AIツール入門講座
「TETDM」

砂山　渡
広島市立大学

本日の内容

- テキストデータマイニングのための統合環境:TETDM (Total Environment for Text Data Mining)

  10:00-12:00

  – テキストマイニングとは？
  – 統合環境TETDMの基本的な使い方

  13:00-16:00

  – １つのツールを利用した分析
  – 複数のツールを利用した分析

  16:10-17:00

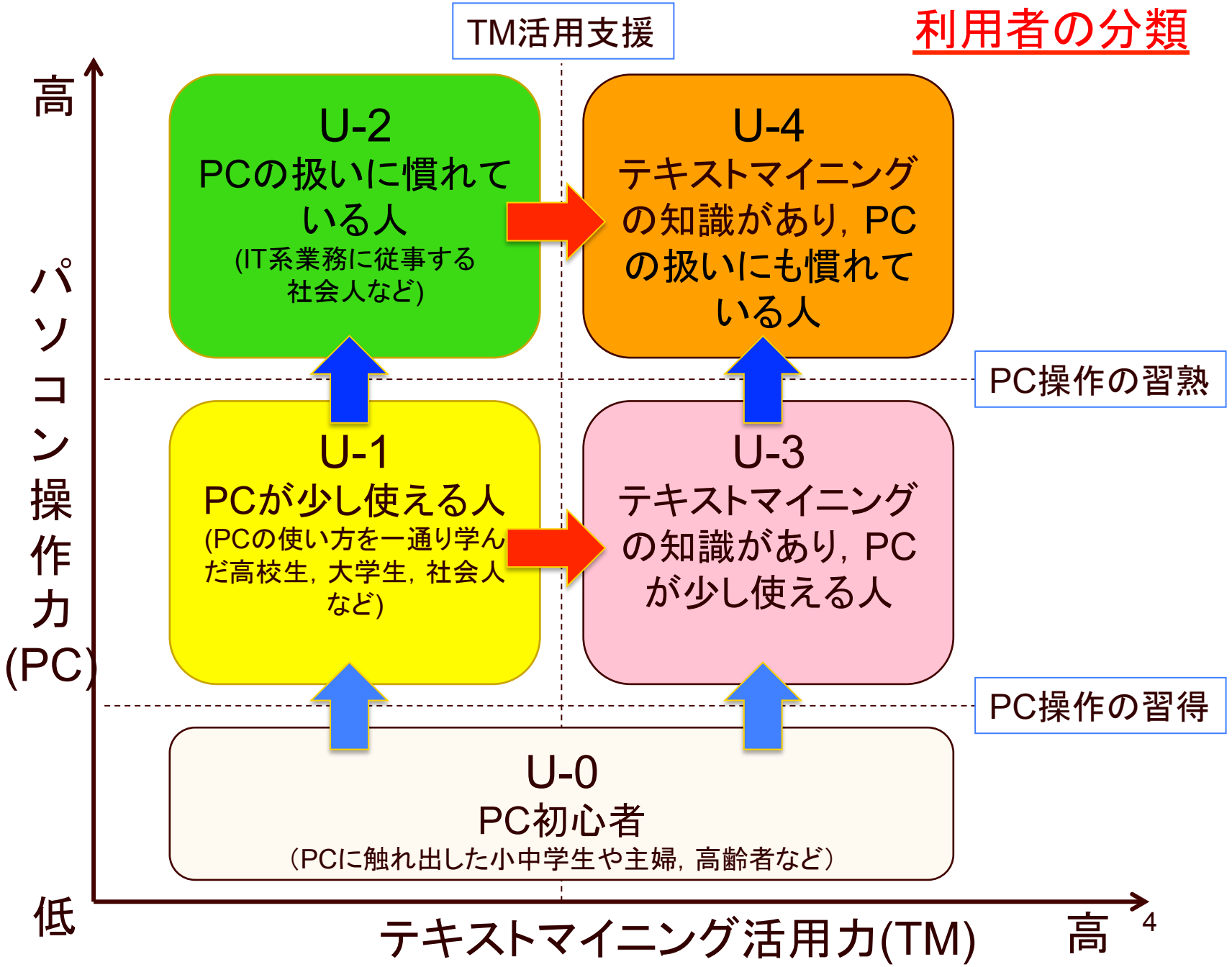
  – 自作作文の分析
  – テキストデータの分析

準備

- TETDM-0.54(お試し版)のダウンロード
  http://www.sys.info.hiroshima-cu.ac.jp/people/sunayama/tetdm-0.54.zip

- 本スライドの資料PDFのダウンロード
  http://www.sys.info.hiroshima-cu.ac.jp/people/sunayama/AItoolSeminarPart1.pdf
- http://www.sys.info.hiroshima-cu.ac.jp/people/sunayama/AItoolSeminarPart2.pdf

