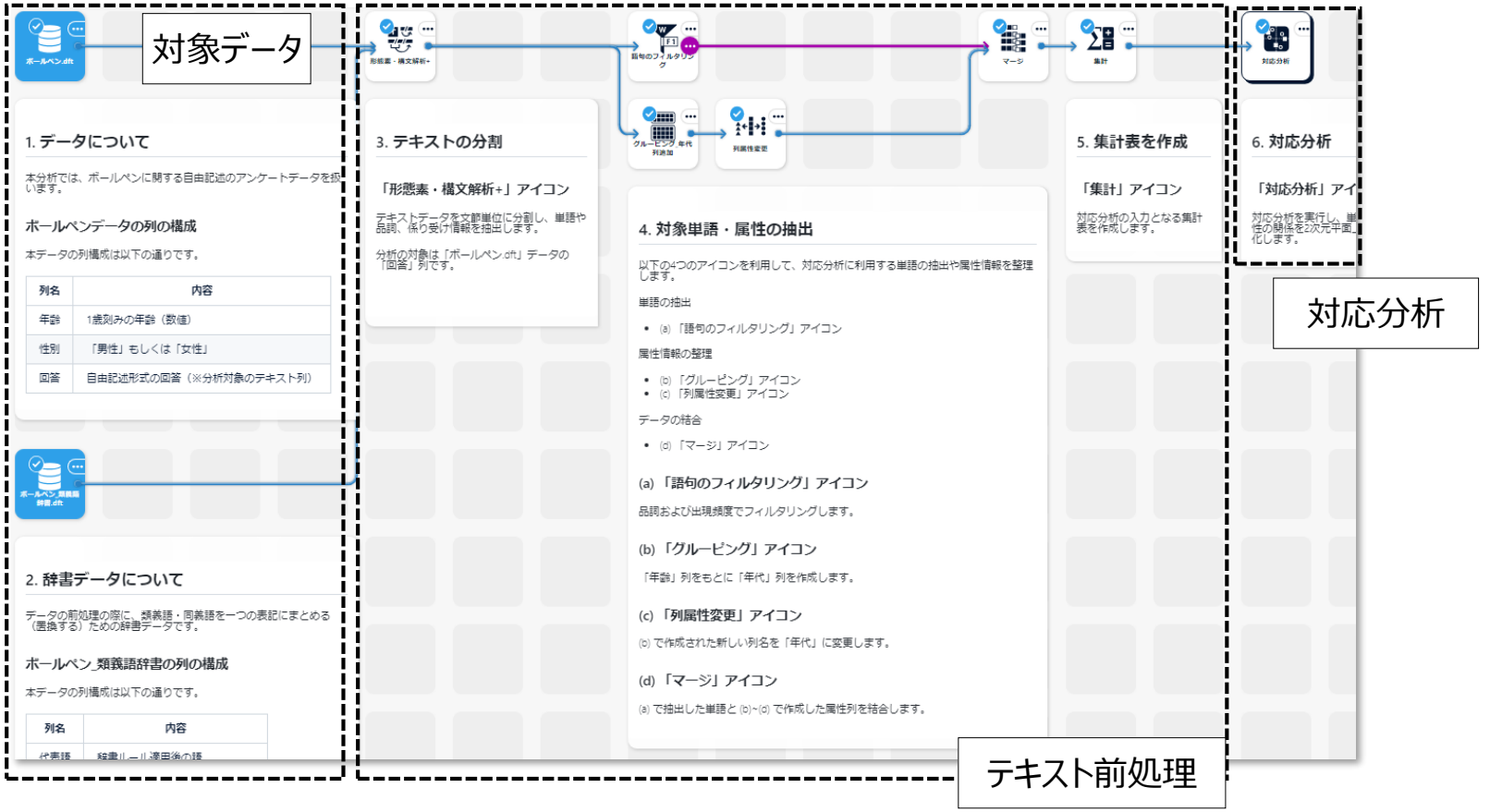
プロジェクトの解説

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。分析対象の単語と属性を紐づけ、単語と属性の関係を可視化します。

