



NTT data

NTT DATA Mathematical Systems Inc.

テキストマイニング関連ソリューションのご紹介 ～導入から活用までのトータルサポート～

数理システムユーザーコンファレンス2019
株式会社NTTデータ数理システム
tmstudio-info@msi.co.jp

1. テキストマイニング関連ソリューション

- **Text Mining Studio**
- アドオンモジュール
 - TextCutter
 - 英語アドオン **Renewal!**
 - 音声テキストアドオン **2018年リリース!**
 - 類義語アドオン **2019年リリース!**

2. 分析コンサルティングサービスのご紹介

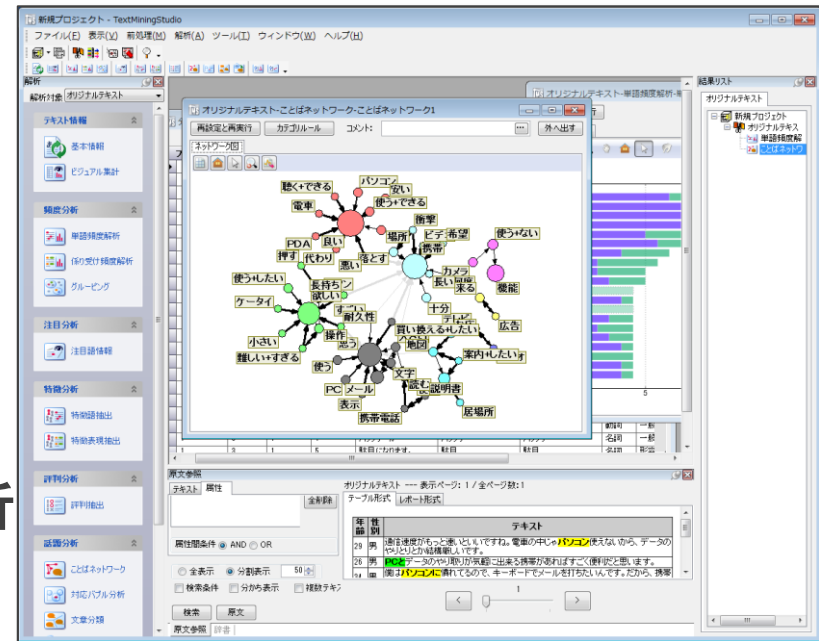
3. セミナーのご案内

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
シームレスな連係



Text Mining Studio (TMS)

