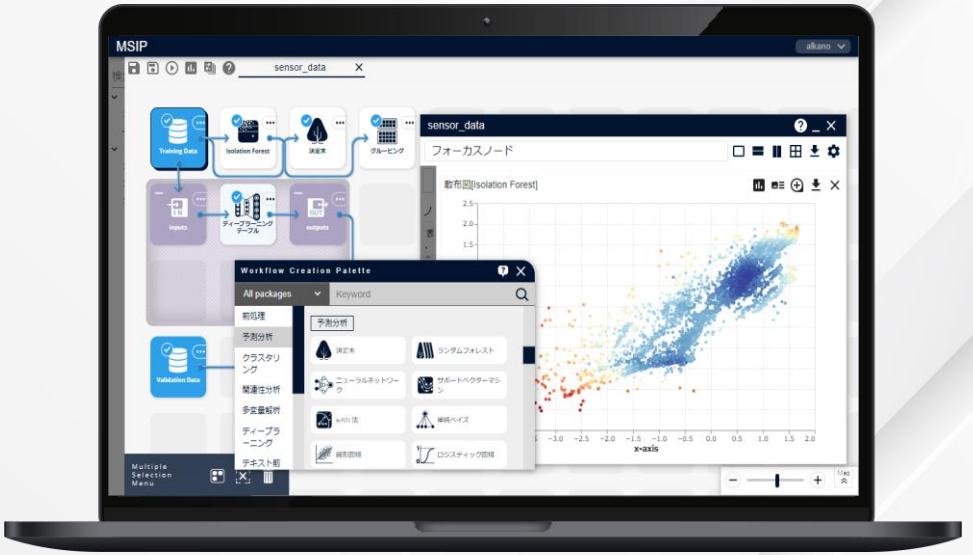
TextExtension

テクニカルサンプルプロジェクト

テキストの話題分析
アソシエーション分析



本資料で解説する「テキストデータのアソシエーション分析」は、2024年6月に、より簡単にご利用いただけるアイコンをご提供予定です。

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

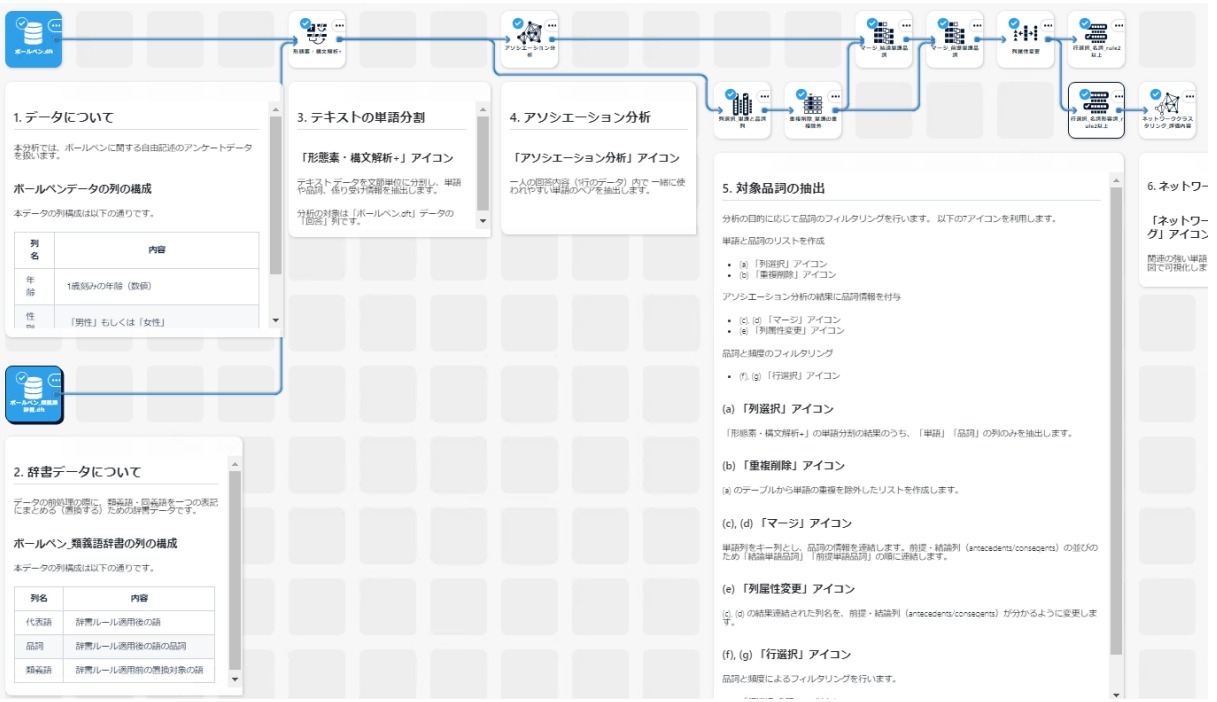
- **テキストデータから話題を抽出したい方**

何をするプロジェクト？

テキストデータの分析を行う際に、どんな単語が出てきているかということだけでなく、どんな話題が語られているかを把握したいということがあります。