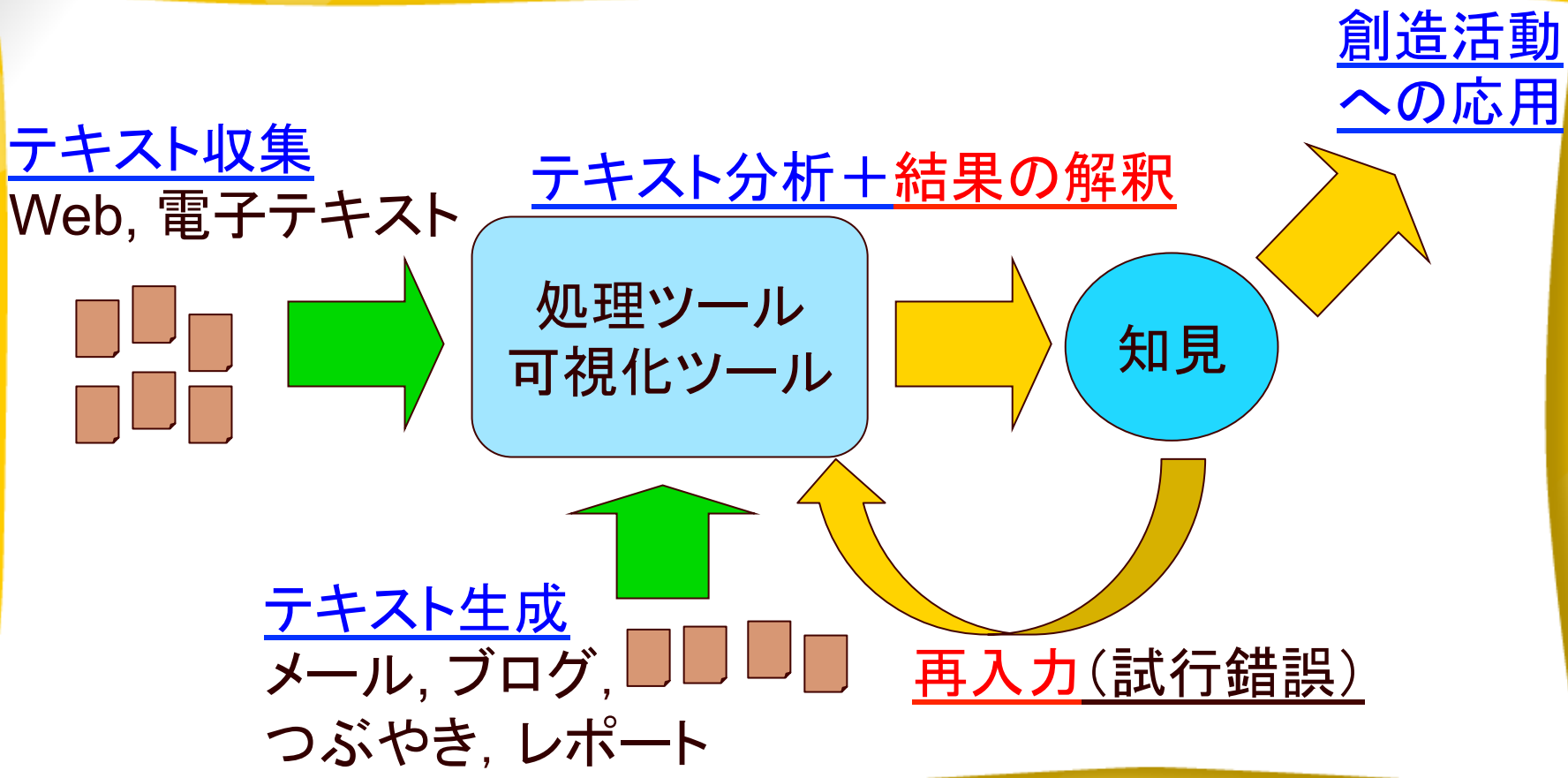
利用者の分類


TM活用支援
高
パソコン操作力(PC)
低
U-2
PCの扱いに慣れている人
(IT系業務に従事する社会人など)
U-4
テキストマイニングの知識があり，PCの扱いにも慣れている人
U-1
PCが少し使える人
(PCの使い方を一通り学んだ高校生，大学生，社会人など)
U-3
テキストマイニングの知識があり，PCが少し使える人
U-0
PC初心者
（PCに触れ出した小中学生や主婦，高齢者など）
PC操作の習熟
PC操作の習得
テキストマイニング活用力(TM)
高

テキストマイニングとは？

- テキストデータ（文章あるいは文章の集合）を処理し，人間が新たな知識を発見する過程（プロセス）
- 「データマイニング」の中で，主にテキストを対象としたもの
  - 文章，段落，文，単語の処理に重点をおいている
- 他に，Webを対象とする「Webマイニング」という言葉もある
  - Webページ内のタグ，Webページ間のリンク，の解析に重点
  - Webページ内のテキストの分析は，テキストマイニングの範疇に入る

テキストマイニングのプロセス



結果の解釈と試行錯誤

- 結果の解釈：テキストの意味や性質を理解すること
- 試行錯誤：得た知見について，根拠や理屈を確認し，新たな知見を得るために再入力を行うこと

意味理解と試行錯誤に集中できる
人間とコンピュータ間のインタラクティブな環境が必要

テキストマイニングのシステム

- 文章自動要約
- 文書集合の自動分類(クラスタリング)
- キーワード抽出
- 属性への分類(文章の著者，年齢，性別，出身地の推定)
- 作文の自動採点

多様な目的があり，目的ごとにシステムが異なる

システムが異なれば，システムごとに，
セットアップ，使い方の学習が必要

意味理解と試行錯誤のためのシステム

- 意味理解：集中して分析したい
  - ツールの組合せなど，分析以外に手間をとられたくない
- 試行錯誤：色々な側面から分析したい
  - 多様なツールで，関係を確認したい

多くのツールがひとつの環境内で動作するとともに
ツール同士が連携して動作することが望ましい

TETDMプロジェクト

- 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を構築

- 電子テキストを扱う多くのユーザの
  創造的活動を支援するツールの提供

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「近未来チャレンジ 」のプロジェクトとして発足

テトリーヌ by えむたこ

TETDMプロジェクトメンバー

- 砂山渡(広島市立大学　大学院情報科学研究科)
- 高間康史(首都大学東京　システムデザイン学部)
- 西原陽子(立命館大学　情報理工学部)
- 徳永秀和(香川高等専門学校)
- 串間宗夫(宮崎大学　医学部附属病院医療情報部)
- 阿部秀尚(文教大学　情報学部)
- 梶並知記(神奈川工科大学　情報学部)
- 松下光範(関西大学　総合情報学部)
- Danushka Bollegala(ダヌシカ ボレガラ)(リバプール大学)
- 佐賀亮介(大阪府立大学　大学院工学研究科)
- 河原吉伸(大阪大学　産業科学研究所)
- 川本佳代(広島市立大学　大学院情報科学研究科)

TETDMチャレンジの内容

チャレンジ1：

幅広い利用者と開発者の参入

チャレンジ2：

モジュール間での相互インタラクションの実現

チャレンジ3：

知識創発のための基盤環境の構築

既存の類似環境との相違点

[1.幅広い利用者と開発者の参入]

