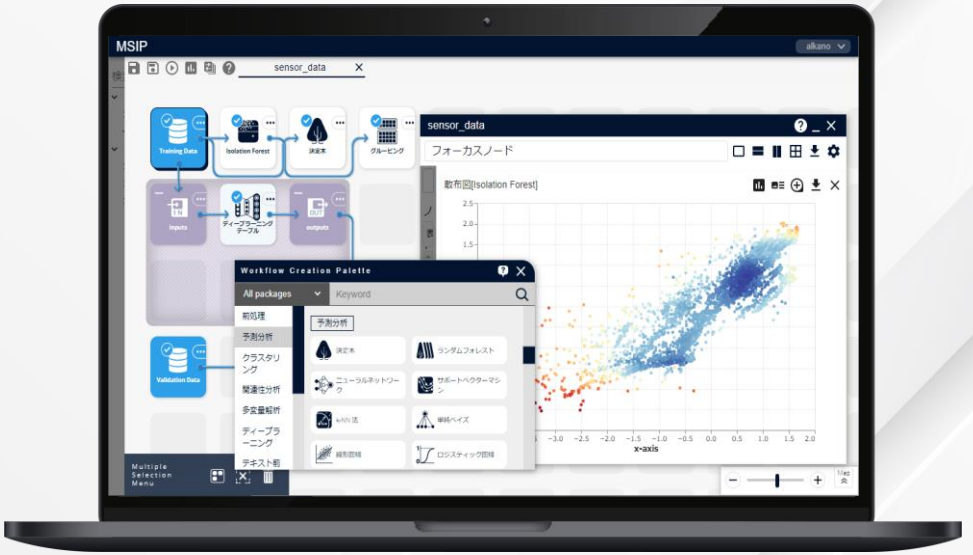
TextExtension

テクニカルサンプルプロジェクト

テキストの話題分析
共起ルールの抽出



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

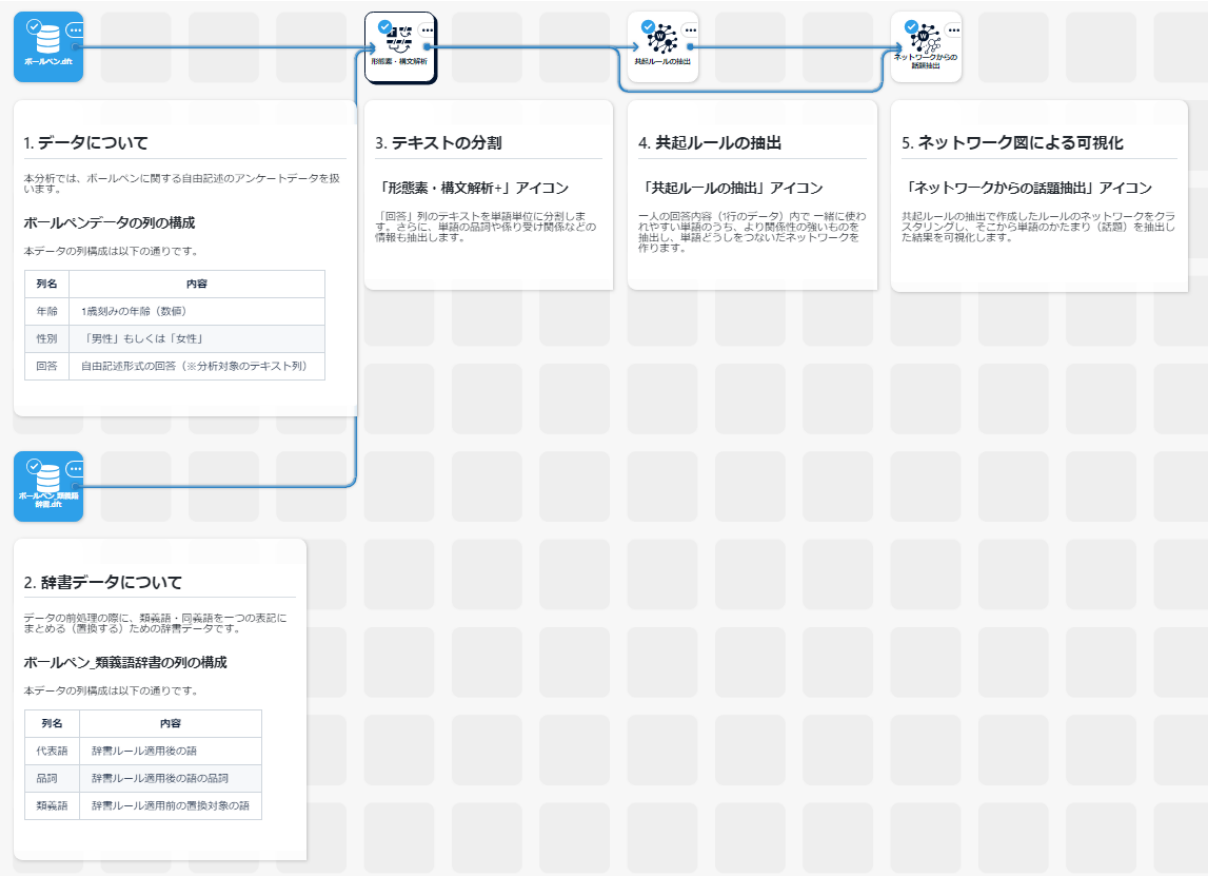
- **テキストデータから話題を抽出したい方**

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「共起ルールの抽出」と、共起ルールが抽出したネットワークを可視化できる「ネットワークからの話題抽出」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、一緒に使われやすい単語のうち、より関係性の強いものを抽出することで、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。共起ルールの抽出では、「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. 共起ルールの抽出