次ページからは各要素を構成するアイコンの中身について説明します。

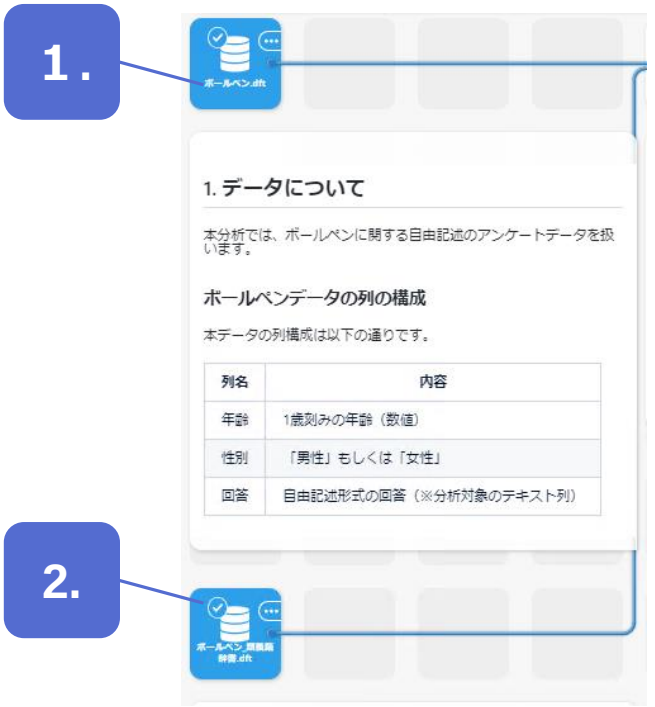


プロジェクト解説 ― 対象データ

1. ボールペン.dft

「ボールペンを選ぶときに重視することは何ですか？」という設問に対する架空の自由記述アンケートデータです。次の3列を含みます。

列名	内容
年齢	1歳刻みの年齢（数値）
性別	「男性」もしくは「女性」
回答	自由記述形式の回答 分析対象のテキスト列



2. ボールペン_類義語辞書.dft

テキストの分割処理の際に、2つ以上の異なる表記の単語を1つの表記にまとめるための辞書データです。

同じ意味の単語を1つの表記にまとめることで、分析結果に表示される単語を整理し、把握しやすい結果を作成することができます。

ボールペン.dft-data　列数: 3 行数: 100

No.	年齢 Integer	性別 Category	回答 String
1	32	女性	手に力が入りにくいので、軽い力で書けるものを買いたいです。
2	18	男性	コンビニで安いのを買ってます。
3	53	女性	ドイツ製のボールペンを使っています。少し値は張りますけど、
4	49	男性	軽い力でサラサラ書けること
5	53	男性	軽さとか、細さとか、スペック的なものよりもフィーリング重視

ボールペン_類義語辞書.dft-data　列数: 3 行数: 4

No.	代表語 Category	品詞 Category	類義語 Category
1	一本	名詞 数詞	1本
2	さらさら	副詞	サラサラ
3	良い	形容詞 一般	よい
4	良い	形容詞 一般	いい