- 日本語を対象として，卑近なデータ，モジュールを扱える
  - 海外のプロジェクトは英語が対象
  - 面白い，斬新な，任意のツールを追加，使用できる

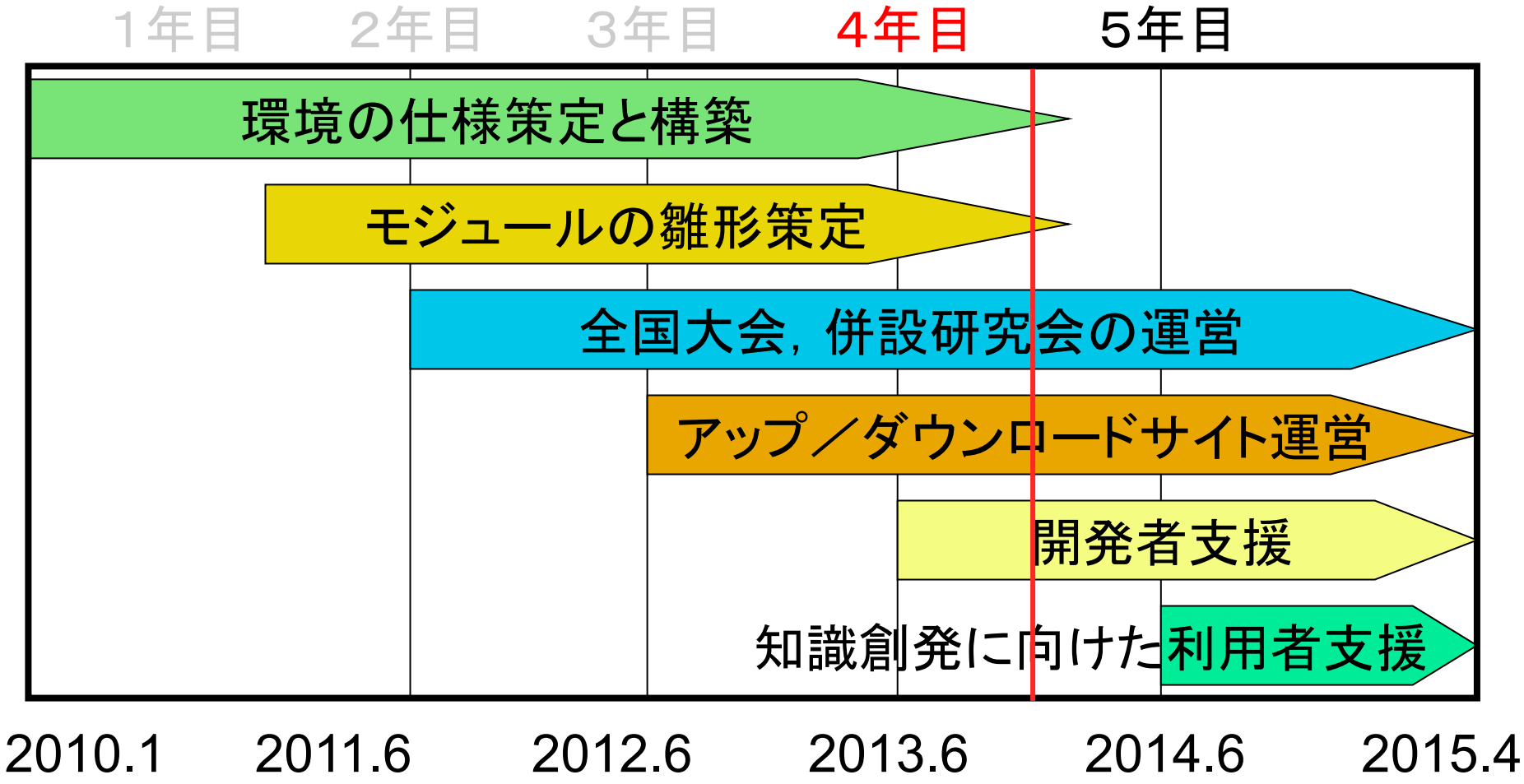
[2.モジュール間での相互インタラクションの実現]

- 独立に作成された複数のモジュールを並列に並べられユーザの操作に対して協調動作，表示が可能
  - 既存研究は1つのグループが作成するシステムの中で協調表示

[3.知識創発のための基盤環境の構築]

