

AIツール入門講座

TETDM: テキストマイニングツール

滋賀県立大学 工学部
砂山渡

1

本日の資料置き場

<http://tetdm.jp/> ->TETDMのイベント->講座当日資料

皆様のタイミングでダウンロードをお願いします

#資料ページへのリンクは本日限りです

2

データとは

⇒ 電子的に収集, 保存, 伝達可能な
数字または文字による情報

ex) 電子メールやSNSのメッセージ
店舗の売り上げや顧客情報

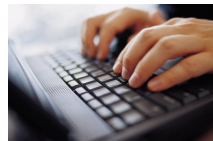
⇒ データの種類

– 数値データ

• 実数値、整数値、カテゴリデータの数値化

– テキストデータ

• 日本語で書かれたデータ、数値データのテキスト化

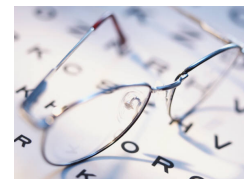


3

データ分析とは

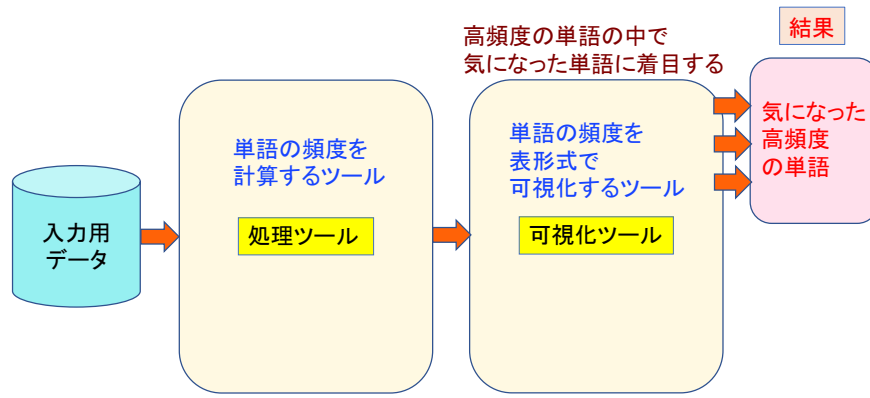
⇒ 意思決定の判断材料とするために
データの背後に潜む知識を得ること

世の中の事象間の
因果関係(または相関)



