MTMineRを用いたテキストマイニング演習

同志社大学　文化情報学研究科
データサイエンス研究室
2018.6.10　MK102

目録

2

目録

3

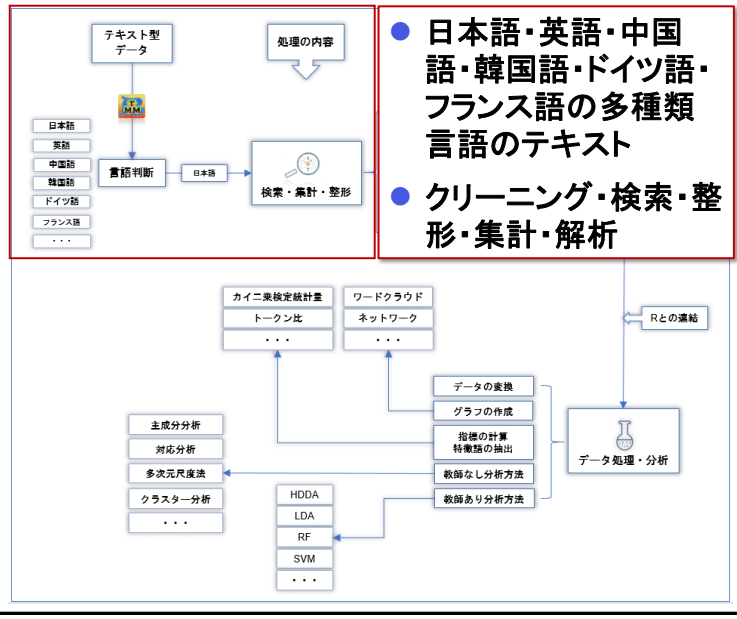
MTMineRの概要

MTMineR(Multilingual Text Miner with R)：

- テキスト型データを構造化して，Rを用いて統計的に分析を行うソフトウェア
- 文学作品・アンケートの自由記述・新聞記事など多種類のテキストの処理やデータの集計，解析等

4

MTMineRの機能

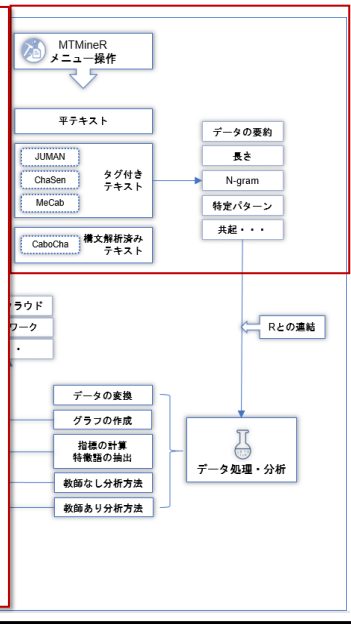


- 日本語・英語・中国語・韓国語・ドイツ語・フランス語の多種類言語のテキスト
- クリーニング・検索・整形・集計・解析

5

MTMineRの機能

- クリーニング・検索・集計など
  - プレーンテキスト
  - 形態素解
  - 構文解析
- 形態素解析器
  - 日本語：JUMAN, ChaSen, MeCab
  - 中国語：NLPIR
  - 英語・ドイツ語・フランス語：TreeTagger
- 構文解析器
  - 日本語：CaboCha

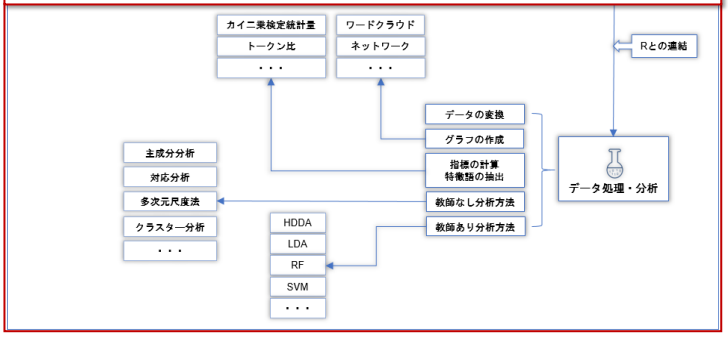


6

MTMineRの機能

データの処理と分析機能

- データの変換
- データの視覚化：ワードクラウド,ネットワーク,折れ線グラグなど
- 語彙の豊富さ指標の計算と特徴語抽出
- 教師なしの分析方法
- 教師ありの分析方法



7

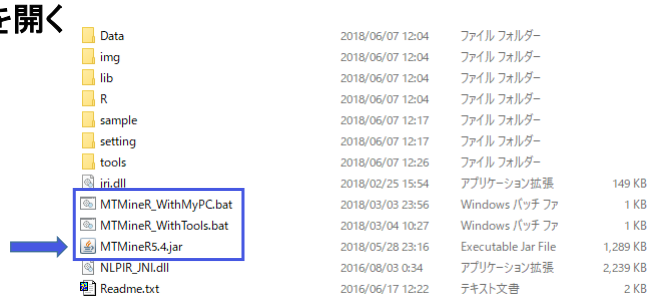
目録

8

データ収集(プレーンテキスト)

- MTMineR5.4_64_20180528.zipファイルを解凍して開く
  - [MTMineR_WithMyPC.bat]
  - [MTMineR_WithTools.bat]
  - [MTMineR5.4.jar]

⇒MTMineRを開く

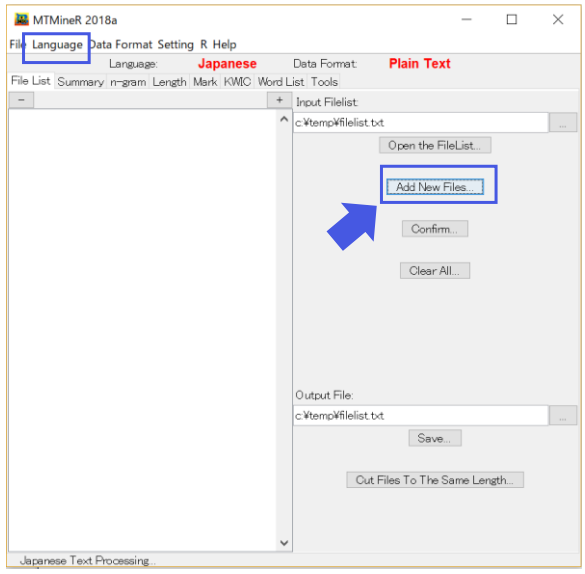


9

データ収集(プレーンテキスト)

- テキストの読み込み
- フォルダ[sample]
  - ⇒ [Japanese]
  - ⇒ [所信表明演説]

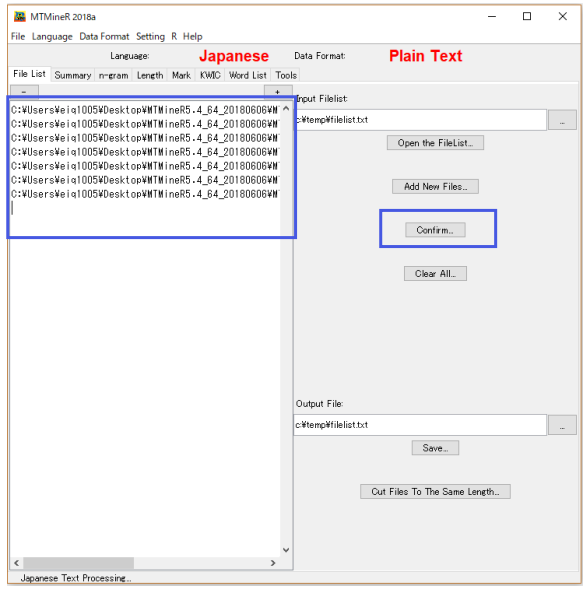
「安倍.txt」「安倍2.txt」
「安倍3.txt 」「菅.txt」
「鳩山.txt」「福田.txt」
「麻生.txt 」「野田.txt 」



10

データ収集(プレーンテキスト)

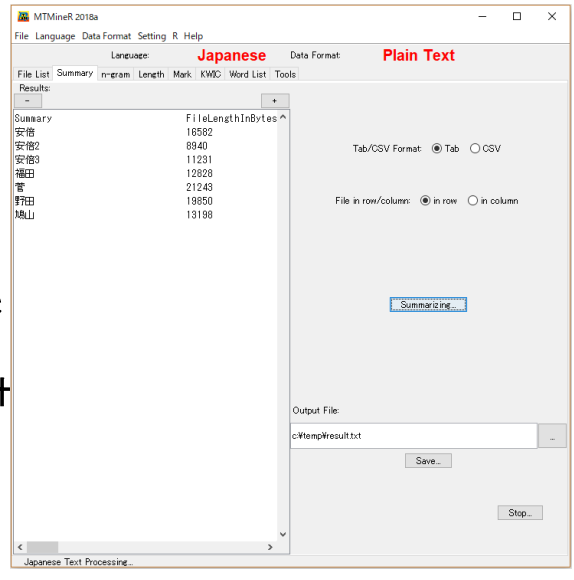
- テキストの読み込み



11

MTMineRの機能(プレーンテキスト)

