



NTT DATA

NTT DATA Mathematical Systems Inc.



Text Mining Studio[®]
新バージョンのご紹介

数理システム新バージョン発表会2021
株式会社NTTデータ数理システム
vmstudio-info@ml.msi.co.jp

1. Text Mining Studio®

–製品概要

–アドオンモジュール

–新機能紹介

2. 動作環境について

3. セミナーのご案内

4. 分析コンサルティングのご紹介

Text Mining Studio®

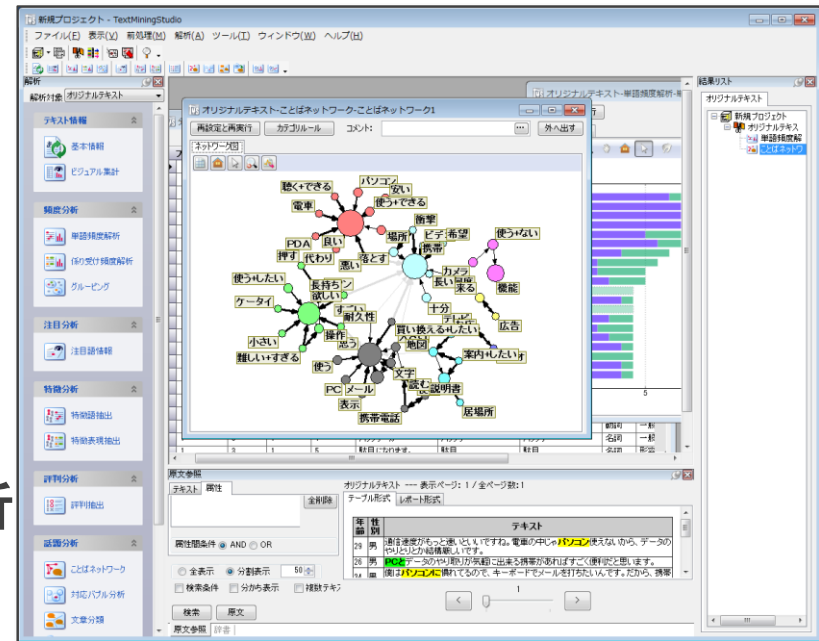
製品概要

Text Mining Studio とは



テキストデータから有益な情報を抽出するための テキストマイニングツール

- **誰にでも高度な**
テキストマイニングを
 - マウスの操作のみで分析が可能
 - 豊富な分析機能と強力なグラフ機能
- より**自由度**の高い分析を
 - カテゴリ機能等「意味」に着目した分析
 - データマイニングツール
Visual Mining Studio® との
シームレスな連携



Text Mining Studio®
(TMS)

Text Mining Studio 利用分野

代表的な利用分野



コールセンター

- ・VOC、QFD活用
- ・マーケティング
- ・離反回避

特許明細書

・技術マーケティング

■特許明細書

知的財産がぎっしりつまった特許明細書を活用した技術マーケティング。技術動向や中心的な特許・技術を把握。パテントマップ作成により、独自の技術開発を実現できます。



医療・看護系

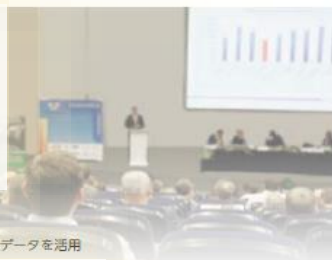
- ・医療サービス向上
- ・医療危険予測

■医療

電子カルテや患者の医療サービスへの満足度の向上を狙います。
医療危険予測：看護師のナレッジを分析、危険につながる経験知を形にすることで、ヒヤリ・ハット (Medical incident) の発目だけでなく、事故そのものの回避につながります。

新聞・雑誌記事

- ・社会動向の把握
- ・海外マーケティング



■学術系

学術研究において重要な論文データを活用

論文の整理や分析を効率化する形式の記載は、その研究成果を裏付ける重要な証拠。証拠の整理の evidence が求められる今、Text Mining Studio は、論文の整理や分析において、期待される大きなメリットです。

論文分析：

論文そのものを分析することにより、全ての論文を読まなくても全体の情報を把握し、読むべき論文の選定などが行えます。数理システムでは JDream11 (論文) データ (※) の販売も行っております。データご購入の方には論文データの分析ノウハウ冊子をプレゼント。

※JDream11は、独立行政法人科学技術振興機構が管理運営する日本最大の科学技術データベースシステムです。

アンケート自由記述文

- ・商品開発
- ・製品弱点の発見



営業・業務日報

・社員育成

や業務日報を活用

社員育成：

優秀な営業担当者のノウハウや、業績貢献に直接している活動を分析することで、現場での意思決定および迅速なアクションを実現、売り上げの底上げを狙います。

ナレッジマネジメント：

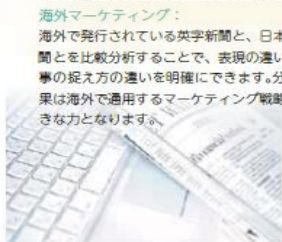
営業現場にてヒアリングした顧客の要望や、製造現場における日々の作業の記録はナレッジとして価値ある情報に昇華させる必要があります。これからの事業展開において、それらをいつでも参照できることが、企業の大きな強みとなります。



分析代行、コンサルティング もご相談ください

■その他

お客様のニーズに合わせた分析手法・ソリューションのご提案をいたします。Text Mining Studio のカスタマイズ、Web ブラウザでの利用、システムのエンジンとしての利用など、お気軽にご相談ください。



海外マーケティング：

海外で発行されている英字新聞と、日本の新聞とを比較分析することで、表現の違いや物事の捉え方の違いを明確にできます。分析結果は海外で適用するマーケティング戦略の大きな力となります。

SNS分析

- ・評判分析
- ・他社動向把握
- ・広告効果把握

広告効果把握：

広告を打つ前と後の書き込みから、その違いを分析することで、広告効果は一目瞭然。キャッチコピーや広告イメージの浸透状況、起用タレントのイメージ調査など、次回広告時の改善点もみつけることができます。



TextMiningStudio[®]と4つのアドオンツールでテキスト分析をサポートします！

製品名	概要
TextMiningStudio [®]	日本語の分かち書き機能と豊富な分析機能を備え、簡単な操作で本格的なテキストマイニングが行えるツール
TextCutter	テキストを話題毎に精度よく分割し、興味のある話題を抽出するアドオンツール 例：ホテルの口コミを「立地」、「従業員」、「部屋」の話題に分割
英語アドオン	英文の形態素解析、係り受け、連語処理により日本語と同様の分析を行うアドオンツール
音声テキストアドオン	コールセンター、会議での発話内容を音声認識したデータの不要語の削除、まとめあげを行うことで分析精度向上を計るアドオンツール
類似抽出アドオン	分散表現作成： 大規模コーパスから単語の分散表現を作成します 類義語検索： 分散表現から文脈の似た類義語を自動的に抽出します 類似テキストツール： 文章と文章の類似度を算出し、類似 / 非類似の分類を行うラベリング支援ツール

• 分かち書き機能

– 文章を単語単位に区切る機能

◆ 単語自動連結

- 文節単位にまとめ上げを行い、意味のある語を抽出する
3/次元/造形/装置 → 3次元造形装置

◆ 係り受けの抽出

- 単語間の修飾関係を得る



◆ 態度表現(ニュアンス)の抽出

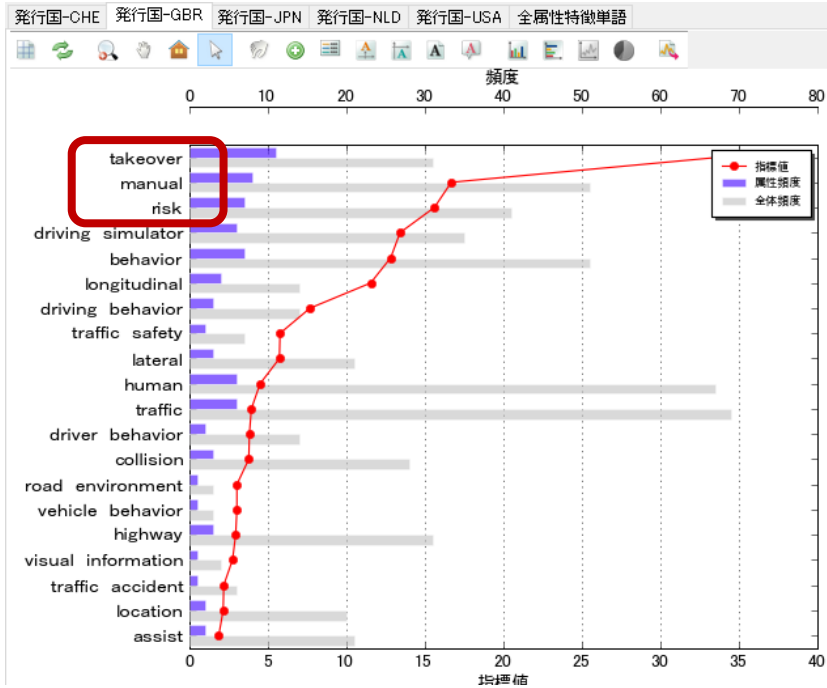
- 要望・否定など文章から読み取れる記述者の主観的な態度・認識を把握
- 「使わない」→「使う + ない(否定)」