4

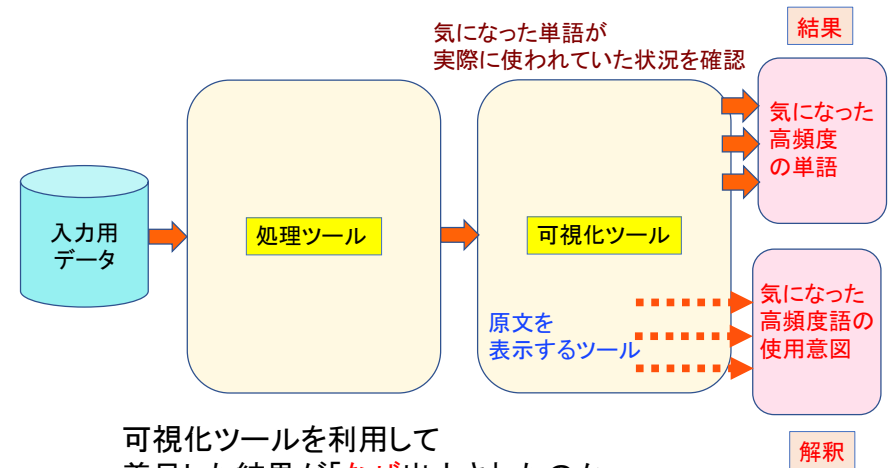
テキストデータの分析では まずは高頻度の単語に着目



処理ツールの動作原理ではなく
「高頻度語は**主題**や**多数意見**を表す」
という意味が重要

5

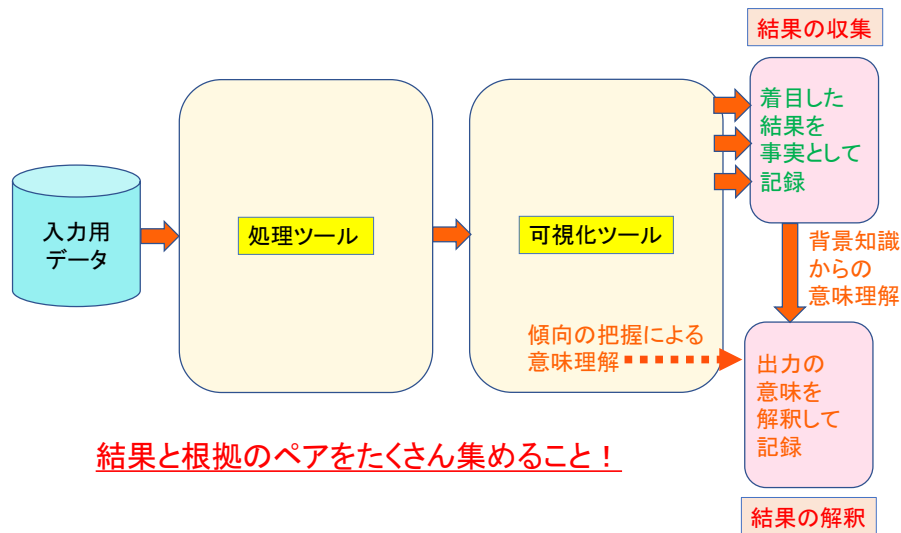
着目した単語の使用傾向を確認



可視化ツールを利用して
着目した結果が「**なぜ**出力されたのか」
出力の意味(根拠)を捉えることが重要

6

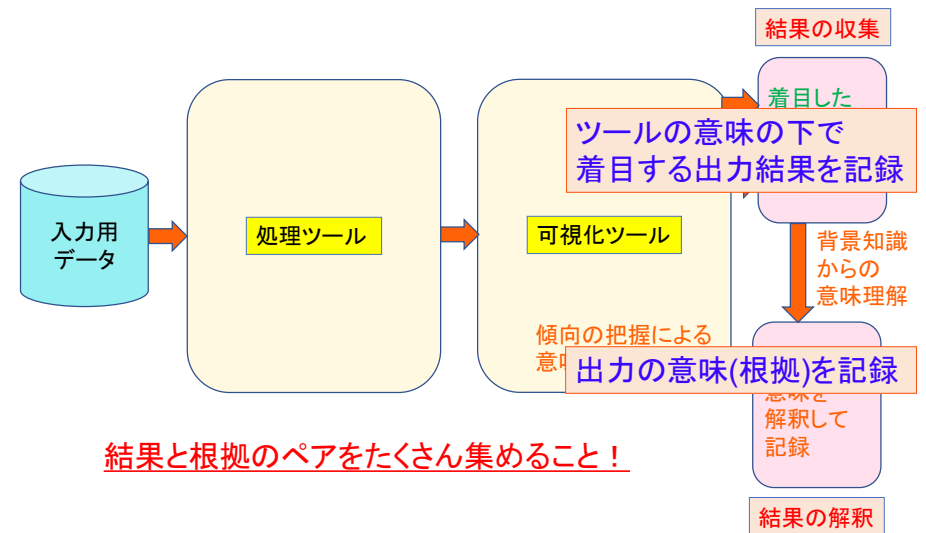
最も端的なデータ分析とは



結果と根拠のペアをたくさん集めること！

7

最も端的なデータ分析とは



結果と根拠のペアをたくさん集めること！

8

テキストマイニングツール TETDM

テキストマイニングのための統合環境 TETDM <http://tetdm.jp/>

The screenshot shows the TETDM web interface with three panels:

- Panel 1: Text Evaluation Application** (Text: urashima.txt2)
 - Text: urashima.txt2
 - Text Basic Data: 5124 characters, 4 segments, 94 words, 115 keywords.
 - Text Evaluation: Topic coherence (63%), Topic consistency (61%), Article structure (89%), Unnecessary expressions (23).
 - Article Structure: Main theme coherence (89%), Main theme consistency (61%), Article structure (89%), Unnecessary expressions (23).
- Panel 2: Text Evaluation (Analysis Results)** (Text: urashima.txt2)
 - Main theme: 浦島 (32)
 - Text: urashima.txt2
 - Text: urashima.txt2
- Panel 3: Text Editor + Text Display** (Text: urashima.txt2)
 - Text: urashima.txt2
 - Text: urashima.txt2

テキスト分析の統合環境 TETDM

- 複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境
- 電子テキストを扱う多くのユーザの創造的活動を支援

2010年度から5年以内に達成する課題として人工知能学会全国大会「近未来チャレンジ」のプロジェクトとして発足、2015年度の大会で卒業認定!

9

TETDMの構成

メニューウインドウ

The screenshot shows the TETDM menu window with various analysis and visualization tools:

- Text [urashima.txt2]
- Text Basic Data: 5124 characters, 4 segments, 94 words, 115 keywords.
- Text Evaluation: Topic coherence (63%), Topic consistency (61%), Article structure (89%), Unnecessary expressions (23).
- Article Structure: Main theme coherence (89%), Main theme consistency (61%), Article structure (89%), Unnecessary expressions (23).
- Text Display: urashima.txt2

パネル: 処理ツールと可視化ツールのペアがセットされる

12

TETDMの導入

- ⇒ OSに依存せずダウンロードのみでインストール完了
 - TETDMはJava言語で記述されている
 - Javaの実行環境のみインストールが必要
- ⇒ 動作環境
 - メモリ1.5GB以上（バージョン4.0から）
 - 横方向の解像度1600pixel以上推奨

13

TETDMの起動

- 以下のファイルをダブルクリック
 - TETDM-4.00.jar
 - TETDM.bat (Windows)
 - TETDM.command (Mac)
- コマンドプロンプト/ターミナルでコマンド入力

```
java -Xmx1536 -jar TETDM.jar text/urashima.txt
```
- 以下のファイルに入力テキストファイルをドロップ
 - TETDM.bat (Windows)
 - TETDM.app (Mac)

14

TETDMの起動モード

- スーパーライトモード
 - 基本的なデータ分析を試したい人向け
- ライトモード
 - 入力単語の設定をして分析したい人向け
- 通常モード
 - 利用するツールの組み合わせを自分で考えて分析したい人向け
- 拡張モード
 - さまざまなツールからデータ分析の結果を幅広く集めたい人向け



熟練度に応じた
起動モードを用意

15

TETDMへのテキスト入力

- ⇒ コピー＆ペーストによる入力
 - 処理ツール「エディタ」があるパネルにコピーして「保存＋実行」
- ⇒ ファイルによる入力
 - 通常モードか拡張モードでメニューウィンドウの「ファイル」ボタンから
 - 日本語文字コードはShift-JISかEUC
 - オプションでUTF-8を選択可能



16

TETDMの操作方法の習得

- TETDMサイトの「TETDMGUIDE」を参照する
- 「キャラクターアシストチュートリアル」を実施する
- TETDMサイトの利用者向け情報を確認する
- TETDMのツールセミナーに参加する



デモンストレーション

電子カルテの分析事例

- 宮崎大学医学部附属病院との共同研究では新人(経験年数3年以下)の看護師とベテラン(同6年以上)の看護師のカルテを比較して分析

新人とベテランのそれぞれでよく使われやすい単語に着目

新人作成カルテの特徴

| 特徴 | 協力者 | 着目単語 |
|-----------------------|------------|---------------------------------------|
| あいまいな言葉が多い | A,B,D, E,H | キロ, サイド, クリア, ルート, 上記, BP, Bp |
| 観察, 測定したものをそのまま記載している | B,D,F | kg, Wt, CV, BP, SpO2, 低下, カニニューラ, トイレ |
| 具体的でない・詳細が記入されていない | C,D,E, G | トラブル, 不良, 少量, 安定 |

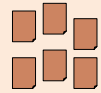
ベテラン作成カルテの特徴

| 特徴 | 協力者 | 着目単語 |
|-------------------------|--------------|-----------------------------------|
| 他者とのつながり, 医師とのやりとり | A,D,E F,G | 依頼, 当直, 医師, 母親, 家族, 報告, Dr. |
| 状態を表す記載 が多い | A,B | 発熱, 閉眼, 排尿 |
| 管理・安全など症状 以外の事が書いてある | D,F,G | 管理, 確認 |

