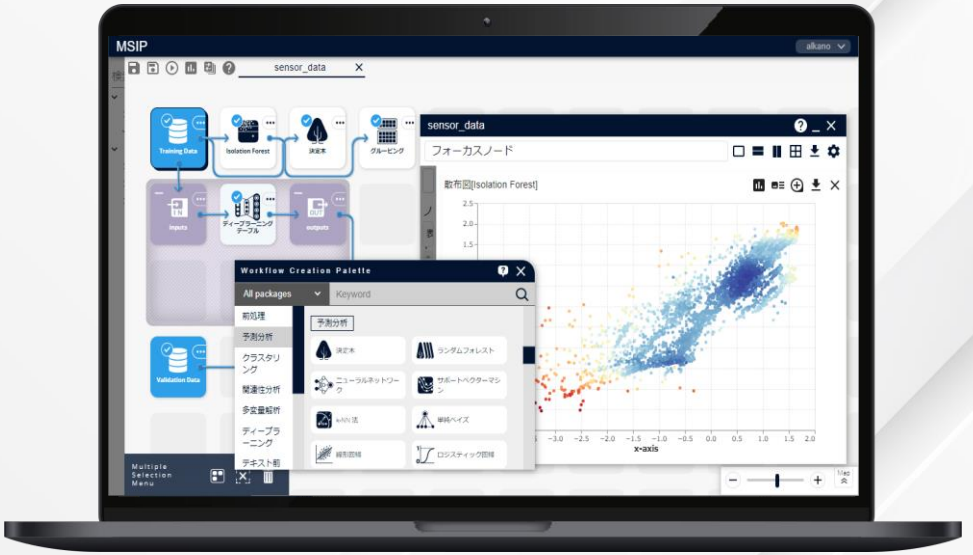
TextExtension

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

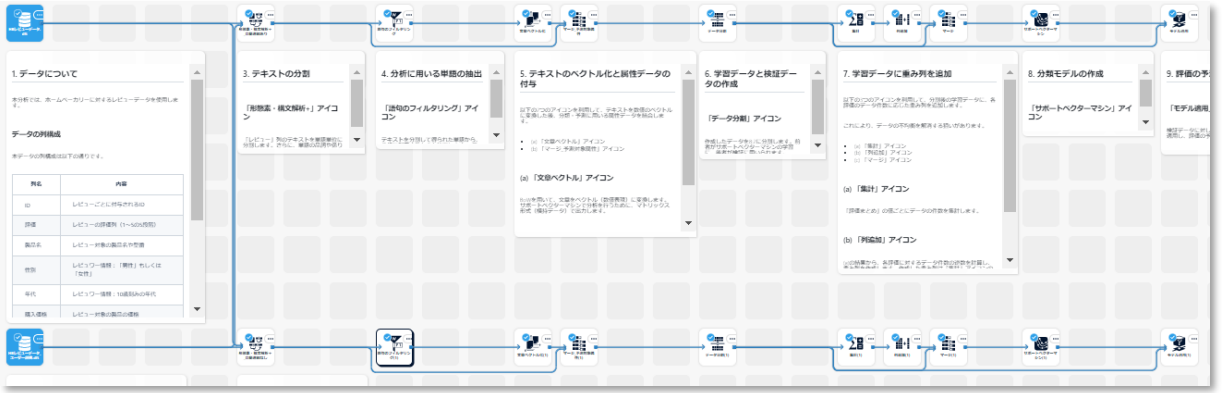
このプロジェクト について

こんな方におすすめします

- テキストデータや属性データを利用して、予測モデルを作成したい方
- 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、テキストデータを利用して、機械学習の有名な手法であるサポートベクターマシンで分類・予測モデルを作成する一連の流れを紹介します。
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となりますので、これを応用することで様々な機械学習手法を用いてテキストデータを扱うことができます。
また、付随する属性データも機械学習で利用し、分析することができます。