- **File List**:
  テキストの読み込み
- **Summary**:
  テキストの要約
- **N-gram**:
  n-gramデータの集計
- **Length** :
  文・段落・リズムの長さ
- **Mark**:
  特定文字・記号の集計
- **KWIC**:
  クウィック検索
- **Tools**:
  テキストの整形



12

KWIC（クウィック検索）

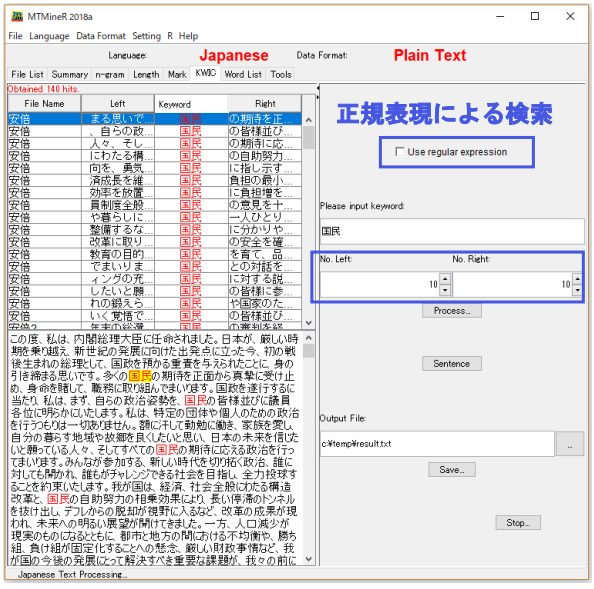
- **KWIC:**

(Keyword in Context)

指定したキーワードの前後の文脈を原文から抽出

「No.Left」「No.Right」

前後の文脈長さの指定



13

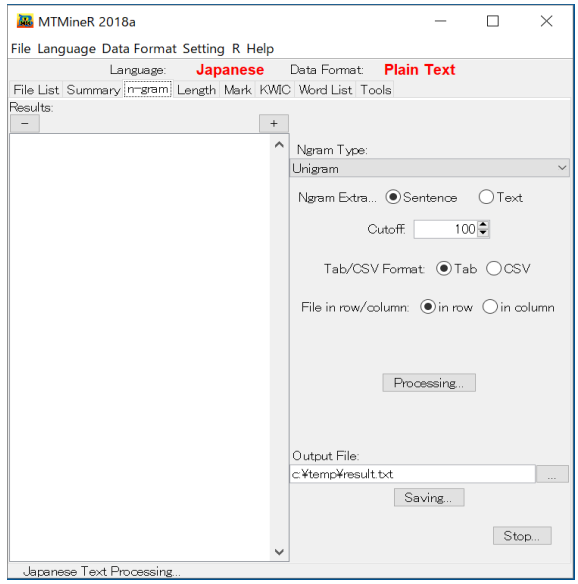
N-gram

- **n-gram**
  - **テキストにおけるある言語単位（文字や形態素、品詞など）が1単位または2単位，3単位などN単位が隣接して生じる言語単位**
  - **Unigram(n=1)，Bigram(n=2)，Trigram(n=3)**
    Fourgram(n=4)，Fivegram(n=5)・・・
  - **文字・記号**
  - **形態素**
  - **品詞**
  - **文節**
  - **など**

14

N-gram

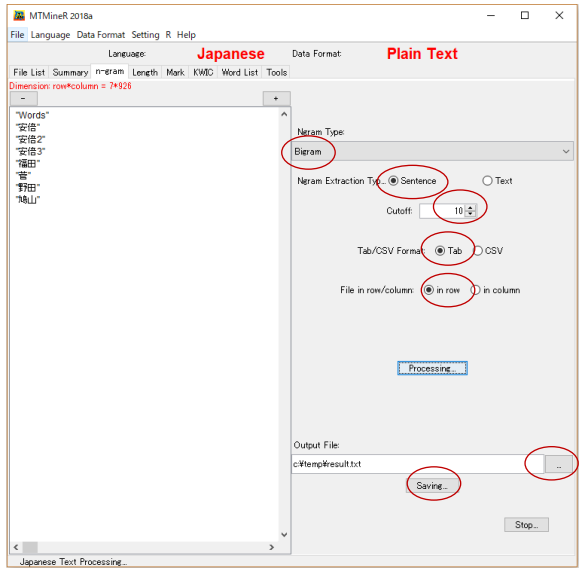
- 文字単位のn-gram
- Ngram Type(n=1,…,6)
  - Unigram(n=1)
    ⇓
  - Sixgram(n=6)
- Ngram Extraction Type
  - Sentence：文ごと
    - 「。_文頭文字」×
  - Text：文章ごと
- Cutoff（閾値）：
  データセットのサイズのコントロール



15

N-gram

- 文字単位のbigram
  ステップ
  1. Bigram(n=2)
  2. Sentence
  3. Cutoff(10)
  4. Tab Format
  5. File in row
  6. Processing



16

目録

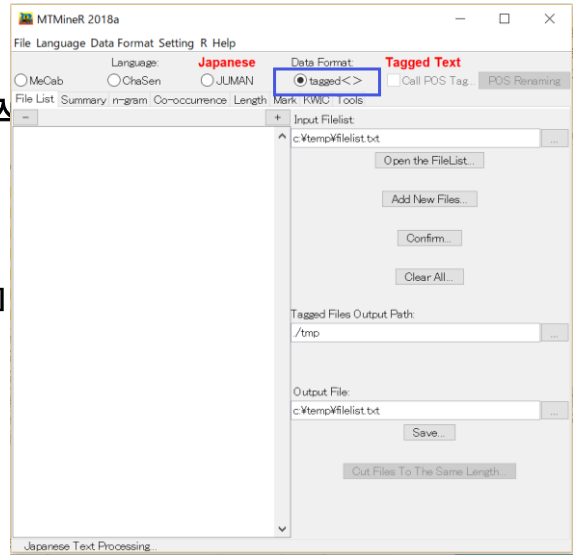
17

タグ付きテキストについて

- Format⇒Tagged Text
- 整形したタグ付きtext
  - [tagged]⇒テキスト読み込み⇒データ集計
- 形態素解析済みtext
  - MeCabで解析⇒[MeCab]
  - ChaSenで解析⇒[ChaSen]
  - JUMANで解析⇒[JUMAN]

  ⇒テキスト読み込み⇒データ集計
- プレンテキスト

  ⇒テキストを読み込む⇒形態素解析⇒データ集計



18

日本語の形態素解析

ステップ：

① 形態素解析器を選ぶ

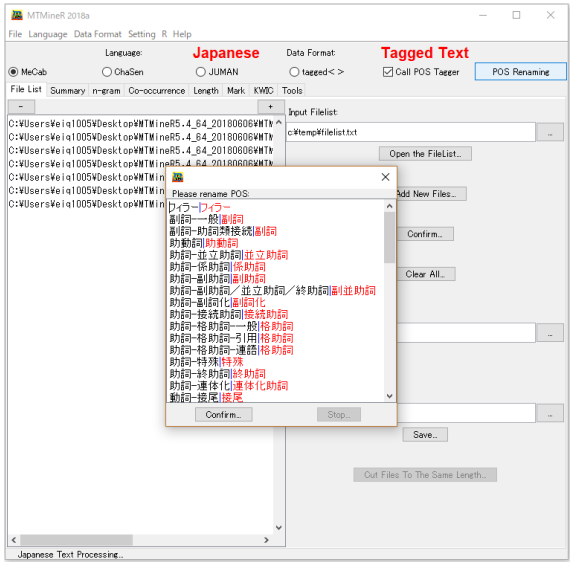
(MeCab, ChaSen, JUMAN)

② 「Call POS Tag」☑

③ 「POS Processing」

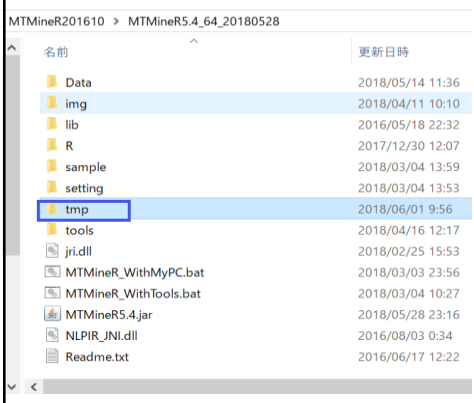
④ Tag名前の変更(必要な場合)