データ分析による 意思決定プロセス

生じやすいデータ分析プロセスへの誤解1

蓄積データ



手元に蓄積されてきたデータを
ぽんと分析ツールに入力したら
すぐに使える知識を出してくれるんじゃないの？

データ処理 データ可視化

分析処理ツール

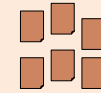
コンピュータ任せでは有効な知識は得られない



とりあえずツールを導入しよう！
ツールを触る担当者を
一人くらいつけとけば何とかなるよね

生じやすいデータ分析プロセスへの誤解2

蓄積データ



データ分析ツールって
数学とかの専門知識がないと
使えないでしょ

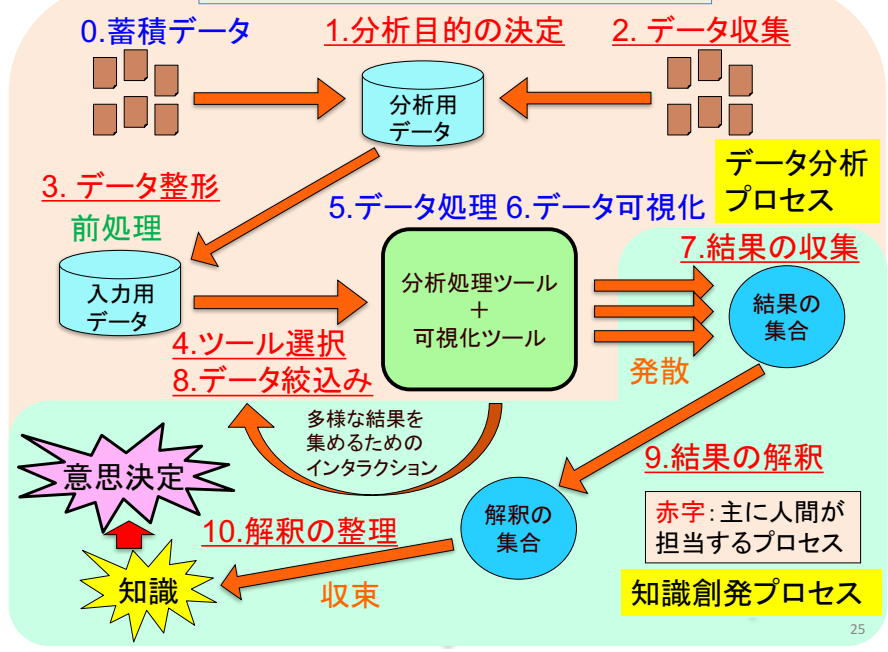
データ処理 データ可視化

データを活用したい人が分析しないと
有効な知識は得られない



データ分析は専門家に外注して
結果だけ送ってもらおう

データ分析による意思決定のプロセス

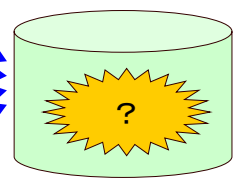


TETDMの適用範囲

- データ分析による意思決定のプロセスの4.ツール選択 から10.解釈の整理 までを幅広くカバーしている
 - 2.データ収集 3.データ整形 も一部カバーしている
- 分析対象は日本語テキストデータとする
 - 数値データ、英語データにも一部対応している

0.蓄積データ

- 手元の業務データ
- オープンデータ
- 紙データの電子化
- 音声のテキスト化

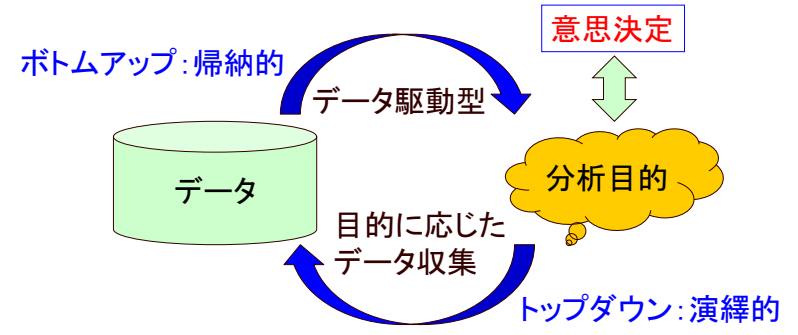


これらのデータを分析すれば「何か」新しい事実がわかるかもしれない

「何か」とは何か？

1.分析目的の決定

- データが先か目的が先か



分析の目的と分析指針

⇒ 主要な知識の獲得

- 主要なデータ
- メジャーなツール
- 堅実な解釈

現状を把握したい



新しいアイデアを出したい

⇒ 新しい知識の獲得

- 幅広いデータ
- マイナーなツールを含む
- 論理の飛躍を含む解釈

29

2. データ収集

⇒ データの領域

- 獲得したい知識を潜在的に含むか

⇒ データの粒度

- 欲しい知識の詳細さに応じた細かさがあるか

⇒ データの量

- 得た結果を一般的に適用できるか

⇒ データの質

- データに偏りがいないか

分析の目的に応じて
収集するデータを吟味

30

データの価値の見積もり

⇒ データの価値

- 潜在的にデータが含むと推定される知識の価値

⇒ 価値推定のための仮説(知識)

- 「データ」から得られると推定される知識の「仮説」

ex) データ: 「年齢情報を含む顧客の購買履歴」

→ 仮説: 「特定の年代の人に、よく売れている商品」

仮説を得て価値を見積もるためには
分析手法の(定性的な)意味理解も必要

31

3. データ整形

⇒ 数値データのデータ整形

- CSV形式のファイルを用意
- 欠損値の補完/補間

-分析ツールを用いるため
-分析結果の質を高めるため

⇒ テキストデータのデータ整形

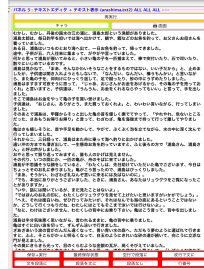
- 日本語文字コードの統一
- 句点(。)の挿入
- 分析に不要な単語(ストップワード)の除去

32

TETDMで可能なデータ整形

⇒ 処理ツール「テキストエディタ」によるデータ整形

- 空行で段落に
- 改行で文に
- 文を段落に
- 段落を文に
- 段落なし



文や段落(セグメント)の区切りを自動的に生成

33

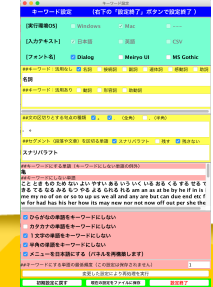
TETDMの起動確認

35

TETDMで可能なデータ整形

⇒ 「キーワード設定」によるデータ整形

- 分析対象品詞の指定
- 分析対象外の単語(ストップワード)の指定
 - ひらがな、カタカナ、1文字、半角文字の除外指定
 - 単語指定除外
 - 除外指定の例外の指定
- 分析対象単語の最低頻度の指定