「共起ルールの抽出」アイコンで、同一行内あるいは同一文内で同時に使われやすい共起関係にある単語のうち、より関係性の強いものを抽出します。

ここでは以下の方針で共起ルールの抽出の設定を行っています。

- 広く話題をとりたい
  「行単位の共起ルールを抽出する」を選択する

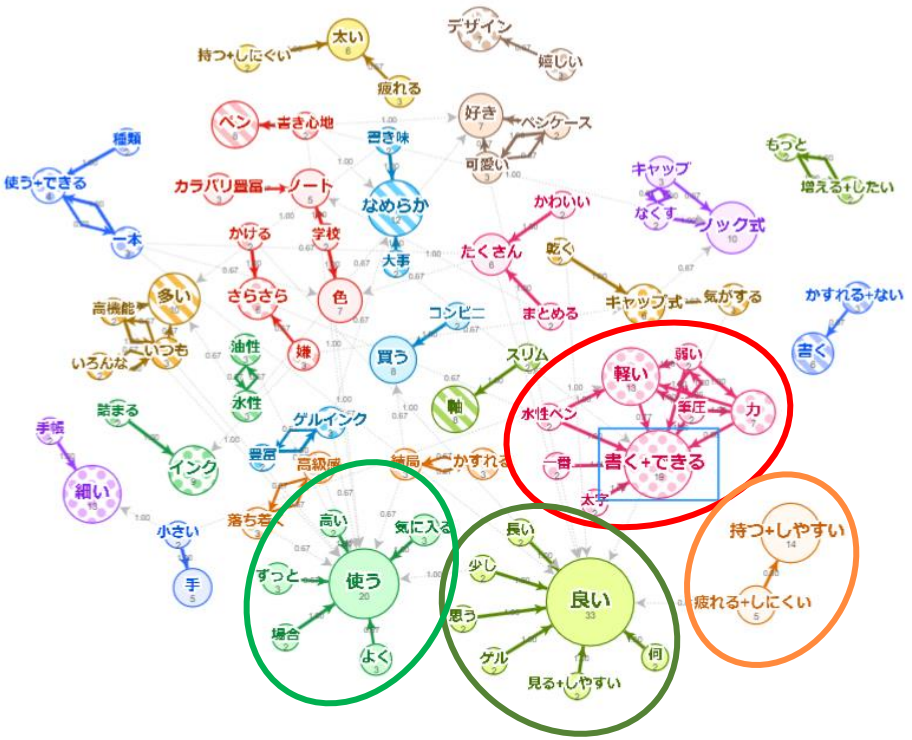
- 共起関係の強さを表す指標値は「信頼度」でランク付けを行う
  「信頼度」とは、共起する単語をそれぞれ「前提単語」「結論単語」として、「前提単語」があるときに「結論単語」がある割合を算出した値

プロジェクト解説 ―話題の可視化

5. ネットワーク図による可視化

「ネットワークからの話題抽出」アイコンを用いて、共起ルールの抽出で作成した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 –共起ルールの抽出の結果①