- **Visual Mining Studio®**の機能を利用した 15 種類の分析メニュー
頻度解析、特徴語抽出、ことばネットワーク、評判分析、文章分類、グルーピング、時系列分析...
 - 各分析結果をグラフィカルに表示
 - きめ細かい分析に対応するためのパラメータ設定

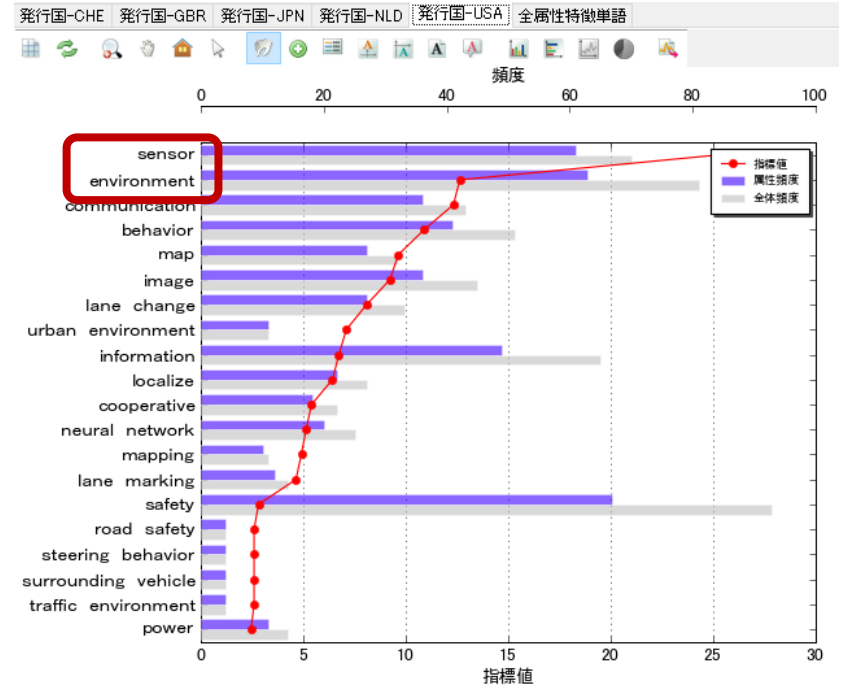
- 説明しやすい分析結果への対応
 - 原文に沿った分かち書き情報の提示
 - 技術資料のご提供

- 自社開発によるサポート体制
 - 熟練技術者によるサポート
 - 年数回のバージョンアップ

「自動運転」に関する論文分析例： 発行国ごとの技術の特徴 特徴語抽出



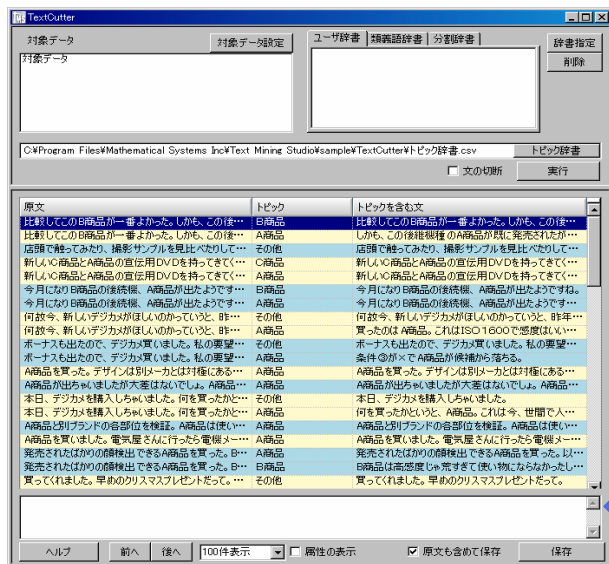
GBR では、自動運転の手動・自動の切り替えに関する takeover と、risk に関する論文の投稿が特徴的



USA では、周囲環境を認識するセンサーに関する論文が特徴的 communication などの単語もあり、機器に関する技術の投稿が特徴的と考えられる

Text Mining Studio®
アドオンモジュール
TextCutter

雑多なテキストをトピック毎に分割するアドオンツール



新機種 XXX-01 に乗り換えました。デザインがやや厚ぼったいかな、と思いましたが機能的にはかなりイイ感じです。

機種変手続きした後に高校時代の友人と落ち合って焼肉屋へ。ホルモンの種類が多くて満足でした。

新機種
の話題

カット！



飲食店
の話題

当社オリジナル！！

新機種 XXX-01 に乗り換えました。デザインがやや厚ぼったいかな、と思いましたが機能的にはかなりイイ感じです。

機種変手続きした後に高校時代の友人と落ち合って焼肉屋へ。ホルモンの種類が多くて満足でした。

TextCutterでトピック分割したデータ分析では
精度が大幅増加

60%→85%

アドオンモジュール: TextCutter

(例)ホテルの口コミサイト

「場所は良いし、受付混むのもまあいいけど、部屋がキレイではない。だから嫌。」

一般的な (TMS含め) テキストマイニングツール 「グルーピング」機能の場合	立地	受付	部屋	その他
場所は良いし、受付混むのもまあいいけど、部屋がキレイではない。	✓	✓	✓	
だから嫌。				✓

Text Mining Studio(TMS)の TextCutterの場合	立地	受付	部屋	その他
場所は良いし、	✓			
受付混むのもまあいいけど、		✓		
部屋がキレイではない。			✓	
だから嫌。			✓	

キーワードがない文章でも直前の文のトピックに振り分けられる！

Text Mining Studio®

アドオンモジュール

英語アドオン

英文テキストを分析可能にするアドオンツール

大幅に機能強化しました

1) 英語分かち書きエンジンの刷新

- 分かち書き処理の高速化(30分→3分)

データサイズ	処理時間(旧版)	処理時間(現行版)	速度比
20KB (特許要約 30件分)	100秒	30秒	3.3倍
200KB (特許要約 300件分)	15分	1分	15倍
2.3MB (特許要約 2,700件分)	60分	5分	12倍
23MB (特許要約 27,000件分)	10時間	40分	15倍

7並列で動作 PC スペック: CPU Core i7 クラス、16GBメモリ
このスペックでの日本語の分かち書きは1分 1MB程度

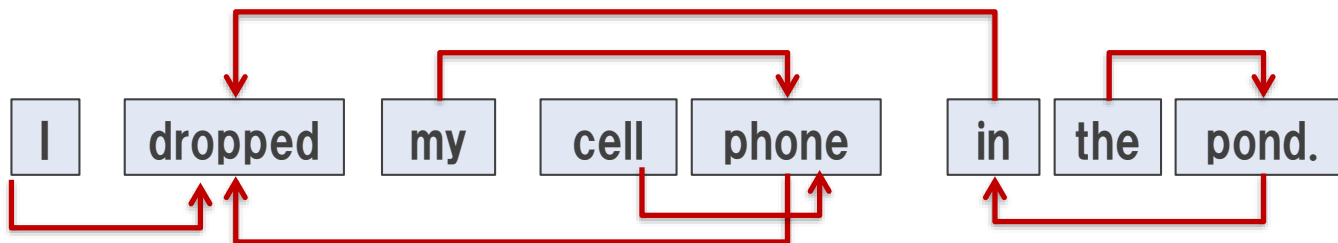
2) 分析精度向上のためのアプローチ

- 自動連結機能
 - 文章構造を考慮した単語のまとめ上げ
- 連語登録支援機能
 - 連語ランキングを自動提示

自動連結機能により、自然な係り受けを取得できるようになりました

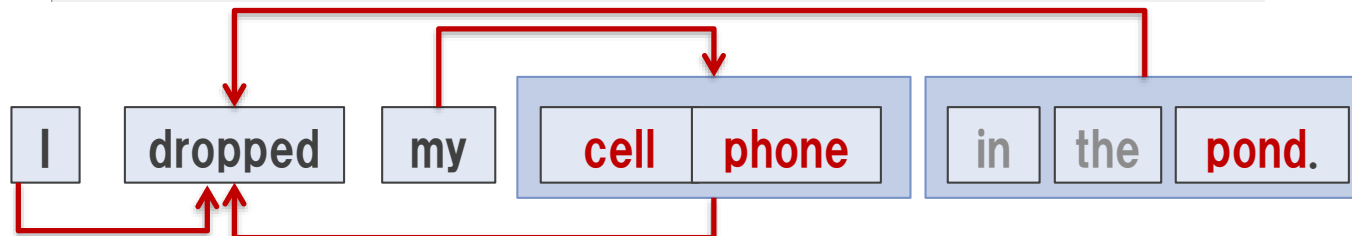
旧版

ファイルID	行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳細	係り先
1	1	1	1	I	I	I	代名詞		2
1	1	1	2	dropped	drop	drop	動詞	一般	-1
1	1	1	3	my	my	my	代名詞		5
1	1	1	4	cell	cell	cell	名詞	一般	5
1	1	1	5	phone	phone	phone	名詞	一般	2
1	1	1	6	in	in	in	前置詞		2
1	1	1	7	the	the	the	決定詞		8
1	1	1	8	pond	pond	pond	名詞	一般	6
1	1	1	9	.	.	.	記号	一般	2