プロジェクトの解説

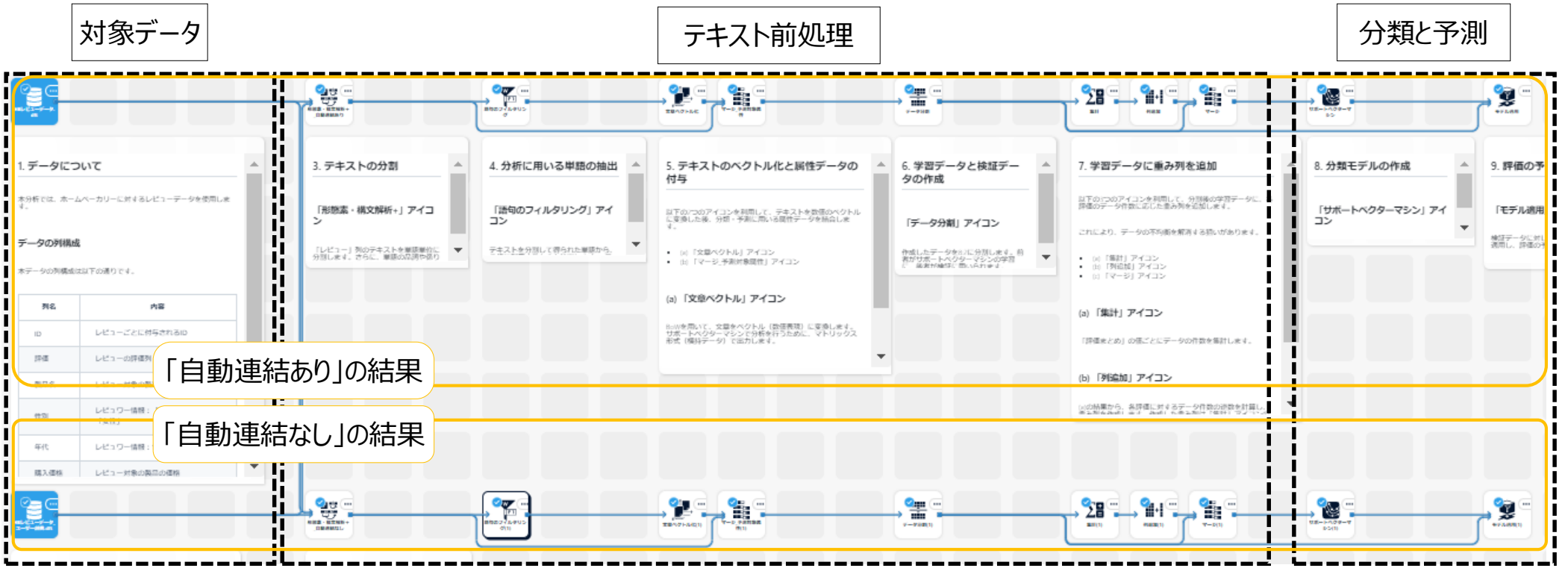
プロジェクト概観

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。

本サンプルプロジェクトでは、テキストの分割の粒度を変えて予測を行っています。「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

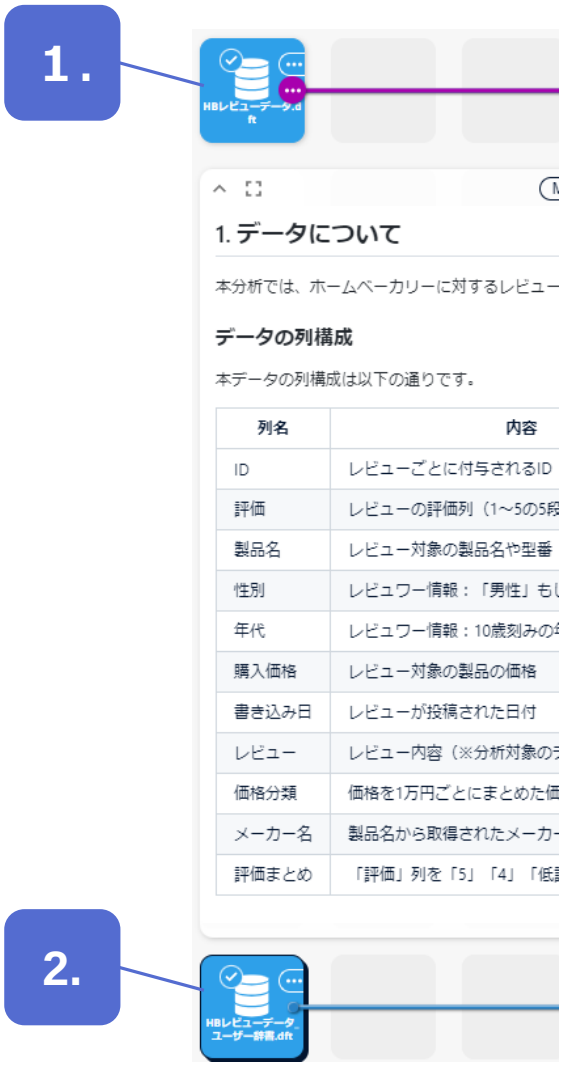
ECサイトで様々なホームベーカリーに対してのレビューをまとめた、仮想の口コミデータです。MSIPの上では、csv形式のデータをdft形式に変換し、シナリオ編集エリア上に配置して使用します。1行が1レビューに対応します。