- 処理＋可視化に加えて「解釈」「創発」を支援(予定)
  - 既存研究は処理結果を提示するところまでに主眼がおかれている

TETDMのスケジュール

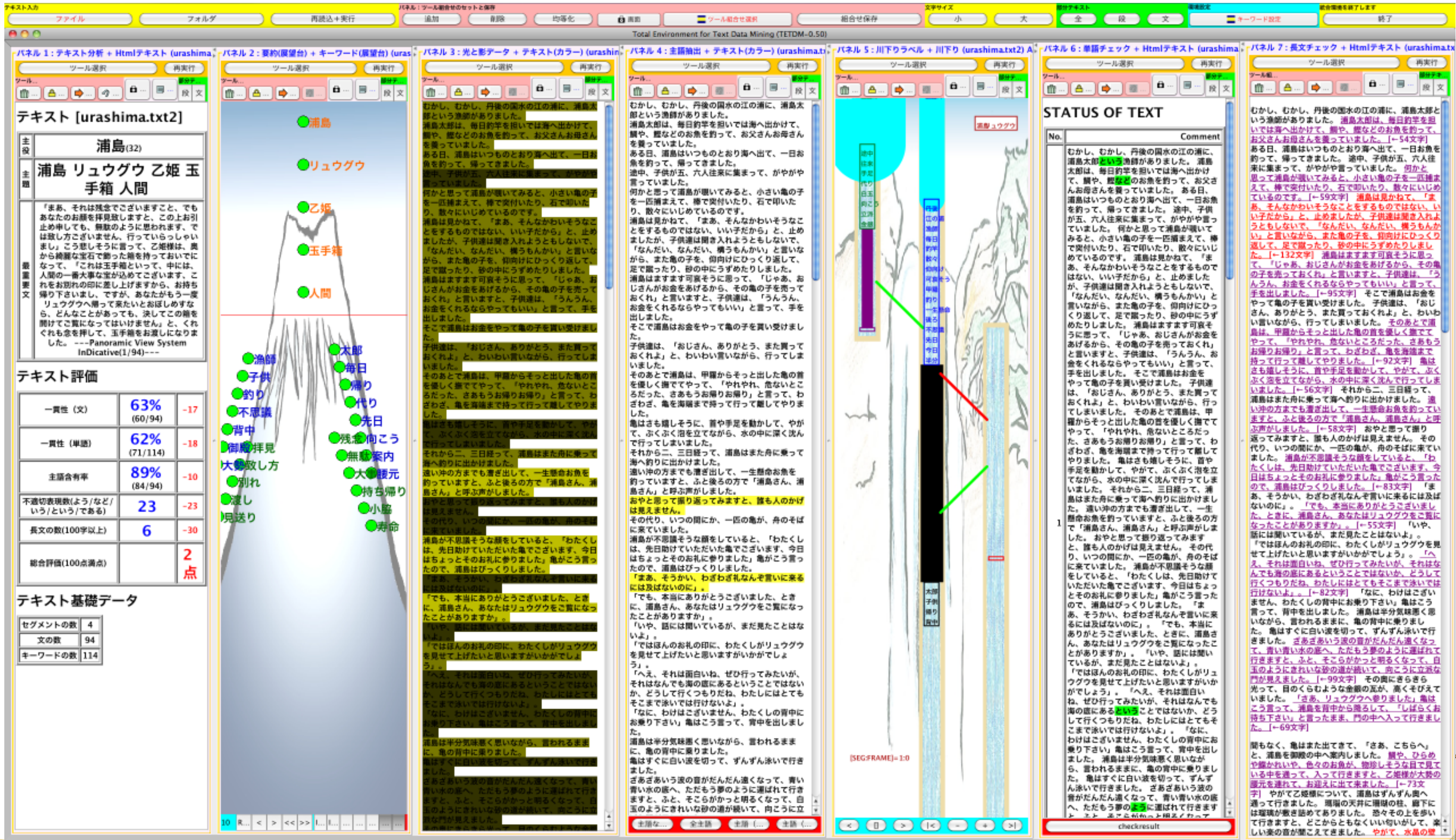


学会活動

- 人工知能学会全国大会
「近未来チャレンジ」セッション 2014年5月
(http://www.ai-gakkai.or.jp/jsai2014/)

- 人工知能学会，インタラクティブ情報アクセスと可視化マイニング研究会 2014年3月
(http://must.c.u-tokyo.ac.jp/sigam/)

TETDM


Total Environment for Text Data Mining (TETDM-0.50)
テキスト [urashima.txt2]
テキスト評価
テキスト基礎データ
STATUS OF TEXT

TETDMの活用場面

[利用者（プログラミングをしない人）]

- 卒論指導，学会発表原稿の作成支援
- レポートの評価支援
- メール，ツイッター，ブログ，電子掲示板のまとめや分析
- Web検索結果の要約や分析

[開発者（プログラミングする人）]

- 研究のためのシステム構築
- 授業や研究室における javaプログラミングの演習

統合環境のライセンス

1. TETDMプログラムは無保証です．完全自己責任で！

2. TETDMプログラムを販売したり，不特定多数の人がアクセスできる環境へのアップロードはダメ！

3. 個人使用や企業内での使用であれば，自由にTETDMプログラムを変更，追加して著作権を持てます

- 詳細なご利用条件は，TETDMサイト内の記述をご確認下さい

チュートリアル

- MISSION 0: TETDMの基礎知識

- クリア条件
  - 説明を読む
  - 説明を読んで質問に答えて正解する
  - [クリア条件]として書かれている操作を行う

チュートリアル

- MISSION 0: TETDMの基礎知識

用語-1

- 「ウインドウ」:画面中央のウインドウ
- 「メニューウインドウ」:画面上部のウインドウ
- 「パネル」:ウインドウ内の1つの区画
- 「ツール」:パネルにセットする
  「処理ツール」または「可視化ツール」

用語-2

- 「テキスト」: TETDMに入力するテキスト
- 「テキストデータ」: 前処理後のテキスト
- 「キーワード」:「キーワード設定」で、
  指定した品詞の単語
- 「セグメント」:「段落」または「文章」
- 「文」: 句点で区切られた文字列

利用手順

- 1) TETDMを起動する
- 2) 入力テキストを読み込む
- 3) パネルにツールをセットする
- 4) 処理結果を確認しながら
  テキストを分析する
- 5) TETDMを終了する

TETDMの入力テキスト

- テキスト: 日本語で書かれた文の集合

- 文の終わりに句点（全角）「．」「。」がある*
  *キーワード設定で指定の文字に変更可能
- 段落の終わりに「スナリバラフト」がある*
  *キーワード設定で指定する任意の文字列に変更可能
- 日本語文字コードは，SHIFT-JISかEUC*
  *UTF-8は使えない．文字コードの自動判別は難しくjavaの問題

TETDMの起動方法

- 1) アイコンをダブルクリックする
  （入力ファイルを後で読み込む）
- 2) アイコン上にファイルを
  ドラッグ&ドロップする
- 3) ターミナルまたはコマンドプロンプト
  を利用する(プログラマ向け)
- TETDMをおいているフォルダ名に、
  日本語や半角スペースが含まれていると起動できません

アイコンのダブルクリックによる起動

- TETDMフォルダ内の、以下のいずれかをダブルクリック
  - TETDM.jar
  - TETDM512.jar
  - TETDM1024.jar
- 512,1024は、確保するメモリの大きさ

ファイルのドラッグ&ドロップによる起動

- TETDMフォルダ内の、下記ファイル上に入力テキストファイルをドラッグ&ドロップ
  - TETDM.bat (Windowsの場合)
  - TETDM.app (Macの場合)(顔アイコン)
- いずれも1024MBのメモリを確保して起動

コマンドラインからの起動

- ターミナルまたはコマンドプロンプトで、
  コマンドラインから、以下の方法で起動します
- 入力ファイルがない場合

  >Java –jar TETDM.jar
- 入力ファイルがある場合(textフォルダ内)

