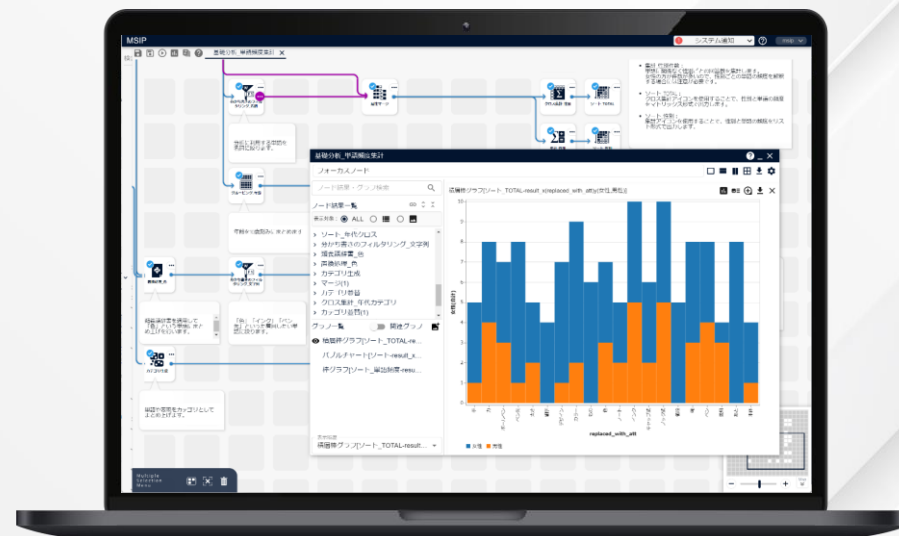


Alkano / BayoLinkS 向けテキスト機能拡張パック

TextExtension

バージョン 1.0.0
機能のご紹介

2024年3月末リリース



株式会社 NTTデータ数理システム

TextExtension 1.0.0 搭載機能のご紹介

前処理

形態素・構文解析+

日本語・英語で書かれたテキストを分析するための、リッチな前処理

- テキストデータの分析に適した単位に分割し、各単語の品詞や文法的な係り受け関係を抽出します
- 面倒な辞書の整備をしなくても、辞書登録なしでちょうどいい単位で単語を抽出する「**自動連結**」機能を搭載しています。
- 「**態度表現**」を自動で取得することで、語尾からとれる否定や要望といったニュアンスを抽出できます

入力

今年こそは東京スカイツリーに行きたい。



「形態素・構文解析+」アイコンで、テキストデータを文節単位に自動分割



出力



前処理

語句のフィルタリング・語句の重み算出

テキストデータから不要な語句を取り除き、重要な語句のみを抽出

- 「語句のフィルタリング」機能を用いると、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できます。
- 「語句の重み算出」機能では、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。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。

「語句の重み算出」で利用可能な指標値

- | | |
|---------------|-----------------------|
| ✓ 補完類似度 | ✓ Fisherの直接確率 |
| ✓ χ^2 乗値 | ✓ Yates補正 χ^2 乗値 |
| ✓ Jaccard係数 | ✓ TF-IDF |
| ✓ Dice係数 | ✓ cosine |
| ✓ 相互情報量 | |

TextExtension 1.0.0 搭載機能のご紹介

前処理

ルールベース文書ラベル付与

日本語・英語で書かれたテキストを分析するための、リッチな前処理

特定の観点や話題でテキストをまとめあげるためのルールを定義し、各文書が各観点を満たすか否かの真偽値表を作成することができます。テキストの概要把握、特徴量作成、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。

入力（テキストデータ）※

- いろんな種類のカラーを使えること。仕事用の油性ペンとプライベート用の4色ペンをいつも持ち歩いています
- ノートをきれいにまとめる用にパステル系のカラーがたくさん入ったものを選んでます。あと細字の方がかわいくていい。
- スムーズに書けるものの方がいいですね。太字のものの方が見やすいです。
- 軽くてサラサラと書ける水性ペンをずっと使っています。