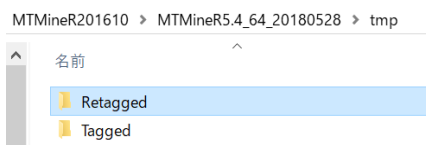
⑤ 「Comfirm」



19

日本語の形態素解析





Tagged：

形態素解析済みデータ

Retagged:

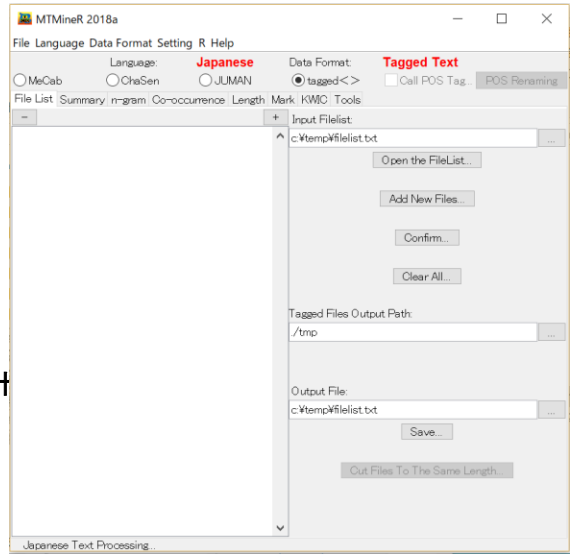
整形したデータ

MTMineRを閉じると，「tmp」は自動的に削除されるため,必要に応じて各自保存する

20

MTMineRの機能(タグ付きテキスト)

- **File List**:
  テキストの読み込み
- **Summary**:
  テキストの要約
- **N-gram**:
  n-gramデータの集計
- **Length**:
  形態素・品詞などの長さ
- **Mark**:
  特定形態素・品詞などの集計
- **KWIC**:
  クウィック検索
- **Tools**:
  テキストの整形



21

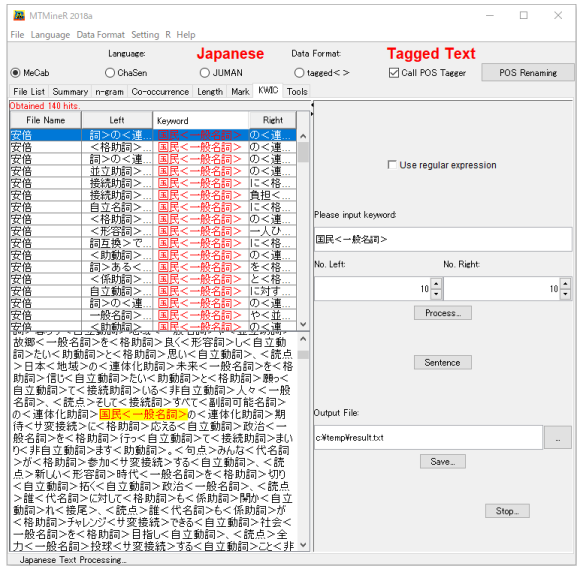
KWIC(クウィック検索)

- **KWIC:**

タグ付きテキストから指定したキーワードの前後の文脈

入力形式：

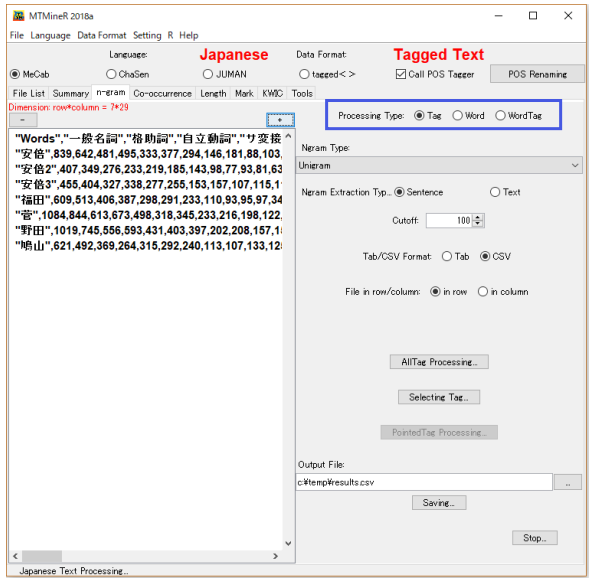
例：国民<一般名詞>



22

N-gram

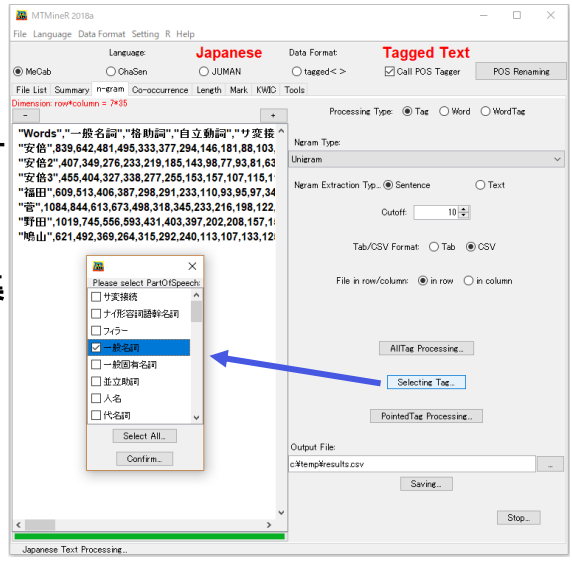
- **Processing Type**:
  データの種類の指定
  - タグ(Tag)
  - 形態素(Word)
  - タグ付き形態素
    (Word Tag)
- **Ngram Type**
  - Unigram(n=1)
    ⇓
  - Sixgram(n=6)
- **N-gram Extraction type**
  - Sentence：文ごと
    - 「句点_品詞」×
  - Text：文章ごと



23

N-gram

- **[All Tag Processing]**
  - すべてのデータを集計
- **[Selecting Tag]⇒[Pointed Tag Processing]**
  - 指定したデータのみ集計



24

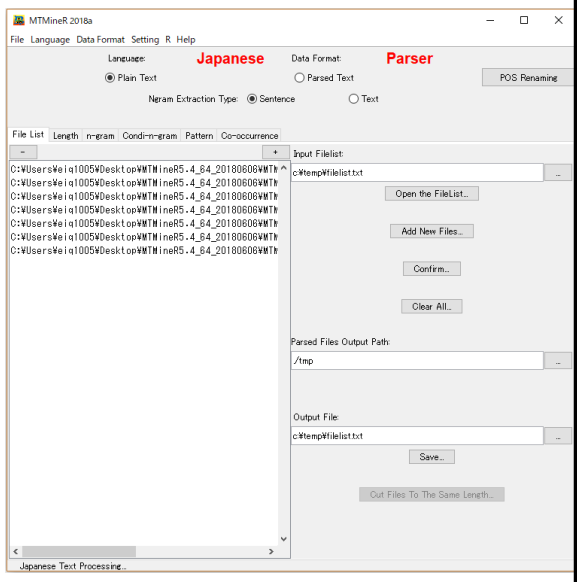
目録

25

構文解析

Data Format⇒Parser

- **Parsed Text:**
  - 整形したテキスト
  - 構文解析済みテキスト
- **Plain Text:**
  - Plain Text
  - 「POS Processing」
- **N-gram**
  - 文節のn-gram
- **Condi-n-gram**
  - 条件付きn-gram
- **Co-occurrence**
  - 文節の共起



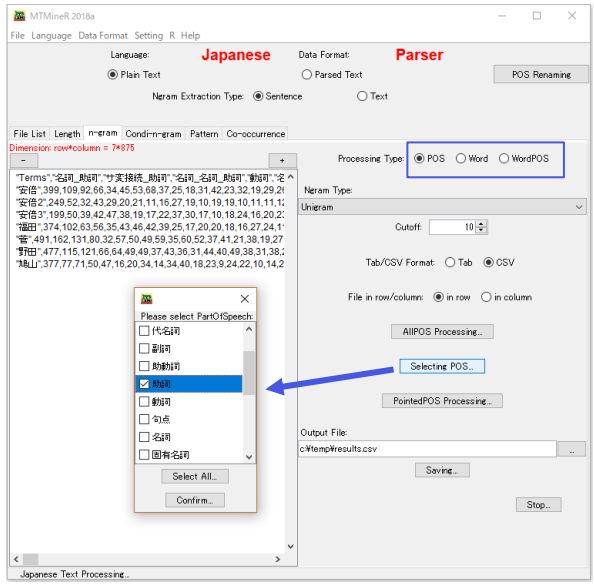
26

構文データの集計

- **N-gram:**
- **Processing Type:**

  データの種類の指定

  - タグ(Tag)
  - 形態素(Word)
  - タグ付き形態素（Word Tag）



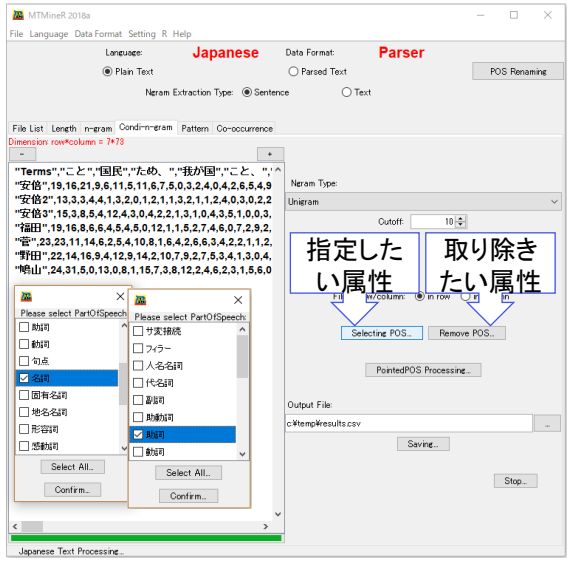
構文データの集計

- **Condi-n-gram:**
- 条件付き文節のn-gram
  指定した属性のn-gram
  の中から一部の属性データを除外する集計方法
- 複合語を含めた形態素より長い単位の語句の集計

