9章 潜在意味解析

9.5 Rによる潜在意味解析

07t4072f　岡崎　駿

語句ー文書行列の作成

使用するライブラリ：RMeCab

実行手順

1. RでRMeCabのライブラリを読み込む
2. LSAのパッケージの内容を変更したLSA.txtを読み込ませる

```
> library(RMeCab)
> source("C:/Users/mikami/Desktop/Program-datamining/chap9/LSA.txt")
> 
```

3. RMeCabのdocMatrixでディレクトリ指定

```
> docterm <- docMatrix("C:/Users/mikami/Desktop/Program-datamining/chap9/Z")
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc01.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc02.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc03.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc04.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc05.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc06.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc07.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc08.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc09.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc10.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc11.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc12.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc13.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc14.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc15.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc16.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc17.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc18.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc19.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc20.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc21.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc22.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc23.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc24.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc25.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc26.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc27.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc28.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc29.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc30.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc31.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc32.txt
file = C:/Users/mikami/Desktop/Program-datamining/chap9/Z/doc33.txt
Term Document Matrix includes 2 information rows!
whose names are [[LESS-THAN-1]] and [[TOTAL-TOKENS]]
if you remove these rows, run
result[ row.names(result) !=  "[[LESS-THAN-1]]" , ]
```

```
> docterm
                 docs
terms             doc01.txt doc02.txt doc03.txt doc04.txt doc05.txt doc06.txt doc07.txt doc08.txt doc09.txt
  [[LESS-THAN-1]]         0         0         0         0         0         0         0         0         0
  [[TOTAL-TOKENS]]       19        14        38        46        48        45        33        29        29
  観察                    1         0         0         0         0         0         0         0         0
  距離                    1         0         0         0         0         0         0         0         0
  周り                    1         0         0         0         0         0         0         0         0
  素直                    1         0         0         0         0         0         0         1         0
  良い                    1         0         0         0         0         1         0         0         0
  そう                    0         1         0         0         0         0         0         0         0
  感じ                    0         1         0         0         0         0         0         0         0
  強い                    0         1         0         1         0         0         0         0         0
  嫉妬                    0         1         0         0         0         0         0         0         0
  勝気                    0         1         0         0         0         0         0         0         0
  深い                    0         1         0         0         0         0         0         0         0
  独占                    0         1         0         0         0         0         0         0         0
  負けず嫌い              0         1         0         0         0         0         0         0         0
  欲                      0         1         0         0         0         0         0         0         0
  こと                    0         0         1         0         2         2         0         0         0
  ため                    0         0         1         0         0         0         0         0         0
  一生懸命                0         0         1         0         0         0         0         0         0
  気持                    0         0         1         0         0         1         0         1         0
  行動                    0         0         1         0         0         0         0         0         1
  人                      0         0         2         2         0         0         0         2         0
  性格                    0         0         1         1         0         0         0         1         0
  誰                      0         0         1         0         0         0         0         0         0
  的                      0         0         1         1         0         0         0         1         0
  普段                    0         0         1         0         0         0         0         0         0
  無い                    0         0         1         0         0         0         0         0         0
  明るい                  0         0         1         0         0         0         0         0         0
  優しい                  0         0         1         1         0         0         0         0         0
  優先                    0         0         1         0         0         0         0         0         0
  落ち着き                0         0         1         0         0         0         0         0         0
  ユーモア                0         0         0         1         0         0         0         0         0
  温かみ                  0         0         0         1         0         0         0         0         0
  間                      0         0         0         1         0         0         0         0         0
  傾向                    0         0         0         1         0         0         0         0         0
```

特異値分解と次元数の選択

- docMatrix内にある分析対象の行列には1つの文書しか現れていない語句が多いため、語句が複数の文書しか現れていないものしか取り扱わないようにする。

```
> docterm2 <- Kyoki(docterm,minDocFreq=2)
```