※実際には、「形態素解析」アイコンまたは「形態素・構文解析+」アイコンで前処理をした後のデータを入力します。

ラベル付与規則

ラベル名: 色

(1) 「カラー」

(2) 「カラバリ」

(3) 「色」

上記のルールが 1 個以上成り立つ場合、ラベル「色」を付与します

ルールを決める

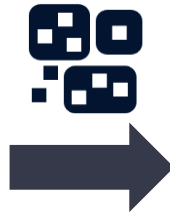
「カラー」「カラバリ」「色」などのキーワードがあれば「色」というカテゴリにまとめ、などのルールを指定します。

をすべて含む

をすべて含む

+ ルール追加

カテゴリ生成



出力（観点表）

テキスト	色	水性_油性
いろんな種類のカラーを使えること。仕事用の油性ペンとプライベート用の4色ペンをいつも持ち歩いています	true	true
ノートをきれいにまとめる用にパステル系のカラーがたくさん入ったものを選んでます。あと細字の方がかわいくていい。	true	false
スムーズに書けるものの方がいいですね。太字のものの方が見やすいです。	false	false
軽くてサラサラと書ける水性ペンをずっと使っています。	false	true

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・観定のキーワードがテキスト中に現れていれば true, なければ false という値の列が追加されます。話題・観定の有無のラベルを付与したデータを作成します。

TextExtension 1.0.0 の特徴

NEW

わかりやすいチュートリアル

テキストデータの分析が初めての方でも、スムーズに分析をはじめていただけるよう、わかりやすいチュートリアルを用意しました。アンケートデータの分析を題材として一連の操作を体験していただくことで、TextExtension の操作に慣れながら、テキスト分析の基本的な考え方を理解していただけます。

TextExtension 1.0.0 ドキュメント

2.3. TextExtension チュートリアル

前へ | 次へ | 索引

前のトピックへ

2.2. TextExtension の機能

次のトピックへ

3. TextExtension リファレンスマニュアル

クイック検索

検索

目次

目次

1. TextExtension マニュアル

1.1. はじめに

1.1.1. TextExtension について

1.1.2. TextExtension マニュアルについて

1.1.3. サポートと制限事項

1.1.4. お問い合わせ

2. TextExtension ユーザーマニュアル

2.1. はじめに

2.1.1. 本マニュアルの構成

2.1.2. 改訂履歴

2.2. TextExtension の機能

2.2.1. テキストマイニング

2.2.2. テキスト分析機能

2.3. TextExtension チュートリアル

2.3.1. 分析の準備とデータの配置

2.3.2. テキスト前処理 (単語分割と情報付与)

2.3.3. 単語頻度の分析

2.3.4. 特徴単語の分析

2.3.5. 観点別のラベルを活用した分析

2.3.6. 係り受けを用いた分析

2.3.7. 最後に

3. TextExtension リファレンスマニュアル

3.1. テキスト前処理アイコン

3.1.1. 形態素・構文解析+

2.3. TextExtension チュートリアル

チュートリアルでは、MSIP 上で TextExtension の基本的な操作を行いながら、MSIP および TextExtension の機能を理解していただくことを想定しています。ここではアンケートの回答結果を用いて、TextExtension の四つのアイコン (「形態素・構文解析+」「語句のフィルタリング」「ルールベース文書ラベル付与」「語句の重み算出」) に関する操作について説明します。

具体的には、ボールペンの新商品を企画する立場を想定し、世間のニーズを把握するために「ボールペンを選ぶときに重視することは何ですか?」というアンケートを実施したと仮定します。アンケートから得られた回答結果のデータを基に、以下の分析を行っています。