- 「共起ルールの抽出」アイコンでは、指標値を信頼度とし、上位 100 件のルールを表示しています。また、ランク付けの結果、同順位のルールが複数存在し指定した件数を超える際に、同順位のルールをすべて出力するように設定しています。ルールが多く、話題の把握がしづらいと感じた際には、上位表示件数を減らしたり、指定件数までしか出力しないように設定したりすることで、出力ルールを減らすことができます。
- ここでは、書きやすさに関して、「水性ペンは軽くてさらさら書ける」や「書き心地がなめらか」という話題があるようです。また、「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。

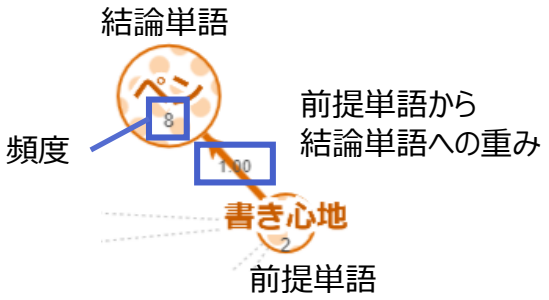
共起ルールの抽出-rules_table　列数: 15 行数: 111

No.	rule-name Category	frequency Integer	from-replaced Category	from-attitude Category	from-pos Category	from-pos_detail Category
65	高い-使う	2	高い	なし	形容詞	一般
66	かわいい-良い	2	かわいい	なし	形容詞	一般
67	かわいい-たくさん	2	かわいい	なし	形容詞	一般
68	書き心地-使う	2	書き心地	なし	名詞	一般
69	書き心地-好き	2	書き心地	なし	名詞	一般
70	書き心地-ノック式	2	書き心地	なし	名詞	一般
71	書き心地-ペン	2	書き心地	なし	名詞	一般
72	書き心地-なめらか	2	書き心地	なし	名詞	一般
73	かすれる-結局	2	かすれる	なし	動詞	一般
97	気がする-良い	2	気がする	なし	動詞	一般
98	気がする-キャップ式	2	気がする	なし	動詞	一般
99	嬉しい-デザイン	2	嬉しい	なし	形容詞	一般
100	かすれる+ない-書く	2	かすれる	否定	動詞	一般
101	ずっと-使う	2	ずっと	なし	副詞	
102	油性-使う	2	油性	なし	名詞	一般
103	油性-水性	2	油性	なし	名詞	一般
104	疲れる-太い	2	疲れる	なし	動詞	一般
105	いつも-使う	2	いつも	なし	副詞	
106	いつも-多い	2	いつも	なし	副詞	
107	いつも-いろんな	2	いつも	なし	副詞	
108	いつも-高機能	2	いつも	なし	副詞	
109	結局-買う	2	結局	なし	名詞	副詞可能
110	結局-使う	2	結局	なし	名詞	副詞可能
111	結局-かすれる	2	結局	なし	名詞	副詞可能