このプロジェクトでは**同時に出現する(共起する)単語同士**を抽出する「アソシエーション分析」と、ネットワークを構成して可視化できる「ネットワーククラスタリング」機能を組み合わせて、単語のかたまり(クラスタ)を表示し、話題を把握します。

共起関係は係り受け関係よりも広い関係の単語を抽出できます。また、SNSなど助詞が省略されがちな短い文章でも単語間の関係を抽出できるため、幅広いテキストデータに適用可能な分析です。



プロジェクトの解説

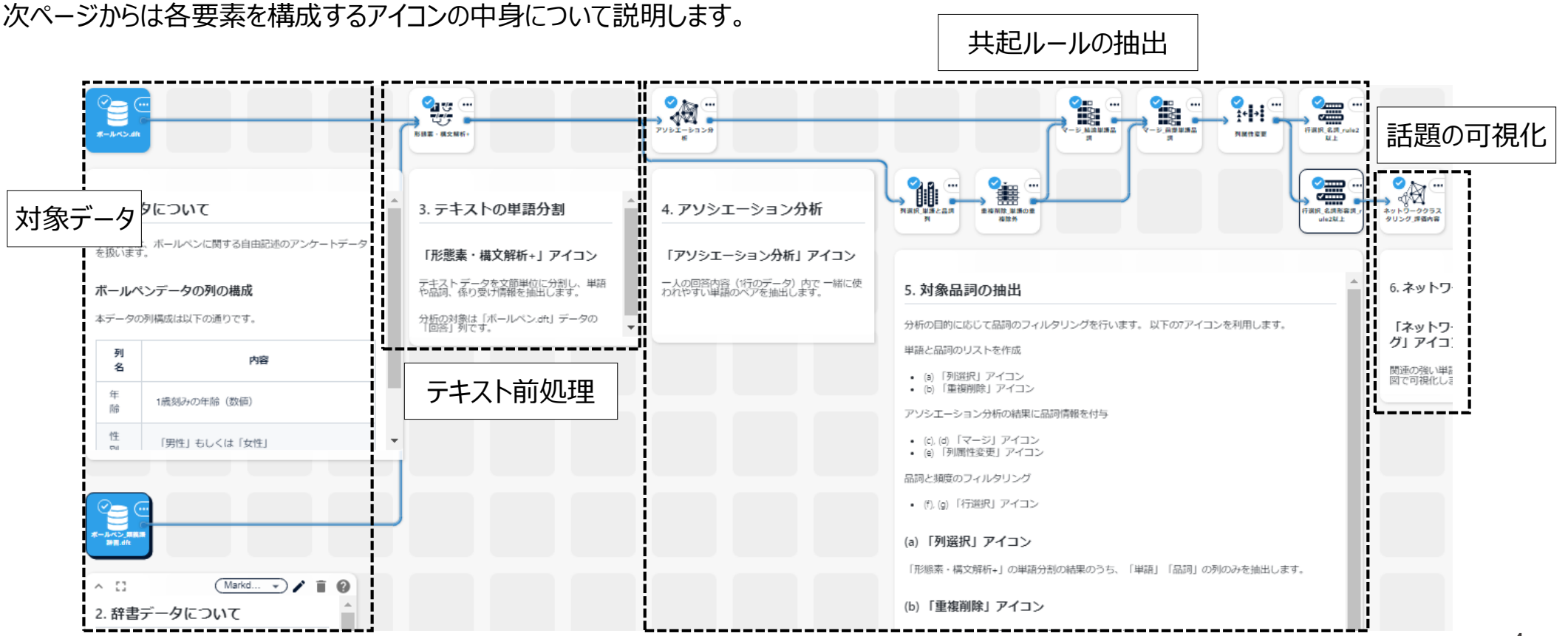
プロジェクト概観

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4つの要素に分けられます。

本サンプルプロジェクトでは、共起関係にある名詞と名詞を抽出することで、ボールペンの機能と性質の関係について把握し、名詞と形容詞・形容動詞を抽出することで、ボールペンに対する評価や印象を表す単語の関係を視覚的に確認します。