(a) データの全体像を大まかに把握する 「単語頻度の分析」

(b) 性別による意見の違いを明らかにする 「特徴単語の分析」

(c) 観点ごとに意見をまとめて集計する 「観点別のラベルを活用した分析」

(d) データの内容を振り返る 「係り受けを用いた分析」

アンケート回答データ

単語頻度の分析

特徴単語の分析

観点別のラベルを活用した分析

係り受けを用いた分析

チュートリアルでは、最終的に下記のようなワークフローを構築します。

3.1.1. 文書のアップロードと書庫の作成

3.1.2. 書庫の確認

3.2. テキスト前処理アイコン

3.2.1. テキスト抽出

3.2.2. 形態素・構文解析+

3.2.3. 語句のフィルタリング

3.2.4. ルールベース文書ラベル付与

集計を実行する

MSIP の「集計」ノードを用いて、「形態素・構文解析+」の結果を出力します。

分析処理ノードを配置する

「テキスト前処理 (単語分割と情報付与)」で配置した「形態素・構文解析+」ノードに対して、「集計」ノードを追加します。

1. 「形態素・構文解析+」ノードの [...] にマウスカーソルを合わせます。

MSIP

検索

共有ワークスペース

テキスト分析ワークスペース

テキスト分析プロジェクト

テキスト分析ワークフロー

→ ノードメニューが表示されます。

メモ

ノードの上で右クリックしても、ノードメニューを表示できます。

2. [...] をクリックします。

MSIP

検索

共有ワークスペース

テキスト分析ワークスペース

テキスト分析プロジェクト

テキスト分析ワークフロー

ステップバイステップで
ていねいに操作を解説

豊富な図解で
わかりやすく解説

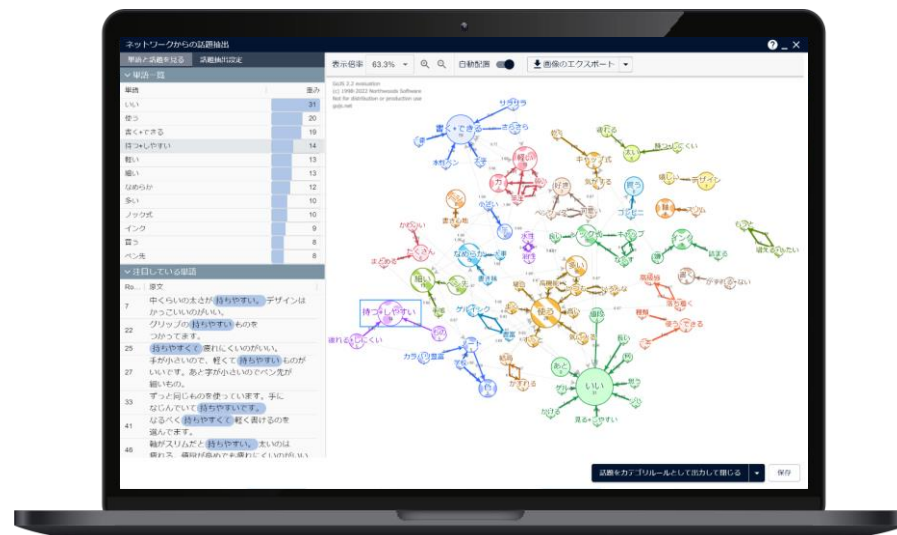
Alkano / BayoLinkS 向けテキスト機能拡張パック

TextExtension

バージョン 1.1.0 新機能のご紹介

2024年6月末リリース予定

2024年6月末リリース予定機能としてご紹介している機能は、すべて開発途中のものであり、今後予告なく変更される可能性があります。



株式会社 NTTデータ数理システム

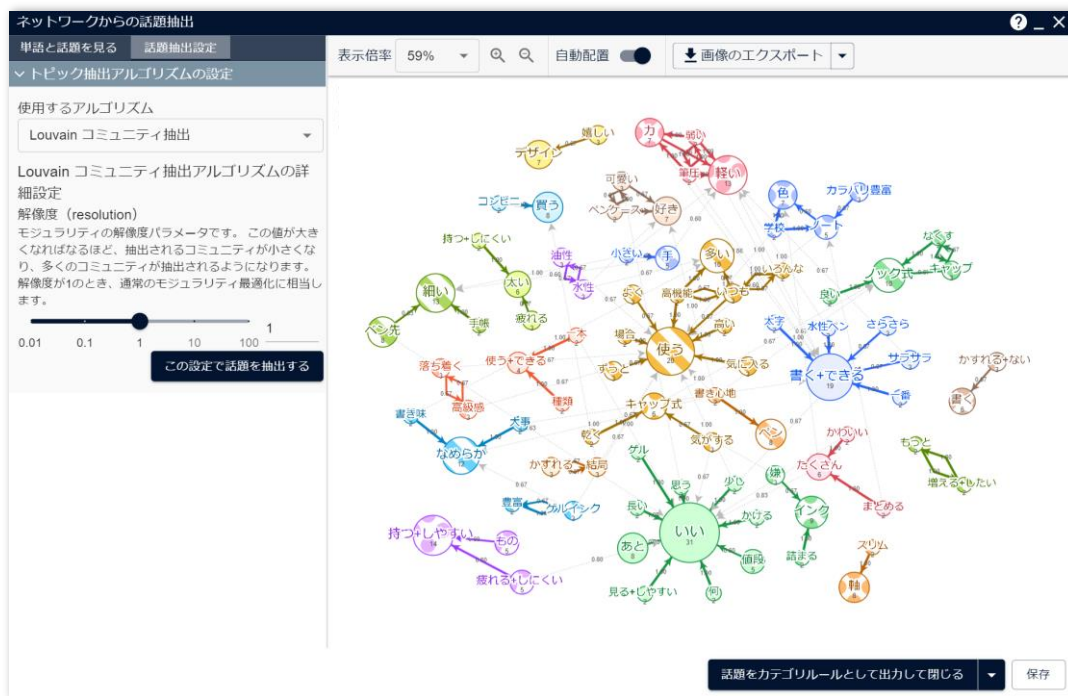
TextExtension 1.1.0 新機能のご紹介

2024年6月末リリース予定

共起ネットワーク分析

文書中の話題をグラフィカルなネットワーク図で把握

文単位に共起しやすい単語を、共起ネットワーク図としてグラフィカルに表示して整理できます。さらに、ネットワークをクラスタリングして色分けして表示することで、文書内にある話題を直観的に把握できます。