現行版新機能

ファイルID	行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳細	係り先
1	1	1	1	I	I	I	代名詞		2
1	1	1	2	dropped	drop	drop	動詞	一般	-1
1	1	1	3	my	my	my	代名詞		4
1	1	1	4	cell phone	cell phone	cell phone	名詞	一般	2
1	1	1	5	in the pond	pond	pond	名詞	一般	2
1	1	1	6	.	.	.	記号	句点	2



連語らしさをランキング表示

現行版新機能

連語の重要度が高いものから表示

登録支援

連語抽出設定

品詞設定

連語全般
品詞を制限せず連語候補を抽出する
例) at least, one or more

名詞系
名詞として使われる連語候補を抽出する
例) New York City
City of New York

数字のみの単語を含む連語を抽出する

単語数

2 単語 以上

4 単語 以下

からなる連語を抽出する

単語検索(スペース区切り)

検索単語 AND検索 OR検索

全選択 全解除

	選択	連語	品詞	指標値	頻度	単語数
1	<input checked="" type="checkbox"/>	unmanned aerial vehicle	名詞 一般	1062.43	534	3
2	<input checked="" type="checkbox"/>	aerial vehicle	名詞 一般	716.54	721	2
3	<input checked="" type="checkbox"/>	air vehicle	名詞 一般	167.90	170	2
4	<input type="checkbox"/>	unmanned vehicle	名詞 一般	77.05	78	2
5	<input type="checkbox"/>	unmanned aerial vehicle body	名詞 一般	72.00	24	4
6	<input type="checkbox"/>	unmanned air vehicle	名詞 一般	54.67	29	3
7	<input type="checkbox"/>	wing unmanned aerial vehicle	名詞 一般	54.00	18	4
8	<input type="checkbox"/>	rotary wing vehicle	名詞 一般	44.00	24	3
9	<input type="checkbox"/>	flight vehicle	名詞 一般	43.77	45	2
10	<input type="checkbox"/>	vehicle body	名詞 一般	38.17	43	2

連語抽出 ファイル出力 全13270件 辞書作成 閉じる

402件表示しました

Text Mining Studio®

アドオンモジュール

音声テキストアドオン

コールセンター、会議での発話内容を音声認識したデータを まとめあげ、不要語などを削除するアドオンツール

新規プロジェクト* - 音声テキストアドオン

ファイル データ加工 環境設定 ヘルプ

AmiVoice® から取込

データ加工
分析に用いる列を選択する
列の選択

不要なテキストを行ごと
削除する
削除表現

行を指定して複数の行を
1行に連結する
行の連結

他のファイルから
マスターデータを読み込む
マスター結合

実行結果の出力
最後の実行結果を出力する
ファイル出力

実行結果リスト

AmiVoiceから取込

再取得
↓
1 列の選択
再設定
ファイル出力

2 削除表現
再設定
ファイル出力

3 行の連結
再設定
ファイル出力

会話識別子	発話者
1	20170802120000 OP
2	20170802120000 CU
3	20170801120000 CU
4	20170801120000 OP
5	20170731120000 CU
6	20170731120000 OP
7	20170731120000 CU
9	20170729120000 OP
10	20170729120000 CU
11	20170728120000 CU
12	20170728120000 OP
13	20170727120000 CU
14	20170727120000 OP
15	20170727120000 CU
16	20170726120000 CU
17	20170725120000 CU
18	20170725120000 OP

表示行数: 98

AmiVoice® からの検索・取込

A	B	C
1	会話識別子	発話者
2	20170802120000	OP
3	20170802120000	CU
4	20170802120000	OP
5	20170802120000	CU
6	20170802120000	OP
7	20170802120000	CU
8	20170802120000	OP
9	20170802120000	CU
10	20170802120000	CU
11	20170802120000	OP
12	20170802120000	CU
13	20170802120000	CU
14	20170802120000	CU
15	20170802120000	CU

不要な表現の削除

行ID	テキスト名	削除
1	1 発話内容	お電話ありがとうございます
2	2 発話内容	もしもし。
3	3 発話内容	はい。
4	4 発話内容	コバヤシショウケンさん。
5	5 発話内容	はい。コバヤシ保険でござい
6	6 発話内容	それでお宅はどこ横浜。
7	7 発話内容	こちらは東京本社のコール
8	8 発話内容	あそうですか。
9	9 発話内容	はい。
10	10 発話内容	ちよとお尋ねしますけどね
11	11 発話内容	はいどうぞ。
12	12 発話内容	そのまーあの一解約申し込

削除

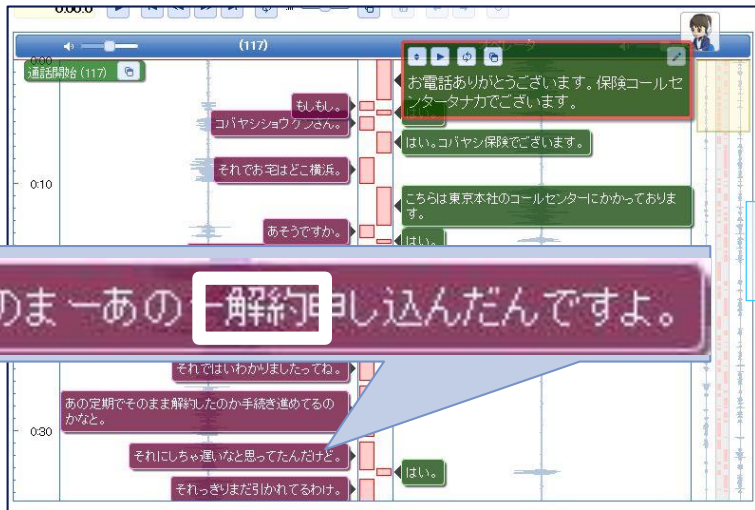
発話内容のまとめあげ

会話識別子	発話者	発話内容
20171128130CU		それでお宅はどこ横浜。そのまーあの一解約申し込んだんです。もう3ヶ月か4ヶ月ぐらい前に、あの定期でそのまま解約したのか手続き進めてるのかなと。それにしちゃ遅いなどぶの中それしか入ってなくて置いていかれただけなので、どれを提出するのかしらとかね。読めばいいんだらうけども。そうですね。その方が助かり。もしあれだったら担当の人書いて入院給付金を請求するにあたって。本人からえっともう亡くなるんですよ。えっと入院給付金を請求するのは、あくまでも法定相続人になるっていうふうに関心して。そ
20170802120CU		
20170801120CU		