例：

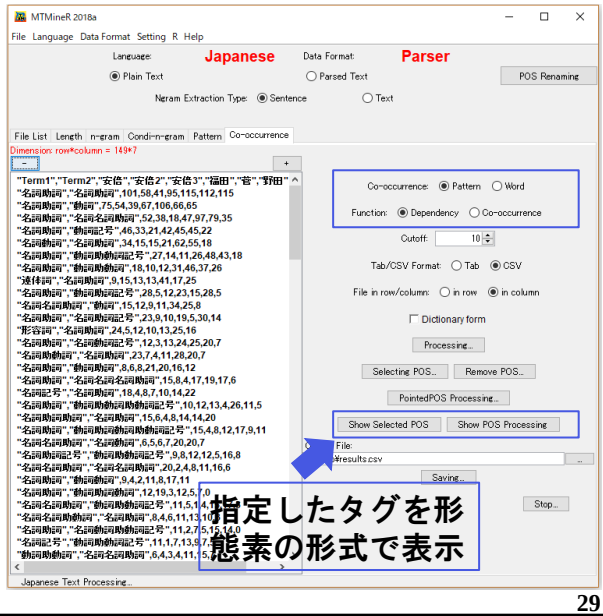
- [Selecting POS]
  「名詞」
- [Remove POS]
  「助詞」

⇒「国際社会」「社会保障」「経済成長」のような複合語



構文データの集計

- **Co-occurrence:**
  文節の共起
- **集計形式：**
  ① Pattern
  ② Word
- **共起抽出方法：**
  ① **Dependency**:
  係り受け先を考慮
  ② **Co-occurrence**:
  係り受け関係を考慮せず



29

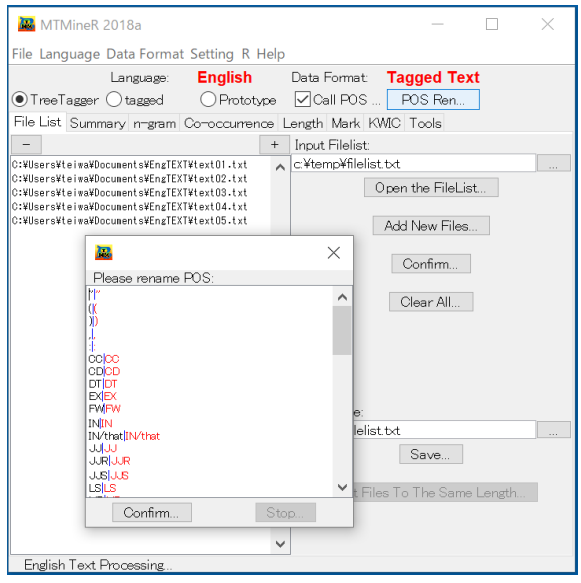
英語の形態素解析

Language⇒English Data

Format⇒Tagged Text

ステップ：

① TreeTagger
② 「Call POS Tag」
③ 「POS Processing」
④ Tag名前の変更（必要な場合）
⑤ 「Comfirm」



30

英語の形態素解析

- Prototypeを選択しない
  [単語/タグ]
- Prototypeを選択
  [単語の原型/タグ]

単語	タグ	単語の原型	
133	CD	@card@	
years	NNS	year	
ago	RB	ago	
,	,	,	
Joseph	NP	Joseph	
Hardy	NP	Hardy	
Neesima	NP	Neesima	
broke	VVD	break	
new	JJ	new	
ground	NN	ground	
in	IN	in	
Japanese		JJ	Japanese
education		NN	education
and	CC	and	

31

目録

- MTMineRの概要
- データ集計
  - 平テスト
  - 形態素解析
  - 構文解析
- Rによる分析
  - 特徴抽出
  - ワードクラウド，ネットワーク分析
  - 教師なし分析
  - 教師あり分析

32

特徴抽出

●**特徴抽出**
2群及び多群のものに関してグループ間に差がある
ものを抽出をすることでそれぞれの特徴をみる

●**演習**
MTMineRの中にある三島由紀夫と川端康成の作品
について、一般名詞の特徴を考察してみる

〇フォルダ[sample]⇒[Japanese]⇒[川端・三島]

33

特徴抽出

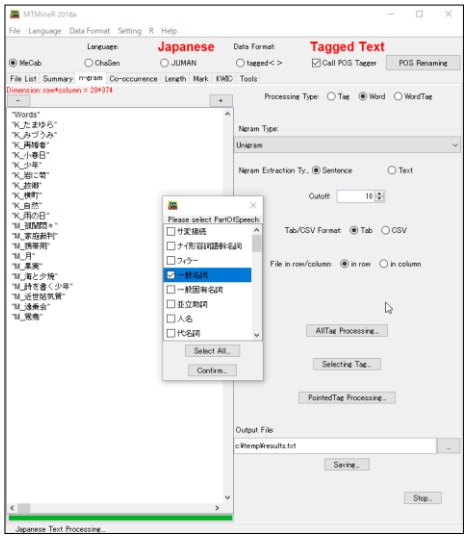
●**データ集計・準備**
（1）一般名詞の集計
（2）集計データをRに読み込み
＊（3）度数データを相対頻度データに変換

●**特徴量抽出**
Chi-square Test（カイ二乗統計量）
Likelihood Ratio Test（尤度検定統計量）
Mahalanobis' Distance （マハラノビス距離）
Mean Accuracy （RFのMean Decrease Accuracy）
Mean Gini(RFのMean Decrease Gini)
Kruskal-Wallis （クラスカル・ウォリス検定統計