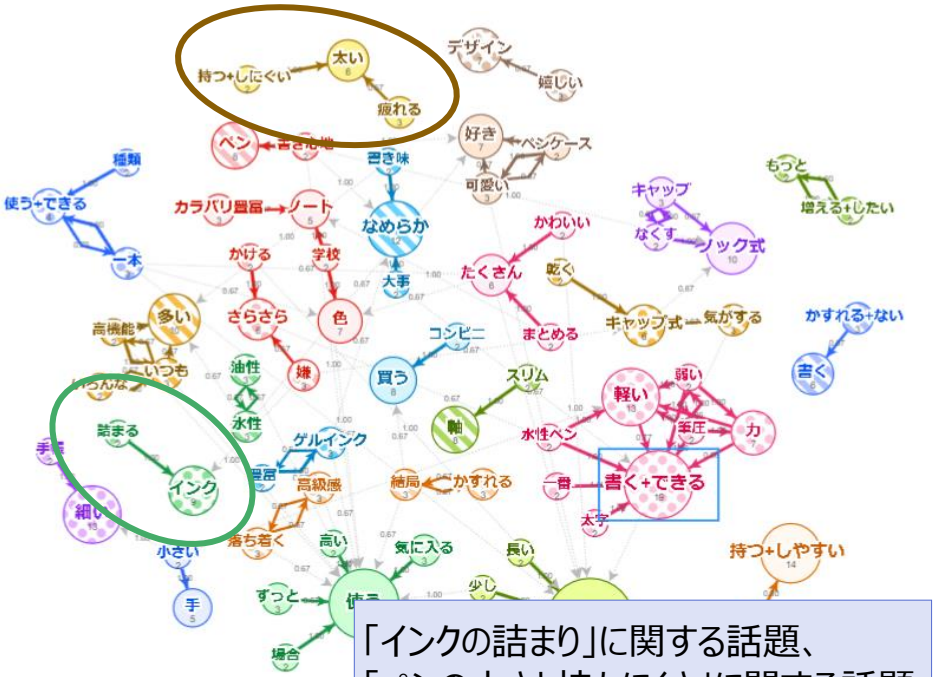
同順位のため、
100 位以降も出力されている

アウトプット –共起ルールの抽出の結果②

- 「共起ルールの抽出」アイコンでは、抽出されたルールのリストを、テキストネットワーク形式で出力します。テーブルやグラフではないため可視化画面で中身を開くことはありません。ネットワークを可視化するには、「共起ルールの抽出」アイコンから「ネットワークからの話題抽出」アイコンを繋ぎ、「ネットワークからの話題抽出」アイコンの画面上で確認できます。
- 前提単語や結論単語、それらの共起関係と重み情報を用いてネットワークを作成しています。



- 同じ色・模様の単語は、共起関係にある単語を表し、共起しやすい語のクラスタ（塊）なので、一つの話題を表していると考えられます。このように、ネットワーク図からテキスト中にどのような話題が存在するのかを把握できます。



「インクの詰まり」に関する話題、
「ペンの太さと持ちにくさ」に関する話題
などがあることが分かる

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

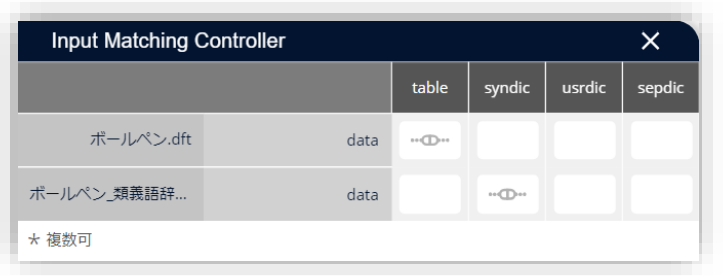
アイコン – 形態素・構文解析+

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、単語の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、類義語辞書を「syndic」、ユーザー辞書を「usrdic」、分割辞書を「sepdic」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**