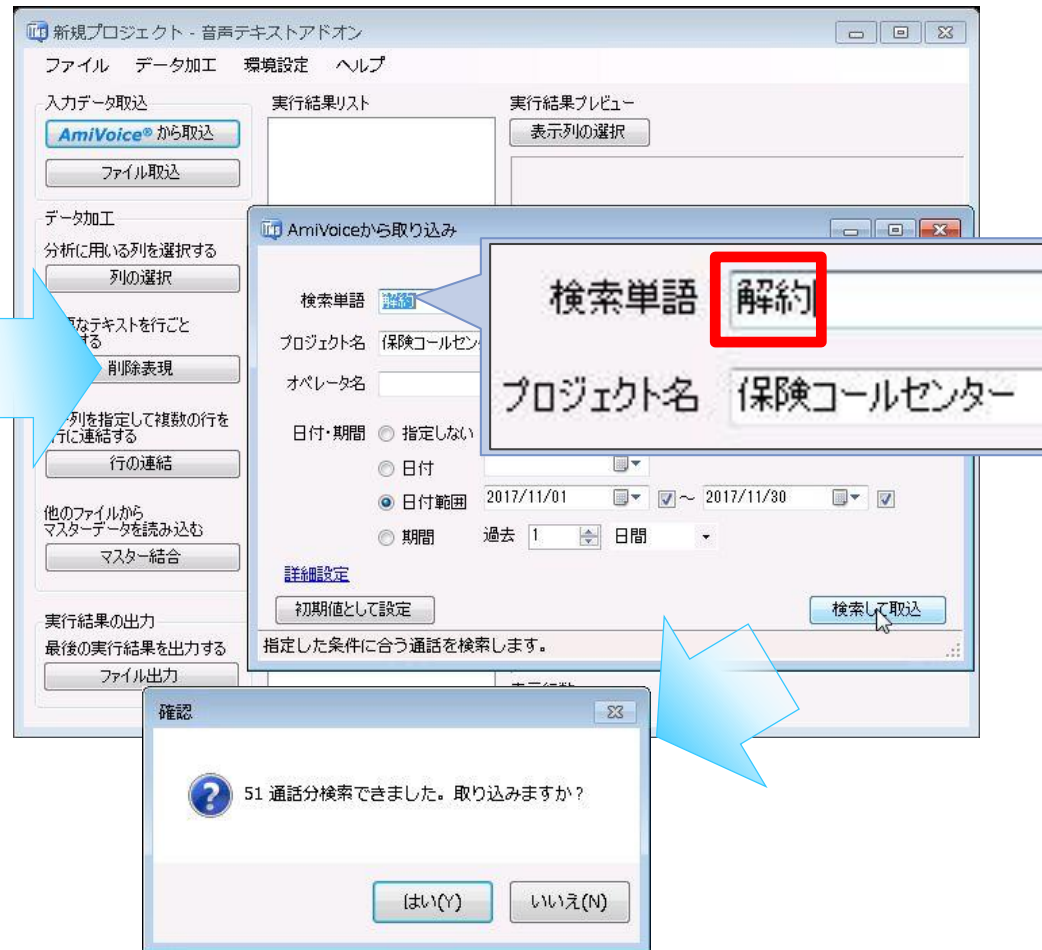
音声テキストアドオンで
まとめあげたデータを**TMS**で分析

音声テキストアドオン: *AmiVoice*[®] からの検索・取込

音声認識ソフト *Amivoice*[®] から取り込める機能をご用意
キーワード、期間などを指定し、簡単にデータが取り込めます



Amivoice Speech Visualizer



音声テキストアドオン： 不要な表現の削除

不要表現は典型的な表現を事前にご用意
導入からすぐにご利用いただけます

通話毎、発話者毎など行のまとめ上げを設定可能 通話時間の合計なども適切に算出することができます

元データ

会話識別子	発話者	発話内容	発話時間(ミリ秒)	文字数	会話時間
20170802120000	OP	お待たせいたしました。	1040	11	217.0002
20170802120000	CU	はい。	336	3	217.0002
20170802120000	OP	はい。えー確認いたしまし	7792	52	217.0002
20170802120000	CU	え。	336	2	217.0002
20170802120000	OP	私忘れていたかもしれませ	2992	18	217.0002
20170802120000	CU	ふとの中それしか入ってな	4136	28	217.0002
20170802120000	OP	さようございましたか。	1136	12	217.0002

キー列を指定して複数の行を
1行に連結する

行の連結

連結後データ

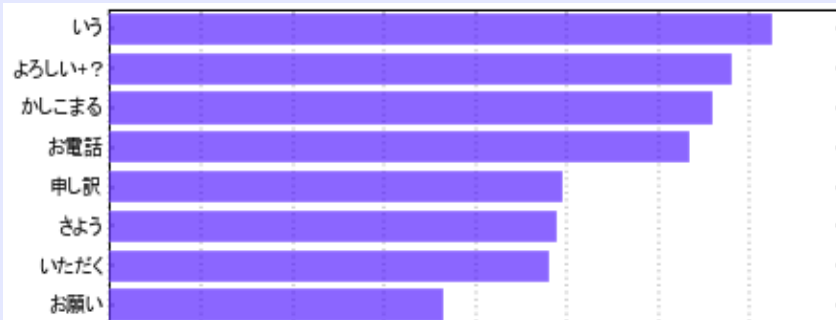
会話識別子	発話者	発話内容	発話時間(ミリ秒)	文字数	会話時間
20170802120000	CU OP	…お待たせいたしました。はい。はい。えー確認いたしました所あの書類は同意書を作成はさせていただいたんですがもしかしたら担当者がですね。え。私忘れていたかもしれないので一度。ふとの中それしか入ってなくて置いていかれただけなので…	184912	1370	217.0002
20170801120000	CU OP	…お電話ありがとうございます。小林保険コールセンタータナカでございます。そうですね。はい。入院給付金を請求するにあたって。はい。本人からえっともう亡くなってるんですよね。さようございますか。はい。…でございますね。えっとー入院給付金を請求するのは…	144976	1086	177.7004

通話毎のまとめ

合計処理

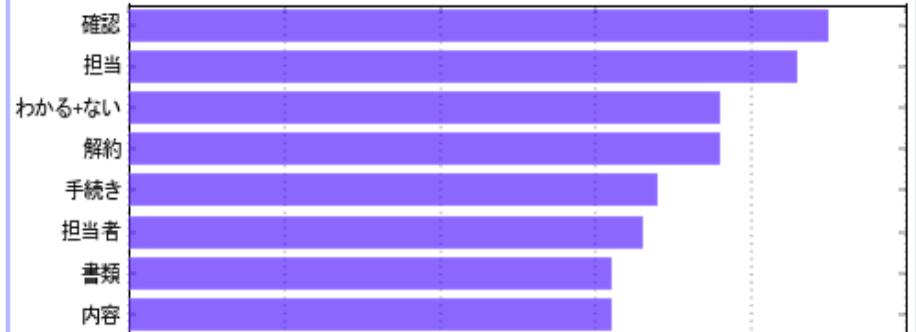
単語頻度解析の結果比較 :

整形前



定型文句や当たり前の単語ばかり

整形後

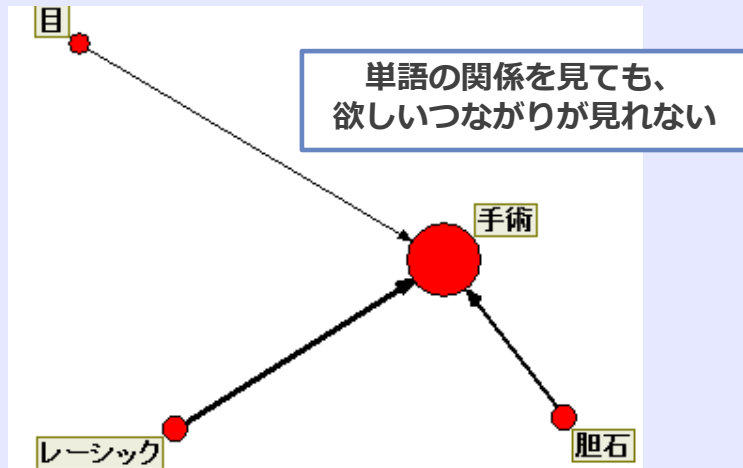


重要な単語が
分析結果に表れる

不要表現の削除により、
定型文句、相槌のような不要単語が削除され、
重要な単語を分析結果に表示することができます

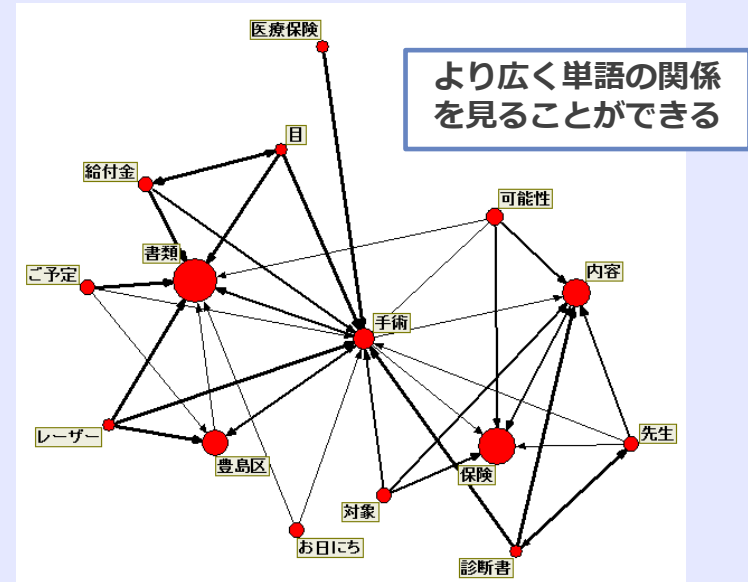
ことばネットワークの結果比較 :

整形前



会話識別名	テキスト
20170723120000	そうです。胆石の手術で。
20170723120000	かしこまりました。胆石の手術は何日ぐらいにされるかお決まりでございませうか。
20170723120000	あと胆石の手術なんです。
20170723120000	例えば胆石だけを取る手術ですか、あと、
20170717120000	えっとレーザーの手術を受けるんですよ。
20170717120000	レーザーの手術でございませうね。はい。
20170717120000	はい。それでこちらレーザーの手術をされたということですが。
20170717120000	でまこちらレーザーの手術ということですが。
20170629120000	あの目をねちょっと手術したんですよ。
20170629120000	目割がされて手術なさったということですね。かしこまりました。

整形後



給付。ついて伺いたいという伺いたいんですけども。はい。どのようなことでしょうか。えー過去にもあるんですけども。えー主人の。目の。えっとレーザー上こちら受けたんですけども。給付金の対象にはまできていましたかなりますよね。はい。では今回あの一えと再度ですね。レーザーの凝固術を。されるご予定で。そうですね。ではえっと一すすお済みいただいている手術が給付金の対象になるかどうかの確認でございますね。というか

不要表現の削除と行のまとめにより、より広範囲の単語の関係を把握することができます

Text Mining Studio®

アドオンモジュール

類似抽出アドオン

2020年夏
Renewal!!