34

特徴抽出

（1）一般名詞の集計



・解析器選択：Mecab

・「Call POS Tagger」にチェック

・「POS Renaming」をクリック

・処理の種類：Word

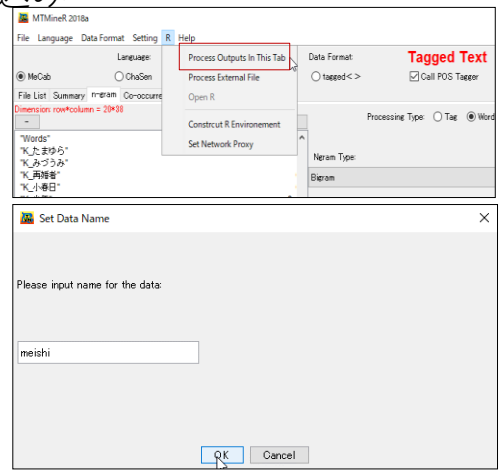
・cutoff 値設定：10

・集計対象タグの選択：一般名詞

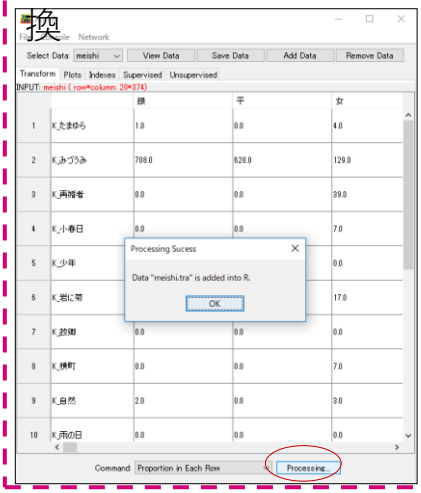
35

特徴抽出

（2）集計データをRに読み込み



*（3）度数データを相対頻度データに変換



36

特徴抽出

メニュー［Indexes] ⇒［Features］

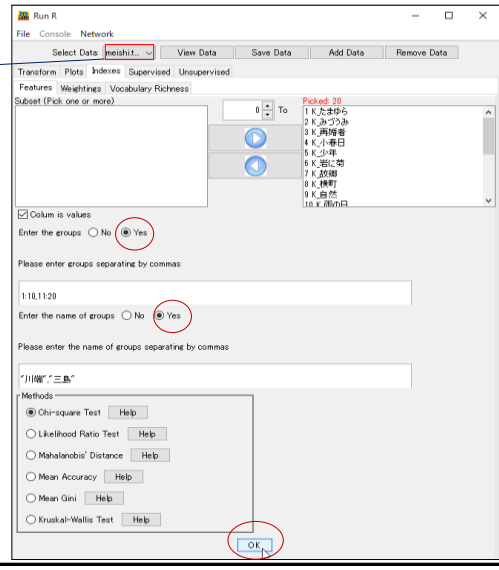
データ選択「meishi

チェックを入れ，個体を指定

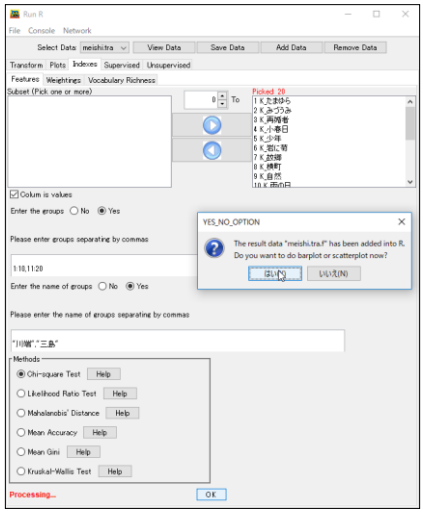
グループを指定

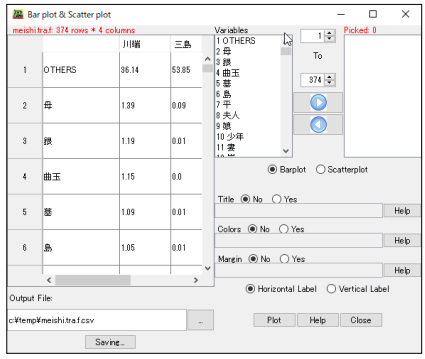
グループの名前を記述

手法選択

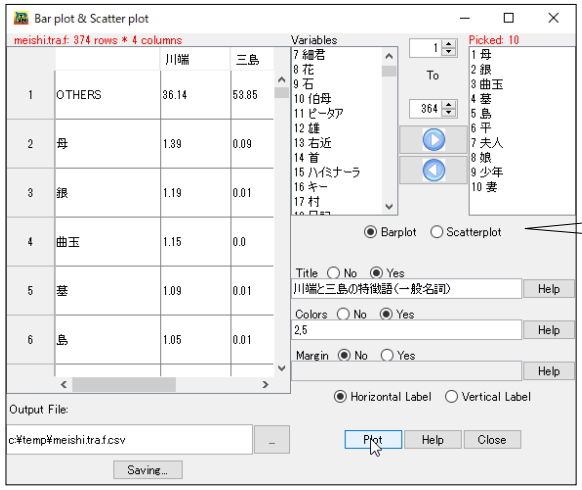


特徴抽出





特徴抽出



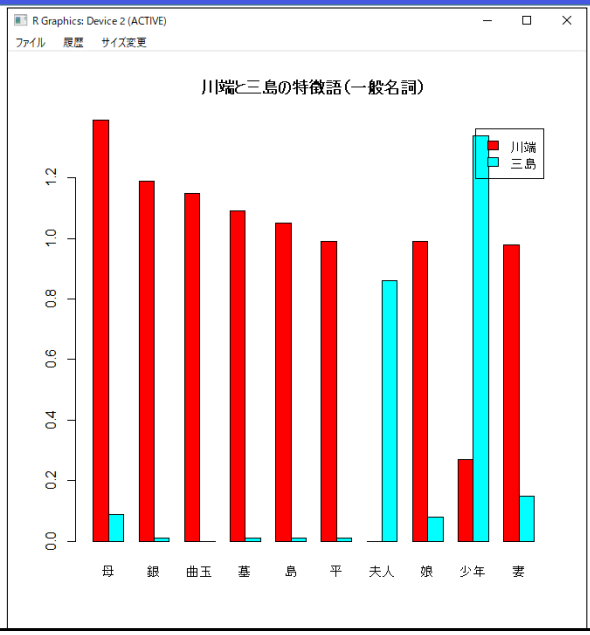
変量指定

Barplot:棒グラフ
Scatterplot:散布図

グラフのタイトル，色，グラフ周辺のマージン，ラベルの文字列の方向を指定

39

特徴抽出



・川端の特徴

母，銀，曲玉，墓，平

娘，妻

・三島の特徴

夫人，少年

40

目録

41

ワードクラウド

- **ワードクラウド**

  文章中で出現頻度が高い単語をその頻度に応じた大きさで図示する方法

- **演習**

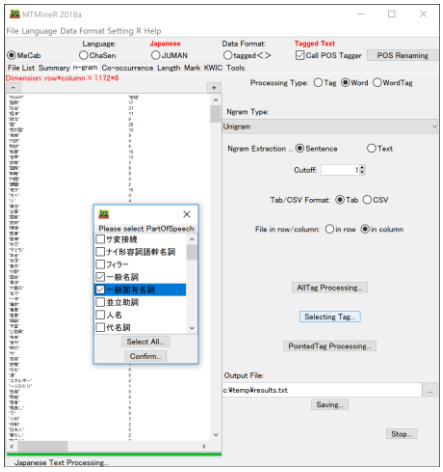
  MTMineRの中にある各首相の所信表明演説文についてワードクラウドで考察せよ.

  ○フォルダ[sample]⇒[Japanese]
  ⇒ [所信表明演説]の「安倍.txt」「福田.txt」「麻生.txt 」
  「安倍.txt」「福田.txt」「麻生3.txt 」

42

ワードクラウド

（1）名詞の集計



・解析器選択：Mecab

・「Call POS Tagger」にチェック

・「POS Renaming」をクリック

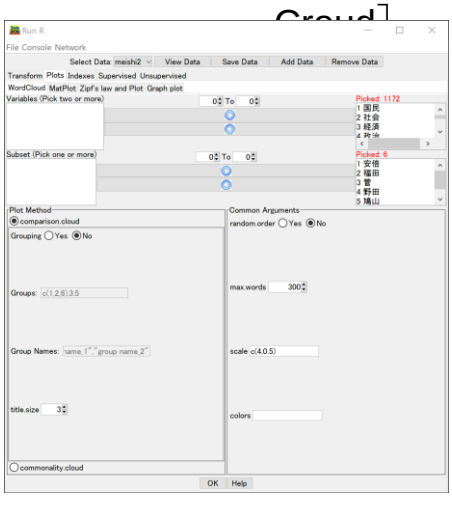
・処理の種類：Word

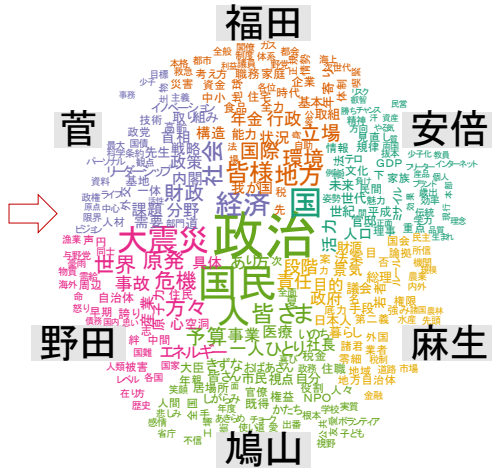
・cutoff 値設定：1

・集計対象タグの選択：一般名詞,

ワードクラウド

（2）集計データをRに読み込み　→メニュー　［plot] ⇒［Word Cloud］





ネットワーク分析

- **ネットワーク分析**
  さまざまな対象を点と線からなるネットワークで表現
  構造的な特徴を探る

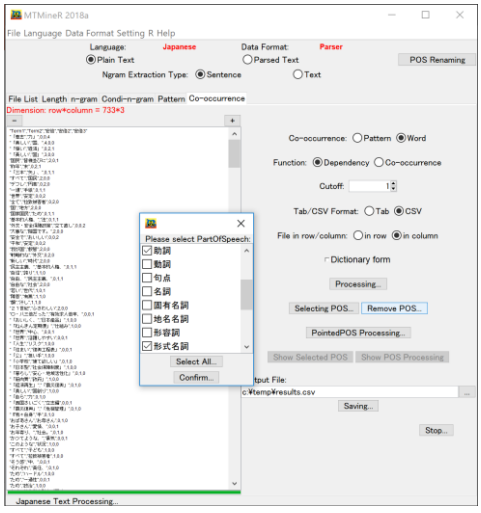
- **演習**
  MTMineRの中にある安倍総理の所信表明演説文
  についてネットワーク分析を行う