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

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

2. ボールペン_類義語辞書.dft

テキストの分割処理実行時に、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_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、を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。

【前処理としてテキストの分割のみを行い、フィルタリングを用いない理由】
多くの場合、テキストデータの前処理として、品詞や頻度によるフィルタリングを行いますが、今回は行いません。
事前にフィルタリングを行う場合には、該当する単語が含まれない行の情報がすべて除外されてしまいます。アソシエーション分析では、「全データ数（全行数）」を考慮した値が算出されるため、行の情報が除外された状態では結果が変化してしまいます。今回は「全文章数」を一定にするために、事前のフィルタリングなどの前処理は行いません。

プロジェクト解説 ―共起ルールの抽出

4. アソシエーション分析

「アソシエーション分析」アイコンで、同一回答内で同時に使われやすい共起関係にある単語のペアを抽出します。

ここでは以下の方針でアソシエーション分析の設定を行っています。

- 単語群による話題を把握したい
  対象は「replaced」列 (replacedは活用処理・置換処理された単語で、分析単位に適する)

- 一人ずつの回答ごとに、一緒に使われやすい単語群を抽出したい
  一人の回答は行単位で記載されているため、「オプション設定」のキー列の設定を「RowID」とする

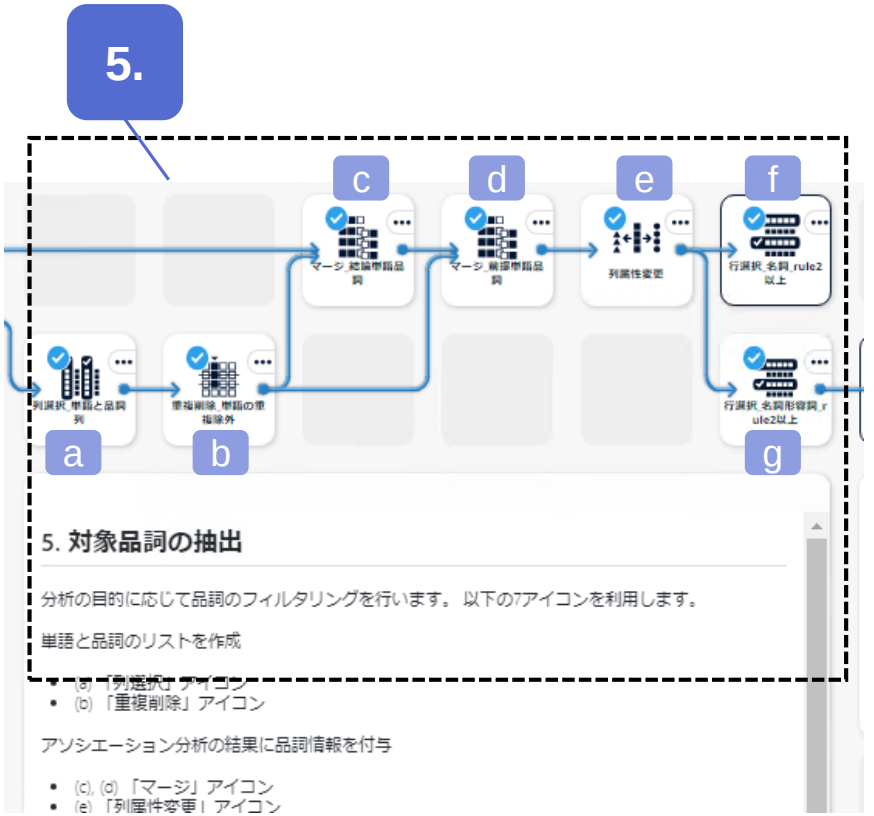
プロジェクト解説 ―共起ルールの抽出

5. 対象品詞の抽出

アソシエーション分析の結果は、前提・結論単語ともに様々な品詞を含み、そのままでは解釈がしにくいことが多いです。そのため、分析の目的に応じて、品詞によるフィルタリングを行います。