34

TETDMの起動確認

⇒ お手元のPCでTETDMが起動することを確認してください

- 起動できない場合
 - Javaインストール(パスの設定含む)の確認
 - TETDMフォルダの場所の確認
 - コマンドラインから起動してエラーを確認

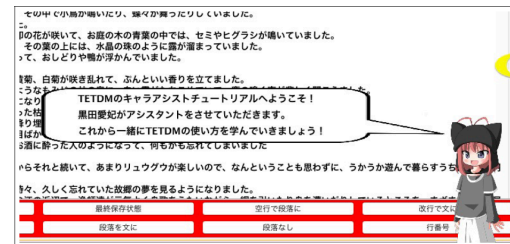
36

TETDMにおける データ分析の基本操作

37

キャラクターアシスト チュートリアル

- ⇒ キャラクターがTETDMの使い方やデータ分析の手順を説明
 - クイズや操作課題も出題される



ゲームモードを進めると
キャラクターを変更可能

38

キャラクターアシスト チュートリアルの構成

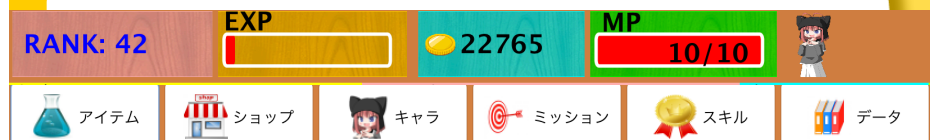
- ⇒ TETDMの操作方法
 - 操作の説明
 - TETDMの機能の説明と操作方法
 - ツールの説明
 - 処理ツールの説明と操作方法
- ⇒ テキストマイニングの流れ
 - データ分析の手順に応じた使い方

39

ゲームモード

- ⇒ TETDMの利用意欲を高めるためのモード
 - 経験値を獲得するとランクが上がる
 - コインを獲得するとキャラクターが獲得できるクイズや操作課題も出題される
 - 経験値やコインを獲得するためのミッションも用意されている

学生には好評



40

チュートリアル

- ⇒ スーパーライトモード(50分程度)
 - TETDMの操作方法
 - 操作の説明(15分)
 - ツールの説明(25分)
 - マイニングの流れ(10分)
- ⇒ スーパーライトモードがすべて終わった人は次のモードに進んでください



41

データ分析ツールと データ分析スキル

43

お昼休み

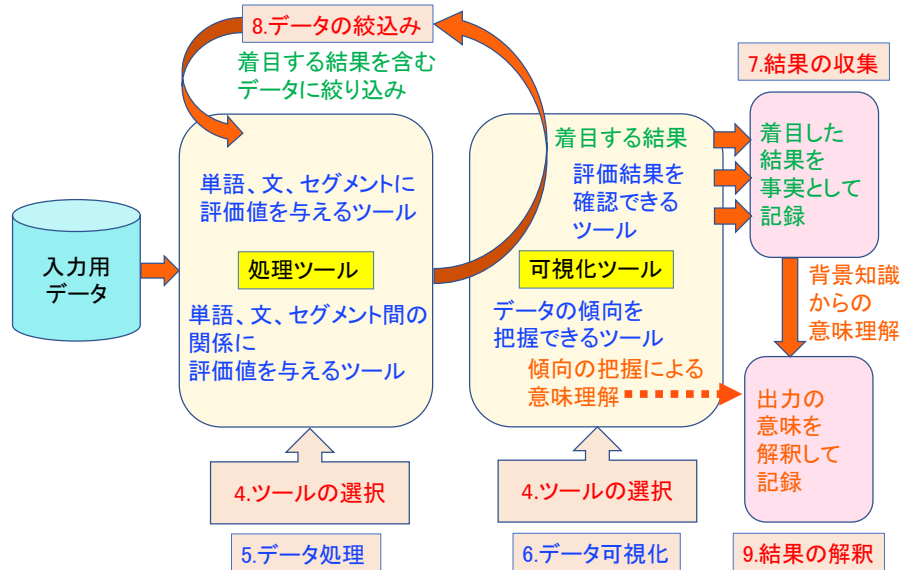
42

データ分析スキルとは

- ⇒ データ分析プロセスを理解して一連のプロセス全体を独自に実行できるスキル
 - 簡単にでも**全ての手順を理解している**ことが大事
 - ツールや分析手法に関する知識が多いに越したことはないが**焦らずに1つずつ使えるツールを増やしていけば良い**

44

4. ツール選択 から 9. 結果の解釈 の詳細



45

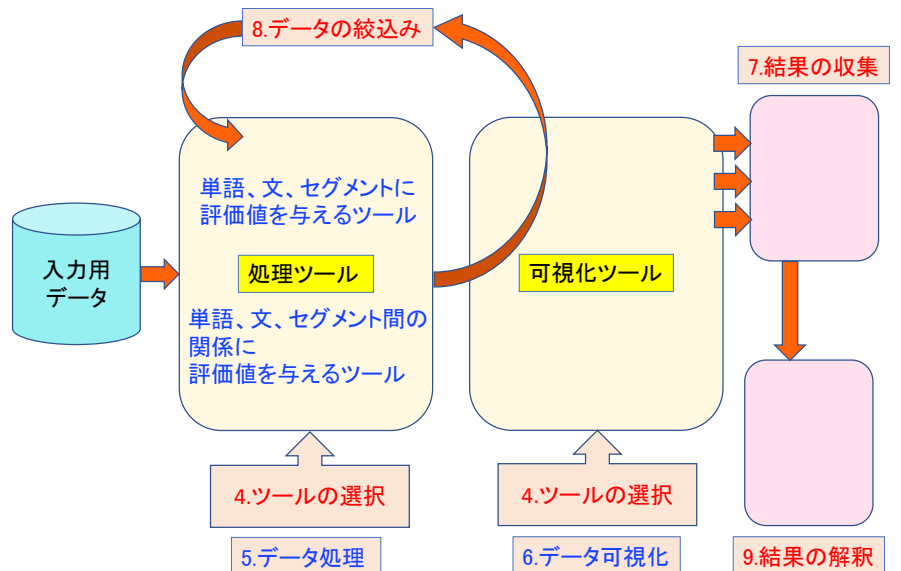
5. データ処理

ツールの処理は、以下の6つに大別される

- 単語の評価
- 文の評価
- セグメント(段落, 文章)の評価
- 単語間の関連度評価
- 文間の関連度評価
- セグメント間の関連度評価

46

5. データ処理



47

セグメント評価(セグメントにフォーカス)

セグメント情報まとめ
レポート評価(結果+意見)
コミュニケーション要約

文評価(文にフォーカス)

文情報まとめ 文評価まとめ 文章要約(展望台)
意見理解文抽出 主題関連文評価(光と影)
単語抽出(文章評価用) 単語冗長文抽出
類似文抽出 失礼単語抽出
長文抽出 主語抽出 主題関連語評価(川下り)

単語評価(単語にフォーカス)

単語情報まとめ 単語抽出 TFIDF
単語頻度リスト 文内キーワード抽出(Yahoo)
専門用語抽出(FLR) 専門用語抽出(C-Value)
専門用語ハイライト 文内キーワード抽出(頻度)