主な機能

✓ネットワーク図の出力

表示しているネットワーク図は、png形式で出力することができます。

✓ ラベル付与ルールの作成

クラスタリング結果は「○○という単語が現れていたら、●●という話題を表すラベルを付与する」という形式で、ラベル付与ルールを作成することができます。

作成したラベル付与ルールは、「ルールベース文書ラベル付与」アイコンに入力することで、文書にラベルを付与することができます。

抽出したラベル付与 ルールを適用

： (1) 目「豊富」をすべて含む

∴ (2)  「ゲルインク」 をすべて含む

上記のルールが 1 個以上成り立つ場合、カテゴリ「ゲルインク」を付与します。

[13] ラベル名: 詰まる

… (1)  「インク」 をすべて含む

… (2) 目「詰まる」をすべて含む

(3) 「嫌」をすべて含む

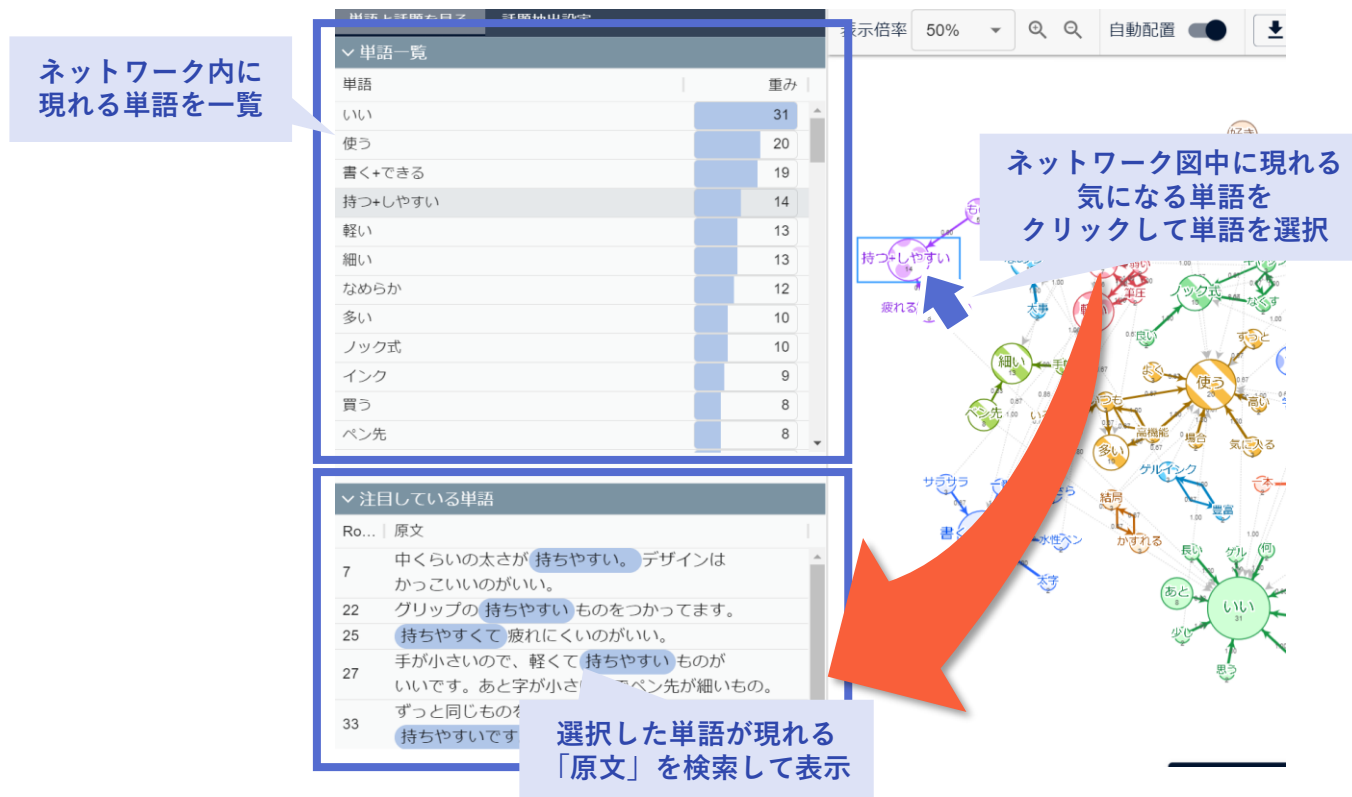
上記のルールが 1 個以上成り立つ場合、カテゴリ「詰まる」を付与します。

※掲載している画面は開発途中のものであり、今後予告なく変更される可能性があります。

共起ネットワーク分析 2024年6月末リリース予定

特徴 結果の解釈をサポートする単語一覧表や、原文検索機能を搭載

ネットワーク図に現れる単語を、頻度情報とともに一覧表示するため、どの単語が高頻度なのが一目でわかります。また、ネットワーク図に現れる単語が、原文（元データ）の中でどのように使われているかを確認できます。



※掲載している画面は開発途中のものであり、今後予告なく変更される可能性があります。

TextExtension 1.1.0 新機能のご紹介

2024年6月末リリース予定

新機能

Word／PowerPoint／PDF 文書取り込み

表形式以外の非定型文書も、簡単にテキスト分析可能に！

これまでMSIPへの入力が面倒だった、Word文書、PowerPointプレゼンテーション、PDF文書、テキストデータを直接MSIPへ取り込むことができます。
入力してテキストデータを抽出し、AlkanoやBayoLinkS、TextExtensionによる分析に適したテーブル形式に変換し、そのまま分析アイコンに入力できます。