ここでは、(a)単語と品詞のリストを作成し(b)重複を削除して、(c,d)「マージ」アイコンでアソシエーション分析の結果に品詞列を追加し、(e)列名を修正して、最後に(f,g)「行選択」アイコンを利用して次の2パターンの品詞のフィルタリングを行っています。

- 同時に語られやすい機能・性能を把握する
  前提・結論ともに「名詞」の単語
- 評価の対象と評価の内容を把握する
  前提：名詞、結論：形容詞・形容動詞



プロジェクト解説 ―話題の可視化

6. ネットワーク図による可視化

「ネットワーククラスタリング」アイコンを用いて、関連の強い単語同士を1クラスタ（塊）とするようなネットワーク図を描画します。

話題を構築する単語群がひと塊のネットワークで描画されるため、視覚的に話題を把握することが可能です。

アウトプットの説明

アウトプット –名詞どうしのアソシエーション分析の結果

- 「行選択_評価項目_rule2以上」アイコンでは、アソシエーション分析の結果のうち、前提・結論単語ともに名詞であり、共起ルール数が2以上の単語のペアを抽出しています。名詞に絞ることで、レビュー内で語られる項目（機能や性能など）の関連の強いものを抽出することができます。
  さらに、より関係の強い単語から確認できるよう、「confidence」列は降順、と「antecedents」列は昇順になるように行の並べ替えを行っています。
- ここでは、「書き心地」を気にしている人は「ノック式」も気にしていることが多い、ということが分かります。書き心地に関する機能改善を行うのであれば、「ノック式のボールペン」についても考慮したほうが良いかもしれません。
- 機能改善を検討する際に、単一の機能だけでなく複数の機能向上を目指すべき、などの検討を行うことができるようになります。

行選択_名詞_rule2以上-result　列数: 12 行数: 13

No.	antecedents Category	consequents Category	confidence Float	support Float	lift Float
1	リフィル	交換	100.000000	2.000000	50.000000
2	交換	リフィル	100.000000	2.000000	50.000000
3	学校	色	100.000000	2.000000	14.285714
4	学校	ノート	100.000000	2.000000	20.000000
5	書き心地	ペン	100.000000	2.000000	12.500000
6	書き心地	ノック式	100.000000	2.000000	10.000000
7	筆圧	力	100.000000	2.000000	14.285714
8	キャップ式	ノック式	71.428571	5.000000	7.142857
9	キャップ	ノック式	66.666667	2.000000	6.666667
10	一本	色	66.666667	2.000000	9.523810
11	水性	油性	66.666667	2.000000	22.222222
12			7	2.000000	22.222222
13			0	3.000000	8.571429