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。

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。

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（1～5の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
価格分類	価格を1万円ごとにまとめた価格帯
メーカー名	製品名から取得されたメーカー名
評価まとめ	「評価」列を「5」「4」「低評価」にまとめた列



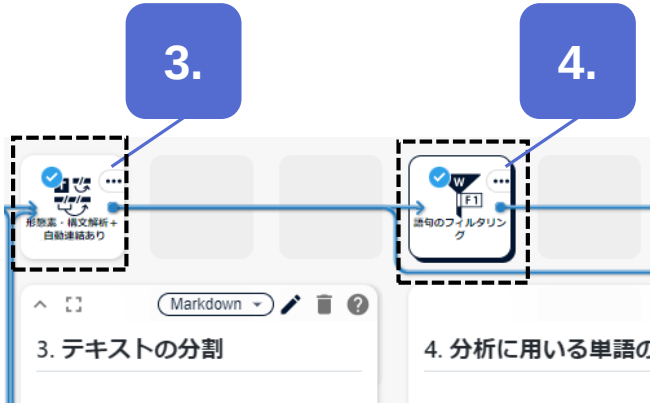
HBレビューデータ_ユーザー辞書.dft-data　列数: 2 行数: 2

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

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、テキストデータを文節単位に分割し、単語や品詞、係り受け情報を抽出します。詳細は補足情報の『テキストのベクトル化: 「自動連結あり」と「自動連結なし」の違い』をご参照ください。



今回は、「HBレビューデータ.dft」データの「レビュー」列に入っているテキストデータが分割の対象です。

4. 分析に用いる単語の抽出

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、その中でベクトルの次元数を調整するため、頻度2~100、文字数2以上の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