テキスト評価

テキスト評価(分析結果まとめ)
主題語含有率 漢字の割合

文書整形

テキストエディタ 単語置換
セットセグメント テキスト2分割
段落並び替え

セグメント間, 文間, 単語間の関連度評価

単語間関連度 関連単語情報
関連文・関連セグメント情報
段落順序評価(トップダウン)
テキスト分類(再帰的クラスタリング)
テキスト間類似度(独自性)
段落間木構造(類似度, トップダウン)
段落間ネットワーク 段落間類似度表示
(類似度順, ばねモデル)

処理ツール一覧

ツール選択支援

英文読解アプリケーション
タイピングアプリケーション
テキスト評価アプリケーション
テキスト集合評価アプリケーション
文章推敲アプリケーション
プロフィールチェック

その他

形態素解析(Igo) 英文音読 なし
形態素解析サンプル(Igo)
辞書(オンライン) 辞書再構築
タイピング フォーカス情報
プログラムチェック サンプル1
データ送信テスト サンプル2
連動データ確認(フォーカス) 48

Ver.4.0時点

* 茶色は可視化ツール

情報アクセス

Webブラウザ
URLアクセス
2ちゃんアクセス

4. ツールの選択

処理ツール

- ⇒ 確認したい情報を評価するツールを選択
 - 多数意見, 少数意見
 - 全体的な単語のor特定の単語集合の使用傾向
 - 指定条件に当てはまる文章
- ⇒ ツールの切り替えにより異なる視点からの結果を収集
 - 使えるツールの種類が多いほど多様な結果を集められる

49

処理ツールの信頼度

- ⇒ 客観的な根拠資料の有無
 - 学術論文として公開されているか
 - 客観的な精度評価の結果があるか
 - ⇒ 直感的な出力の妥当性の有無
 - 入力と出力の対応関係に違和感はないか
- どんなツールを使う場合でも
出力結果を100%信頼できることはない
→使用するツールの信頼度と組み合わせて結果を取り扱う

50

結果の信頼度と分析の目的

- ⇒ メジャーなツールを用いると
 - 結果の信頼度は高くなる
 - 当たり前or既知の事実が多くなる
 - 少数意見、例外が無視されやすい
 - ⇒ マイナーなツールを用いると
 - 結果の信頼度は必ずしも高くない
 - 可能性を検討するための多様な結果を集められる
- 単語間関連速度:
空港-飛行機(90%)
男性-スカート(10%)
- 現状を把握したい
- 失礼単語抽出:
失礼になる可能性
がある単語を抽出
- 新しいアイデアを出したい

51

処理ツールの処理内容の理解

処理ツールの意味理解

- ⇒ 処理ツールの処理
 - 入力テキストを評価
 - ⇒ 処理内容の理解
 - 評価する計算式やルールを理解
- 複雑な計算式を理解できなくても、定性的に評価アルゴリズムを理解することができれば出力の意味が理解可能

処理内容の理解は、ツールの信頼度の把握にもつながる

52

処理ツールの意味理解の具体例

- ⇒ 文章要約(展望台)
 - 文章の中の主題を表す単語と重要文を抽出
- ⇒ 主題関連文評価(光と影)
 - 主題を表す単語に関係する文を抽出
- ⇒ 主題関連語評価(川下り)
 - 主題を表す単語に関係する単語を抽出
- ⇒ 意見理解文抽出
 - 意見の理解を助ける文を抽出

多くの場合
この程度の
理解で十分

53

処理ツールのアルゴリズムの例

意味によるアルゴリズムの説明

- ⇒ 文章要約(展望台)
 - 文章中の高頻度語を含む文によく使われている単語を主題を表す単語(観点語)として抽出
 - 観点語を含む文にしか使われていない単語(特徴語)を抽出して、観点語と特徴語を含む文を重要文として抽出

中身を簡単に理解したい場合
この程度の理解で十分
ただし情報が提供されるとは限らない

54

処理ツールのアルゴリズムの例

数式によるアルゴリズムの説明

- ⇒ 文章要約(展望台) このレベルの理解は基本不要
 - S: 観点語(頻度上位5個がSの初期単語集合)
key2の評価値が高い上位5個の単語でSを更新、を繰り返し更新されなくなった時のSを主題を表す単語(観点語)として抽出
 - 観点語Sを使ってkey3の評価値を計算して特徴語を抽出
 - 各文が含む単語にkey2とkey3の評価値を与えてその合計を文の評価値として、評価値の高い文を重要文として抽出

$$key2(w) = \prod_{s \in S} \left(\frac{n(w \cap s)^2}{n(w)n(s)} + 0.8 \right)$$

+0.8は共起頻度が0になった時に0が掛けられるのを避けるため。
(単語評価値順位には影響がない)

$$key3(w) = \prod_{s \in S} \left(\frac{n(w \cap s)}{n(w)} + 0.8 \right)$$

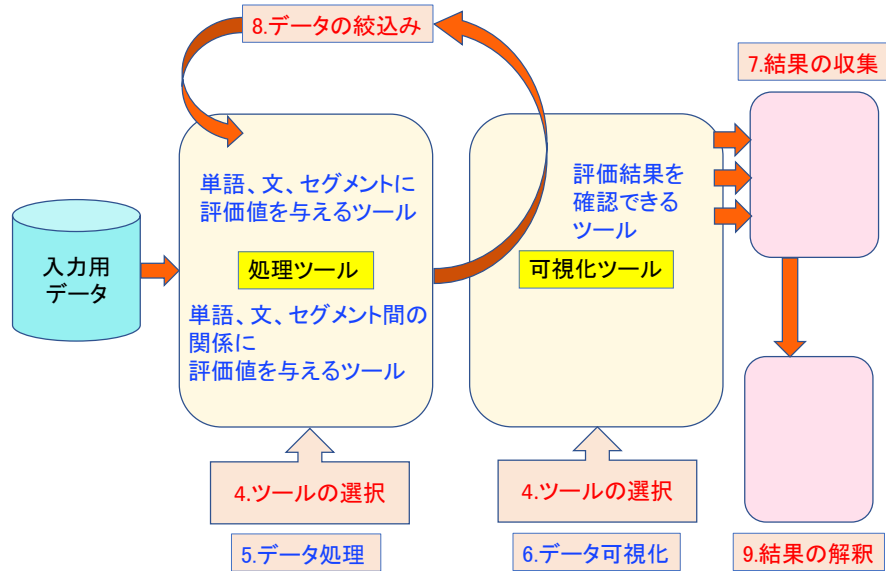
55

6. データ可視化

- ⇒ 可視化ツールは、以下の4つに大別される
 - データ表示フォーマット(一次元データ表示)
 - 関連度表示(二次元データ表示)
 - テキスト表示フォーマット(テキスト表示)
 - 特定処理ツールの結果の可視化