  〇フォルダ[sample]⇒[Japanese]
  ⇒ [所信表明演説]の「安倍.txt」「安倍2.txt」「安倍3.txt 」

45

ネットワーク分析

（１）名詞，形容詞の共起関係の集計



・Data Format：Parser

・Plain Text を選択

・ 「POS Renaming」をクリック

非自立名詞を形式名詞と変更

・処理の種類：Word

・cutoff 値設定：1

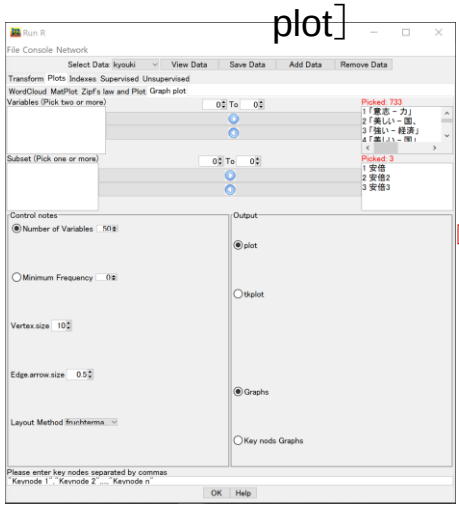
・集計対象タグ：名詞，形容詞

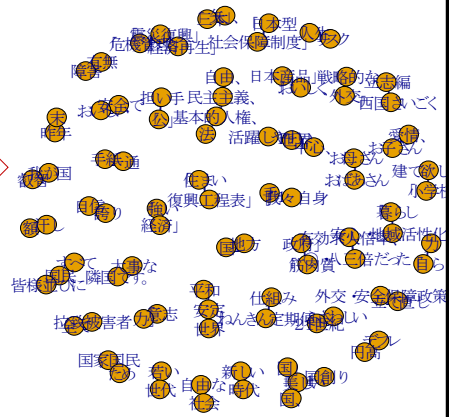
・集計除外タグ：助詞，形式名詞

46

ネットワーク分析

（2）集計データをRに読み込み　→メニュー　［plot] ⇒［Graph plot］





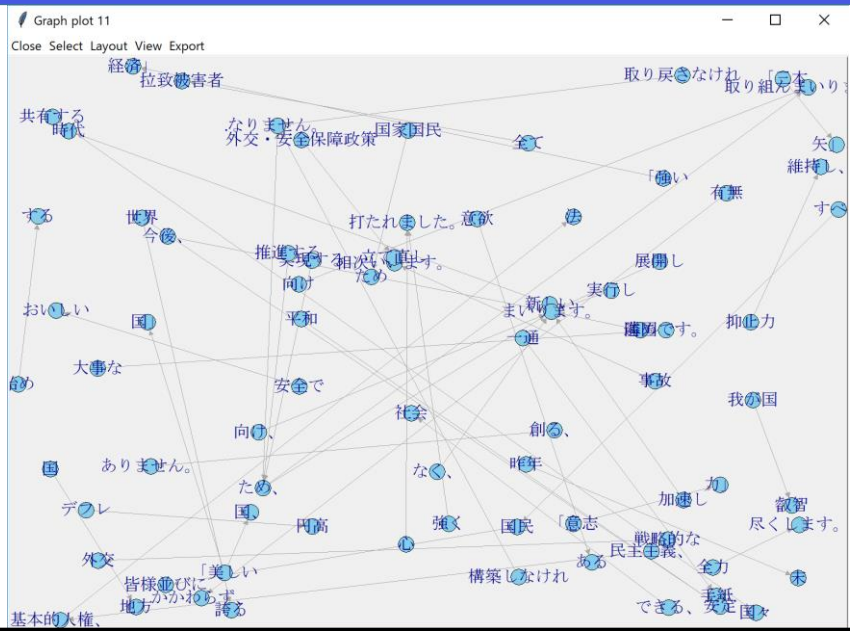
ネットワーク分析(tkplot)

OR Console

1. データを変換
2. 色の指定：“red”,” blue”等
3. 図を出力

```
> abe <- graph.data.frame(kyouki[1:50,])
> V(abe)$color <- “skyblue”
> tkplot(abe)
```

ネットワーク分析(tkplot)



49

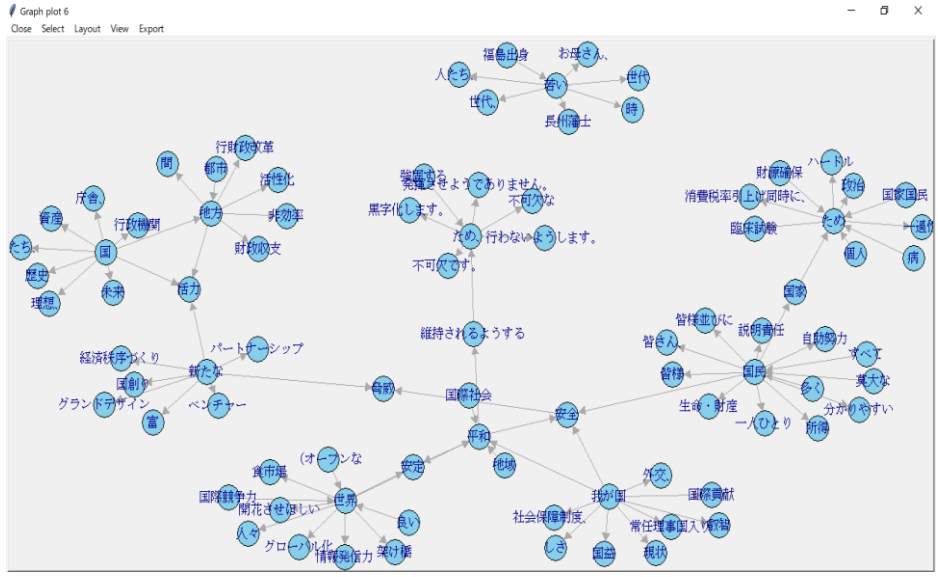
ネットワーク分析(次数が高いもの)

OR Console

```
>ng <- neighborhood(abe,1,sort.list
      (degree(abe),decreasing=TRUE)[1:10])
>z.sub <- c(ng[[1]])
>for(i in 2:10){
      z.sub<-c(z.sub,ng[[i]])
          }
>zisuu <- induced.subgraph(abe,z.sub)
>zisu <- simplify(zisuu)
>tkplot(zisu)
```

50

ネットワーク分析(tkplot)



51

目録

- MTMineRの概要
- データ集計
  - 平テスト
  - 形態素解析
  - 構文解析
- Rによる分析
  - 特徴量抽出
  - ワードクラウド，ネットワーク分析
  - **教師なし分析**
  - 教師あり分析

52

教師なし分析

●教師なし分析
目的変数なしの分析方法

●演習
MTMineRの中にある三島由紀夫と川端康成の作品について教師なし手法を用いて分析せよ

〇フォルダ[sample]⇒[Japanese]⇒[川端・三島]

53

教師なし分析

●データ集計・準備
（1）データの集計
（2）集計データをRに読み込み
＊（3）度数データを相対頻度データに変換

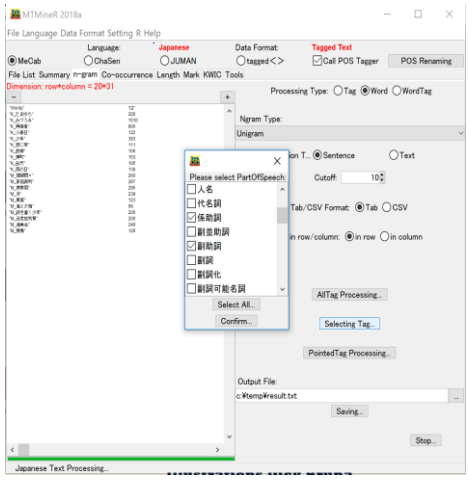
●分析手法
Principal Components Analysis （主成分分析）
Correspondence Analysis （対応分析）
Hierarchical Clustering （階層的クラスター分析）
K-means Clustering （K-meansクラスタリング）
Multidimensinal Scaling （多次元尺度法）
Partitioning Around Medoids （k-medoids 法）
Latent Dirichlet Allocation （潜在的ディリクレ