- 分かち書き機能

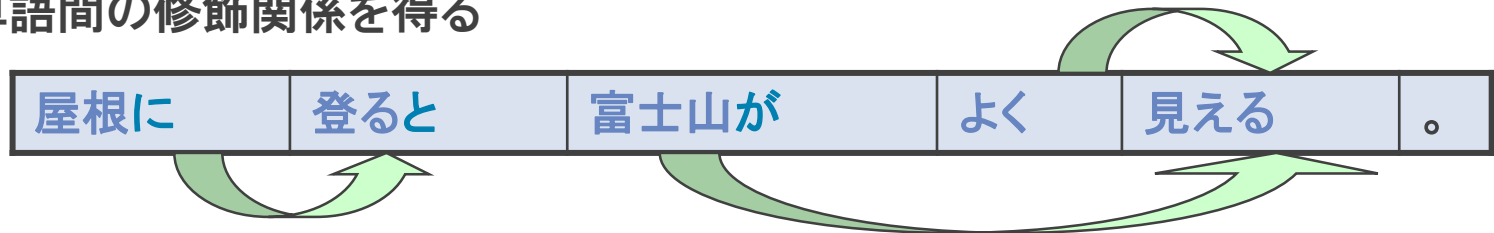
- 文章を単語単位に区切る機能

- ◆ 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)」

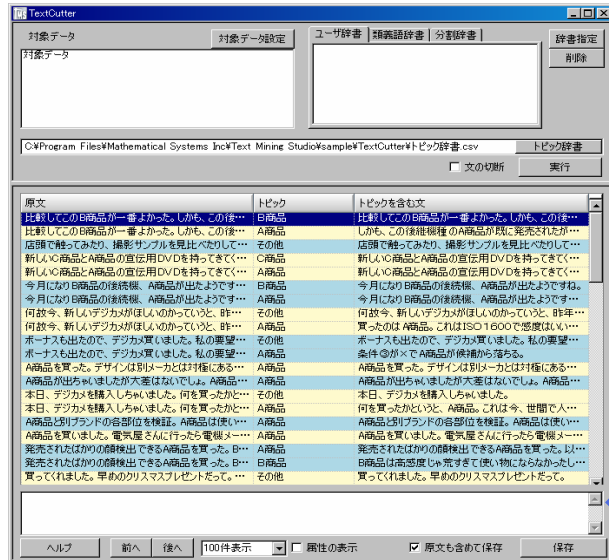
Text Mining Studio の特徴2

- **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

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！

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

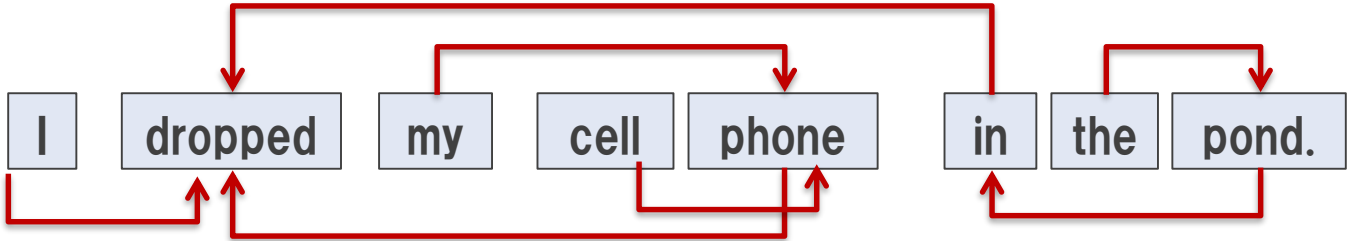
2) 分析精度向上のためのアプローチ

- 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

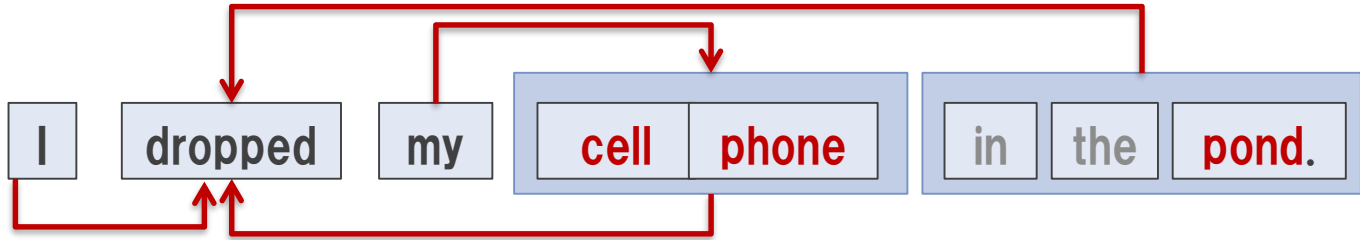
旧版

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

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 vehicle

☒ 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	☒	unmanned aerial vehicle	名詞 一般	1062.43	534	3
2	☒	aerial vehicle	名詞 一般	716.54	721	2
3	☒	air vehicle	名詞 一般	167.90	170	2
4	☐	unmanned vehicle	名詞 一般	77.05	78	2
5	☐	unmanned aerial vehicle body	名詞 一般	72.00	24	4
6	☐	unmanned air vehicle	名詞 一般	54.67	29	3
7	☐	wing unmanned aerial vehicle	名詞 一般	54.00	18	4
8	☐	rotary wing vehicle	名詞 一般	44.00	24	3
9	☐	flight vehicle	名詞 一般	43.77	45	2
10	☐	vehicle body	名詞 一般	38.17	43	2

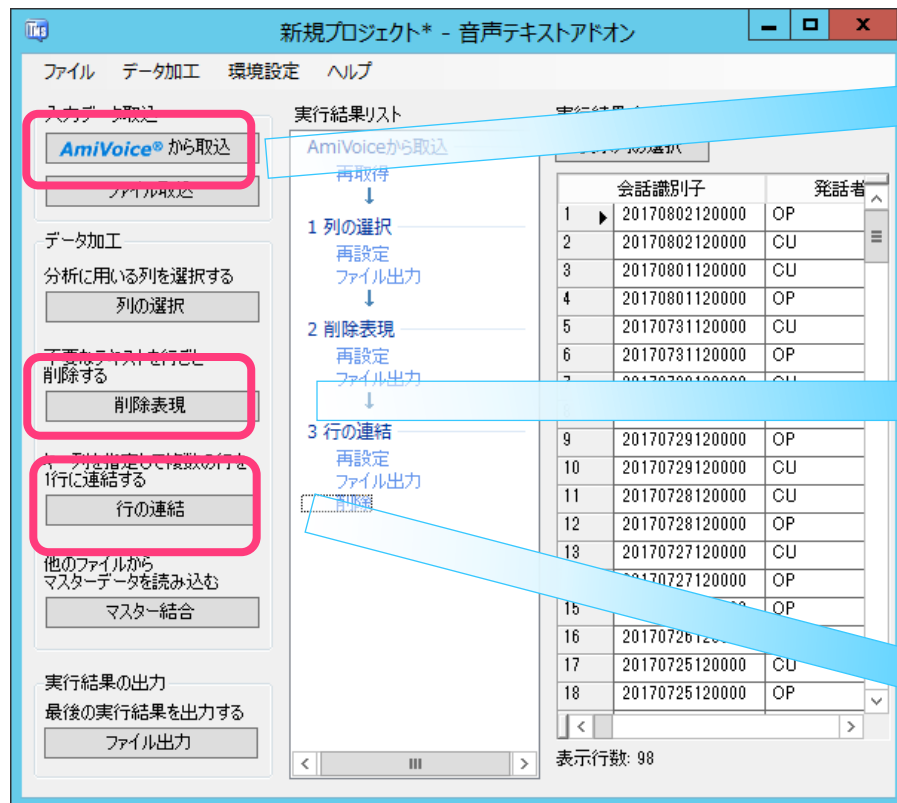
連語抽出 ファイル出力 全 13270 件 辞書作成 閉じる

402 件表示しました

連語の重要度が高いものから表示

指標値	頻度	単語数
1062.43	534	3
716.54	721	2
167.90	170	2
77.05	78	2
72.00	24	4
54.67	29	3
54.00	18	4
44.00	24	3
43.77	45	2
38.17	43	2