主な機能

1. 分散表現(次頁説明)作成機能

- テキストの「文脈」を学習し、各単語のベクトルを作成する
Deep Learner 新機能 **Word Embedding** を利用

2. 類義語検索機能

- 単語分散表現を利用し、“似た単語”を探す
⇒テキスト分類の際に重要となる類義語を効率よく収集できます
- **TMS**の各分析結果と連係
- 大規模コーパス分散表現
⇒一般的な“似た単語”を探す

1. 類似テキストツール

- 文章と文章の類似度を算出し、類似 / 非類似の分類を行うラベリング支援ツール
- 特許・論文の先行技術調査、コールセンター・アンケートのカテゴリ化

Word Embedding (単語分散表現)

- それぞれの単語に対する N 次元のベクトル (N はユーザが指定)
- 似た文脈の単語 → 似たベクトル となるように学習
 - ✓ 分布仮説「意味の似ている単語は類似した文脈に出現する」
 - ✓ 大量の「文」からなるデータを用意する
 - ✓ 文を単語に分割し、単語をニューラルネットワークの入力とする
 - 例: 「自分(単語)の左右 m 個の単語達から自分を予測(CBOW)」
 - ✓ この結果生成される N 次元のベクトルが単語の分散表現
- 類義語の探索などに利用される

夏目漱石は日本の小説家で、帝国大学英文科卒業後、松山中学、第五高校の教師を経て33年ロンドンへ留学。帰国後、東大講師となる。

芥川龍之介は明治の文明開化期など、さまざまな時代の歴史的文献に題材をとり、スタイルや文体を使い分けたたくさんの短編小説を書いた。

ニューラルネットワーク

夏目漱石	0.1	0.2	...	0.2	...
...
小説家	0.3	0.1	...	0.8	...
...		
ゲーテ	0.4	0.3	...	0.1	...
...		
小説	0.2	0.6	...	0.7	...
...					...

分散表現

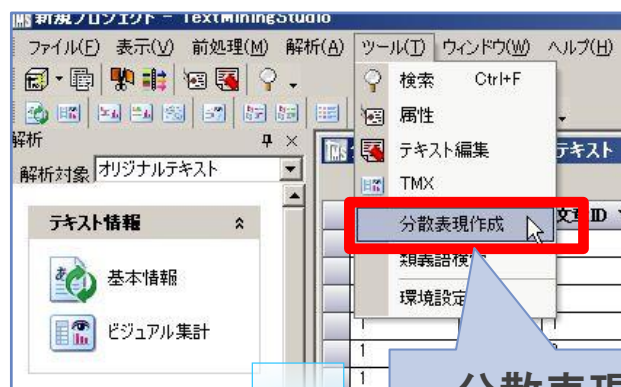
小説家

作家	0.7
詩人	0.65
文学者	0.5
夏目漱石	0.4
ゲーテ	0.3

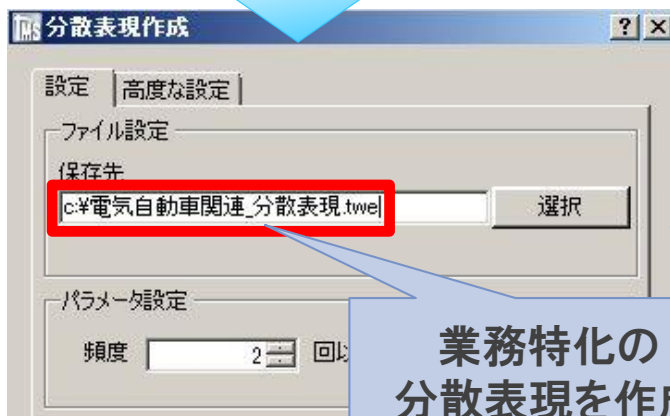
類似度

類似抽出アドオン - 分散表現作成機能

TMSに入力したデータから特許、製造など
業務領域に特化した分散表現を構築できます

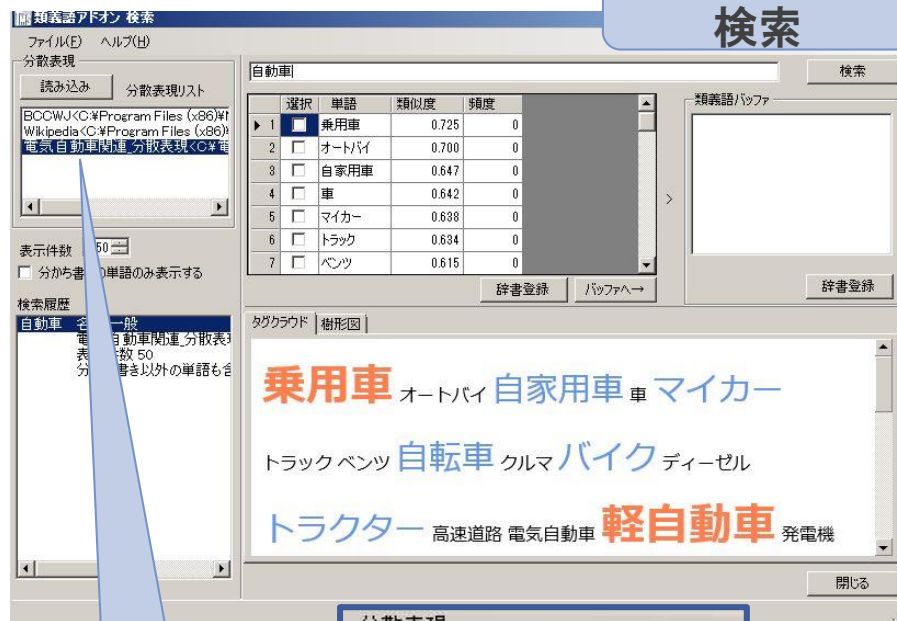


分散表現作成

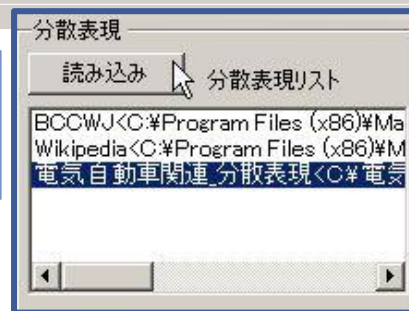


業務特化の
分散表現を作成

「自動車」で
検索



読み込んで
検索



TMSの各分析結果から簡単に類義語を検索&登録 類義語辞書、グルーピング機能へ類義語を簡単に登録

TMS

類義語の検索

類義語アドオン

辞書登録

選択	単語	類似度	頻度
<input checked="" type="checkbox"/>	首相	0.714	0
<input checked="" type="checkbox"/>	総理	0.700	0
<input type="checkbox"/>	内閣	0.666	0
<input type="checkbox"/>	総裁	0.663	0
<input type="checkbox"/>	参議院	0.650	0
<input type="checkbox"/>	小泉総理	0.647	0
<input type="checkbox"/>	国会議員	0.646	0
<input type="checkbox"/>	閣僚	0.632	0
<input type="checkbox"/>	政治家	0.622	0

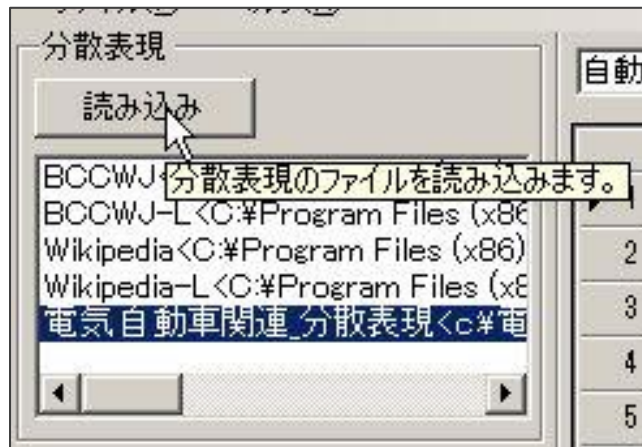
類義語辞書登録

代表語	品詞
コンビニ	名詞 一般
スーパー	名詞 一般
スーパーマーケット	名詞 一般
売店	名詞 一般
総理大臣	名詞 一般
首相	名詞 一般

類似抽出アドオン - 大規模コーパスから作成した分散表現

業務に特化しない、一般的な類義語を抽出するために
大規模コーパスから作成した分散表現を用意

- BCCWJ (国立国語研究所: 現代日本語書き言葉均衡コーパス)
- Wikipedia (フリー百科事典)



複数の分散表現を
切り替えて検索

類似義語アドオン 検索

ファイル(F) ヘルプ(H)

分散表現

読み込み 分散表現リスト

BCCWJ<C:\Program Files (x86)\Ma
Wikipedia<C:\Program Files (x86)\M

表示件数 50

分かち書きの単語のみ表示する

検索履歴

自動車 名詞 一般
電気自動車関連_分散表現
表示件数 50
分かち書き以外の単語も含む

自動車

選択	単語	類似度	頻度
<input checked="" type="checkbox"/>	乗用車	0.725	0
<input type="checkbox"/>	オートバイ	0.700	0
<input type="checkbox"/>	自家用車	0.647	0
<input type="checkbox"/>	車	0.642	0
<input type="checkbox"/>	マイカー	0.638	0
<input type="checkbox"/>	トラック	0.634	0
<input type="checkbox"/>	ベンツ	0.615	0