前提単語　結論単語　信頼度

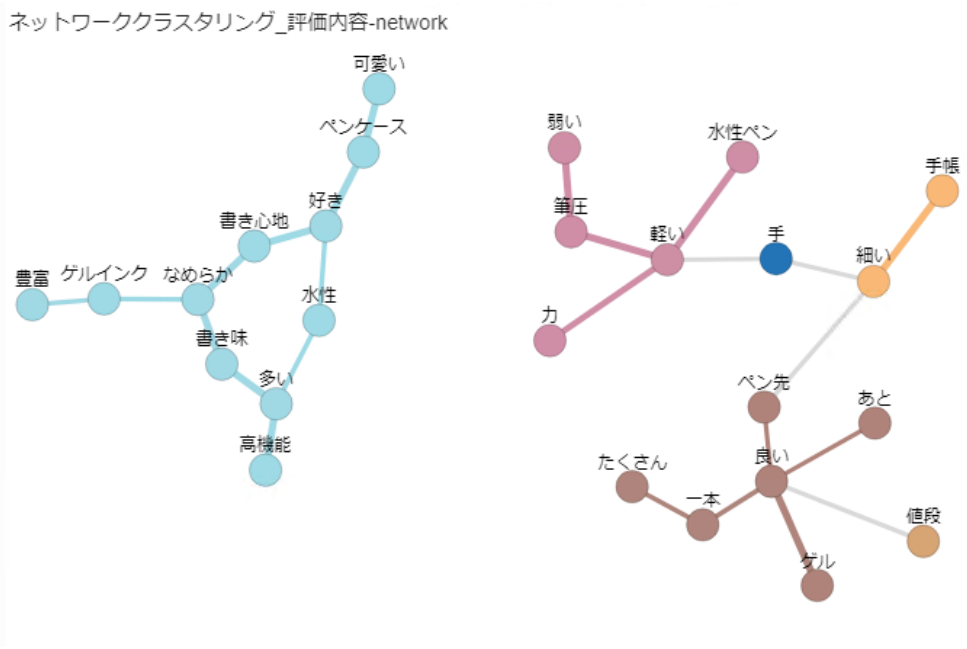
アウトプット – ネットワーク図による可視化

「ネットワーククラスタリング」アイコンは、アソシエーション分析の結果の「信頼度」を単語間の関連の強さとし、この関連の強さをリンクとするネットワークのクラスタリングにより、関連の強い単語群をひとかたまりとする話題を抽出します。

いくつのクラスタリングに分けるか、は試行錯誤が必要な部分で、

- 単独のノードが多くない
- 解釈のしやすさ
- 説明のしやすさ

などがクラスタ数を決定する目安となります。



テキストデータ全体でどのような話題があるのかを視覚的に把握することができます。

「書き味」や「書き心地」の「なめらか」さや、「筆圧」の「軽い」ボールペンが言及されていることが分かる。

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

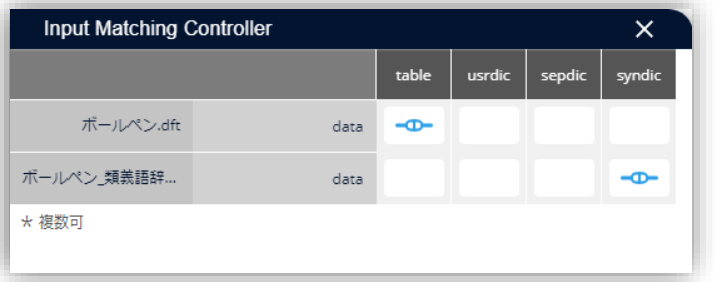
アイコン – 形態素・構文解析+

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、分割処理の対象の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列のみ指定が可能です。

アイコン – アソシエーション分析①

インプット設定

アソシエーション分析を適用するテーブルの設定を行います。

テキストの分割処理結果の単語表である「result」テーブルを指定します。

変数選択

- **分析対象列**

  関連を見たい単語の列を指定します。ここでは「replaced」を選択します。

アイコン – アソシエーション分析②

オプション

キー列・親子関係の設定

- **キー列**

  指定した値が同じ行を同一レコードとして扱います。

  「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 – 列選択_単語と名詞列

インプット設定

アソシエーション分析実行時に落とされてしまう、単語の品詞情報を保持します。対象テーブルは「result」です。

選択列

保持する情報を絞り込みます。今回は「単語」と「品詞」情報を保持するため、「replaced」列と「pos」列を抜き出します。

不要な列を除くことで、以降の分析フローにおいて、データ量が増大することを防ぐことができます。

アイコン – 重複削除_単語の重複除外

キー列

データの重複を除外します。テキストの分割処理結果から抜き出した単語と品詞のリストは、単語の情報が重複するため、「replace」「pos」のペアの重複を除外します。

