

R プログラミング

tm を使う (WindowsXP 上の R)

次のパッケージをインストールしておく。

- RMecab , tm , Snowball
 - 後 , Snowball には , Java が必要なため , Weka というソフトウェアと同梱されている Java をインストールする。
 - 2 つのソフトウェアは以下からダウンロード。
 - ソフト名 : weka-3-6-1jre.exe
 - 入手先 : <http://www.weka-jp.info/>

資料 (txt 形式) を tm を使って計算する。

- (1) 変数にテキストファイル・データを取り込む
- ただし , この代入式は , マニュアルにかかれたものをそのまま使用しています。その為 , 最終的には , リファレンスを見てテキストファイルの取り込みディレクトリを変更するようして下さい。

```
> txt <- system.file("texts", "txt", package="tm")
```

- C:\Program Files\R\R-2.9.2\library\tm\texts\txt に入っているすべてのテキストファイル・データを変数 " txt " に代入されることを意味しています。
- よって , 当初 , サンプルプログラム " ovid*.txt " という 5 つのテキストファイルが存在しますが , 今回は , 調べたいテキストファイルと入れ替えます。
- 今回は , 35 個の txt データを入れます。

メモリの拡張を宣言するコマンドを実行して下さい。

```
> old.op <- options(max.print=999999)
```

作業内容は , 次の通りです。

```
# sink コマンドを使って , 出力結果をテキストファイルに出力します。
> sink(file ="ovid.txt", split="true")

# テキストファイルを " txt " 変数に代入します
> txt <- system.file("texts", "txt", package="tm")
> (ovid <- Corpus(DirSource(txt), readerControl=list(language="eng")))
A corpus with 35 text documents

# ここまでで変数 " txt " にテキストファイルが取り込まれました。
# その後 , 変数 " ovid " に tm で利用しやすいような形式のテキストデータの塊 ( コーパス ) が作成されます。

# inspect 関数を利用して内容を確認します。
# テキストファイル 1 ~ 3 5 個のテキストファイル・データが表示されます。
> inspect(ovid[1:35])
> sink()
# 以上でファイルに記録されました。
```

コーパスになったデータのうち , 空白やSTEMMINGを行う

```
# tmMap コマンドを利用して空白や英語に関係ないものを取り除く ( かなり時間がかかります )
> ovid2 <- tmMap(ovid, stripWhitespace)
> ovid3 <- tmMap(ovid2, removeWords, stopwords("english"))
> ovid3stm <- tmMap(ovid3, stemDoc)

# sink コマンドを使って , 出力結果をテキストファイルに出力します。
```

```
> sink(file ="ovid3stm.txt")
> inspect(ovid3stm[1:35])
> sink()

# 共起関数で計算しやすいようにする
> ovid3stm_dtm <- DocumentTermMatrix(ovid3stm)

# sink コマンドを使って、出力結果をテキストファイルに出力します。
> sink(file ="ovid3stm_dtm_fft.txt",split="true")
> findFreqTerms(ovid3stm_dtm,5)
> findAssocs(ovid3stm_dtm,"proprietor",0.97)[1:35]
```