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



AmiVoice® からの検索・取込

会話識別子	発話者	発話内容
20170802120000	OP	お待たせしました。
20170802120000	CU	はい。
20170802120000	OP	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んですがもしもかんたんな担当者ですか。
20170802120000	CU	え。
20170802120000	OP	私忘れていたかもしれませんので一度。
20170802120000	CU	ふと中それしか入ってなくて置いていかれただけなので。
20170802120000	CU	さようでした。
20170802120000	CU	それを提出するのかもしれない。
20170802120000	CU	読めばいいんだらうけども。そうですね。その方が助かり。
20170802120000	OP	読めばいいんだらうけども。ええではですね。同意書もあのご持参させていただいてえー一度ご訪問の上ご説明させていただいてよろしいですか。
20170802120000	CU	そうですね。その方が助かり。
20170802120000	CU	かしこまりました。
20170802120000	CU	もしあれだったら担当の人書いてもらってできないんですかね。
20170802120000	CU	何か全然見えてないですか。なんかちょっと私なんか押しなんです。

不要な表現の削除

行ID	テキスト名	発話内容
1	1	発話内容
2	2	発話内容
3	3	発話内容
4	4	発話内容
5	5	発話内容
6	6	発話内容
7	7	発話内容
8	8	発話内容
9	9	発話内容
10	10	発話内容
11	11	発話内容
12	12	発話内容

削除

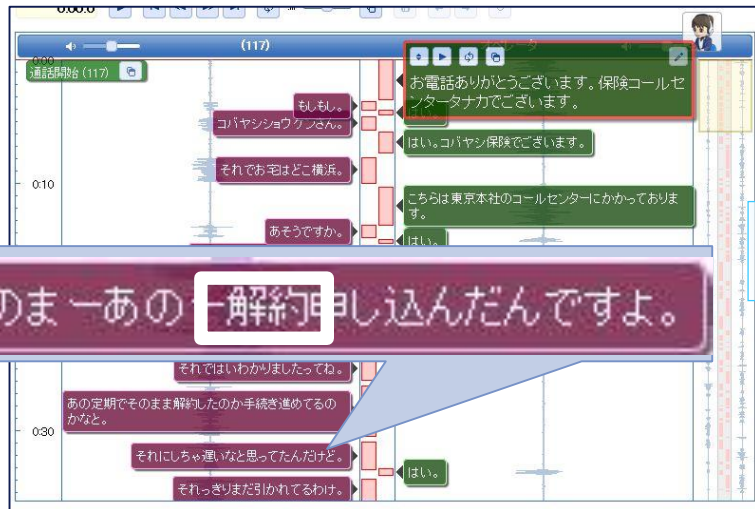
発話内容のまとめ上げ

会話識別子	発話者	発話内容
2017112813	CU	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。それにしちゃ遅いかなと。
2017080212	CU	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。それを提出するのかもしれない。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もらってできないんですかね。
2017080112	CU	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ってるんですね。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というふうに関わってますが。