辞書登録 バックファへ

類義語バックファ

辞書登録

タグクラウド 樹形図

乗用車 オートバイ 自家用車 車 マイカー

トラック ベンツ 自転車 クルマ バイク ディーゼル

トラクター 高速道路 電気自動車 軽自動車 発電機

開じる

類似抽出アドオン - 2019年度にリリースした機能

- 分散表現作成機能**大幅高速化**

計算ロジック見直しにより、**お手軽に**分散表現作成が実行可能に

- 複数の**データを結合**して分散表現作成が可能に

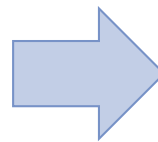
【分析対象以外の大規模データ & 分析対象のデータ】で学習

対象データが少なくてもその他の大規模データを加味した
より自然な類義語が検索可能に

- 類義語検索結果にユーザの**フィードバック**反映機能を追加

	選択	単語	類似度	頻度	類似	非類似
1	<input type="checkbox"/>	格安	0.679	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	安上がり	0.674	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	割高	0.653	0	<input type="checkbox"/>	<input checked="" type="checkbox"/>

ユーザが類似、非類似を選択



	選択	単語	類似度	頻度	類似	非類似
▶ 1	<input checked="" type="checkbox"/>	格安	0.762	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	リーズナブル	0.735	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	安上がり	0.731	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>

検索結果にフィードバック

より**意図に沿った**類義語が表示されるよう、**調整可能**に

文章間の類似度を算出し、文章を 類似 / 非類似 で分類し、ラベリングを支援するツールです。

ターゲット設定

- 耐久性
 - ベース文
 - 電池が長持ちする。耐...
 - 耐久性がある。
 - 壊れない。壊れにくい。
 - ベース文の追加
 - 類似度
 - set1(手法:SWEM 平均, 単位:行, 単...)
 - 閾値設定
- PC家電連携
 - ベース文

	閲覧	コメント	set1_耐 久性_類 似度 ▼1	set1_耐 久性_類 似度 _rank	set1_耐 久性	set1_PC家 電連携_類 似度
1	詳細表示	私のすぐ壊れる。壊れないのが欲しい。	0.85	1	<input checked="" type="checkbox"/>	0.37
2	詳細表示	水に濡れても壊れないようにして！アドレス帳が消えて...	0.78	2	<input checked="" type="checkbox"/>	0.38
3	詳細表示	仕事で携帯を使っているので、簡単に壊れたりするの...	0.70	3	<input checked="" type="checkbox"/>	0.55
4	詳細表示	象が踏んでも壊れない携帯を希望。	0.68	4	<input checked="" type="checkbox"/>	0.57
5	詳細表示	この前トイレに落としたら、動かなくなった。買ってから1...	0.61	5	<input checked="" type="checkbox"/>	0.36
6	▶ 詳細表示	携帯を落として壊してしまったことが何度もあります。ど...	0.60	6	<input checked="" type="checkbox"/>	0.53
7	詳細表示	もっと操作を簡単に。ボタンを押すのも疲れる。	0.55	7	<input type="checkbox"/>	0.58
8	詳細表示	画面もボタンも字が小さいし、操作が難しくすぎて年寄...	0.55	8	<input type="checkbox"/>	0.59
9	詳細表示	携帯でサッカー観戦したりする時に、電池の消耗が早...	0.54	9	<input type="checkbox"/>	0.63
10	詳細表示	パソコンに差し込んですぐに使えるといいですね、そうす...	0.54	10	<input type="checkbox"/>	0.59
11	詳細表示	もっと電池の持ちを長くしてほしいですね。				

携帯を落として壊してしまったことが何度もあります。どんな衝撃にも耐えられるくらいの携帯をお願いします。

総行数: 101 / 170

複数の比較対象となる文章を設定できる

比較対象となる文章毎に類似度の表示と類似/非類似のラベリングが可能。

類似度に寄与する単語を確認

Text Mining Studio®

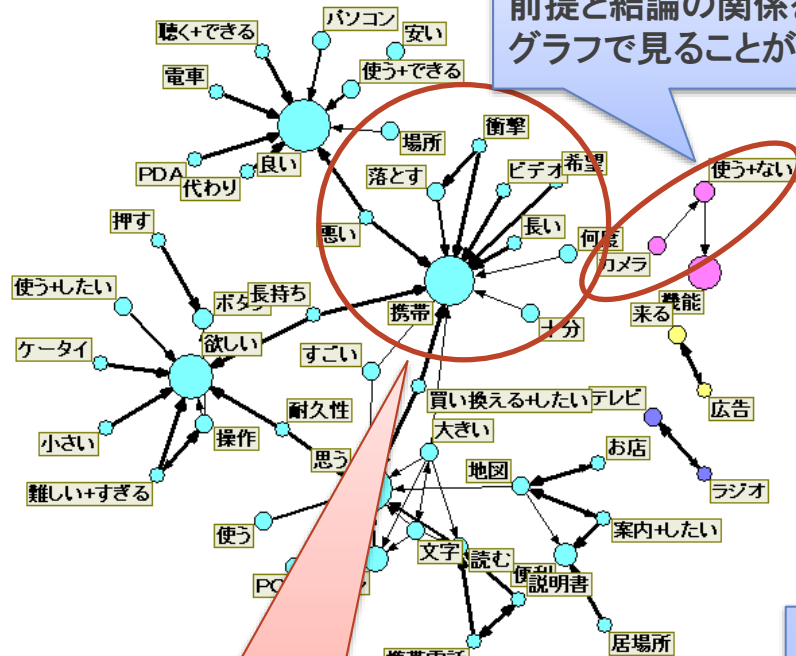
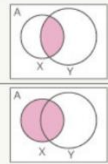
新機能

ことばネットワーク・注目語情報にJaccard係数追加

ことばネットワーク結果比較

【既存】信頼度

$$\frac{|X \cap Y|}{|X|}$$

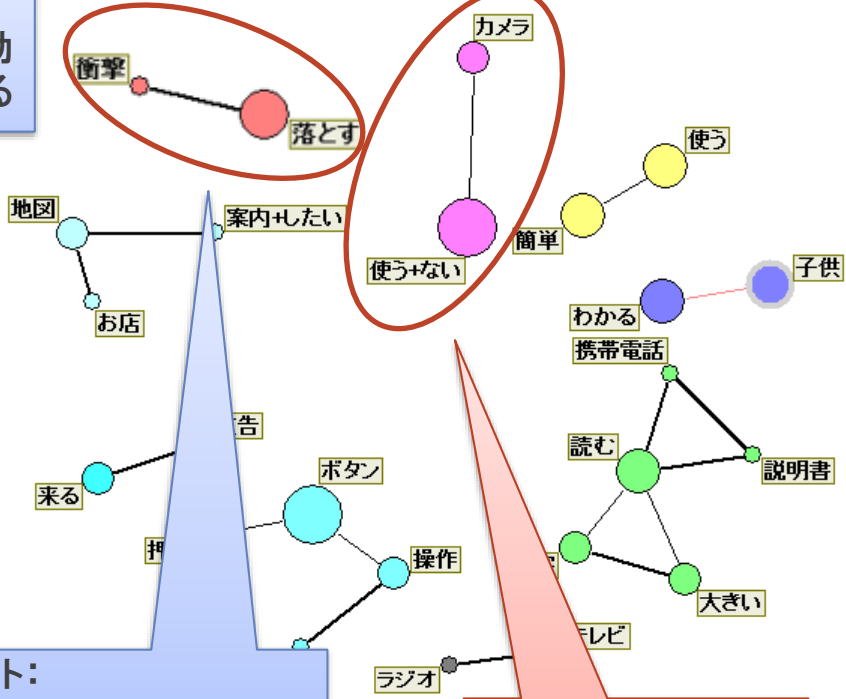
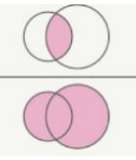


メリット：
前提と結論の関係を有効
グラフで見ることができる

デメリット：
高頻度単語にエッジが集中しやすく、クラスタ(話題)が大きくなりやすい

【New】Jaccard係数

$$\frac{|X \cap Y|}{|X \cup Y|}$$



メリット：
クラスタ(話題)が分かれやすい。高頻度単語にエッジが集中しない

デメリット：
前提と結論の関係を
見るができない

- 特徴語抽出の指標値にTF-IDF、Jaccard係数を追加
- 英語アドオンの文分割機能を改修
 - 処理速度の向上
 - 技術文書などで小数点の部分で文が区切れていたケースなどが改善

修正前	修正後
They all deliver 7200 RPM performance in a standard 3. / 5-inch form factor.	They all deliver 7200 RPM performance in a standard 3.5-inch form factor.
You can get yours online at geekstuff4u. / com for only ¥3,500.	You can get yours online at geekstuff4u.com for only ¥3,500.

