3B3-NFC-01b-2

R によるテキストマイニング用 TETDM モジュール開発

Format guideline for manuscripts of JSAI 20XX for MS Word

徳永 秀和
Tokunaga Hidekazu

香川高等専門学校
Kgawa National College of Technolog

TETDM which is the environment which can interlock various text-mining tools is developed. On the other hand, the method of performing fundamental text mining using RMeCab is opened to books or Web in large numbers. Furthermore, the statistical processing and data mining which are used for text mining are offered in large numbers as a package of R. Therefore, we develop the module which uses R and RMeCab by TETDM. This paper explains R,RMeCab,rJava and module linkage of TETDM.

1. はじめに

様々なテキストマイニングツールを連動して利用できる環境として TETDM の開発が進んでいる[砂山 2013]. また, 基本的なテキストマイニングを RMeCab を利用して行う方法が, 書籍や Web 上に多く公開されている. さらに, テキストマイニングで多く用いられる統計処理やデータマイニング処理はRのパッケージとして提供されている. そこで TETDM においても R と RMeCab を利用できると有用である. R,RMeCab,rJava, TETDM のモジュール連動について簡単に解説し, TETDM での R の利用方法について説明する.

2. Rによるテキストマイニング

統計解析言語R はデータ解析・グラフィックス環境を備えたオープンソースのプログラミング言語である. Rには, 様々なデータマイニング処理がパッケージとして提供されており, 簡単に使用することができる. 最近は, Rに関する書籍やWeb上での情報も多くなり, 初心者にも利用しやすい状況になっている.

テキストマイニングにRを用いる事例を, ESTRELA に連載された”フリーソフトによるデータ解析・マイニング”と, 書籍[石井 2008]より要約すると以下のようになる.

- テキストから集計したデータの平均、分散, 四分位数などの記述統計量を計算して, ヒストグラムやひげ図などを表示する.
- パッケージ igraph を用いて, 語の共起関係をネットワークマップデータに変換してグラフィック表示する.
- 文節の長さの分布のモデルを推定するために, 最尤推定量, 情報量基準を計算する.
- 1 文当たりの読点の数の平均値の区間推定や仮設検定, 割合の検定を行う.
- 2作品に対して, 読点をどの文字の後に打っているかの差を調べる, 分割表によるカイ 2 乗独立性検定, フィシャーの直接確率計算.
- 2値のクロス表による, マクネマーの検定.
- 複数テキストからのカイ 2 乗値を用いた特徴語の抽出.
- 複数テキストを語の出現頻度ベクトル表現し, 主成分分析, パッケージ MASS を利用した対応分析.
- 性別, 出身地, 気遣い度が語尾に与える影響を調べるような多項ロジットモデルを, パッケージ nnet を用いて解析する.
- 複数テキストを語の出現頻度ベクトル表現し, k-means 法, ウォード法などのクラスター分析を行う.
- 複数テキストを語の出現頻度ベクトル表現し, 決定木, サポートベクトルマシン, ニューラルネット, 自己組織化マップなど機械学習を用いた分類を行う.

連絡先:徳永秀和, 香川高等専門学校 機械電子工学科,
〒761-8058 高松市勅使町355, tokunaga@t.kagawa-nct.ac.jp

3. RMeCab によるテキストマイニング

Rの統計処理とデータマイニング処理を行うための入力データを作成するためのパッケージがRには提供されている. 代表的なパッケージとして RMeCAb がある. それ以外にも tm, RCaBoCha や Yahoo!の日本語解析 API を使う YjdnJlp などがある.

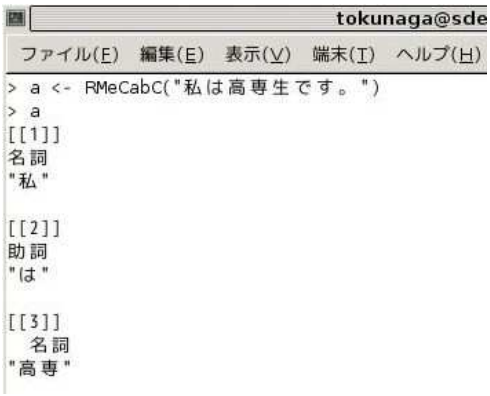

図1 リストを返す RMeCab の関数

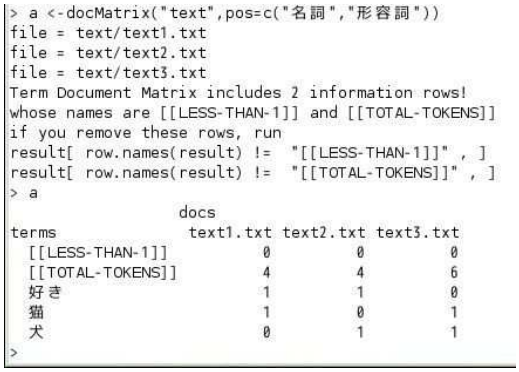

図2 データフレームを返す RMeCab の関数

RMeCab は，R から日本語の文章やファイルを指定して MeCab に解析させ，その結果を R で標準的なデータ形式に変換して出力させるプログラムである．図1はリスト形式を返す RMeCabC( )関数の例である．図2はデータフレーム形式を返す docMatrix( )関数の例である．R のデータフレームとは，複数の異なるデータ型のベクトルを1にまとめたもので，ラベル（列名）を付け，列名を使って要素にアクセスできる．RMeCab の主な関数を表1に示す．