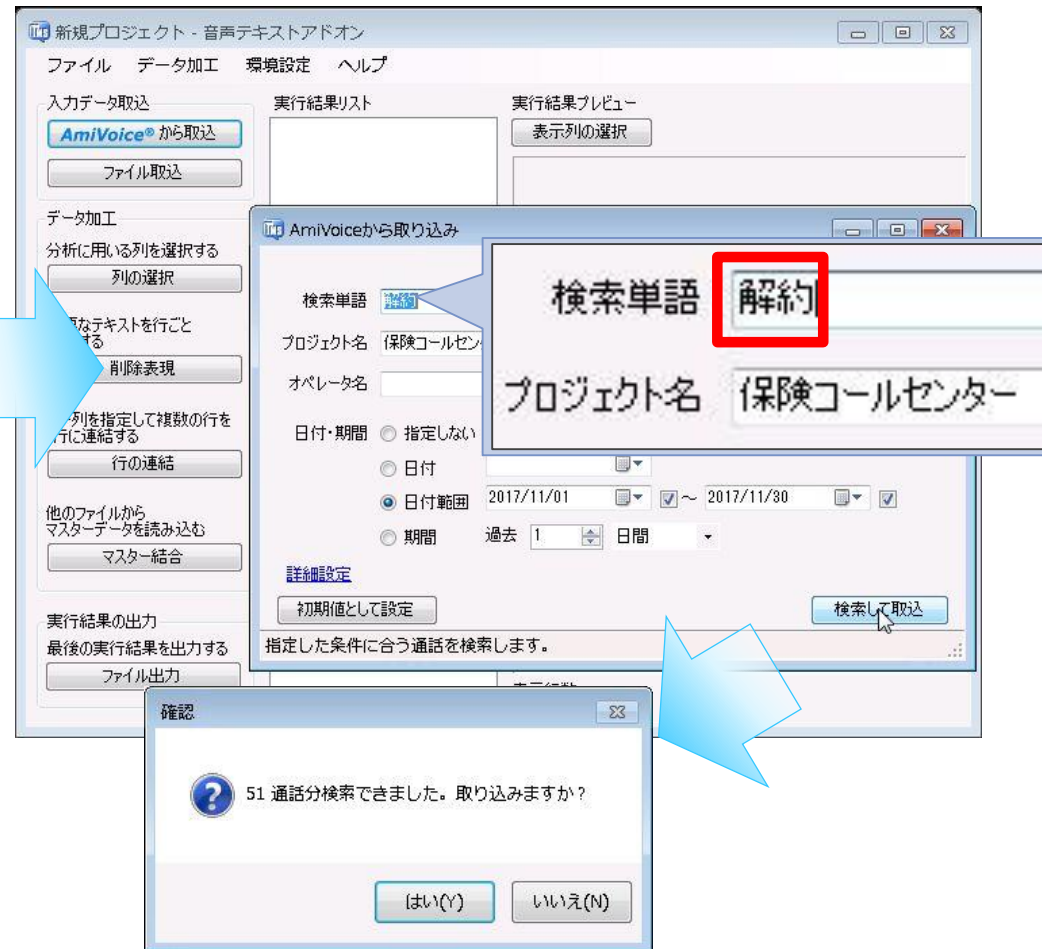
音声テキストアドオンで
まとめ上げたデータを**TMS**で分析

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

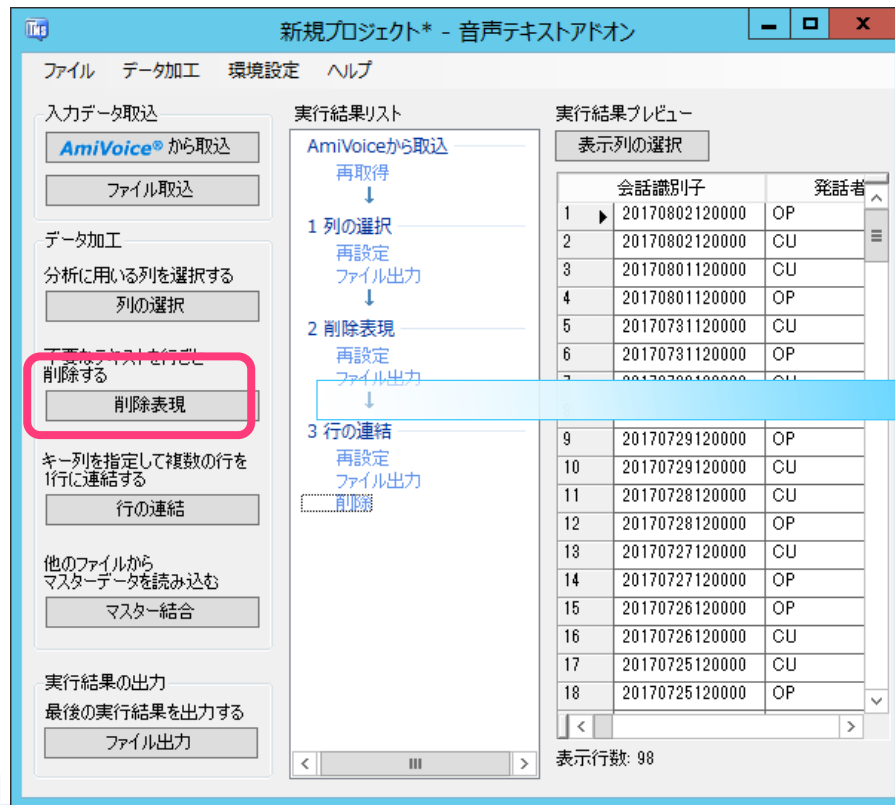


Amivoice SpeechVisualizer



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



削除判定結果プレビュー

全体行数: 4058 削除行数: 2157 削除率: 53.15%

行ID	発言者	発言内容
1	1	お待ちいたしました。
2	2	お待ちいたしました。
3	3	お待ちいたしました。
4	4	お待ちいたしました。
5	5	お待ちいたしました。
6	6	お待ちいたしました。
7	7	お待ちいたしました。
8	8	お待ちいたしました。
9	9	お待ちいたしました。
10	10	お待ちいたしました。
11	11	お待ちいたしました。
12	12	お待ちいたしました。
13	13	お待ちいたしました。
14	14	お待ちいたしました。
15	15	お待ちいたしました。
16	16	お待ちいたしました。
17	17	お待ちいたしました。
18	18	お待ちいたしました。
19	19	お待ちいたしました。
20	20	お待ちいたしました。
21	21	お待ちいたしました。
22	22	お待ちいたしました。
23	23	お待ちいたしました。
24	24	お待ちいたしました。
25	25	お待ちいたしました。
26	26	お待ちいたしました。
27	27	お待ちいたしました。

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

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

元データ

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

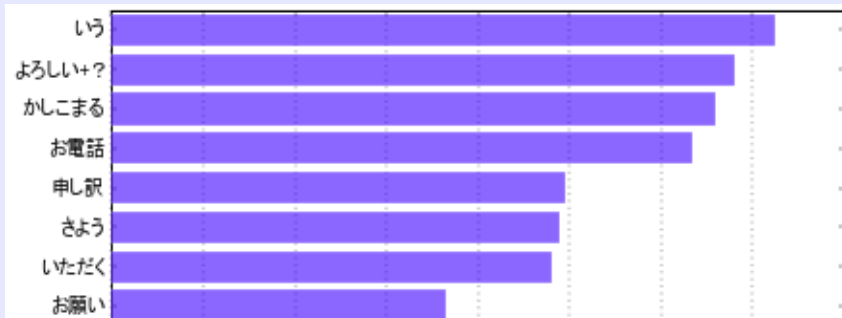
会話識別子	発話者	発話内容	発話時間(ミリ秒)	文字数	会話時間
20170802120000	CU OP	… お待たせいたしました。はい。はい。えー確認いたしました所あの書類は同意書を作成はさせていただいたんですがもしかしたら担当者がですね。え。私忘れていたかもしれないので一度。ふとの中それしか入ってなくて置いていかれただけなので…	184912	1370	217.0002
20170801120000	CU OP	… お電話ありがとうございます。小林保険コールセンタータナカでございます。そうですね。はい。入院給付金を請求するにあたって。はい。本人からえっともう亡くなってるんですよね。さよ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