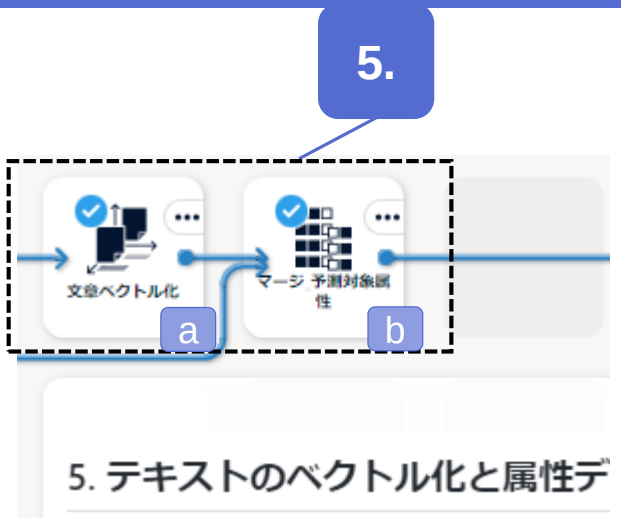
5. テキストのベクトル化と属性データの付与

テキストを数値のベクトルに変換した後、分類・予測に用いる属性データを結合します。

(a) ここでは、BoW（Bag of Words）を用いて、「どの単語が何件出現しているか」という数値のベクトル表現を獲得します。

また、分析アイコンであるサポートベクターマシンの入力として利用するために、横持データであるマトリックス形式のベクトル表現を作成します。

(b) (a)の結果に、分類・予測に用いたい属性を紐づけます。属性データには、サポートベクターマシンの予測対象（目的変数）である「評価まとめ」が含まれます。



「文書ベクトル化」アイコンで単語列に指定した「replaced」列の単語が列名となり、その単語の出現回数がその列の値になります。

プロジェクト解説 ― テキスト前処理 学習データと検証データ

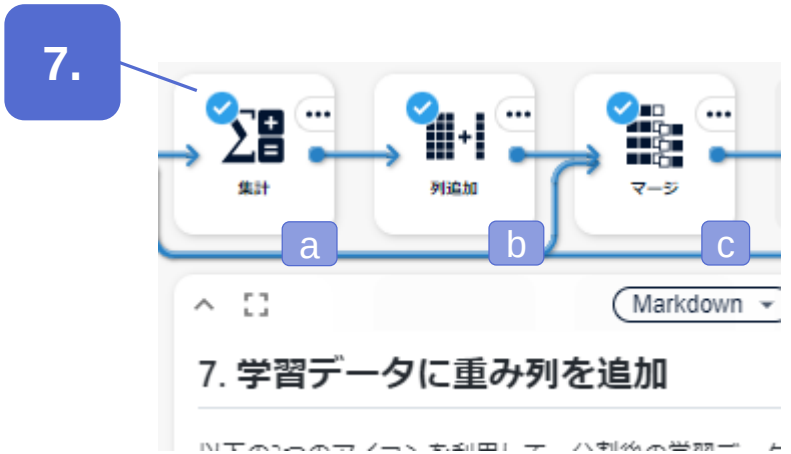
6. 学習データと検証データの作成

学習データと検証データを8:2に分割します。データ分割の比率は、用いるデータ件数や問題設定により、おおよそ5:5～9:1の範囲で違いはありますが、7:3や8:2が多く用いられる傾向にあります。

プロジェクト解説 ― テキスト前処理 学習データと検証データ

7. 学習データに重み列を追加

分割後の学習データに、各評価のデータ件数に応じた重み列を追加します。これにより、データの不均衡を解消する狙いがあります。
重み列は、各評価の件数を「集計」アイコンで集計し、その逆数を重み値として採用することで、サポートベクターマシンの重み列指定に利用することができます。



(a) 「評価まとめ」の値ごとにデータの件数を集計します。
(b) (a)の結果から、各評価に対するデータ件数の逆数を計算し、重み列を作成します。作成した重み列は「集計」アイコンの結果に追加されます。
(c) (a)と(b)の結果を結合し、学習データに重み列を追加します。

プロジェクト解説 ― 分類モデルの作成

8. 分類モデルの作成

評価を予測する予測モデルを構築します。予測モデルには決定木やランダムフォレスト、ディープラーニングなど様々な手法がありますが、今回はデータ数が多くないため、ある程度の件数でも精度が見込めるサポートベクターマシンを採用しています。

プロジェクト解説 ― 予測

9. 評価の予測

サポートベクターマシンで構築したモデルを検証データに適用し、予測値を出力します。今回はデータ分割後の検証データを利用していますが、（同じ形式の）新規の口コミデータがある場合、それを入力として評価値の予測を行うことができます。

アウトプットの説明

アウトプット（サポートベクターマシン）

「サポートベクターマシン」アイコンの結果を確認すると、predicted_value、評価まとめ、元データの順で列があることがわかります。

「predicted_value」列が予測モデルにより予測された値（予測値）、「評価まとめ」列が元データにあった値（正解値）です。

学習データに対するモデルの適用なので、全体として正しく予測できていることがわかりますが、28,29 行目では、誤った予測がされているデータがあることも確認できます。

サポートベクターマシン-result　列数: 1,083 行数: 290

No.	predicted_value Category	評価まとめ Category	RowID Integer	もともと Integer	焼き立て Integer	オーブン Integer
15	5	5	349	0	1	0
16	5	5	348	1	0	0
17	低評価	低評価	347	0	0	0
18	4	4	346	0	0	0
19	4	4	344	0	0	0
20	5	5	343	0	0	0
21	5	5	342	0	0	0
22	5	5	341	0	0	0
23	5	5	340	0	0	0
24	5	5	339	0	0	0
25	5	5	338	0	0	0
26	5	5	337	0	0	0
27	5	5	336	0	0	0
28	4	5	335	0	0	0
29	4	5	334	0	0	0
30	5	5	333	0	0	0
31	低評価	低評価	332	0	0	0
32			329	0	0	0

予測値　正解値

アウトプット（モデル適用）

「モデル適用」アイコンの結果を確認すると、predicted_value、元データの順で列があることがわかります。「predicted_value」列は予測モデルにより予測された値（予測値）です。「サポートベクターマシン」アイコンと異なり、未知の新規データに対して予測を行っているので、正解値の列はありません。