文書データ群から書庫作成

ドラッグアンドドロップで簡単にアップロードできます。複数の文書をまとめた書庫という形式で保存されます。

分析に適した表形式へ

No.	Filetype	Filename	Page ID	Text Type	Text
73	PPTX	新機能紹介.pptx	6	main	主な機能
74	PPTX	新機能紹介.pptx	6	main	文書中の図表をグラフィカルなネットワーク図で把握
75	PPTX	新機能紹介.pptx	6	main	2024年6月末リリース予定
76	PPTX	新機能紹介.pptx	6	main	
77	PPTX	新機能紹介.pptx	6	main	
78	PPTX	新機能紹介.pptx	6	main	ネットワーク図の出力 表示しているネットワーク図は、png
79	PPTX	新機能紹介.pptx	6	main	
80	PPTX	新機能紹介.pptx	6	main	抽出したラベル付与ルールを適用
81	PPTX	新機能紹介.pptx	6	main	※掲載している画面は開発途中のものであり、今後予告なく変
82	PPTX	新機能紹介.pptx	7	main	
83	PPTX	新機能紹介.pptx	7	main	
84	PPTX	新機能紹介.pptx	7	main	選択した単語が現れる「原文」を検索して表示
85	PPTX	新機能紹介.pptx	7	main	
86	PPTX	新機能紹介.pptx	7	main	共有ネットワーク分析
87	PPTX	新機能紹介.pptx	7	main	
88	PPTX	新機能紹介.pptx	7	main	2024年6月末リリース予定
89	PPTX	新機能紹介.pptx	7	main	特徴
90	PPTX	新機能紹介.pptx	7	main	結果の解釈をサポートする単語一覧表や、原文検索機能を搭載
91	PPTX	新機能紹介.pptx	7	main	ネットワーク図に現れる単語を、関連情報とともに一覧表示
92	PPTX	新機能紹介.pptx	7	main	
93	PPTX	新機能紹介.pptx	7	main	ネットワーク図の出力 表示しているネットワーク図は、png