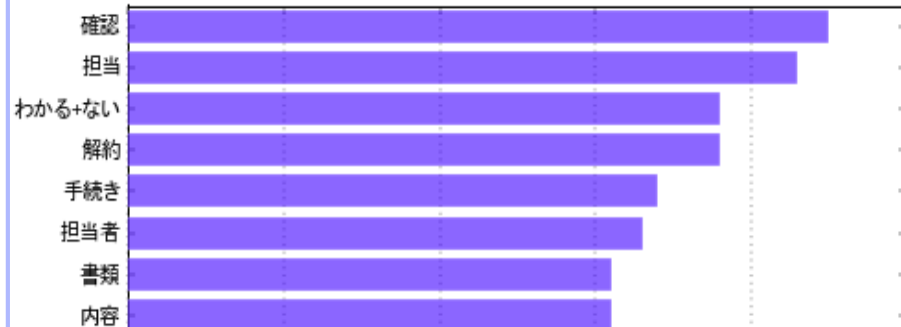
単語頻度解析の結果比較：

整形前



定型文句や当たり前の単語ばかり

整形後

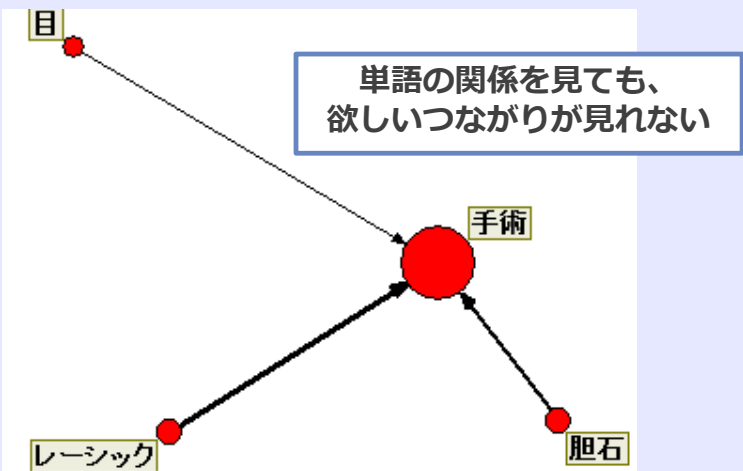


重要な単語が
分析結果に表れる

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

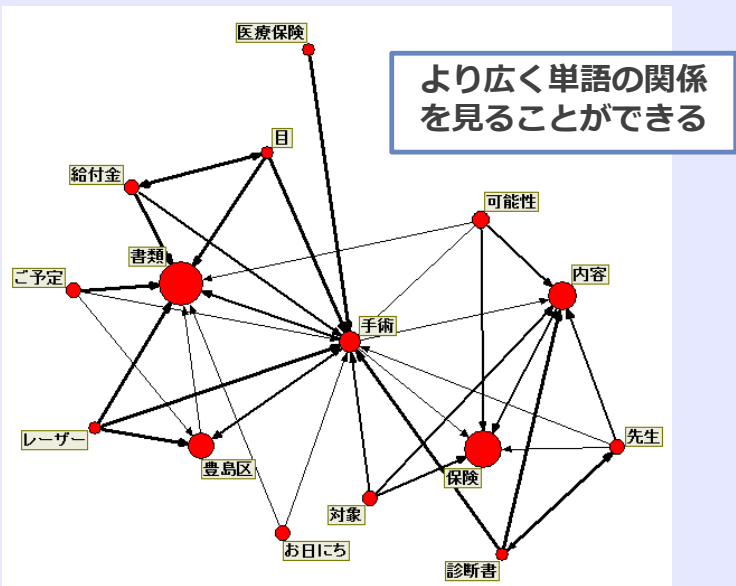
ことばネットワークの結果比較：

整形前



会議識別名	テキスト
20170728120000	そうです。胆石の手術で。
20170728120000	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 いますか。
20170728120000	あと胆石の手術なんです。
20170728120000	例えば胆石だけ取る手術ですか、あと、
20170717120000	えっとレーシックの手術を受けるんですよ。
20170717120000	レーシックの手術でございませぬ。はい。
20170717120000	はい。それではこちらレーシックの手術をされたということですが。
20170717120000	ではこちらレーシックの手術ということでございます。
20170629120000	あの目をねちょっと手術したんですよ。
20170629120000	目がね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だなっていませんでしたね。はい。では今回あのーえと再度ですね。レーザの凝固術を。されるご予定で。そうですか。ではえっとー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

2019年春
リリース!!