重複行のうち残す行数

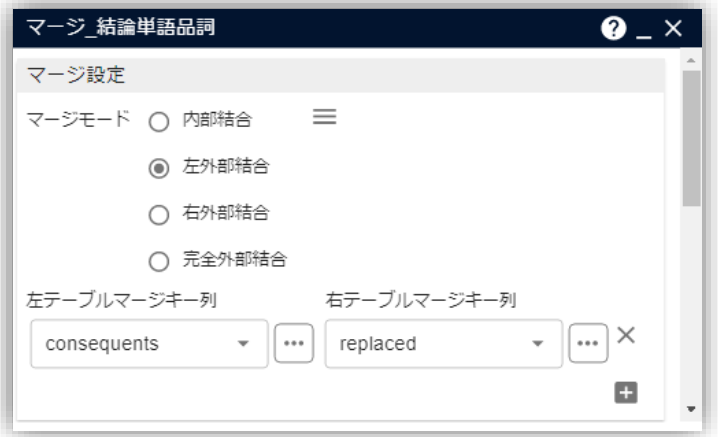
重複行のうちデータに保持する行数を指定します。ここでは「1」と設定します。

アイコン – マージ_結論単語品詞

インプット設定

アソシエーション分析と重複削除（単語と品詞のペア）の結果を紐づけます。

マージ設定



- **マージモード**

  紐づけを行う際の結合方法を指定します。ここでは、「アソシエーション分析」結果（インプット設定「left」）はすべて保持しますが、「重複削除」結果（インプット設定「right」）は該当する情報のみを保持するため、「左外部結合」を指定します。

- **左テーブルマージキー**

  インプット設定の「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「consequents(結論)」を指定します。

- **右テーブルマージキー**

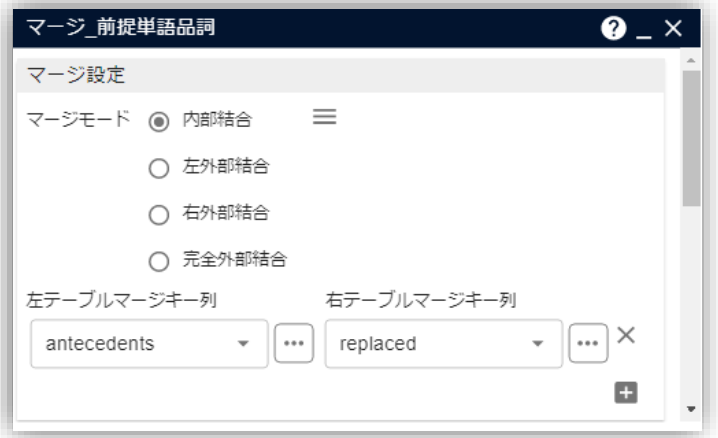
  インプット設定の「right」で指定したテーブルにおいて紐づけのキーとなる列を指定します。ここでは「replaced」を指定します。

アイコン – マージ_前提単語品詞

インプット設定

アソシエーション分析と重複削除（単語と品詞のペア）の結果を紐づけます。

マージ設定



- **マージモード**

  紐づけを行う際の結合方法を指定します。ここでは、「アソシエーション分析」結果（インプット設定「left」）と「重複削除」結果（インプット設定「right」）の両方に該当する情報のみを保持するため、「内部結合」を指定します。

- **左テーブルマージキー**

  インプット設定の「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「antecedents(前提)」を指定します。