54

教師なし分析

●**データ集計・準備**

（1）助詞のunigramの集計



・解析器選択：Mecab

・「Call POS Tagger」にチェック

・「POS Renaming」をクリック

・処理の種類：Word

・cutoff 値設定：10

・集計対象タグの選択：助詞全般

主成分**分析**

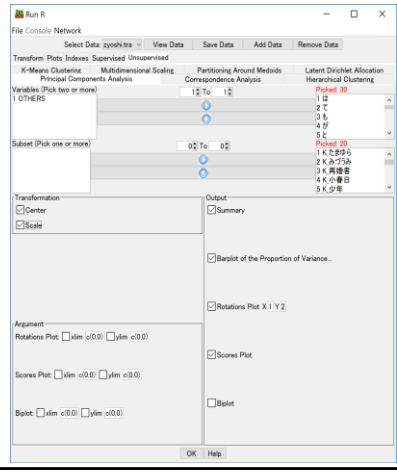
●主成分分析

多次元のデータをデータの損失をなるべく少なくし，低次元に縮約，データの概要を把握

○量的データを用いる

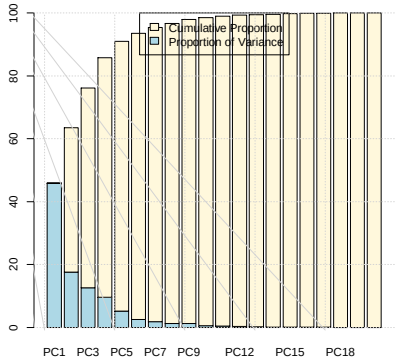
→相対度数データ

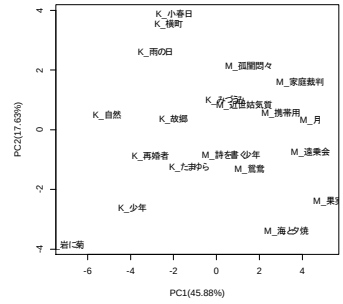
Centerのみにチェック
→分散共分散行列
Center とScale両方にチェック
→相関行列

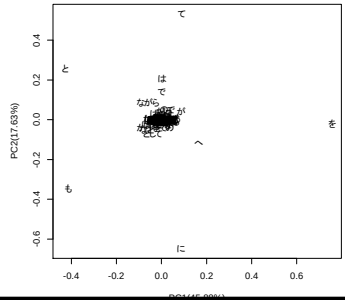


主成分**分析**

- 主成分分析
  第2主成分までのプロット



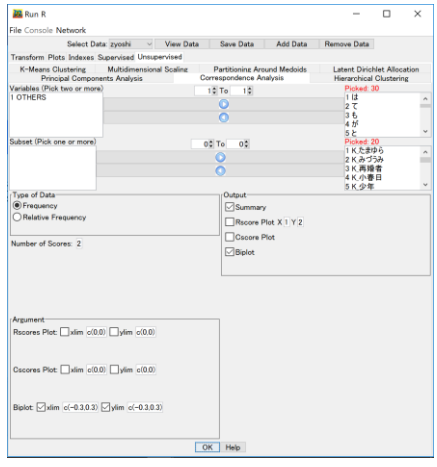




57

対応分析

- 対応分析
  多次元のデータをデータの損失をなるべく少なくし，低次元に縮約，データの概要を把握
  ○質的データを用いる
  →度数データ



58

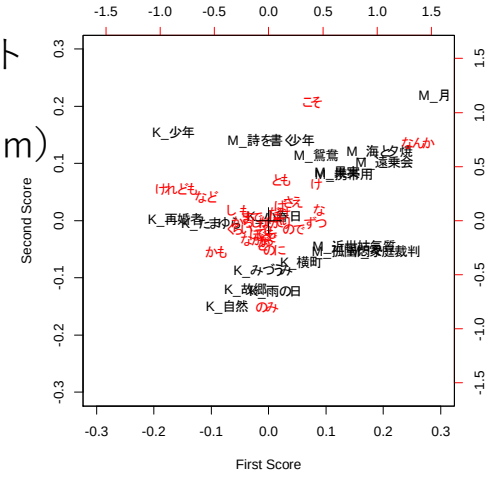
対応分析

●対応分析

第2軸までのバイプロット

赤：変数（助詞のunigram）
黒：作品

類似しているものは
近くに配置



59

階層的クラスター分析

●階層的クラスター分析
　異なる性質が混ざった集団から，互いに似た性質のものを集めクラスターを作成

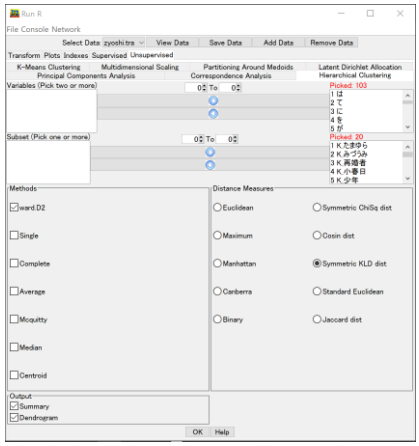
●距離
Euclidian，Maximum
Manhattan，Canberra
Binary
Symmetric Chisq dist
Cosain dis
Symmetric KLD dist
Standard Euclidian
Jaccard dist

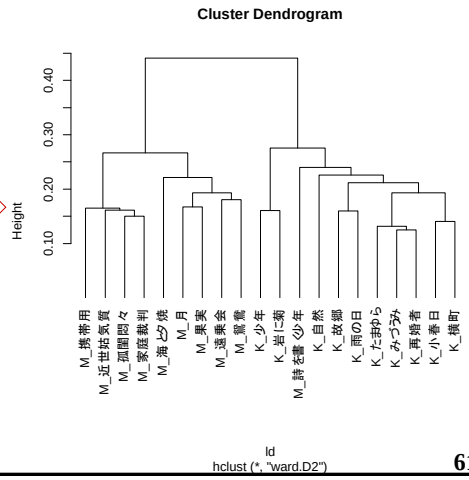
●結合方法
ward.D2
Single
Complete
Average
Mcquitty
Median
Centroid

60

階層的クラスター分析

●階層的クラスター分析



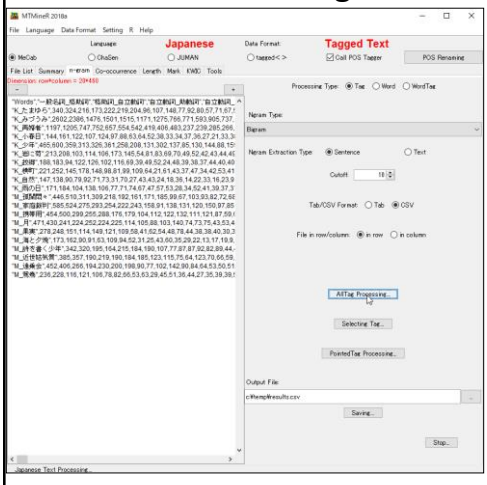


61

教師なし分析

●**データ集計・準備**

（1）品詞タグのbigramの集計



・解析器選択：Mecab

・「Call POS Tagger」にチェック

・「POS Renaming」をクリック

・処理の種類：Tag

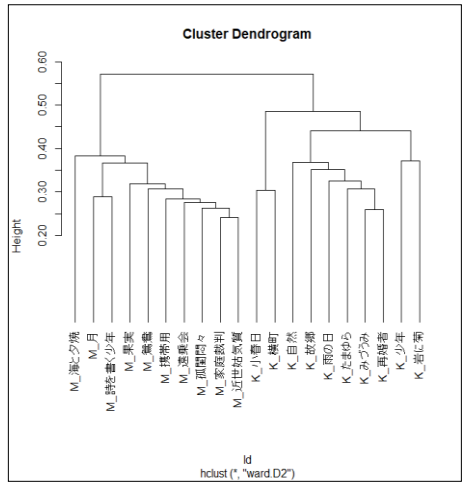
・cutoff 値設定：10

・集計実施（All Tag Processing）

62

階層的クラスター

●階層的クラスター分析



63

非階層的クラスター分析

●非階層的クラスター分析

異なる性質が混ざった集団から，互いに似た性質のものを集めクラスターを作成

→階層構造を持たず，あらかじめいくつのクラスターに分けるか設定

例）今回は川端と三島の2群であるため，2クラスターと設定

分析手法

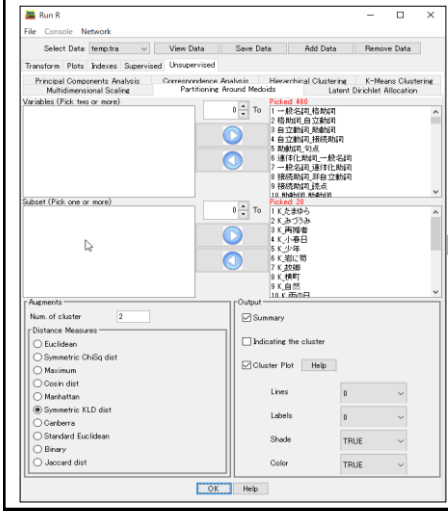
○K-means Clustering　（K-meansクラスタリング）

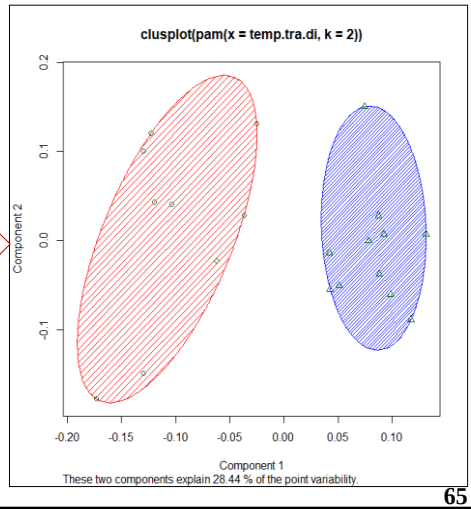
○Partitioning Around Medoids　（k-medoids 法）