今回のデータの場合、データ分割により交差検証の形で分析フローが構成されているため、「モデル適用」アイコンの代わりに、「予測精度検証」アイコンを利用して予測モデルの精度を確認することができます。

モデル適用-result　列数: 1,081 行数: 73

No.	predicted_value Category	RowID Integer	もともと Integer	焼き立て Integer	オーブン Integer	故障 Integer
1	5	4	0	0	0	0
2	5	6	0	0	0	0
3	低評価	7	0	0	0	0
4	5	9	0	1	0	0
5	5	10	0	0	0	0
6	5	13	0	0	0	0
7	5	21	0	0	0	0
8	5	27	0	0	0	0
9	4	30	0	0	0	0
10	5	43	0	0	0	0
11	5	46	0	0	0	0
12	5	52	0	0	0	0
13	4	53	0	0	0	0
14	4	54	0	0	0	0
15	5	56	0	0	0	0
16	4	57	0	0	0	0
17	5	58	0	0	0	0
18	4	62	0	0	0	0

予測値

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

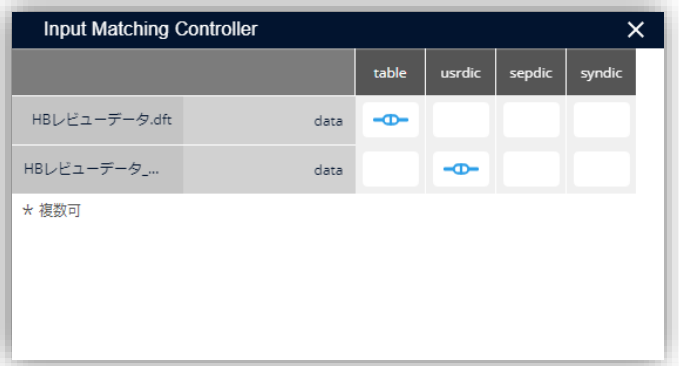
アイコン – 形態素・構文解析+

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、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

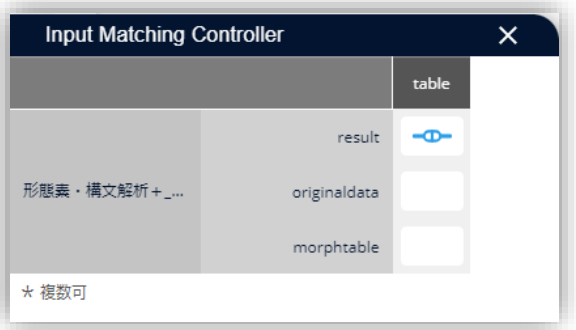
- **テキスト列**

  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。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①