表1　RMeCab の関数

関数	機能概要
RMeCabC	文字列の形態素解析結果を返す
RMeCabText	単一ファイルの形態素解析結果を返す
RMeCabFreq	ファイル内の形態素の頻度表を返す
docMatrix	フォルダ内のファイルのターム・文章行列を作成，tf-idf 値も計算
docMatrixDF	データフレームのターム・文章行列を作成，tf-idf 値も計算
Ngram	N-gram を返す
docNgram	N-gram を文書・ターム行列として返す
collocate	テキストファイルの共起の頻度表を作成
collScores	共起頻度のオブジェクトの共起スコアを計算

4. Java からRの利用

Java から R を利用するツールはいくつかあるが，ここではJRI について説明する．CRAN の rjava パッケージに含まれており，Rから install.packages('rJava')と入力するだけでインストールできる．Java クラスの書き方のを簡単に示す．R との接続は「Rengine」というクラスが受け持つことになり、値の受け渡しは assign()メソッドで、R 内部のコマンド実行は eval()メソッドで行う。簡単なサンプルプログラムを図3に示す．R の変数 a に値9の代入された整数配列を代入し，R の sqrt( )関数で計算し，結果を REXP オブジェクトに代入した後，整数型に変換して出力している．REXP には，asDoubleArray()，asDoubleMatrix()，asList()などのメソッドにより様々なRの関数が返すデータ型への変換が可能である．

```
Rengine engine =
        new Rengine(new String[ ]{"–no-save"},  false, null);
engine.assign("a",new int[ ]{9});
REXP result = engine.eval("sqrt(a)");
System.out.println(result.asInt());
engine.end();
```

図3　rJava による R の実行

5. TETDM モジュール

TETDM 統合環境は，図4のような構成となっており，開発言語は Java である．簡単な処理の流れは次のようになる．解析対象のテキストを統合環境処理部に入力する．テキストは形態素解析され，テキストマイニングに必要な様々なデータを取り出すインタフェースを持つ Java のインスタンス"TextData"を生成する．パネル内のマイニングモジュールが TextData インスタンスと連動制御処理部の機能を利用して処理を行い，可視化 IF モジュールにデータを渡して表示する．

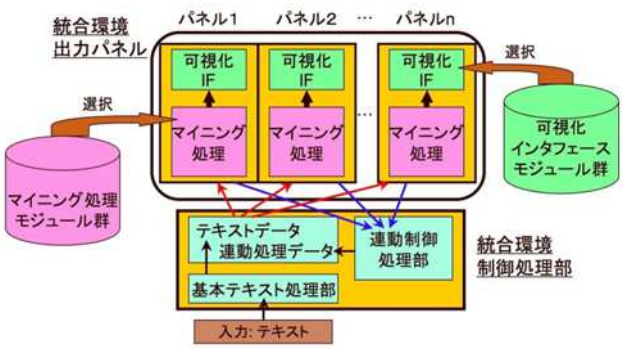

図4　TETDM 統合環境

マイニング処理モジュールが，他のマイニング処理モジュールの処理結果のデータを利用する仕組みを図5に示す．処理を要請するモジュールは，モジュール ID とオプション番号により欲しいデータ型に対応したメソッドを実行する．結果として整数の配列が欲しい場合は，int[] getDataIntegerArray(int getModuleID, int dataID);を実行する．処理を要請されるモジュールは，オプション番号とデータ型に対応したデータを，void setDataIntegerArray(int dataID, int data[]); を実行し作成する．

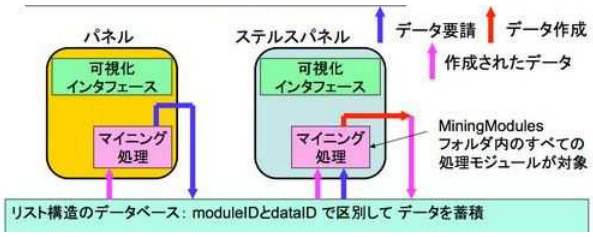

図5　モジュール連携

6. R と RMeCab の処理モジュール

4章の rJava を利用して Java で R と RMeCab を実行し，処理結果データを取得することができる．5章の TETDM のモジュール間連動の機能を用いることにより，他のモジュールにおいて簡単に R と RMeCab により処理したデータを利用できる．2章と3章に示したような基本的なテキストマイニング処理をオプション番号に対応して行うモジュールを準備しておくと有用である．

7. おわりに

TETDM から R を利用する利点と方法を説明した．今後，テキストマイニングの基本的処理を簡単に利用できる R と RMeCab を用いたモジュールの開発を行っていく．

参考文献

[砂山 2013] 砂山渡，高間康史，西原陽子，徳永秀和，串間宗夫，阿部秀尚，梶並知記: テキストデータマイニングのための統合環境 TETDM の開発，人工知能学会論文誌，Vol.28，No.1,2013.

[石井 2008] 石井基広:Rによるテキストマイニング入門，森北出版，2008