類義語アドオン

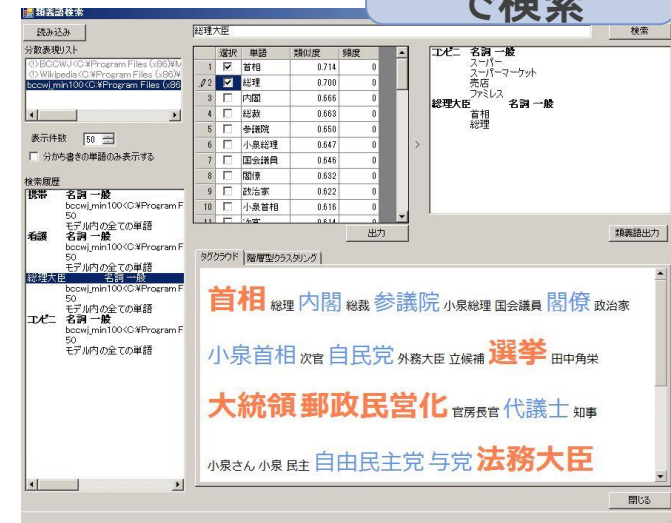
1. 分散表現(次頁説明)作成機能

- 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

「総理大臣」
で検索



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	...
...					...

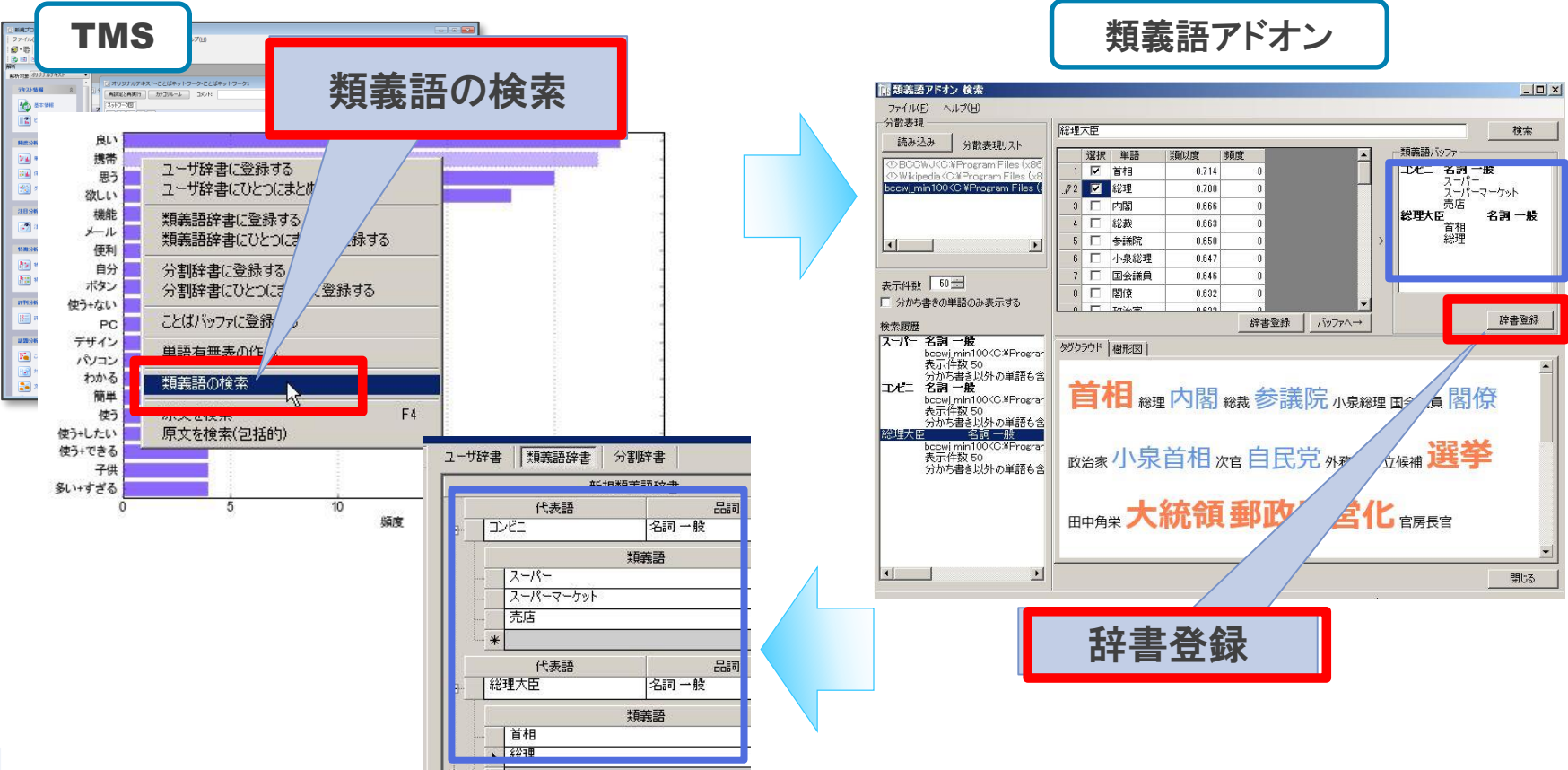
分散表現

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

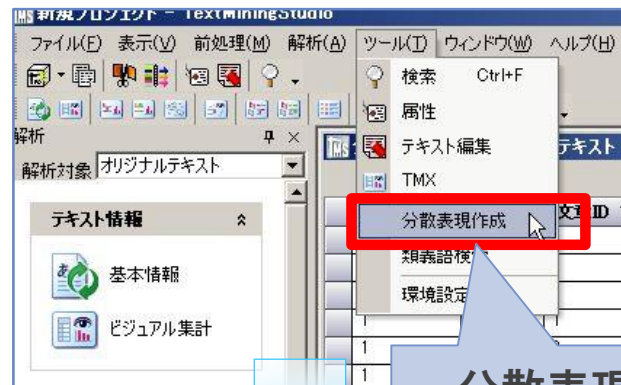
類似度

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

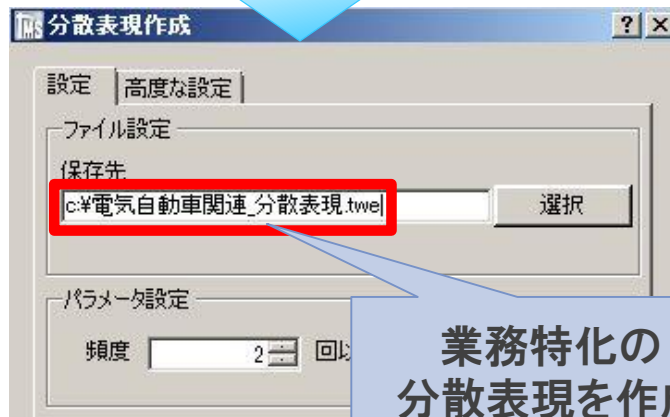


類義語アドオン – 分散表現作成機能1

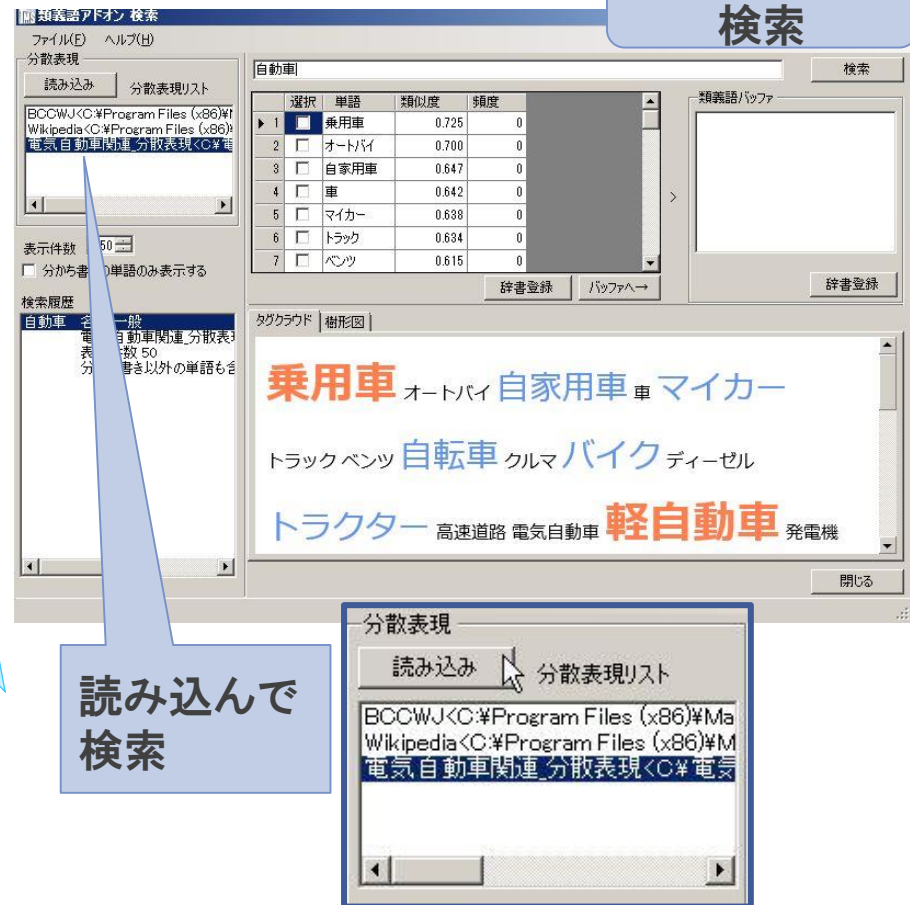
**TMSに入力したデータから特許、製造など
業務領域に特化した分散表現を構築できます**



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**

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



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

テキスト分類型の問題解決