  単語の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 – 共起ルールの抽出①

基本設定

- **共起ルールを抽出する単位**

  共起の単位を、行もしくは文のいずれかから選択します。ここでは行単位での抽出を指定します。

- **品詞および品詞詳細が異なる語を、異なる語とみなす。**

  同一表記でも、語句で品詞および品詞詳細が異なる単語を、違う単語とみなす場合にチェックをいれます。

- **態度表現が異なる語を、異なる語とみなす。**

  態度表現も考慮して、ルールの抽出を行うかどうかを指定します。

アイコン – 共起ルールの抽出②

ルールの抽出設定

- **指標値 信頼度(%) が上位 100 件のルールを表示する。**

  特定の指標値でランク付けを行い、ランク上位の共起ルールを抽出します。ここでは信頼度でランク付けを行い、上位 100 件の抽ルールを抽出するよう指定します。

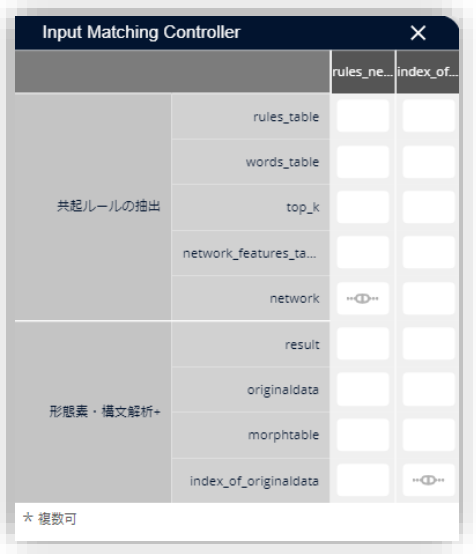
- **指定された件数を超えていても、同順位のものまではすべて出力する。**

  同順位のルールが複数存在し指定した件数を超える際に、同順位のルールをすべて出力するよう指定します。

アイコン – ネットワークからの話題抽出

「共起ルールの抽出」アイコンの結果を入力とし、画面を起動するだけで、ネットワーク図を描画します。

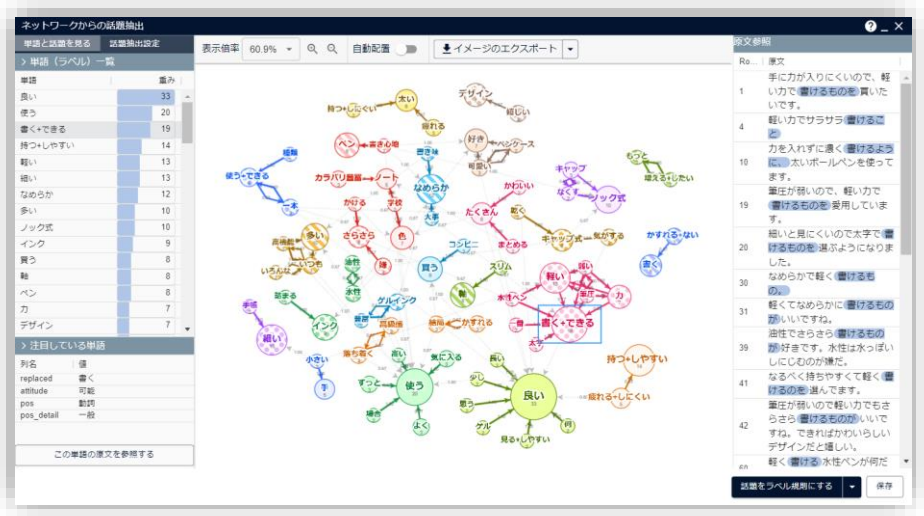
ネットワークを可視化した結果を見ながら、ノードの位置を移動させたり、話題の分割方法を変更したりと、設定画面上でインタラクティブに操作することができます。



インプット設定

「共起ルールの抽出」アイコンの結果のネットワークと「形態素・構文解析＋」アイコンの結果のインデックスの設定を行います。「共起ルールの抽出」アイコンの結果だけでネットワーク図を描画することはできますが、「形態素・構文解析＋」の結果も入力とすることで、原文を参照できるようになります。

原文を参照するには、「単語（ラベル）一覧」の単語もしくはネットワーク図のノードをクリックし、「注目している単語」エリアにある「この単語の原文を参照する」ボタンを押下すると、画面右側の「原文表示一覧」エリアに「注目している単語」を含む原文を表示します。