56

6.データ可視化



57

テキスト表示フォーマット

テキスト表示 テキスト表示(カラー)
 テキスト表示 (HTML)
 テキスト表示(選択)
 テキスト表示(2テキスト)
 ファイル内容表示

データ表示フォーマット

円グラフ 棒グラフ(横向き)
 折れ線グラフ
 折れ線グラフ(表示調整可)
 レーダーチャート
 順位表示(評価値順)

前処理結果を利用した可視化

段落間木構造(トップダウン)
 段落間類似度表示
 段落間ネットワーク(類似度順)
 段落間木構造(類似度)
 段落間ネットワーク(ばねモデル)

データ選択支援

表形式表示
 キーワード選択(フォーカス指定)

特定処理ツールの結果の可視化

キーワード表示(展望台)
 キーワード集合比較
 主題関連語表示(川形式)
 単語間関係(関連度)
 文内キーワード表示(タグクラウド)
 専門用語表示
 段落間ネットワーク(評価値順)
 テキスト分類表示(地図形式)
 段落並び替え

その他

連動データ表示(フォーカス)
 サンプル表示1 なし
 サンプル表示2
 タイピング問題(数字)
 タイピング問題(色)

Ver.4.0時点

可視化ツール一覧

58

4.ツールの選択

可視化ツール

- ⑤ 処理ツールの出力を可視化できるツールを選択
 - 処理ツールが「一次元の数値データ」を出力する場合
「一次元の数値データ」を可視化するツールを選択可能
- ⑤ 処理ツールが複数の出力を行う場合
表示内容を可視化ツールで選択
 - 表示スペースが限られているため
分析しやすいパネル構成を検討して選択

ツール選択時に組合せ可能な
処理ツールと可視化ツールのペアが明示される

59

可視化ツールが行う処理

可視化ツールの意味理解

- ⑤ テキストによる結果の可視化
 - 抽出された一部のテキスト, 文, 単語のハイライト
 - ⑤ 数値による結果の可視化
 - 数字を用いたグラフ
 - 数字の大きさを反映した可視化
 - ⑤ 独自の可視化フォーマット
- 出力形式を理解して処理結果の把握と意味理解につなげる

60

チュートリアル

⑤ ライトモード、通常モード、拡張モード

– TETDMの操作方法

- 操作の説明(15分、20分、10分程度)
- ツールの説明(5分、10分、15分程度)

⑤ 上記が全て終わった人は
マイニングの流れに進んでください



61

データ分析と知識創発

62

7.結果の収集

思考の発散

⑤ データの背後に隠された知識(因果関係や相関)を導く手がかりを集める

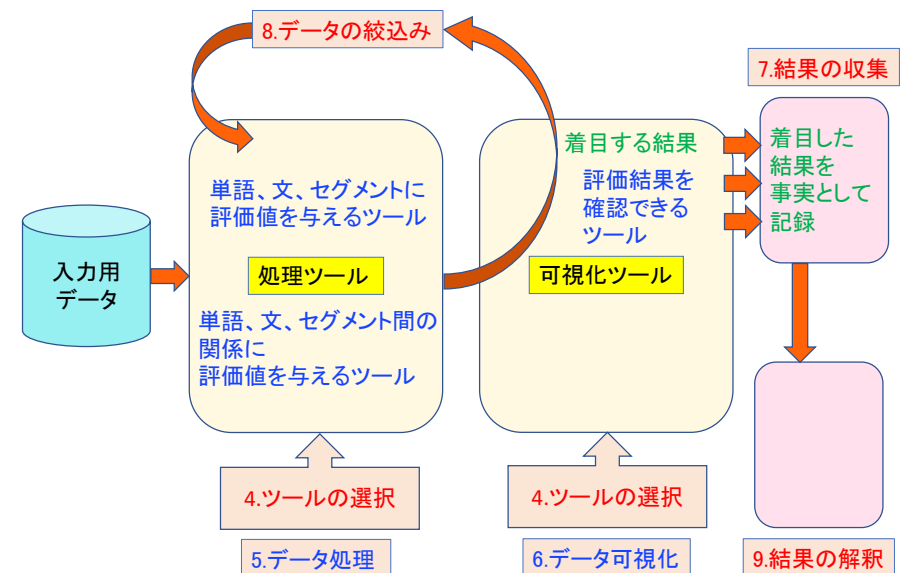
⑤ 多くの有効な手がかりを集めたい

– 多くの試行錯誤が必要

- ツール、処理方法、可視化方法、の切り替え
- どこに着目すればよいか
- 集めた手がかりをどのように記録するか

63

7.結果の収集



64

結果の収集における着目点

－ ずれからの知識発見

- ある基準データ(たとえば平均値)からのずれに意味を見出す
 - 平均気温
 - 平均睡眠時間
 - 確率の期待値
 - データの傾向(平均)を知ってそこからのずれを探る



普通or通常とは違う点はないか
→ 探偵のように違和感に気づけることが大事

65

結果と解釈の登録インターフェース

「結果」と「解釈」をその場で入力できる

結果と解釈の登録 1 / 60

- パネル 1: テキスト評価アプリケーション + テキスト表示 (HTML) (urashima.txt2)
- パネル 2: テキスト評価 (分析結果まとめ) + テキスト表示 (HTML) (urashima.txt2)
- パネル 3: テキストエディタ + テキスト表示 (urashima.txt2)

「！」追加

<結果> (必須)

<解釈> (必須)

[重要度] 5 4 3 2 1

[登録] 終了

[解釈登録手順]

- どのパネルの結果について登録するかを選択する (複数選択可能)
- 「！」追加ボタンで表示される「！」を、気になる結果の上に置く (任意)
- 気になる結果の内容 (事実) を具体的に、<結果>のところに書く
- 結果の意味するところ、結果から言えることを、<解釈>のところに書く
- 解釈の[重要度]を選択、5(特に重要)-4(重要)-3(まあ重要)-2(あまり重要でない)-1(重要でない)
- [登録]ボタンを押して、解釈を登録する (結果、解釈、重要度、画面が保存されます)

* <結果>と<解釈>は、自分以外の人にも伝わるように具体的に書くのがポイントです
* 登録した解釈は、メニュー知識発見の「一覧」、「開発」ボタンから確認、操作できます