Text Mining Studio®

動作環境について

動作環境について

以下の動作環境でご利用になれます。

Microsoft® Windows®

8.1/10

Server2012/Server2012R2/Server2016/Server2019

(※すべて日本語版 OS に限ります)

(※2020年2月リリース以降、下記5製品は 64bitOSのみのサポートとさせていただきます)

	製品名	備考
1	TMS	
2	TextCutter	
3	英語アドオン	メモリ 4GB以上
4	音声テキストアドオン	
5	類義抽出アドオン	メモリ8GB以上

セミナーのご案内

セミナーへのご参加をメンバー一同お待ちしております

日程	TMS関連セミナー
2022/01/11(火)	13:30 - 15:30 【Webinar】TMS 紹介セミナー TMSの機能紹介、事例紹介、製品デモ
2022/02/09(水)	13:30 - 15:30 【Webinar】TMS 紹介セミナー TMSの機能紹介、事例紹介、製品デモ
2022/02/10(木)	10:00 - 17:00 【Webinar】アカデミックコンファレンス アカデミック関係の方から一般企業の方までご参加いただけます。
2022/03/04(金)	13:30 - 15:30 【Webinar】TMS 紹介セミナー TMSの機能紹介、事例紹介、製品デモ
随時開催中！ お問合せください	TMS 体験ハンズオンセミナー（事前インストール不要） 実際に TMS をご操作いただき、テキストマイニングをご体験頂きます

セミナー後の個別相談につきまして：

日頃お客様がご業務、ご研究で扱われているデータを当社の技術スタッフが拝見し、どのような分析が可能か、回答を差し上げます。

<https://www.msi.co.jp/tmstudio/seminar.html>

分析コンサルティングサービスご紹介

分析ツールを最大限ご活用いただき有益な結果を導き出していくために、当社は分析のコンサルティングサービスもあわせて提供させていただいており、好評を得ております。

実際のデータを目の前にして、当社スタッフとお客様とでアウトプットを作り上げていきます。



当社スタッフが！

- データの素性
 - どのようにして集めたデータなのか？
 - **誰がどういうタイミング**で記録したデータなのか？
- データ量と性質
 - 行数，列数，平均的な文字量，利用可能な属性
 - テキストの**文体**
 - 統制されている，くだけている，ほぼ言い切り...
 - 逐次増加していくデータの場合、
月あたり・週あたり の増加量
- 目的
 - 意識されている**お困りごと**は？
 - 複数のステークホルダーへの配慮
 - 目指す**アウトプットイメージ**はあるか？
 - 検証したい**仮説**はあるか？

テキストマイニングで何をを目指す？



	インプット	アウトプット
テキスト分類型	テキスト	それぞれのテキストに対するラベル
サマリー型	テキスト	理解の助けとなる集約情報
発見型	テキスト (+しばし 仮説)	明確ではない
検索型	テキスト+クエリ	検索結果

動機：テキストを分類したい。仕分けしたい。

ご意見
色がこれだけ揃っていると便利に使える。
カラーバリエーションがたくさんあってうれしい。
値段に見合った性能だと思う。
とにかく安い！買いためています。
何といっても書きやすさはピカイチ。
サラサラ書けて、書き出しも線が切れない。すごい。

インプット情報

ご意見
色がこれだけ揃っていると便利に使える。
カラーバリエーションがたくさんあってうれしい。
値段に見合った性能だと思う。
とにかく安い！買いためています。
何といっても書きやすさはピカイチ。
サラサラ書けて、書き出しも線が切れない。すごい。

アウトプット情報

ご意見区分
色
色
値段
値段
書きやすさ
書きやすさ

分析ツールによる処理

テキストの内容から適切な「意見のラベル」を与える

アプローチ

手法	メリット	デメリット
<u>ルールベース</u>	1件1件の分類根拠が明確、 人手で微調整やメンテナンス可能	人手で分類ルール構築の必要あり
<u>機械学習</u>	ラベル付与済みデータ (学習データ) から、 ラベル付与基準を自動的に獲得	大量の学習データが必要、 一般に1件1件の分類根拠は言葉 では説明できない

モデル

動機：テキストを分類したい。仕分けしたい。

カテゴリ
色
値段
書きやすさ

ご意見	ご意見区分
色がこれだけ揃っていると便利に使える。	色
カラーバリエーションがたくさんあってうれしい。	色
値段に見合った性能だと思う。	値段
とにかく安い！買いためています。	値段
何とんでも書きやすさはピカイチ。	書きやすさ
サラサラ書いて、書き出しも線が切れない。すごい。	書きやすさ

分類させたい
カテゴリ群は
決まっているか？

分類済のデータが
存在しているか？

分類の
アプローチは？

NO

YES

YES

NO

機械学習

テキスト(単語)を説明変数、
ラベルを目的変数として
判別のモデルを構築

ルールベース

まず傾向把握を行い、
それに基づいて
カテゴリ群を決定

人手でラベル付けを行って
分類済データを準備、また
カテゴリの意味合いから
類推してルールの雛形作成

ラベル属性毎の単語の
出現状況の違いを把握し、
それに基づいてルール作成

テキスト分類型の問題解決：ルールベース

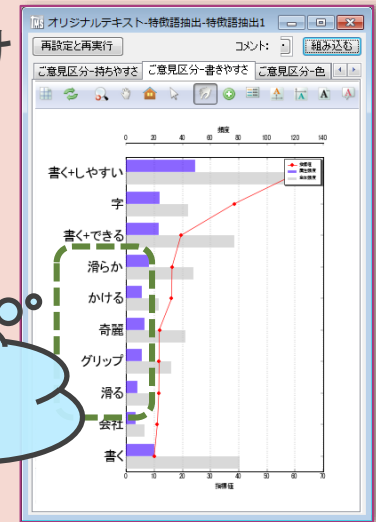
- Text Mining Studio®の機能 **グルーピング** を活用して実現する。

ご意見	ご意見区分
色がこれだけ揃っていると便利に使える。	色
カラーバリエーションがたくさんあってうれしい。	色
値段に見合った性能だと思う。	値段
とにかく安い！買いためています。	値段
何とんでも書きやすさはピカイチ。	書きやすさ
サラサラ書いて、書き出しも線が切れない。すごい。	書きやすさ

ラベル付きデータを
TMSに取り込み

どんな単語や係り受け
がラベルの違いに
効いているか、
集計や特徴分析を
用いて把握する。

「滑らか」や「滑り」も
『書きやすさ』を表す言葉
と考えてよいのでは？



- 取りこぼしをなくそうと
ルールを過度に拡張すると、
適合率が低下する (**ハズレも増える**)
- かといってルールを絞り過ぎると
再現率が低下する (**取りこぼし増える**)

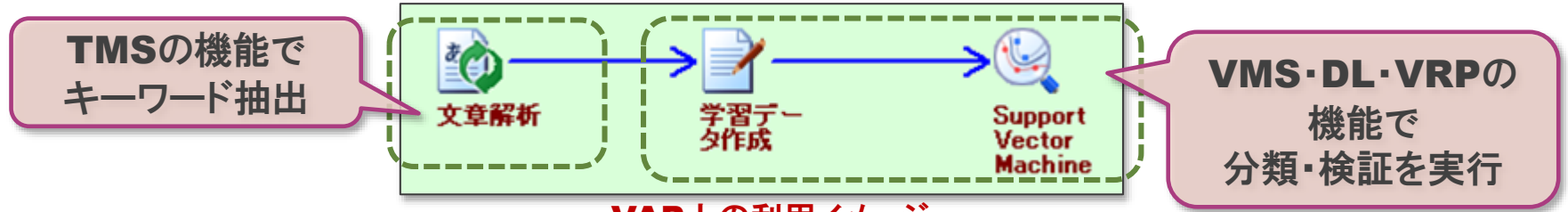
実用的なレベルを探っていく

ルールの定義

テキスト分類型の問題解決：機械学習

- データマイニングツール **Visual Mining Studio**
- ディープラーニングツール **Deep Learner**
- Rユーザ向け分析プラットフォーム **Visual R Platform**

といった製品を、分析基盤**VAP**の上であわせて利用して実現する。



VAP上の利用イメージ

- テキスト情報の数値化が必要、
こういった自由な整形も**VAP**上で実行可能

目的変数

予測すべき情報・予測の対象

Comment	評価
画像はいいですが、楽しみが少なすぎます。戦闘やダンジョンを楽しみたかった!ストーリーもなんだかなあ…。	不評
物語が実に良くできている。映像は綺麗だし、良いキャラが多い。戦闘システムもなかなか良い。買って損はありません。	好評
RPGとしては不出来だが、ゲームとしては最高のクオリティだと思う。ロード時間の短さ、ムービーやフィールドの美しさは、他のソフトを軽く凌駕する。	好評

説明変数

予測の手掛かりとする情報

文章をベクトル化

評価	最高	爽快	良い	気に入る	応じる	面白くない	初心者	つまらない	楽しい+ない	自己満足	...
不評	0	0	1	0	0	0	0	0	0	0	0
好評	0	0	2	0	0	1	0	0	0	0	0
好評	1	0	0	0	0	0	0	0	0	0	0