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。

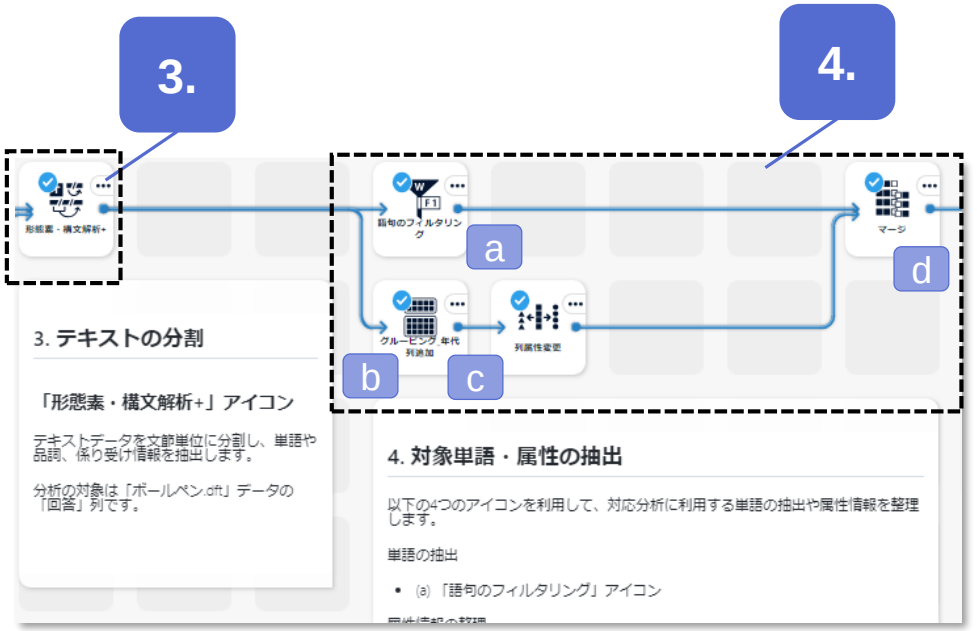
分析の対象は「ボールペン.dft」データの「回答」列です。

4. 対象単語・属性の抽出

対応分析に利用する単語の抽出や属性情報を整理します。

(a) 分割された単語のうち、「語句のフィルタリング」アイコンにて、名詞や形容詞など単語そのもので意味を持つ情報に絞って分析を進めます。あわせて、頻度の小さい単語や頻出単語の除外も有効です。

また合わせて把握したい属性情報も整理します。ここでは、「年代」の傾向をみるために、(b)「グルーピング」アイコンを用いて「年齢」列から10歳ごとにまとめた「年代」に相当する列を作成し、(c)「列属性変更」アイコンで列名を「年代」に変更して、(d)「マージ」アイコンで単語と紐づけています。



プロジェクト解説 ― テキスト前処理

5. 集計表の作成

対応分析の入力とするため、属性値ごとの単語の利用頻度を集計します。対応分析としてはリスト形式、マトリックス形式どちらも入力として設定することが可能ですが、ここではリスト形式の集計表を作成しています。

【集計表の使い分け】

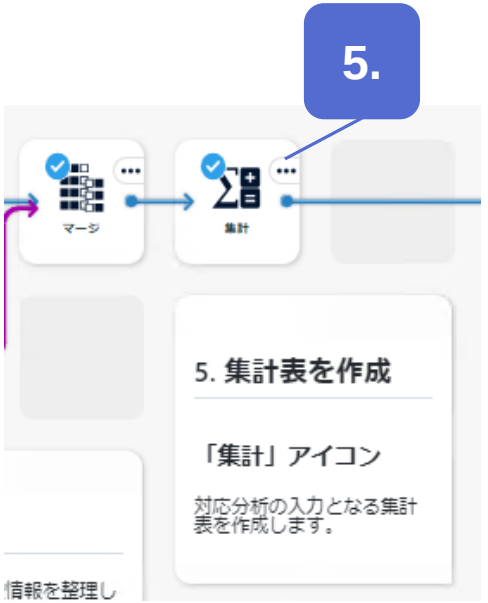
テキストデータの分析において、様々な場面で単語の集計表を利用します。リスト形式、マトリックス形式のどちらを利用するかの目安は、利用する分析手法の仕様に従う他、以下のようなものがあります。

1. 利用できるデータ量

   リスト形式は集計した結果のみを保持するため、情報がコンパクトにまとまっています。一方マトリックス形式は、値を持たないデータについては0という値で埋めるためデータ量が増大する可能性があります。