- **右テーブルマージキー**

  インプット設定の「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「replaced」を指定します。

アイコン – 列属性変更

対象列

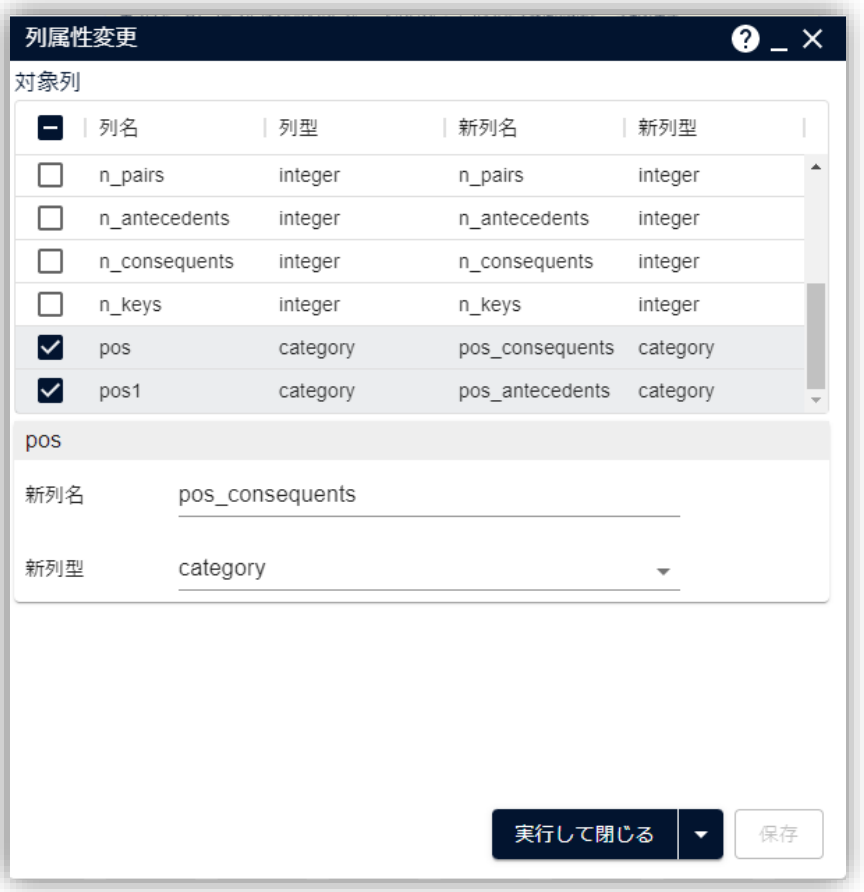
列名を変更するために以下の設定を行います。

- **pos**

  新列名を「pos_consequents」、新列型を「category」に設定します。

- **pos1**

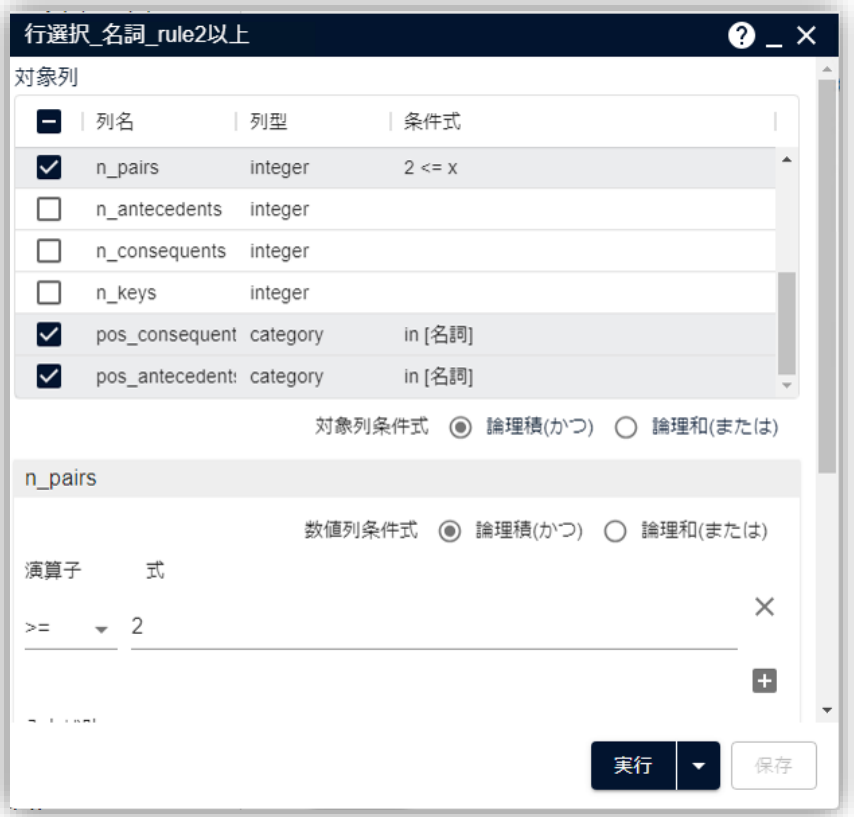
  新列名を「pos_antecedents」、新列型を「category」に設定します。