66

8. データ絞り込み

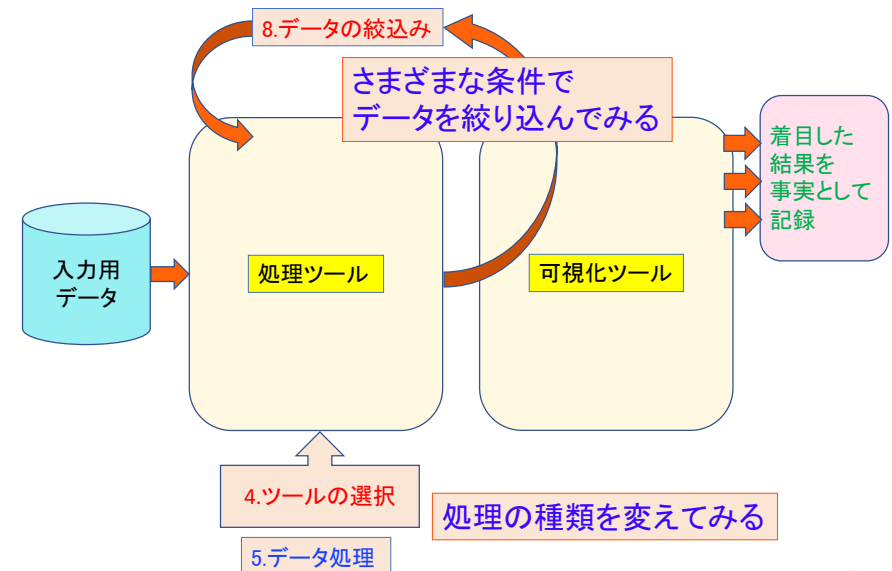
- 着目する情報によるデータの絞り込みが可能
 - 着目する単語を含むセグメント
 - 着目する文を含むセグメント
 - 指定の評価値や属性値をもつセグメント
- 着目している情報を明示するウィンドウを用意

データ (セグメント) の「絞り込み」条件と件数

| | |
|---------------------|-------|
| 表形式表示 で選択した 単語を含む | 3 乙姫 |
| 表形式表示 で選択した 単語を含む | 2 玉手箱 |
| → 表形式表示 で選択した 単語を含む | 1 寿命 |

