新機能紹介

NTT DATA
Trusted Global Innovator

テキスト処理機能ベータ版

のご紹介

株式会社NTTデータ数理システム

テキスト処理（ベータ版）とは

テキスト処理（ベータ版）は当社製品の分析プラットフォームMSIP上でテキストデータを分析するための機能群です。

こんな機能をご提供いたします！（※2023/3 リリース予定）

1. 自由記述の文章を機械的に扱えるようにする「分かち書き」や各種分析機能に適用するための「フィルタリング」などの**前処理機能のアイコン群**
2. MSIP上でテキスト処理を行う作業をサポートする**テクニカルサンプルプロジェクト**

こんなことが実現できます！

1. アンケートデータからよくあるご要望を抽出する
2. 不具合情報やニュース記事などの分類・予測を行う（Alkano連携）

利用イメージ

テキストデータを「分かち書き」「フィルタリング」したうえで、集計等の基礎分析や統計解析・機械学習の分析手法を組み合わせて利用することで、様々な目的に沿った分析をしていただけます。

前処理機能のアイコン群　テキストカテゴリ化

キーワードをもとに、見たい観点や話題でテキストをまとめて、観点・話題の0/1表を作成します。

■ テキストの概要把握
■ 各種機械学習の入力データ（説明変数の利用）
■ 分類・予測分析における教師データの作成

などに有効にご活用いただけます。

▼ テキストデータ

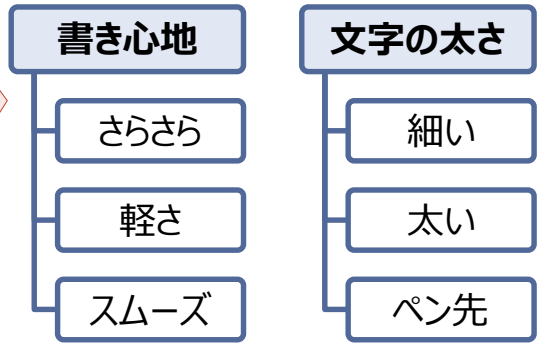
：
- 軽くさらさらかけるといいな
- 書きやすさとかペン先のスペックよりピンときたもの
- グリップもペン先も細目のものが好みです。

：

テキスト	書き心地	文字の太さ
軽くさらさらかけるといいな	1	0
軽さとかペン先のスペックよりピンときたもの	1	1
グリップもペン先も細目のものが好みです。	0	1

▼ ルール

ルールを決める
「さらさら」「軽さ」などのキーワードがあれば「書き心地」というカテゴリにまとめる、などのルールを指定します。



指定したルールに沿って単語やテキストをまとめる

話題・観点のキーワードがテキスト中に現れていれば 1, なければ 0 という値の列が追加されます。話題・観点の有無、すなわちカテゴリデータを作成します。

テクニカルサンプルプロジェクト

Alkano のワークフローは、自分で分析を組み立てていかなければならないので大変そう、というご心配のお声をよく聞きます。そこでテキストデータの分析を行う**分析ワークフローのサンプル**とそのプロジェクトの**解説資料（pdf）**をご用意いたしました。

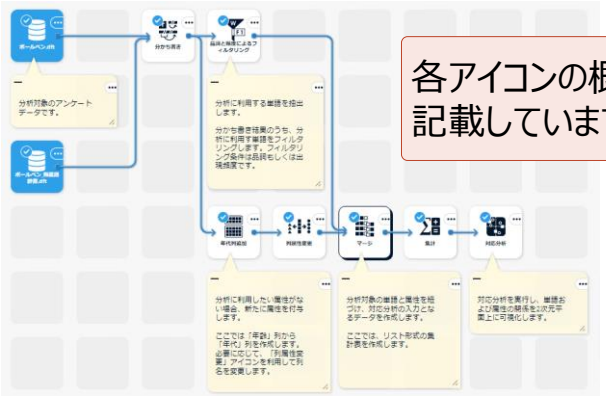
こんな方にお勧めです！

- Alkano (MSIP) を使ったテキスト分析をはじめて行う方
  - Alkano(MSIP)で数値・カテゴリデータと合わせて、テキストデータの分析を始めたい方
  - これまで Text Mining Studio を利用されていて、分析フローを組み立てるのが初めての方

テクニカルサンプルプロジェクトの特徴

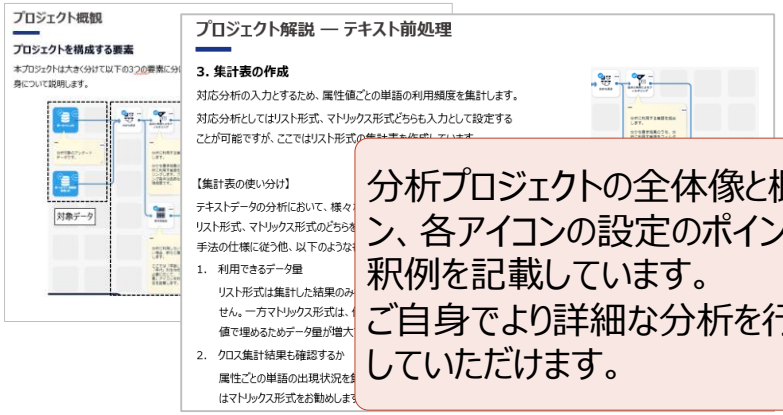
- サンプルデータをお手元のデータに差し替えてご利用いただくことで、データ分析を始めやすい
- 説明資料には分析・設定のポイントを明記しているので、お手元での試行錯誤のポイントが分かりやすい

▼ 分析フローのプロジェクト



各アイコンの概要をメモに記載しています。

▼ 分析フローの解説資料



分析プロジェクトの全体像と概要、活用シーン、各アイコンの設定のポイントや結果の解釈例を記載しています。
ご自身でより詳細な分析を行う際の参考にしていただけます。

前処理機能のアイコン群一覧

分かち書き

テキスト（文章）を単語や文節単位に分割し、品詞や係り受け（単語の意味的なつながり）の情報を解析します。
人が見て解釈できる単位での文節を判定する「自動連結」機能や、否定や要望など記述者の主観等を表現する「態度表現」を付与する機能を有しています。

分かち書き結果のフィルタリング

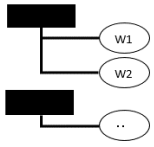
品詞や単語の出現頻度、文字列や文字数によるフィルタリングを行い、分析に有用な単語を抽出します。
【フィルタリング例】

品詞フィルタ：名詞や動詞など、単体で意味を持つ単語を抽出します。
頻度フィルタ：頻度2以上かつ頻度上位5件を除外など、程よくまんべんなく使われている単語を抽出します。
文字列フィルタ：製品名の一部など、「この文字列を含む単語」とターゲットを絞ることができます。
文字数フィルタ：3文字以上と設定し、複合語などより「専門用語」らしい単語を抽出します。

重み算出

属性（テキストに紐づく付加情報）と組み合わせ、属性ごとの単語の重み（重要度）を算出します。
属性ごとに重みが大きな単語を見ることで属性の傾向を把握することができます（TMS特徴分析結果相当）。
さらに重みの値を各種機械学習の入力（説明変数）として分類・予測に利用いただくことも可能です。

テキストカテゴリ化



テキストデータ内に出現する単語や係り受け表現のキーワードをもとにテキストデータのカテゴリ化（グループ分け）を行います。
※詳細は次ページに記載します

NTT DATA

Trusted Global Innovator