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。

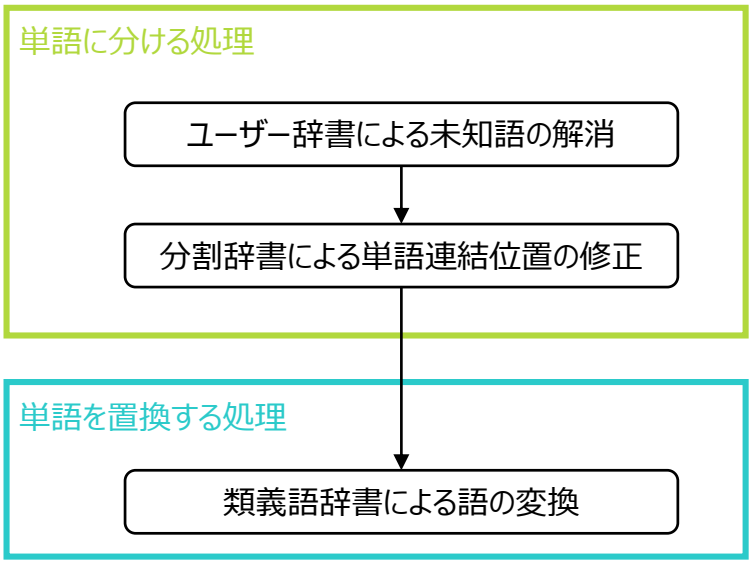
分割辞書

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。

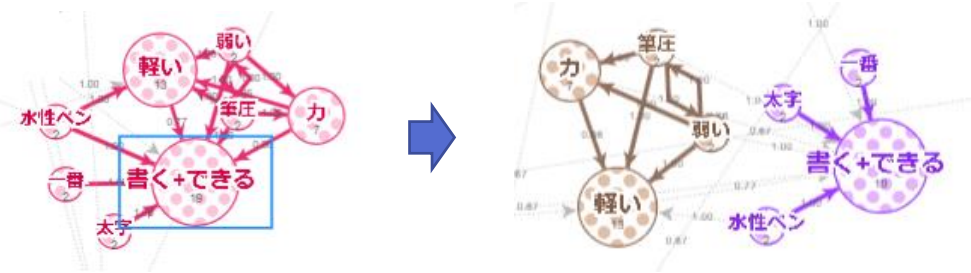
これらの辞書はテキストの分割処理が行われる際、右図のような流れで用いられます。



ネットワークからの話題抽出のネットワーク図

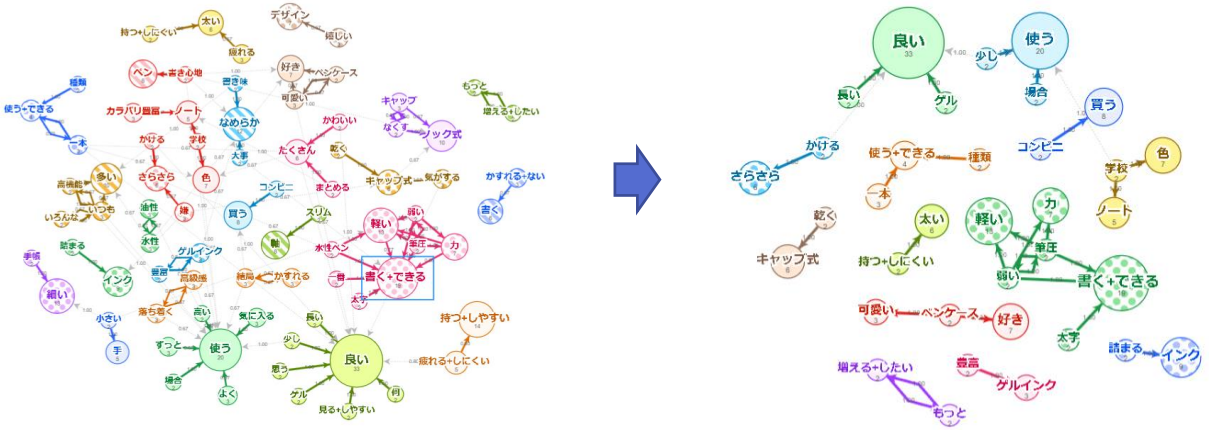
大きいネットワークが作成され、結果の解釈が難しい場合には、次のような方法で話題を分割することができます。

- 「いい」や「使う」など頻度の大きい単語は除外する
- 話題抽出のアルゴリズムで話題の分割を調整する



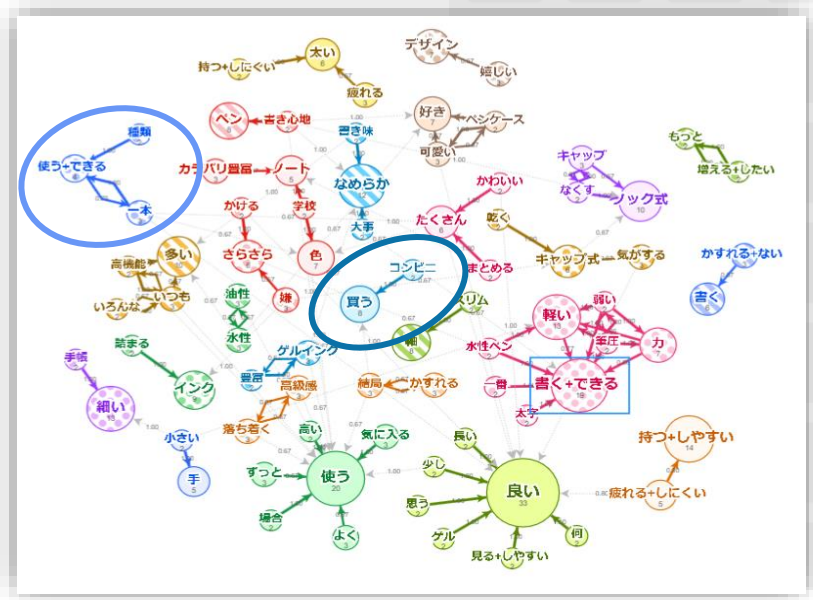
また、ノード数が多い場合には、次のような方法でノード数を減らすことができます。

- 「共起ルールの抽出」アイコンの指標値の上位出力件数を減らす
- 同順位のルールが複数存在している場合でも、指定した件数を超えるルールは出力しない



ネットワークからの話題抽出の結果

- 「ネットワークからの話題抽出」アイコンでは、得られた話題ごとに、どの単語がどの話題に含まれるかというルールをラベル付与規則として出力することができます。その規則を「ルールベース文書ラベル付与」アイコンの入力として指定することで、テキストデータに話題ラベルを付与することができます。



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、MSIP1.10.0 および Alkano1.4.0、TextExtension1.2.0 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
WEB: https://www.msi.co.jp/solution/analytics/index.html

株式会社 NTTデータ数理システム