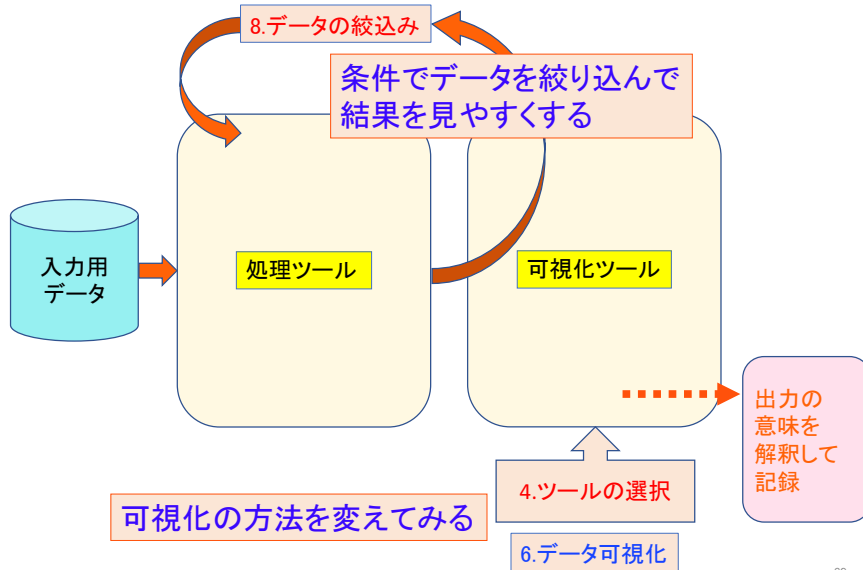
67

8. データ絞り込み → 結果の収集



68

8. データ絞り込み → 結果の根拠を得る



69

9. 結果の解釈

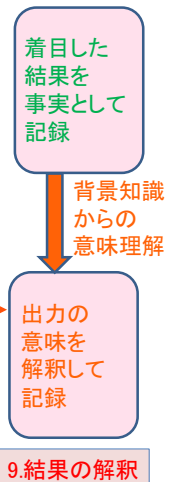
着目した結果が「なぜ出力されたのか」
出力の意味(根拠)を捉える

1. 可視化ツールを利用して
出力データの傾向*2をもとに解釈する
2. データについての
背景知識*1をもとに解釈する

傾向の把握による
意味理解

*1データの解釈には、データとして与えられていない知識が必要
データ分析の専門家が持ち合わせているとは限らない

*2傾向を把握(何が出力されているかを理解)するためにも
データについての背景知識が必要となることがある



70

分析結果の意味づけによる解釈

⇒ 解釈の必要性

- 「結果」はデータが示す事実や傾向
- 「結果」に「意味」を与えることで
意思決定に必要な「知識」のかけらにする

⇒ データとして与えられなかった知識の補完

- データの領域(ドメイン)の背景知識
- 一般的に既知の知識

71

分析結果の意味づけによる解釈

⇒ 結果の妥当性の確認

- 結果を100%は信頼しない

⇒ 背景知識による結果の解釈

- データが対象とする領域(ドメイン)の知識を活用

⇒ データの可視化による解釈

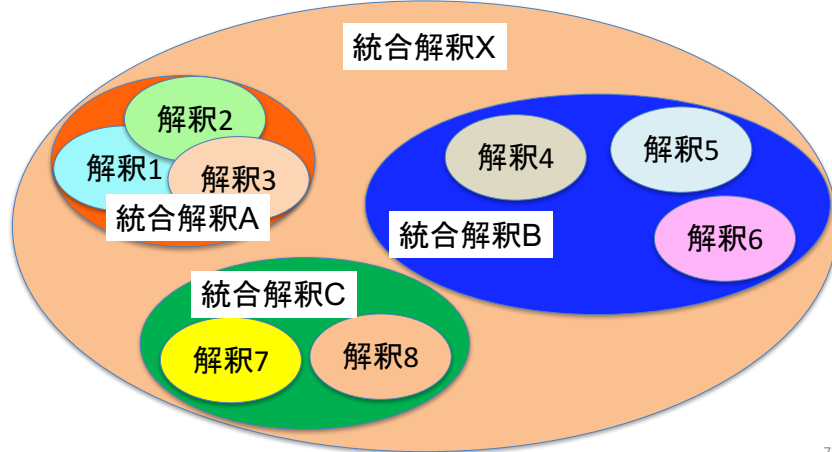
- 分析前の原文(元データ)を手掛かりに
背景知識を呼び起こす

72

10. 解釈の整理

思考の収束

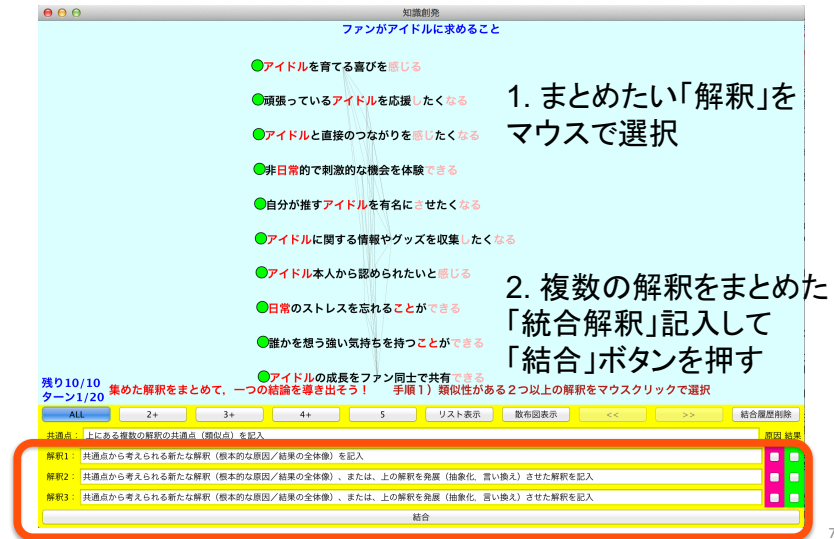
集めた解釈を再帰的に1つに統合して
データからわかることをまとめ、汎用的な知識につなげる(帰納推論)



73

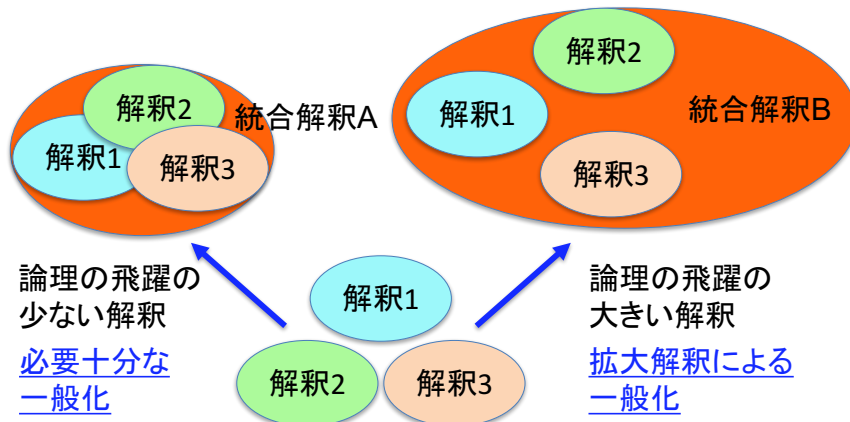
知識創発インタフェース

集めた「解釈」を再帰的に統合できる



74

分析結果の統合と論理の飛躍



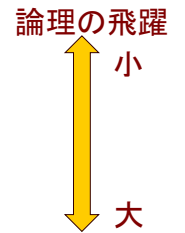
現状を把握したい

新しいアイデアを出したい

75

解釈の統合の例

- 解釈1: 「安いみかんがよく売れている」
- 解釈2: 「安いいちごがよく売れている」
- 解釈3: 「リンゴがよく売れている」
- 統合解釈1: 「安い果物がよく売れている」
- 統合解釈2: 「果物がよく売れている」
- 統合解釈3: 「ビタミンCがよく売れている」
- 統合解釈4: 「デザートがよく売れている」
- 統合解釈5: 「甘いものがよく売れている」



論理の飛躍には背景知識も必要

76

チュートリアル

- ⇒ ライトモード、通常モード1、通常モード2
 - **マイニングの流れ**
 - 10分、30分、40分程度
- ⇒ 上記が全て終わった人は旧チュートリアルへ
 - 一部が最新バージョンに対応していない可能性があります



77

実データによる データ分析実践

78

結果収集に向けた着眼の観点

基本的なポイント

- ⇒ **傾向の把握**
 - 単語の**頻度**情報、**関連度**情報
 - 文の**評価**情報、文間の**関連度**情報
 - セグメントの**評価**情報、セグメント間の**関連度**情報
- ⇒ **特徴の把握**
 - **気になった**単語、文、セグメントによるデータ絞込み

セグメント:
段落または文章

79

結果収集に向けた着眼の観点

レビュー分析のためのポイント

- ⇒ **傾向の把握**
 - レビュー全体を特徴付ける**頻度が高い**単語を見つけてその単語がレビュー内でよく使われた理由を調べる
 - [高評価/低評価]のレビューを特徴付ける**頻度が高い**単語を調べ、その使用理由を調べる
- ⇒ **特徴の把握**
 - **独自の意見**を述べているレビューを探す
 - 低頻度の単語の中から気になる**単語**を見つけてその単語がレビュー内で使われた理由を調べる
 - [高評価/低評価]レビューの中で[ネガティブ/ポジティブ]な**単語**を見つけ、その使用理由を調べる

80

データ分析課題

- ⇒ 各レビューにおいて高評価、低評価の理由をまとめ次の開発に向けたアイデアを提案してください

<分析手順>

1. 自由に分析
2. 「結果」と「解釈」の収集
3. 知識創発
4. アイデアの提案

81

分析結果の共有と創発

82

分析結果の共有

- ⇒ 以下の点について皆様方と情報共有
 - 分析結果と解釈
 - 知識創発の結果(アイデア)

83

協調的な創発

- ⇒ 複数人によるより多くの解釈を元に創発
- ⇒ 複数人による創発結果を統合して創発

cf) グループディスカッション

一人10個の解釈 × 10人分 → 100個の解釈から創発

↓
創発結果 × 10人分 → 10個の創発結果から創発

cf) メンバーの意見をリーダーが集約

84

まとめ

- ➡ 意思決定のためのデータ分析プロセスの全体像と各プロセスの意味について説明した
- ➡ テキストマイニングツールTETDMを用いて各プロセスを実現する方法について学んだ
- ➡ データ利活用の基盤となるツールとしてTETDMを活用いただければ幸いです

85

書籍化計画

「フリーソフトTETDMで学ぶ実践データ分析(仮)」

– コロナ社より2019年度内に出版予定

- 本ツール講座の内容を詳細に補完する内容
- データ分析による意思決定のプロセスを具体的に解説
- TETDMを利用して各プロセスを実際に試す
章末問題によりポイントを押さえたスキル獲得が可能

86