2. クロス集計結果も確認するか

   属性ごとの単語の出現状況を集計し、クロス集計結果を確認するにはマトリックス形式をお勧めします。



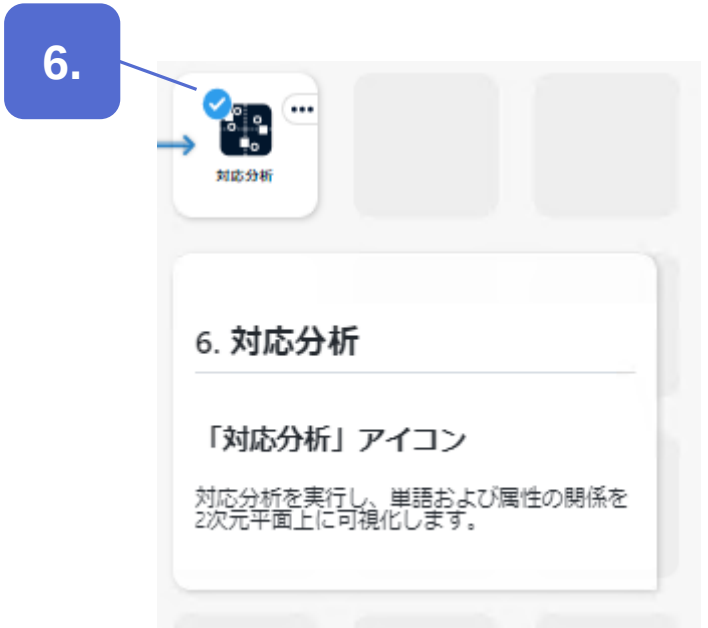
プロジェクト解説 ―対応分析

6. 対応分析

関係性を見たい単語列、属性列、および頻度集計結果の列を指定して対応分析を実行することで、単語・属性値の関連性を表すポジショニングマップを作成します。

「対応分析」アイコンの結果のうち「score_plot」に単語と属性値のスコアをもとに散布図が描画されます。

「score_plot」上で単語の使われ方が似ている属性は近くに配置され、それらの近さ遠さから属性どうしの関係を表すことができます。

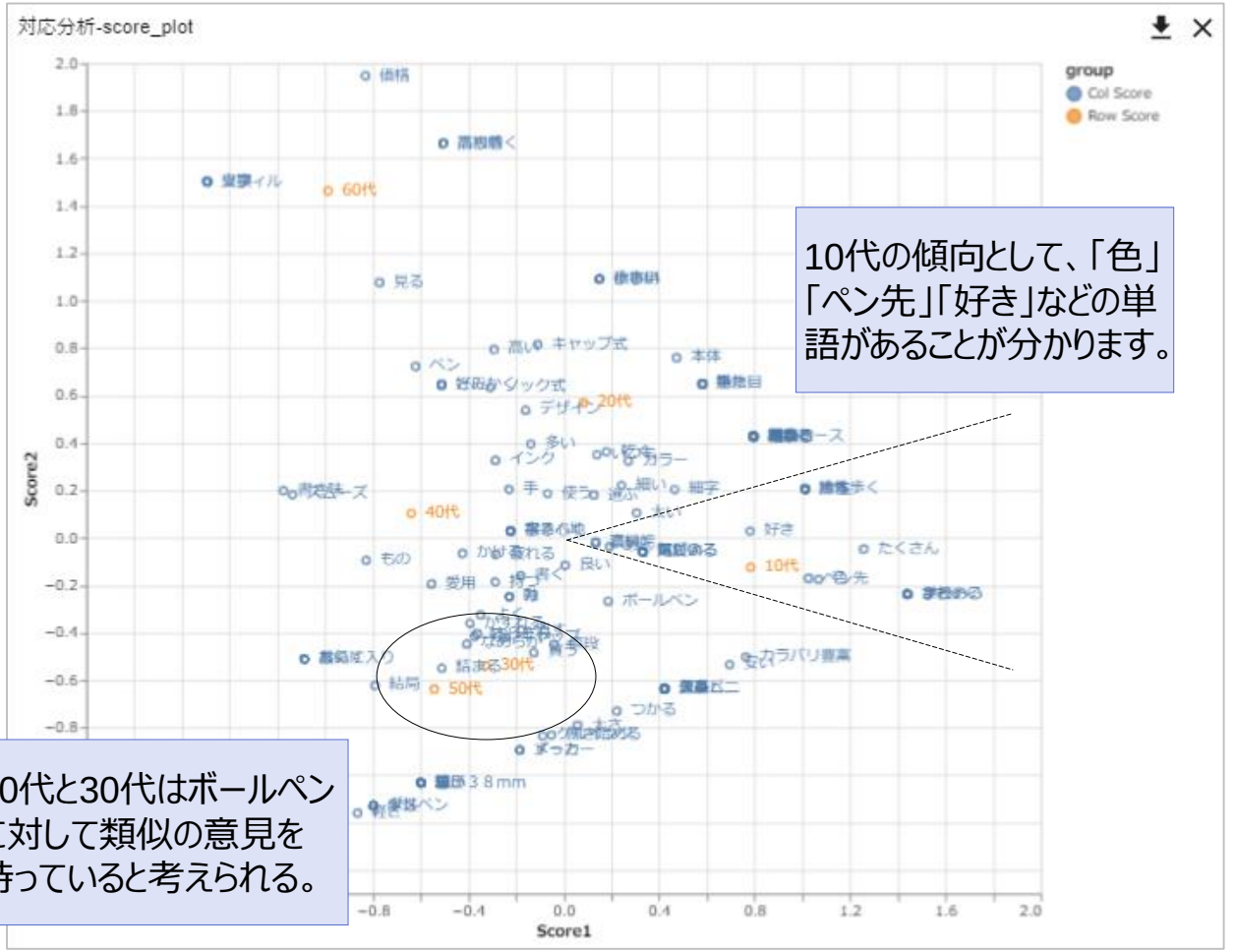


アウトプットの説明

アウトプット

「対応分析」アイコンの結果のうち、「score_plot」では2次元平面上に単語と属性がプロットされ、関係の近さ・遠さを確認することができます。近い位置にある単語同士・属性値同士が関係が近い、即ち関連が強いとみなせるため、類似した属性の傾向や、複数の単語群からなる話題の把握が可能です。

単語と属性の関係は、2次元平面上の近さ遠さではなく、**原点から同じ向きのエリアにプロットされているかどうか**で把握することができます。



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

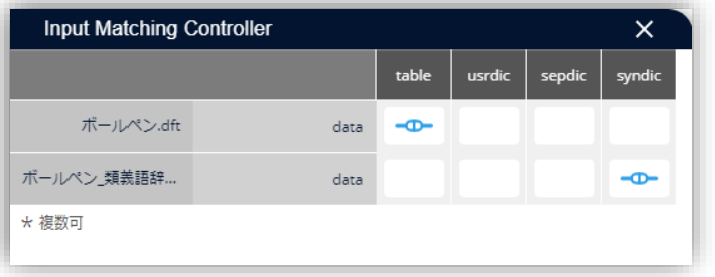
アイコン – 形態素・構文解析+

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、分割処理の対象となるテキスト列を含むデータを「table」、ボールペン_類義語辞書を「syndic」に指定します。

辞書はそれぞれ、ユーザー辞書を「usrdic」、分割辞書を「sepdic」、類義語辞書を「syndic」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。