  >Java –jar TETDM.jar text/urashima.txt
- 確保するメモリの量を指定する場合

  >Java -Xmx512m –jar TETDM.jar text/urasima.txt

ファイル読み込み

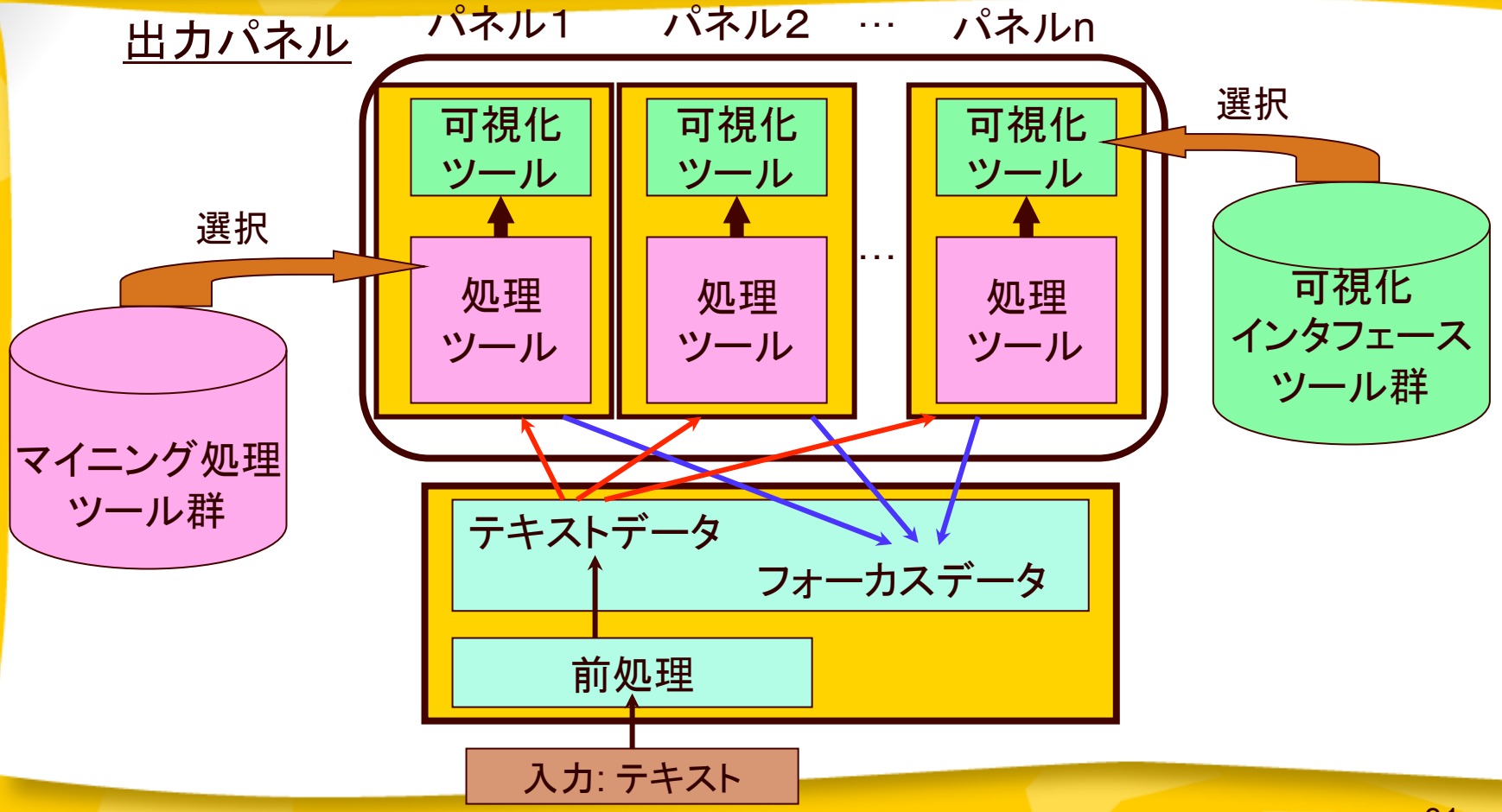
- 1) メニューウインドウ左端の「ファイル」ボタンをクリック
- 2) 表示されるウインドウ内で入力するファイルを選択

- 入力ファイル名、入力ファイルがあるフォルダ名に日本語や半角スペースが含まれていると入力できません

キーワード設定

- メニューウインドウ内の「キーワード設定」ボタンを押すと設定開始
  - 「キーワード」（処理対象とする品詞）の指定
  - 「文」「セグメント」の区切り文字の指定
  - キーワードにしない単語の指定
- 「変更した設定により再処理を実行」ボタンで変更した設定による再処理を実行

TETDMの構成


出力パネル
パネル1
パネル2
…
パネルn
可視化
ツール
可視化
ツール
可視化
ツール
選択
選択
処理
ツール
処理
ツール
処理
ツール
…
可視化
インタフェース
ツール群
マイニング処理
ツール群
テキストデータ
フォーカスデータ
前処理
入力: テキスト

TETDMの前処理

- 形態素解析：文章の単語への切り分け
  - 形態素解析器Igo: 解析結果はほぼMecab互換
- 単語の出現位置,頻度の取得
  - 文章を、文と段落に切り分け、
    各単語の出現する文、段落、とその頻度を計算
- 単語間、段落間の関連度の取得
  - 同じ段落に出現する単語情報から関連度
    （cos類似度）を計算

前処理後のテキストデータ

- 入力テキストの原文（段落，文ごとに区切る）
- 出現単語リスト（段落，文ごと）
- 単語種類数（段落，文ごと）
- 単語の出現頻度（総頻度，段落頻度，文頻度）
- 各単語の出現情報（出現する段落，文）
- インタフェース上で
  ユーザが注目している情報（段落，文，単語）(フォーカスデータ)

複数テキストを入力した場合は，
1つのテキストを段落として扱う

チュートリアル

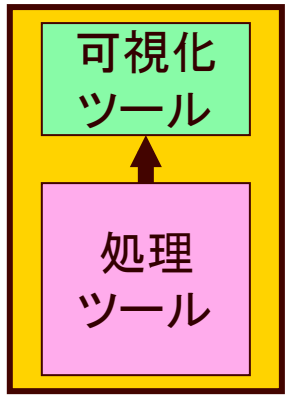
- MISSION 1: ツールの選択とセット
- MISSION 2: ツールの使い方の概要

パネル

- 「追加」「削除」：パネル数の増減(メニューウインドウ)
  - 追加：右端にパネルを追加
  - 削除：右端のパネルを削除
  - パネル数の最大は10
- パネル内の「削除」：そのパネルの削除
- 「均等化」：パネルの幅をそろえる(メニューウインドウ)
  - パネルの幅は、パネル間のバーで調整可能