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 –語句のフィルタリング②

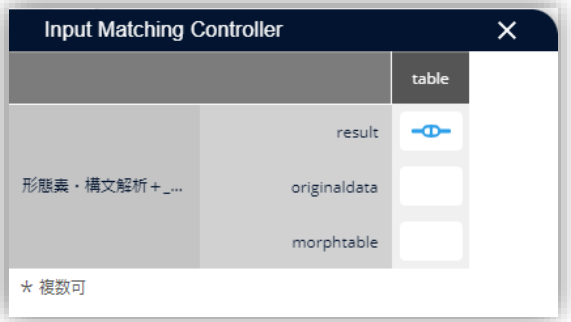
文字数フィルタ

- **対象列**

  文字数を指定したい単語列を指定します。ここでは「replaced」列を選択します。

- **最小文字数を設定**

  対象列のうち、指定した文字数以上の単語を抽出します。ここでは、2文字以上の単語を抽出します。



アイコン – 文章ベクトル化①

変数選択

- **単語列**

  ベクトル化の対象とする単語列を指定します。ここでは、類義語辞書を適用した後の単語の列である「replaced」列を選択します。

- **キー列**

  ベクトルを生成するための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②

モデルの選択

「形態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。

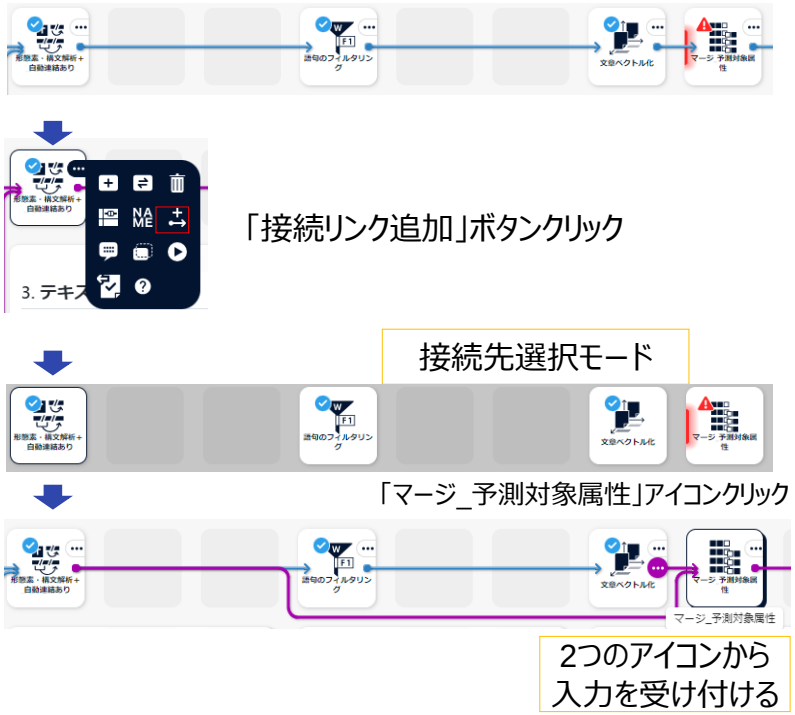
サポートベクターマシンの入力はマトリックス形式のため、ここでは「マトリックス形式」を選択します。

アイコン – マージ_予測対象属性①

接続リンクを追加

既存のノードへ接続リンクを追加するには、ノードメニュー内の「接続リンク追加」ボタンをクリックします。クリックすると、ワークフロー画面が接続先選択モードとなり、接続可能なノードがフォーカスされた状態になります。この状態で接続先のノードをクリックすることで、そのノードへの接続リンクが新しく追加されます。

このようにして、「マージ」アイコンに2つの入力を設定します。



インプット設定

ベクトル化したテキストデータと、分類・予測対象である目的変数を紐づけます。

「文章ベクトル化」アイコンの結果テーブルである「result」を左テーブルとして「left」に、「形態素・構文解析+」アイコンの結果のうち、オリジナルテーブルである「original_data」を右テーブルとして「right」に指定します。

アイコン – マージ_予測対象属性②

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**

  紐づけを行う際の結合方法を指定します。ここでは、「文章ベクトル化」結果（インプット設定「left」）全体に、元のテキストデータの属性データテーブル（インプット設定「right」）の該当する情報のみを紐づけるため、「左外部結合」を指定します。