ファイル名やページ番号などをまとめた表形式にまとめることができます。
このままテキスト前処理アイコン群へつなげて分析をすることが可能です。

主な機能

✓ 詳細なテキスト抽出設定

ファイル中の改行や空白などをどのように扱うか、詳細に設定できます。文書の特徴に応じて、分析しやすいデータを柔軟につくることができます。

PPTX 設定オプション

☒ 改行を削除する

☒ 空白を削除する

☒ ノードから抽出する

☒ 表から抽出する

抽出単位: ファイル単位

連結文字: なし

DOCX 設定オプション

☒ 改行を削除する

☒ 空白を削除する

☒ コメントから抽出する

☒ ヘッダーから抽出する

☒ フッターから抽出する

☒ 表から抽出する

抽出単位: ファイル単位

連結文字: なし

※掲載している画面は開発途中のものであり、今後予告なく変更される可能性があります。

TextExtension 1.1.0 新機能のご紹介

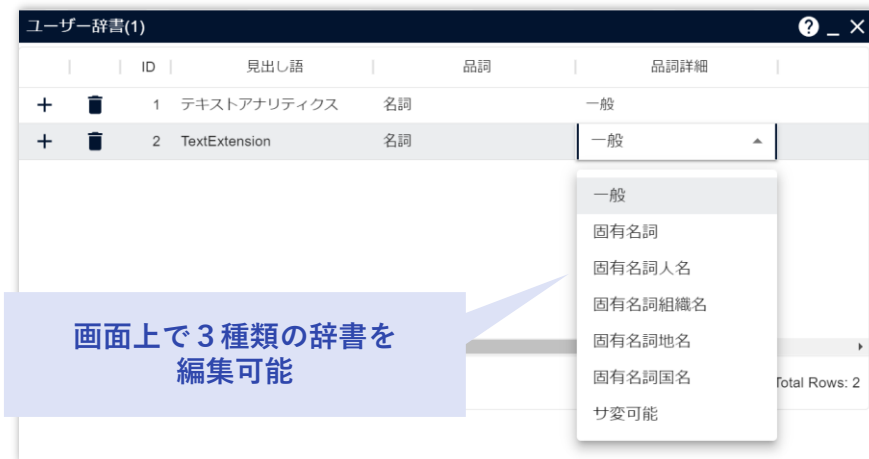
2024年6月末リリース予定

UP

辞書作成機能

MSIP の画面上で「形態素・構文解析+」アイコンに入力する辞書を直接作成・編集できます。

また、外部ファイル（json形式）として作成した辞書を、直接取り込み、さらに画面上で編集をすることもできます。



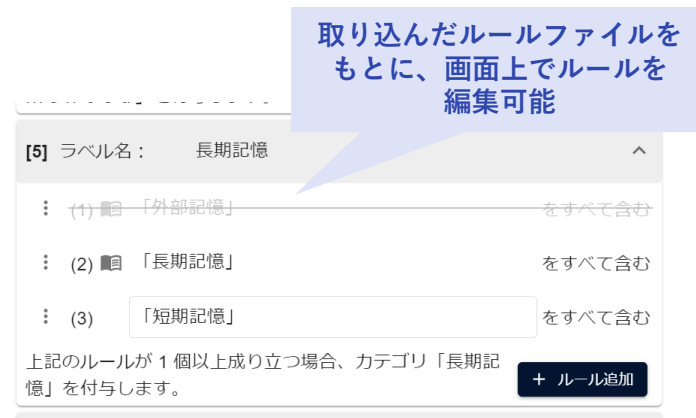
UP

ラベル付与ルールを読み込み

「ルールベース文書ラベル付与」アイコンでも、外部ファイル（json形式）として作成したルールファイルを取り込み、画面上で編集ができます。



外部ファイルとして記述した
ルールファイルを
アップロード・読み込み



※掲載している画面は開発途中のものであり、今後予告なく変更される可能性があります。

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/alkano/top.html>

株式会社 NTTデータ数理システム

NTT DATA NTT DATA Mathematical Systems Inc.