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、 <u>ラベル付与基準</u> 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

テキスト分類型の問題解決

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
何といって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？

NO

まず傾向把握を行い、
それに基づいて
カテゴリ群を決定

YES

分類済のデータが
存在しているか？

NO

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

YES

分類の
アプローチは？

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

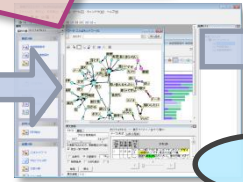
ラベル属性毎の単語の
出現状況の違いを把握し、
それに基づいてルール作成

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。

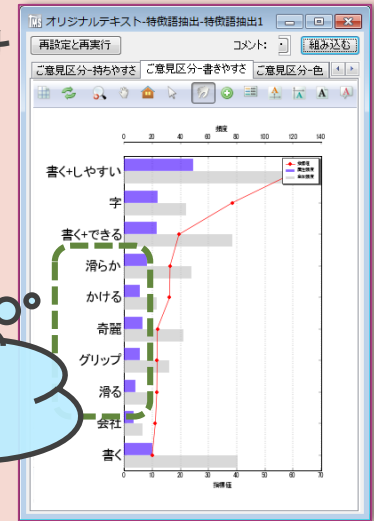
ご意見	ご意見区分
色がこれだけ揃っていると便利に使える。	色
カラーバリエーションがたくさんあってうれしい。	色
値段に見合った性能だと思う。	値段
とにかく安い！買いためています。	値段
何といっても書きやすさはピカイチ。	書きやすさ
サラサラ書いて、書き出しも線が切れない。すごい。	書きやすさ