ケーススタディー その1 : テキスト分類型



メーカー
カスタマーサポート部
A様

コールセンターの対応履歴ログ情報が、蓄積されているのみで全く活用できていない。入力時の**カテゴリ選択が有名無実化**して、みな「その他」ばかり。**カテゴリ再設計**をしたい。さらに**自動的に**カテゴリを振ってくれたらなおよい。

対応日時	2015年 10月 6日
問合せ種別	クレーム
製品種別	
問題種別	<div style="border: 1px solid gray; padding: 2px;">納期 支払い 発送 ...</div> <div style="border: 2px dashed red; padding: 2px; margin-top: 2px;">その他</div>

進め方一例

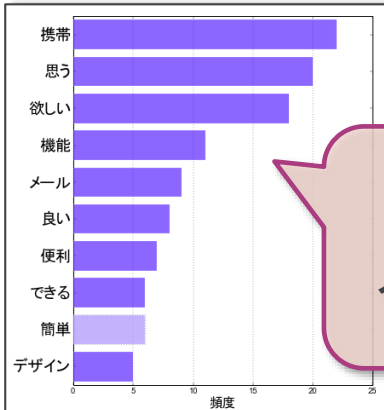
- 正常に機能している入力項目はどれか確認
 - ✓ 記入率はどの程度？
 - ✓ 選択式のカテゴリが、業務の実態に即しているか？
- テキストから「**実態に即したカテゴリ**」を作成
 - ✓ 分析によって問合せ内容の実情を把握
- 分類**ルール**や**モデル**を構築、**自動分類**へ



当社スタッフが！

・TMSで作成できるサマリーの形

単語・係り受けのリストアップ



全体傾向 及び 属性に沿った傾向の把握、
人手で「気になる表現」の
ピックアップ

・ターゲットを絞る

✓ 品詞で絞る

「名詞」だけの設定でものの名前のみを抽出、
形容詞や形容動詞で印象の表現のみを抽出、
etc...

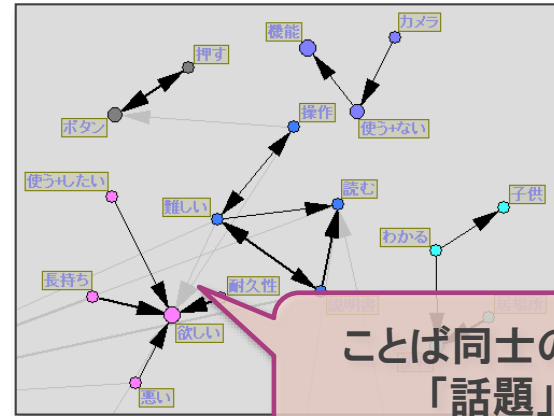
✓ 態度表現で絞る

要望表現、クレーム表現
etc...

・少数意見にあえて着目する

✓ 頻度下位の設定で抽出

共起・ことば同士の関係



ことば同士の固まりから
「話題」を把握

※全データで漠然とした図になってしまう場合は、
層別にフィルタリングしたデータで共起分析を行うと良い

問題解決のためには、
サマリー情報に対して
何らかの解釈を
行うことが不可欠！

運営している施設のアンケート分析を行いたい。
業態の異なるチェーン店舗が複数あり、
問題点や要望点を各店舗毎に、
もしくは全体で把握して
店舗側にフィードバックしたい。
そのほか、何か素敵な分析ができればよい。



サービス業
CS担当
B様

進め方一例



当社スタッフが！

- まずは業態毎の特徴を！
 - ✓ クレーム表現・要望表現を抽出する
- 他の属性との関係をあわせて把握
 - ✓ 例: 利用時間帯、利用頻度……
- 要望と要望の意外な関係は？
- 埋もれた少数意見を掘り起こしてみる



収集したデータは手元にある、
このデータを有効に活用できないだろうか……？

- まずは「**サマリー型**」のアプローチを通じて
データの理解を図り、**課題を探しテーマを決める**。
 - その際の「**気付き**」を大切にする。
 - データに接してはいるが異なる業務の方、立場が異なる方からは異なる「**気付き**」が得られる可能性がある。
- その結果、
サマリー型 の分析の全体把握を推し進める、
それが困難で **テキスト分類型** のアプローチで
分類～整理してはじめて把握できるようになるか、
というパターンが存在する。

データの「1件」とはどういう単位？

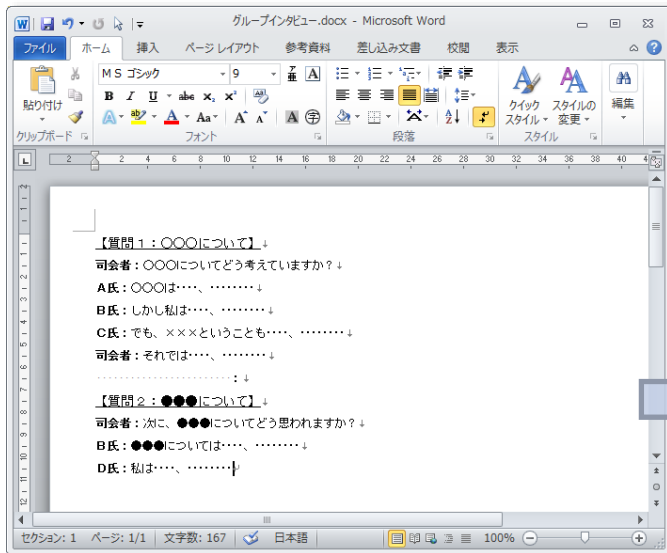
- 1件は、TMS では物理的な「**入力ファイルの1行**」と同じ
- この1件は「**現実世界の何かの事象**」の1件と対応付いている

- 1回 の入電に対する問い合わせ記録
- 1枚 のアンケート用紙
- ブログへの 1回 の投稿
- 1件 の特許文書 etc...

年齢	性別	コメント
38	男	この前トイレに落としたら、動かなくなった。買ってから10ヶ月経ってなかったので、買い換えできなかった。その程度の水には耐えるよ。
17	男	いや料金すまじつつか使います？でも、バカ料金出してほしい
33	女	もっと長く使いたいの、バッテリーが駄目になります。もっと長持ちする携帯が欲しいです。
42	男	テレビ携帯があるんだからラジオが聴きたいから欲しい。でも電車の中でテレビは見たくないが、野球中継が聴きたかったら聴きたい。

この単位が1件であり、
ここではアンケートの1回答

- 何が「1件」になるのか、それが明確ではないケースもある、そういった場合は適切に決定して表形式にデータを整形する必要がある



インタビューの記録

要整形

属性

質問No.	発言者	発言内容
1	司会者	〇〇〇についてどう考えていますか？
1	A	〇〇〇は…、……………
1	B	しかし私は…、……………
1	C	でも、×××という事も…、……………
1	司会者	それでは…、……………
:	:	:
2	司会者	次に、●●●についてどう思われますか？
2	B	●●●については…、……………
2	D	私は…、……………
:	:	:

1件

テキスト



看護学・心理学
研究者
C様

ある療法を受診している患者さんに対する
インタビュー記録を分析したい。
データは**ワード**でまとめた、
この後どうすれば？
患者さんの**心情の**
変化がわかるのか？

```
質問者：
その他、特別な画面有無などを認識させるためのユーザ辞書や、分割辞書などの設定も行
えます。設定後【OK】ボタンをクリックすると、情報保護の処理が開始します。ただし、
情報保護は、分ち書き処理などを伴うため、若干の時間がかかります。その確認の画面
が表示されますので、よろしければ【はい】ボタンをクリックしてください。
Aさん：
図 4-47 時間がかかる旨の確認
終了すると、情報保護画面右側に、保護化したデータが表示されます。実際にマスクされ
たデータを含む行については、背景色がオレンジに変わります。
51
質問者：
入力データ設定に課数ファイルで設定されていた場合には、課数ファイルが同時に処理さ
れます。
Aさん：
情報保護結果がよろしければ、【OK】ボタンをクリックします。すると、入力データ設定
で【OK】ボタンをクリックしたときのように、分ち書きのための画面が表示され、分析
に入っていくことができます。

```

進め方一例

- まずは**データ整形**から！
 - ✓ TMSで分析するのに適したデータの形は？
 - ✓ データの「1件」をどうとらえればよいか？
- **心情の変化**はどこにあわられるか？
 - ✓ カウンセリングの回数による推移、
同一回でも前半・中盤・後半 などの観点などを試行



当社スタッフが！

Text Mining Studio®

特許明細書

NTT DATA 株式会社 NTTデータ 数理システム

Text Mining Studio®は
これ1つでマルチシーンに
対応する 汎用ツールです！

機能

サポート

コスト

営業部 Text Mining Studio®担当

TEL : 03 - 3358 - 6681 FAX : 03 - 3358 - 1727

【URL】 <https://www.msi.co.jp/tmstudio/>

【E-mail】 vmstudio-info@ml.msi.co.jp

医療・看護系

アンケート自由記述文

営業・業務日報

無料体験セミナーお申込み受付中！

コールセンターデータを用いた分析紹介時間をご用意しております。

<https://www.msi.co.jp/tmstudio/seminarRegular.html>

コールセンター

SNS分析

学術系

新聞・雑誌記事