ツールの種類

- マイニング処理ツール
  - 入力として与えられるテキストに関連して行われる処理全般
- 可視化インタフェースツール
  「汎用性を重視して可視化処理のみを実装」
  - マイニング処理の結果を視覚的に出力
  - マイニング処理や可視化の観点を変更するなど利用者が対話的な操作を行う入力インタフェース



パネル

ツールの例

- マイニング処理ツール
  - キーワード抽出
  - 文章要約
  - テキスト分類
  - 一貫性評価

任意の組合せで
ペアを作成可能

- 可視化インタフェースツール
  - テキスト表示
  - 表形式表示
  - グラフ／ネットワーク表示
  - マップ表示

ツールのセット方法

- 1) 各パネル内の
  「ツール選択」ボタンを押して選択
- 2) メニューウインドウ内の
  「ツール組合せ選択」ボタンを押して選択
- 3) 各パネル内の
  「セット」「戻す」ボタンを押す

ツール選択ボタン

- 左右の端部にカーソルを合わせると
  ツール選択のための説明が表示
- 左側の「処理ツール」と
  右側の「可視化ツール」の中から
  1つずつ選んで、マウスでクリックして選択
- 選択済みのツールは濃いオレンジで表示
- 組合せ可能なツールはオレンジで表示

ツール組合せ選択ボタン

- 複数のパネルにツールをセット可能
- ツールにカーソルを合わせると
  ツールの説明が右のエリアに表示
- ピンクの「処理ツール」と
  水色の「可視化ツール」の中から1つずつ選び
  セットしたいパネルの枠内にマウスでドラッグ
- ツール移動中は
  組合せ可能なツールのみ表示

ツールの一括セット

- 各パネル内の「セット」ボタンを押すと
  パネルにセットされている「処理ツール」と
  一緒に使うことが推奨されている
  「処理ツール」「可視化ツール」を一度にセット
- 「戻す」ボタンで、セット前の状態に戻る

ツールの説明の表示

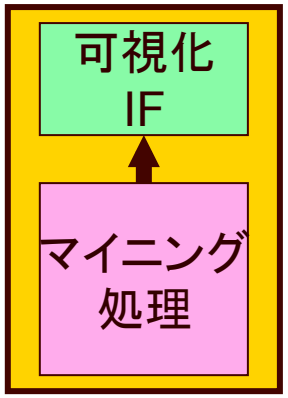
- 1) 各パネル内の「説明」ボタンを押す
  - パネルにセットされている「処理ツール」と「可視化ツール」の説明用ウインドウに説明が表示
- 2) ツール選択時にツールにカーソルを合わせる
  - 説明用のウインドウまたは領域に説明が表示

ツールの組合せ可能性

- 1つのパネル内で組合せ可能な処理ツールと可視化ツール:
  処理ツールの出力するデータを
  受け取り可能な可視化ツール
- 組合せ不能かつエラーが出ている場合
  ツールはセットされない

ツールタイプ

「処理ツール」と「可視化ツール」の組合せにおける組合せの制約の強さ



- シンプル　[単純汎用型](主に「可視化IF」)
  - データの受け渡しがなく，組合せが不要
- プリミティブ　[汎用型]
  - 組合せの指定なしに，１つだけ，データの受け渡しを行う
- セミプリミティブ　[準汎用型]
  - 組合せの指定なしに，データの受け渡しを行う
- 特殊　[専門型]
  - データの受け渡しに，組合せの指定がある

データ型コンバート

- 1つのパネル内で組合せ可能な処理ツールと可視化ツール:
  処理ツールの出力するデータを
  受け取り可能な可視化ツール
- 組合せ不能かつエラーが出ている場合
  ツールはセットされない

データ型コンバート

	優先順位										
送信データ	1	2	3	4	5	6	7	8	9	10	11
boolean	boolean	int	double	String	boolean[]	int[]	double[]	String[]	boolean[][]	int[][]	double[][]
int	int	double	boolean	String	int[]	double[]	boolean[]	String[]	int[][]	double[][]	boolean[][]
double	double	int	boolean	String	double[]	int[]	boolean[]	String[]	double[][]	int[][]	boolean[][]
String	String	int	double	boolean	String[]	int[]	double[]	boolean[]	int[][]	double[][]	boolean[][]
boolean[]	boolean[]	int[]	double[]	String[]	boolean[][]	int[][]	double[][]	String	boolean	int	double
int[]	int[]	double[]	boolean[]	String[]	int[][]	double[][]	boolean[][]	String	int	double	boolean
double[]	double[]	int[]	boolean[]	String[]	double[][]	int[][]	boolean[][]	String	double	int	boolean
String[]	String[]	String	int[]	double[]	boolean[]	int[][]	double[][]	boolean[][]	int	double	boolean
boolean[][]	boolean[][]	int[][]	double[][]	String[]	boolean[]	int[]	double[]	String	boolean	int	double
int[][]	int[][]	double[][]	boolean[][]	String[]	int[]	double[]	boolean[]	String	int	double	boolean
double[][]	double[][]	int[][]	boolean[][]	String[]	double[]	int[]	boolean[]	String	double	int	boolean

変換テーブル

変換前／変換後	boolean	int	double	String
boolean	-	0か1に変換	0.0か1.0に変換	文字列に変換
int	0以外true	-	(double)でキャスト	文字列に変換
double	0以外true	(int)でキャスト	-	文字列に変換
String	""以外true	文字数	文字数	-

変換テーブル+	
変換テーブル+	第一引数(添字)の要素数を 1 にする
変換テーブル+	第一引数(添字)と第二引数(添字)の要素数を 1 にする
変換テーブル+	,区切りで結合
変換テーブル+	,区切りと改行で結合
変換テーブル+	二次元データを一次元に落とす([i][j] -> [k])
変換テーブル+	先頭の要素([0]または[0][0])を取り出す