アイコン – 語句のフィルタリング

インプット設定

「形態素・構文解析+」アイコンの結果のうち分割結果のテーブルである「result」をフィルタリング対象として「table」に指定します。

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**

  頻度を指定して抽出したい単語列を指定します。

- **最低頻度を設定**

  指定した値以上の出現頻度の単語を抽出します。頻度の小さい単語を除外することでノイズを減らします。

- **最高頻度を設定**

  指定した値以下の出現頻度の単語を抽出します。

アイコン – グルーピング_年代列追加

インプット設定

「形態素・構文解析+」アイコンの結果のうち属性情報をもつ「originaldata」をグルーピング対象として「table」に指定します。

対象列

グルーピングの対象となる列を指定します。ここでは「年齢」を指定しています。

年齢

「年齢」列をどのようにグルーピングするかを設定します。10歳刻みにグルーピングする場合、右図のような値を設定します。

- **グループ名**

  グルーピング後の新しい名前を指定します。

- **From**

  各グループ名にまとめる数値の下限値を指定します。

- **To**

  各グループ名にまとめる数値の上限値を指定します。

アイコン – 列属性変更

対象列

列属性変更の対象となる列を指定します。ここでは、「年齢.Grp」を指定します。

年齢.Grp

対象列エリアで「年齢.Grp」行をクリックすると、新列名と新列型の設定エリアが表示されます。

- **新列名**

  新しい列名として「年代」を指定します。

- **新列型**

  新しい列の型を指定します。ここでは元と同じ型の「category」を指定しています。

アイコン – マージ

入力設定

フィルタリング結果と列属性変更結果のデータを紐づけます。

マージ設定

- **マージモード**

  マージ方法を設定します。ここでは、「左外部結合」を選択します。

- **左テーブルマージキー列**

  入力設定の「left」で指定したテーブルにおいて紐づけのキーとなる列を指定します。ここでは「RowID」を指定します。

- **右テーブルマージキー列**

  入力設定の「right」で指定したテーブルにおいて紐づけのキーとなる列を指定します。ここでは「RowID」を指定します。

アイコン – 集計

集計項目

集計項目を設定できます。今回はチェックを入れません。

キー列

カテゴリ値の集計を行うため、キー列の集計機能を利用します。

- **キー列の件数**

  結果のデータフレームに「キー列の件数」列を作成するか指定できます。ここでは、チェックを入れて設定します。

- **キー列の件数割合**

  結果のデータフレームに「キー列の件数割合」列を作成するか指定できます。

- **集計キー**

  集計キーを設定します。ここでは、「年代」と「replaced」を選択します。

アイコン – 対応分析

変数選択

- **行**

  単語列もしくは属性列を1列のみ指定します。

- **列**

  単語列もしくは属性列を1列のみ指定します。

- **頻度**

  集計結果の件数（頻度）列を1列のみ指定します。

分析設定

- **入力データ形式**

  対応分析の入力としているデータの形式を指定します。

  「集計」アイコンの結果を入力とする場合はList形式、「クロス集計」アイコンの結果を入力する場合にはMatrix形式を選択します。

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。

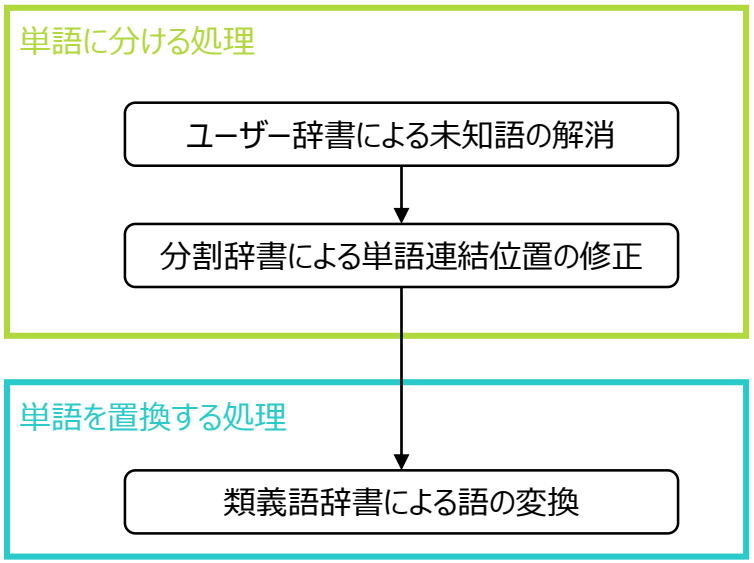
分割辞書

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単語の分割処理を行う際に、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

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が行われる際、右図のような流れで用いられます。



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、MSIP1.9.0 および Alkano1.3.0、TextExtension1.0.0 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
Tel: 03-3358-6681
E-mail: alkano-info@ml.msi.co.jp
WEB: https://www.msi.co.jp/solution/analytics/index.html