- 217語×33人　　→　　73語×33人

特異値分解

- dimReducShare・・・特異値の総合計に対する割合を0.5以上になるまで特異値の累積和を計算し、次元縮小した意味空間を返す

```
> svd.docterm <- svd(docterm2)
> rslt <- dimReducShare(svd.docterm,share=0.5,docterm=docterm2)
> str(rslt)
List of 3
 $ tk: num [1:63, 1:10] 0.0478 0.0737 0.0567 0.1087 0.0265 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:63] "観察" "周り" "素直" "良い" ...
  .. ..$ : NULL
 $ dk: num [1:33, 1:10] 0.0426 0.0517 0.3699 0.329 0.2259 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:33] "doc01.txt" "doc02.txt" "doc03.txt" "doc04.txt" ...
  .. ..$ : NULL
 $ sk: num [1:10] 6.74 4.94 4.62 4.4 3.92 ...
>
```

類似性の計算 I

- myCosine･･･全ての列に関してコサイン計算を行う。

```
> myCosine(a=t(rslt$dk))
              doc01.txt   doc02.txt     doc03.txt    doc04.txt     doc05.txt
doc01.txt  1.000000000 -0.18331428 -0.1021110442 -0.027724107 -0.0525457344
doc02.txt -0.183314283  1.00000000 -0.0578080215  0.184043512  0.0449454383
doc03.txt -0.102111044 -0.05780802  1.0000000000  0.681758975  0.0002858855
doc04.txt -0.027724107  0.18404351  0.6817589745  1.000000000 -0.0393810496
doc05.txt -0.052545734  0.04494544  0.0002858855 -0.039381050  1.0000000000
doc06.txt  0.610001863 -0.14331696  0.1157369097 -0.096231102  0.3298867412
doc07.txt -0.142913266 -0.05836706 -0.0270871057 -0.111159897  0.5891419858
doc08.txt  0.182249889 -0.20975463  0.6930549992  0.471520413 -0.1084214582
doc09.txt -0.113425101 -0.18596398 -0.0235741224  0.173394447 -0.0441819773
```

類似性の計算 Ⅱ

- 浅野くんの好みの女性
  - 「社交性で優しいし、俺のことを優先してくれるけど、意外とクールな面もある人」

```
> unmei <- RMeCabC("社交性で優しいし、俺のことを優先してくれるけど、意外とクールな面もある人")
> tmp <- unlist(unmei)
> lst <- names(tmp) %in% c("名詞","動詞","形容詞")
> unmei <- tmp[lst]
```

- docMatrix()で保持している品詞が「名詞」、「動詞」、「形容詞」だったため、浅野くんの好みを形態素解析する際に、3つ品詞を指定

```
> unmei
   名詞    名詞  形容詞    名詞    名詞    名詞    動詞    動詞    名詞    名詞    動詞    名詞
 "社交"   "性" "優しい"   "俺"   "こと" "優先"    "し" "くれる" "クール"    "面"   "ある"    "人"
```

類似性の計算Ⅲ

- 浅野くんの理想を表す12語が検索対象の73語にあるかどうか確認する

```
> unmei.q <- myQuery(unmei,rownames(rslt$tk))
4        番目の俺          がterm.listに存在しません
6        番目の優先        がterm.listに存在しません
7        番目のし          がterm.listに存在しません
8        番目のくれる      がterm.listに存在しません
10       番目の面          がterm.listに存在しません
11       番目のある        がterm.listに存在しません
 以下にエラー myQuery(unmei, rownames(rslt$tk)) :
> unmei2 <- unmei[-c(4,6,7,8,10,11)]
> unmei.q <- myQuery(unmei2,rownames(rslt$tk))
```

- 4,6,7,8,10,11番目の語がないため、この言葉を除いて分析を行う

類似性の計算Ⅳ

- 潜在意味空間上に浅野くんの理想の人ベクトルを位置づける

```
> unmei.vec <- t(unmei.q) %*% rslt$tk %*% solve(diag(rslt$sk))
> unmei.vec <- as.vector(unmei.vec)
```

- 角度の1番小さい人が浅野くんの運命の人

```
> unmei.cos <- myCosine(a=t(rslt$dk),b=unmei.vec)
> names(unmei.cos) <- as.character(1:33)
> round(sort(unmei.cos,de=T),3)[1:5]
    3    33    22     5    24
0.682 0.605 0.438 0.407 0.394
>
```

- 結果：3番の女の子が運命の人