ツールの組合せの保存

- メニューウインドウ内の「組合せ保存」ボタンを押すと、各パネルにセットされているツールの名前を保存
  - 次にTETDMを起動する際に、保存したツールの組合せで起動します
  - 各パネルの横幅も保存します
  - 組合せを初期状態に戻したいときは「キーワード設定」の「初期設定に戻す」ボタンと「現在の設定をファイルに保存」ボタンを押す

ツールの操作

- ツール(「処理ツール」と「可視化ツール」)には操作が必要なものと、不要なものがあります
  - 1) パネルにツールをセットしただけで
    処理や結果の表示を行ってくれるツール
  - 2) ツール上で何らかの操作を行うことで
    処理や結果の表示を行うツール

ツールの操作の説明

- ツールの操作方法は、次の方法で確認できます
  - 1) 各パネル内の「説明」ボタンを押す
  - 2) 「ツール組合せ選択」によるツール選択時に
    ツールにカーソルを合わせる
    （右側の領域に説明が表示される）

処理ツールの操作

- 操作可能なコンポーネントは
  パネル最下部の（主に）赤い背景のエリアに
  まとめて表示
- コンポーネントの例
  - ボタン
  - プルダウンメニュー
  - テキストエリア

可視化ツールの操作

- 視覚的に表示されているものに対して
  マウスで下記の操作を行います
  - タッチ（カーソルを合わせる）
  - クリック
  - ドラッグ&ドロップ
- 可視化ツール用のボタンがあるものもある
  - 可視化ツール用のボタンは
    処理ツール用のボタンよりも上部に表示される

キーワード設定

- ツールによる処理および処理結果の表示について処理対象とするキーワードを設定
- 設定後は「変更した設定により再処理を実行」で、即座に処理結果の更新が可能

- 「キーワード」の項目では、処理対象となるキーワードとして取り出す品詞を指定

キーワード設定

- 「文の区切りとする句点の種類」:
  文の区切り記号を指定可能
- 「セグメントを区切る単語」:
  段落または文章を区切る単語を指定可能
  - テキスト中の任意の単語を指定可能
  - 「残す」「残さない」のチェックボックスで
    処理の際に区切り単語を使用するかどうかを指定

キーワード設定

- 「キーワードにしない単語」:
  キーワードに指定した品詞の単語でも
  キーワードとして欲しくない単語を指定
- 「ひらがな」「カタカナ」「1文字」の単語指定:
  チェックした表現の単語をキーワードにしない

- 「メニューを日本語にする」:
  ボタンなどの表示の日本語と英語の切り替え

実習 : 基本的な分析

- 1つのツールを利用した分析
  - MISSION 3 – 6:
    - 結果を表示するツール
    - 処理ツールの操作
    - 可視化ツールの操作
    - 処理ツールと可視化ツールの操作
- 複数のツールを利用した分析
  - MISSION 7:複数のパネルを使うツール

ツールを利用した分析 操作なし-1

- 処理ツール「サンプル１」＋
  可視化ツール「テキスト」
- 処理ツール「マイニングなし」＋
  可視化ツール「段落関連度」
- 処理ツール「タグデータ」＋
  可視化ツール「タグクラウド」

ツールを利用した分析 操作なし-2

- 処理ツール「テキスト分析」＋
  可視化ツール「Htmlテキスト」
- 処理ツール「長文チェック」＋
  可視化ツール「Htmlテキスト」
- 処理ツール「アノテーション」＋
  可視化ツール「テキスト（カラー）」

ツールを利用した分析 処理操作-1

- 処理ツール「サンプル2」＋
  可視化ツール「テキスト」
- 処理ツール「川下りラベル」＋
  可視化ツール「川下り」
- 処理ツール「主語抽出」＋
  可視化ツール「テキスト(カラー)」

ツールを利用した分析 処理操作-2

- 処理ツール「タイピング」＋
  可視化ツール「タイピング」
  - マイニングとは異なる処理ツール
- 処理ツール「ソース表示」＋
  可視化ツール「テキスト」
  - プログラマ向け
  - 「README」ボタンは、READMEを作成用のひな形

ツールを利用した分析 可視化操作-1

- 処理ツール「マイニングなし」＋
  可視化ツール「表」
- 処理ツール「マイニングなし」＋
  可視化ツール「ばねモデル」
- 処理ツール「マイニングなし」＋
  可視化ツール「セグメント独自性」

ツールを利用した分析 可視化操作-2

- 処理ツール「マイニングなし」＋
  可視化ツール「セグメント木」
- 処理ツール「マイニングなし」＋
  可視化ツール「トップボトム木」
- 処理ツール「単語間関連度」＋
  可視化ツール「キーワードマップ」

ツールを利用した分析 処理＋可視化操作-1

- 処理ツール「エディタ」＋
  可視化ツール「テキスト」
  - コピー＆ペーストによるテキストの入力にも利用
- 処理ツール「トップダウン段落順序」＋
  可視化ツール「段落並べ替え」

ツールを利用した分析 処理＋可視化操作-2

- 処理ツール「主題語含有率」＋
  可視化ツール「レーダーチャート」
- 処理ツール「国語辞書」＋
  可視化ツール「デュアルテキスト」
  - ネットワークへの接続が必要

ツール間の連動について

- 異なるパネルにセットされたツール同士は独立に動作する
- 連動：複数の異なるパネルにセットされたツール同士が、協調的な動作を行うこと
- フォーカス連動：異なるパネルのツール同士は、共通の注目データ(フォーカスデータ)を参照して動作可能
  - 単語(集合)，段落(集合)，文(集合)

ツールを利用した分析 複数パネル-1

- 処理ツール「アノテーション」＋
  可視化ツール「テキスト」「テキスト（カラー）」
- 処理ツール「光と影データ」＋
  可視化ツール「テキスト（カラー）」
  「スコア分布」「キーワード選択」
  - 複数パネルを用いる場合、パネル間の
    フォーカス連動が実装されていることがある
  - 「キーワード選択」でチェックすると連動

ツールを利用した分析 複数パネル-2

- 処理ツール「関連チェック」＋
  可視化ツール「Htmlテキスト」,
  処理ツール「マイニングなし」＋
  可視化ツール「セグメント独自性」
  「連動可視化メイン」「連動可視化サブ」
  - 「セグメント独自性」で、マウスで触ると連動
  - 「キーワード設定」で動詞,形容詞を追加した上で、
    処理ツール「スコアチェック」＋
    可視化ツール「スコアネットワーク」も追加可能