- **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 – データ分割

分割

データ全体を学習用/検証用 に分割します。

今回は、学習用：検証用＝ 8:2 に設定します。

- **学習用**

  学習データとして保持する割合を指定します。

- **検証用**

  検証データとして保持する割合を指定します。

詳細設定

- **層対象列**

  指定した列の層（種別）ごとに分割を行います。ここでは設定しませんが、値が不均衡な列を層対象列にすることで、元データの分布と同じ分布の学習データを作成します。

アイコン – 集計

インプット設定

重み付け列を作成したいデータを指定します。ここでは学習データに重み付けを行うため「training」テーブルを指定します。

分析設定

指定した列の統計量を計算します。ここではキーに指定した列の件数を求めたいので、集計項目は選択しません。

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**

  ここでは件数を集計します。

集計キー

重み付け対象となる列を指定します。複数列の指定が可能です。

ここでは「評価まとめ」列を指定します。

アイコン – 列追加

列追加

- **追加する列名**

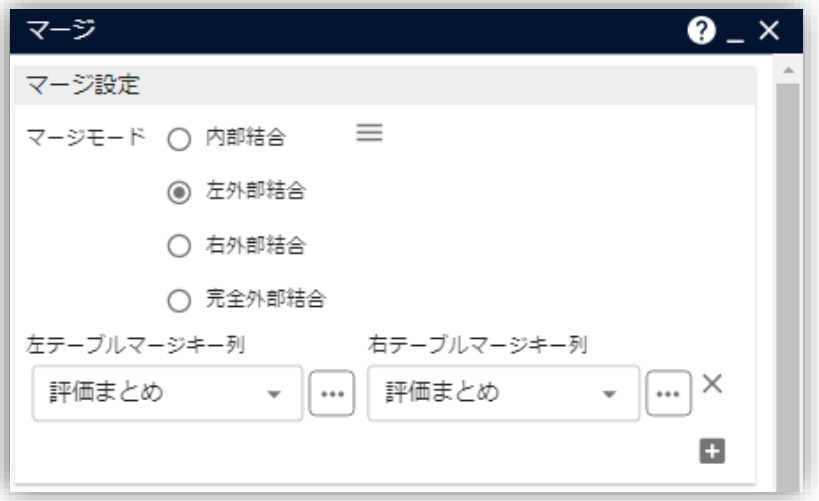
  新しく作成する列名を任意で指定します。

- **計算式**

  新しい列に入力する値を指定します。ここでは「集計」アイコンで得られた各属性の件数列をもとに、件数の逆数を新しく値とするために、計算式に 1/table["件数"] と記述します。

No.	評価まとめ Category	件数 Integer
1	5	181
2	4	58
3	低評価	51

アイコン – マージ



入力設定

学習データと重み列を紐づけます。

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、学習データ（インプット設定「left」）全体に重み列テーブルを紐づけるため、「左外部結合」を指定します。

- **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「評価まとめ」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「評価まとめ」を指定します。

アイコン – サポートベクターマシン

変数設定

- **目的変数**

  分類・予測対象とする列を指定します。ここでは「評価まとめ」を指定します。

- **説明変数**

  分類・予測の説明変数を指定します。ここでは、単語列をすべて選択します。性別や年代といった属性列は選択しません。

SVM

- **重み付け**

  学習データの行ごとに重み付けを行います。ここでは作成した重み列を指定します。

  - 列指定：「重み列」

補足情報

技術的な情報や利用規約について

テキストのベクトル化:「自動連結あり」と「自動連結なし」の違い

サポートベクターマシンで分類するためには、テキストデータを数値データに変換して入力する必要があります。文章を単語に分割して、単語毎の頻度を集計しますが、単語の切れ目が異なると、最終的に得られるベクトルも変わります。「自動連結なし」では、「自動連結あり」に比べて、単語が細かく分割されるため、その分だけ同じ単語を持つ文章が増えて、類似性を表現しやすいベクトルを作成できます。例えば、次の3文に対して、「自動連結あり」と「自動連結なし」でベクトル化すると下記のようになります。「自動連結なし」で作成したベクトルは、「パン」という共通部分がありますが、「自動連結あり」では共通部分がなくベクトルからは類似していると判断が難しくなります。

(A) コッペパンを食べる
(B) カレーパンを作る
(C) メロンパンが好き

自動連結あり

	コッペパン	食べる	カレーパン	作る	メロンパン	好き
A	1	1	0	0	0	0
B	0	0	1	1	0	0
C	0	0	0	0	1	1

自動連結なし

	コッペ	パン	食べる	カレー	作る	メロン	好き
A	1	1	1	0	0	0	0
B	0	1	0	1	1	0	0
C	0	1	0	0	0	1	1

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。

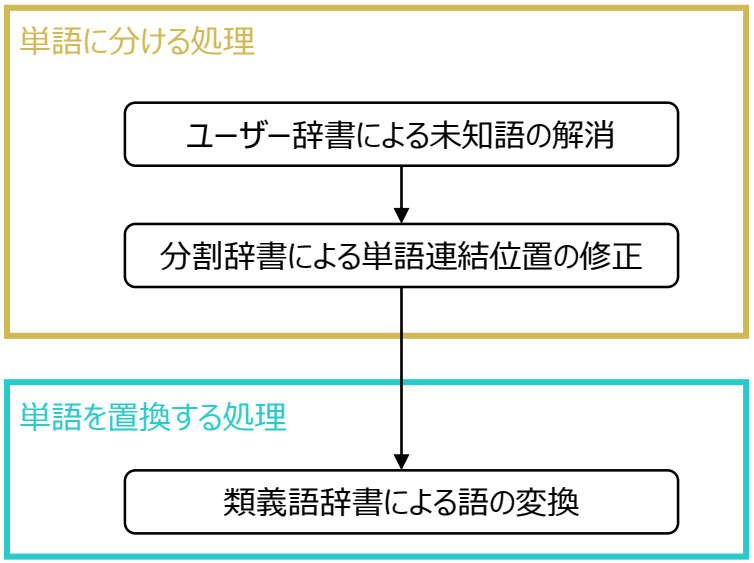
分割辞書

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