アイコン – 行選択_名詞_rule2以上

対象列

条件を指定したい列を選択します。ここでは、ルールの出現回数・前提単語品詞・結論単語品詞の条件を設定します。



- **n_pairs**

  ルールの出現回数を表す「n_pairs」列の条件を指定します。

  ここでは、ルール数2以上の単語を抽出します。

  - 演算子: >=
  - 式: 2

- **pos_antecedents**

  前提単語の品詞を「名詞」のみにします。

  - 演算子: 一致する
  - 式: 名詞

- **pos_consequents**

  結論単語の品詞を「名詞」のみにします。

  - 演算子: 一致する
  - 式: 名詞

アイコン – 行選択_名詞形容詞_rule2以上

対象列

条件を指定したい列を選択します。ここでは、ルールの出現回数・前提単語品詞・結論単語品詞の条件を設定します。

- **n_pairs**

  ルールの出現回数を表す「n_pairs」列の条件を指定します。

  ここでは、ルール数2以上の単語を抽出します。

  - 演算子: >=
  - 式: 2

- **pos_consequents**

  前提単語の品詞を「形容詞」「形容動詞」のみにします。

  - 演算子: 一致する
  - 式: 形容詞・形容動詞

- **pos_antecedents**

  結論単語の品詞を「名詞」にします。

  - 演算子: 一致する
  - 式: 名詞

アイコン – ネットワーククラスタリング_評価内容

変数選択

ネットワークのノード（丸）とエッジ（矢印やリンク）となる列を指定します。

- **ノード1**

  ネットワークのノードとなる列を指定します。ここでは前提単語である「antecedents」を指定します。

- **ノード2**

  ネットワークのノードとなる列を指定します。ここでは結論単語である「consequents」を指定します。

- **エッジ重み**

  ネットワークのエッジの重みとなる列を指定します。ここでは関係の強さを表す指標の一つである「confidence」を指定します。

- **エッジ重みを [類似度/非類似度] とみなす**

  エッジの重みを類似度とみなすか、非類似度とみなすかを指定します。「confidents」は値が大きいほど要素間の関連が強いとみなすため「類似度」とみなします。

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。

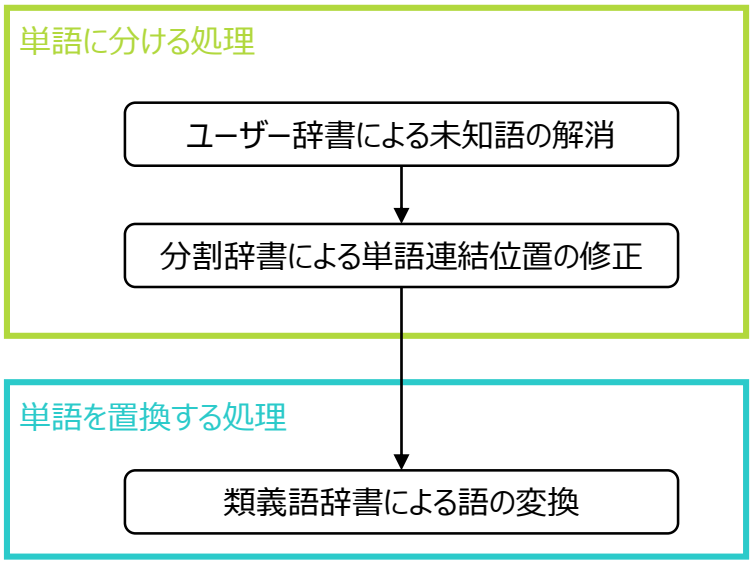
分割辞書

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。



技術情報：アソシエーション分析について

アソシエーション分析は、「こういう前提があるときに、同時にこういう結論が発生する」という物事同士の関連性を、「前提→結論」というアソシエーションルールという形で抽出します。

とくにテキストデータの分析においては、同一文章内で「同時に使われる」＝「共起する」単語の関係に着目し、共起のしやすさである指標値を算出します。共起指標をもとにクラスタリングすることで、共起しやすい単語群の抽出と、それらの単語群からなる話題を把握する分析です。

単語同士の共起だけでなく、「単語」と「属性」の共起を見ることで、属性ごとの傾向を把握することも可能です。

本プロジェクトでは、「アソシエーション分析」アイコンを用いて、共起指標として「信頼度」と呼ばれる値を算出しています。

前提単語w_A に対する 結論単語w_B の信頼度はは以下の形で計算できます。

$$\text{信頼度}(w_A, w_B) = \frac{w_A \text{と} w_B \text{が同時に出現する頻度 (共起)}}{w_A \text{の頻度}}$$

単語w_Aが出現するときに必ず単語w_Bが出現する場合は、信頼度の値は 1 になります。

例えば、次の3文に対して信頼度を計算すると、右のようになります。

・ボールペン 書く
・ボールペン 買う
・サインペン 書く

信頼度(買う, ボールペン) = 1/1 = 1
信頼度(ボールペン, 買う) = 1/2 = 0.5

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