ツールを利用した分析 複数パネル-3

- 処理ツール「再帰的クラスタリング」＋
  可視化ツール「OKmap」「連動可視化メイン」
  - 「OKmap」で、マウスで触ると連動
- 処理ツール「要約（展望台）」＋
  可視化ツール「テキスト」「キーワード（展望台）」
  - 「キーワード（展望台）」で、クリックすると連動

入力テキストの加工

- 処理ツール「エディタ」を使う*起動後真ん中のパネル
  - コピーペーストでテキストを貼付けて，入力テキストにできる
  - 「編集保存＋実行」で編集結果が即座に反映
  - 「改行で句点」「空行で分割」で，文，段落の区切りを自動挿入して再実行
    「タグカット」で段落の区切りを消去
- 可視化ツール「段落並び替え」を使う
  - マウスで段落を並べ替えた後，「並び替え保存」

辞書への単語の登録

- 処理ツール「辞書構築」
  ＋可視化ツール「ファイル」をパネルにセット
- 登録したい単語と、その読みを、
  半角カンマで区切って入力
  →「保存＆辞書構築」ボタンで登録完了
- 「辞書初期化」ボタンで登録削除

分析の実践

- 自作作文の分析
- テキストデータの分析

自作作文の分析（全体構成）

- 文章の主題と一貫性
  - 処理「テキスト分析」＋可視化「Htmlテキスト」→「セット」
  - 処理「要約（展望台）」＋可視化「テキスト」
  - 処理「光と影データ」＋可視化「テキスト(カラー)」
  - 処理「川下りラベル」＋可視化「川下り」
- 段落間の関係と構成
  - 処理「なし」＋可視化「段落関連度」
  - 処理「なし」＋可視化「セグメント木」
  - 処理「なし」＋可視化「トップボトム木」（Ver.0.54用）
  - 処理「トップダウン段落順序」＋可視化「段落並べ替え」

自作作文の分析（表現）

- 主語，曖昧表現，長文の有無
  - 処理「テキスト分析」＋可視化「Htmlテキスト」→「セット」
  - 処理「主語抽出」＋可視化「テキスト(カラー)」
  - 処理「単語チェック」＋可視化「Htmlテキスト」
  - 処理「長文チェック」＋可視化「Htmlテキスト」
- 使用単語の確認
  - 処理「なし」＋可視化「表」
  - 処理「なし」＋可視化「ばねモデル」

テキストデータの分析（全体傾向）

- 段落（テキスト）の分類
  - 処理「再帰的クラスタリング」＋可視化「OKmap」
  - 処理「なし」＋可視化「ばねモデル」→「段落(共起単語)」
- 使用単語の分布と傾向
  - 処理「タグデータ」＋可視化「タグクラウド」
  - 処理「単語間関連度」＋可視化「キーワードマップ」

テキストデータの分析（部分的特徴）

- 部分テキストによる結果の可視化
  - 処理「なし」＋可視化「セグメント選択」
    →段落を選択後，「「段」データ作成」ボタン
    →他のパネルで部分テキストの「段」ボタン

↓

選択した段落のみを対象とした処理結果を生成

↓

全体の結果と，部分の結果を，
複数のパネルに並列に出力

TETDMのまとめ

- 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
  統合環境の構築と提供を目指す

- 多くの開発者によるツール開発と
  多くの利用者による活用
  人々の連携が見込まれる環境を目指したい

今後の展開：知識創発に向けて

1. モジュールの柔軟かつ多様な組合せにより斬新な結果の取得による発想
   - 開発者が想定していなかったモジュールの組合せ
   - 信頼性だけでなく面白さや斬新さを重視したモジュールとの組合せ

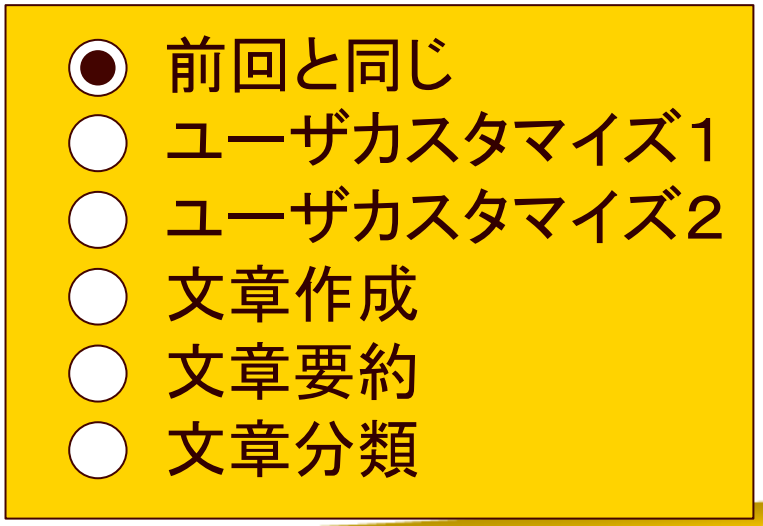
2. データによる処理結果の比較を支援するツールの開発による発想
   - データの量が増えるとデータの絞り込みが不可欠
   - 複数の絞り込みデータ間の比較を促すツール群を開発

ツールの開発と収集

- 電子掲示板の要約、流れ、雰囲気の可視化
- メール、つぶやき、コメントにおける
  失礼表現の抽出
- トップダウン、ボトムアップなどの
  文章構成の作成支援

ツール選択支援

- ツールの組合せを選択して起動:
  - 目的別ツールの組合せ
  - ユーザが登録した組合せ



マイニングのこれから

- 情報や機能，人々が1つにつながっていく時代
- オンラインを活用して他の人の分析情報を知りたい
  - 他の人は，どんなツールをどんな用途に使っているか？
  - 他の人のマイニングの手順は？
  - 他の人はどんな結果をどんな方法で導き出せたか？

協調，協働によるマイニングの発展に向けて

- 利用における協調，協働（利用者間の連携）
- 開発における協調，協働（モジュール間の連動）
- 利用者と開発者の協調，協働（ニーズとシーズのマッチング）

お疲れさまでした.

アンケートにご記入ください.