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

ルール追加

グループ名 頻度 品詞

色	165	グループ
値段	121	グループ
書きやすさ	236	グループ

要素1 要素2 要素3 頻度

書く+しやすい			139
書き味→良い			26
字→書く+できる			15
滑らか→書く+できる			9
良い→書く+できる			9
滑らか→書き味			6
書き味→書く+できる			5
書く+できる→良い			8
字→かける			6
スムーズ→書く+できる			2
書き味→気持ちいい			3
書く+しにくい			5

表示モード切替 モード1 モード2 自動展開

グループ自動生成 属性からのグループ自動作成 フィルタリング条件として登録 書き出し 読み込み 頻度計算 カテゴリへ

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、楽しさが少なすぎます。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ます。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	嫌+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	
好評	0	0	2	0	0	1	0	0	0	0	
好評	1	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	納期 支払い 発送 その他

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

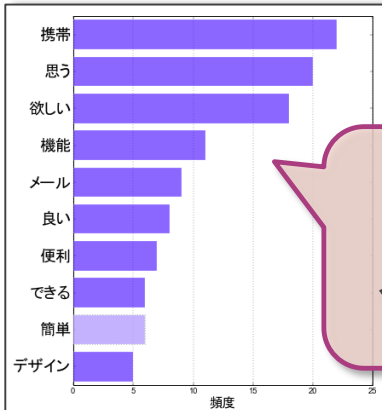


当社スタッフが！

サマリー型の問題解決

・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...

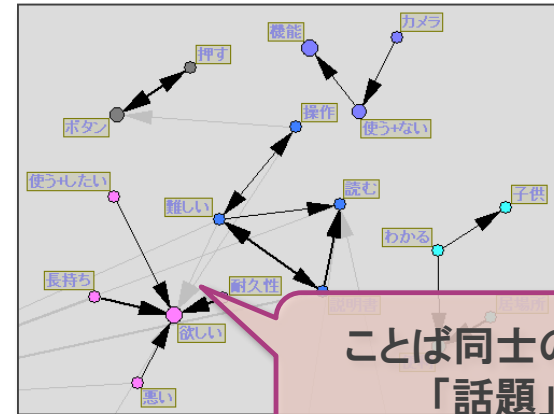
✓ 態度表現で絞る

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！

ケーススタディー その2 : サマリー型

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。

データの整形

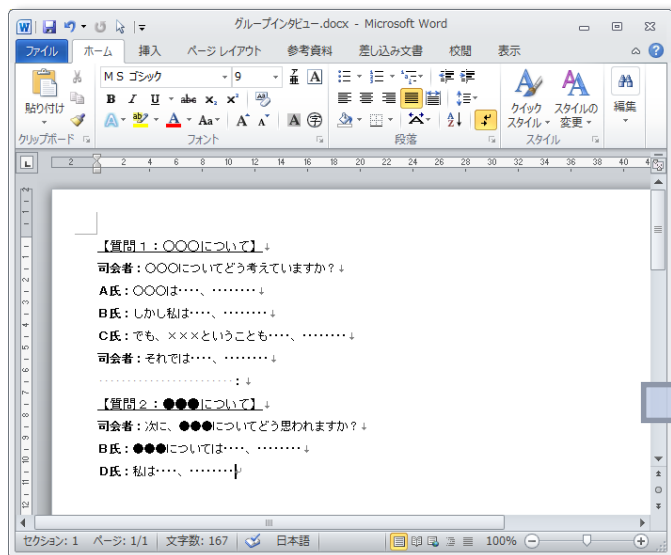
• データの「1件」とはどういう単位？

- 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ろよ。
17	男	いや判るすま、つか使います？でも、バカ判りな感じでほしい
33	女	もっと長く使いたいのに、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ってラジオが壊れてしまったらどうしよう。こんな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。
55	男	まず操作方法がわからない。情けないが頭に全部やっけてもらってる。
49	女	最近老眼になってきて、メールを読むにも一苦労です。文字が大きく表示されるような携帯がほしいです。
46	女	説明書も読んでも、厚くて持ち運びが面倒くさいし、もっとシンプルにしたい。

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	OOOについてどう考えていますか？
1	A	OOOは……、………
1	B	しかし私は……、………
1	C	でも、××××ということも……、………
1	司会者	それぞれ……、………
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

ケーススタディー その3 : アカデミック・発見型



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、特別な固有名称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：「
入力データ設定に保護ファイルで設定されていた場合には、保護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？
- 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！

まずはセミナーへのお申し込みをお待ちしております

TMS無料体験セミナー

- 12/5(木) 名古屋
 - 10:00 ～ 12:00 アカデミックセミナー
 - 看護現場のテキストデータを用いた分析演習
 - 13:30 ～ 16:30 製造セミナー
 - ユーザーの声の分析、自動車不具合データの分析演習
- 12/13(金) 東京
 - 13:30 ～ 17:00 **TMS**体験セミナー
 - ユーザーの声の分析、特許情報、SNSの分析演習

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

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】 tmstudio-info@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