
データの分析と知識発見
Introduction to Data Analysis

今回の構成

テキストマイニングや
そのツールについて知る

RMeCabを用いて
テキストマイニングを行う

講義全体の振り返りを行う

自然言語処理

自然言語 ⟷ プログラミング言語

形態素解析

文をその構成要素である語に分解する

（例）本／を／読ん／だ

構文解析

文をその構成要素である語と語の関係を分析

（例）

形態素解析

形態素解析　文を語（形態素）に分解（分かち書き）

（例）「横浜に行った」

横浜	名詞	固有名詞
に	助詞	格助詞
行っ	動詞	五段活用「行く」の連用形（促音便）
た	助動詞	

絶対値，印刷教材，超伝導／超電導などのような例も・・・

形態系解析ソフトウェア

形態系解析ソフト

辞書や文のルールなどをもとに単語を区切る

主なソフトウェア（フリーで入手できるもの）

- Chasen（茶筅）
- MeCab（和布蕪）
- Juman

など

RMeCabによる形態素解析

形態系解析ソフト

- 形態系解析ソフト
  MeCab（工藤 拓 氏）
- Rで和布蕪を用いる
  インターフェース
  RMeCab（石田 基広 氏）

両方をインストールする

計算

- 文の解析：
  文を形態素に分解
  RMeCabC()
- 文章の解析：
  文章ごとに
  語の頻度を計算
  docMatrix()

青空文庫

青空文庫

- 著作権の保護期間が切れたものを
ボランティアの手で
校正・公開されたもの

- 右の7つの短編を
題材にする

芥川龍之介	トロッコ
芥川龍之介	鼻
芥川龍之介	羅生門
有島武郎	一房の葡萄
梶井基次郎	檸檬
小泉八雲	耳なし芳一
新美南吉	ごんぎつね

RMeCabを用いた形態素解析

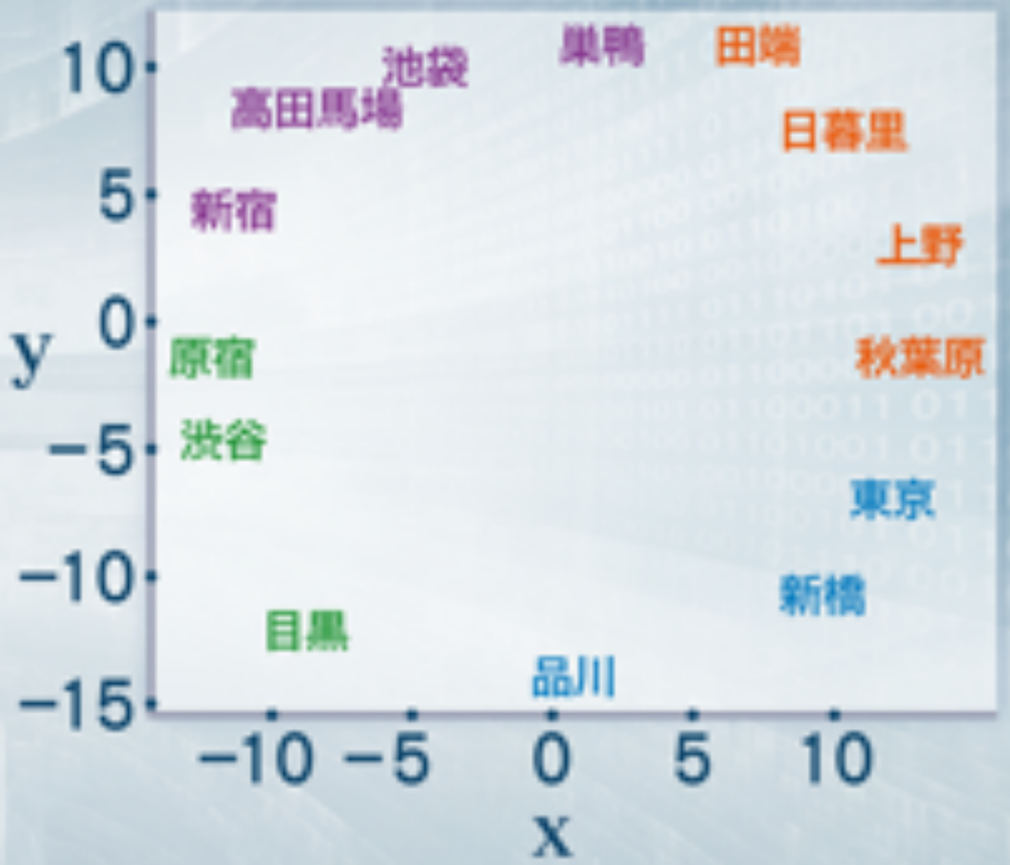
```
> library (RMeCab)
> a1 <- docMatrix(“C15”, pos = c(“連体詞”, “副詞”))
> a2 <- docMatrix(“C15”, pos = c(“連体詞”, “副詞”),2)
```

```
> a3 <- a1[row.names(a1) ! =“[[LESS-THAN-1]]” ,]
> a4 <- a3[row.names(a3) ! =“[[TOTAL-TOKENS]]” ,]
> write.table (a4, “hindo.txt”)
```

多次元尺度法・k-means法

所要時間データ

	1	2	3	…	12	13	14
1. 品川	0	7	12	…	14	10	7
2. 目黒	7	0	5	…	21	17	14
3. 渋谷	12	5	0	…	26	22	19
4. 原宿	14	7	2	…	28	24	21
…				…			
12. 秋葉原	14	21	26	…	0	4	7
13. 東京	10	17	22	…	4	0	3
14. 新橋	7	14	19	…	7	3	0



RMeCabを用いた形態素解析

```
> c1 <- t(a4)
> c2 <- dist(c1)
```

```
> c3 <- cmdscale(c2,eig=T); c4 <- c3$points
> c5 <- kmeans(c4,2); c6 <- c5$cluster
```

```
> plot(c4,xlim=c(-25,30),ylim=c(-12,16))
> text(c4,row.names(c1),col=c6,pos=3)
```