64

非階層的クラスター

●k-medoids 法





65

トピックモデル

●トピックモデル

文章データ群から，各文章の主題（トピック）を
判断するためのモデルを構築
例）ホームラン，キャッチャー　→　「野球」

○LDA(潜在的ディリクレ配分法)

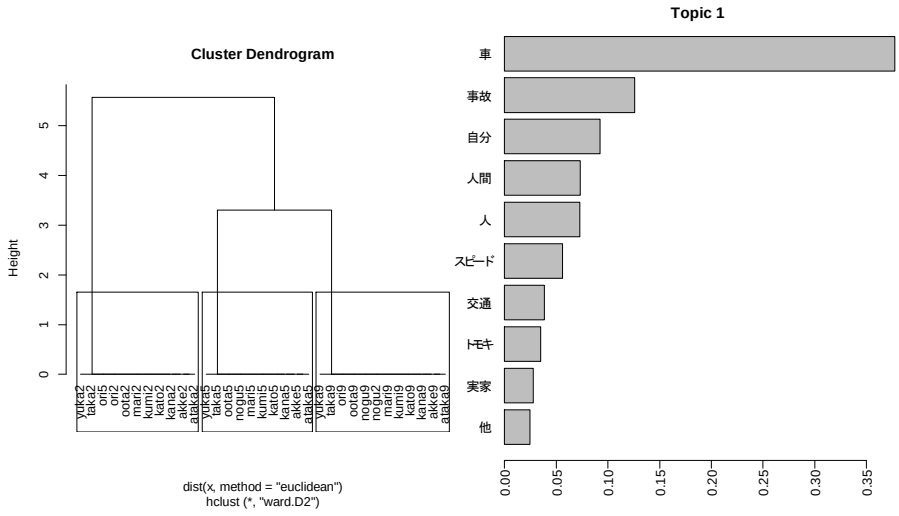
多次元のデータをデータの損失をなるべく少なくし，
低次元に縮約，データの概要を把握

〇フォルダ[sample]⇒[Japanese]⇒[作文]
三つのテーマで書いた11人の作文

66

トピックモデル

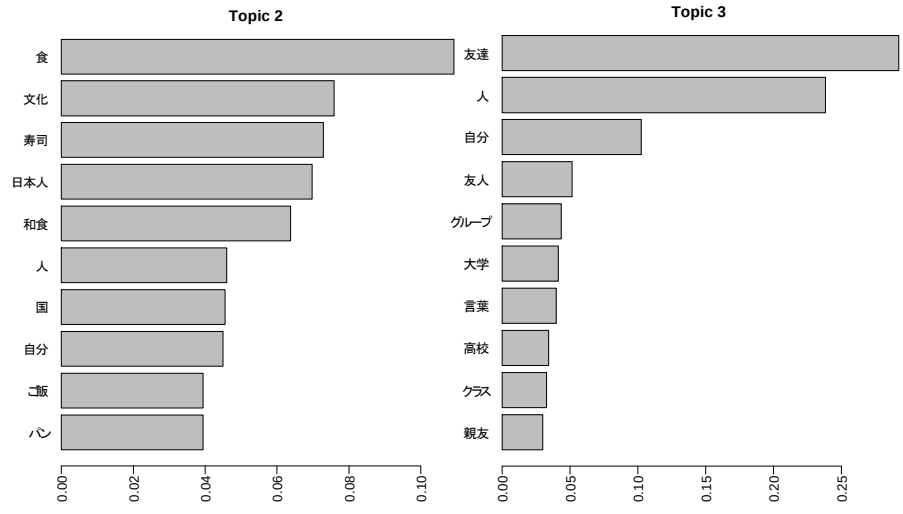
- LDA



67

トピックモデル

- LDA



68

目録

69

教師あり分析

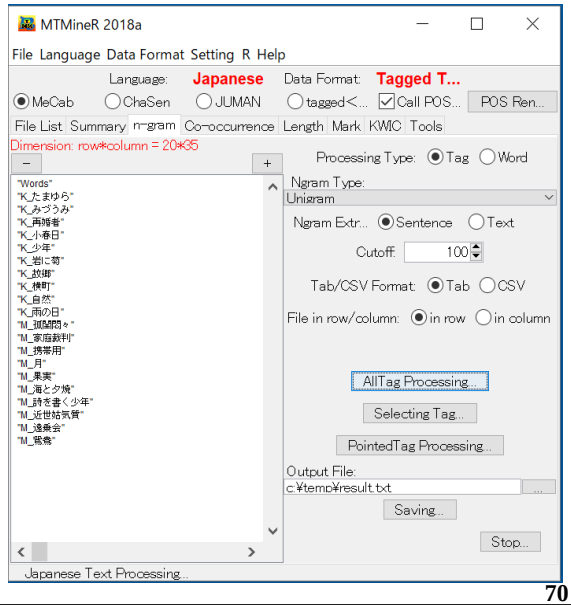
- **データ集計:**

MTMineR → sample

- 中国語
- 英語
- 日本語
  - 川端康成　10編
  - 三島由紀夫　10編
- 韓国語

形態素解析済みのテキスト→Tagのunigram

cutoff=0



70

教師あり分析

●データ変換：

Proportion in Each Row

↓

Processing

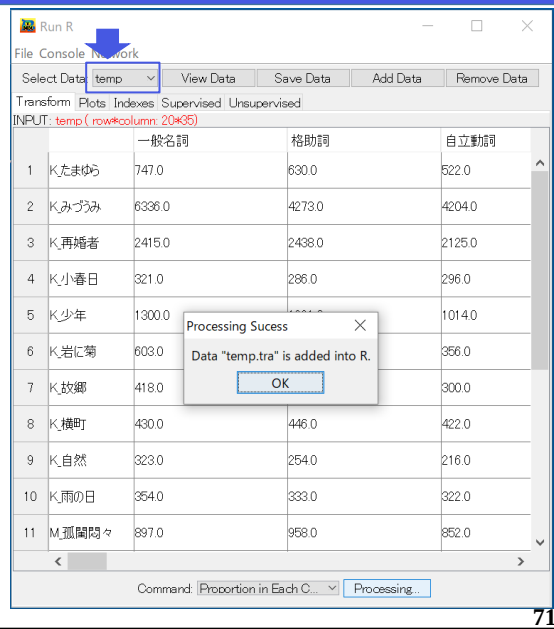
↓

Data “temp.tra” is added into R→OK

↓

Select Data:

[temp.tra]



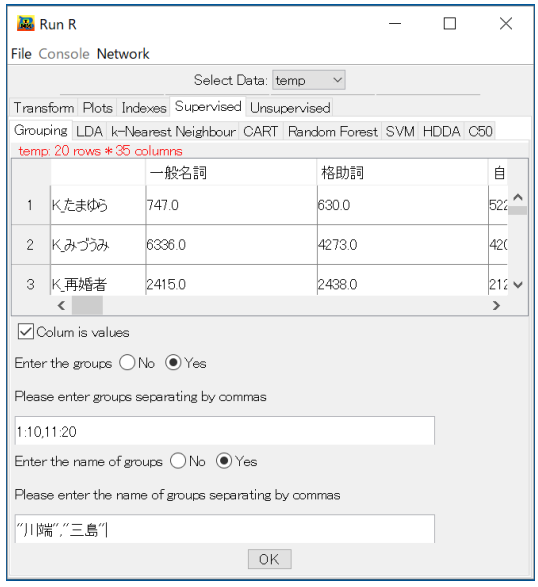
71

教師あり分析

●ラベル設定：

ステップ：

① グループの指定
② ラベルを付ける
③ OK → ラベル付きデータセットが完成
④ 作成したデータセット temp.tra.groupedを[Select Data]で指定



72

教師あり分析

タブSupervisedには教師ありの機械学習法が実装されている

- CART
- C5.0
- k-Nearest Neighbour
- RandomForest
- SVM (support vector machine)
- LDA (Linear Discriminant Analysis)
- HDDA (High-Dimensional Discriminant Analysis)

MTMineRの中の各分析方法における関数およびパラメータはRと同様であるため、必要であれば各パッケージのサイトにご確認ください

73

CART (決定木)

決定木：

テキストの分類や回帰分析を行う機械学習法の一つ，変数を分岐頂点とし，葉を予測値とした樹状構造の統計モデル

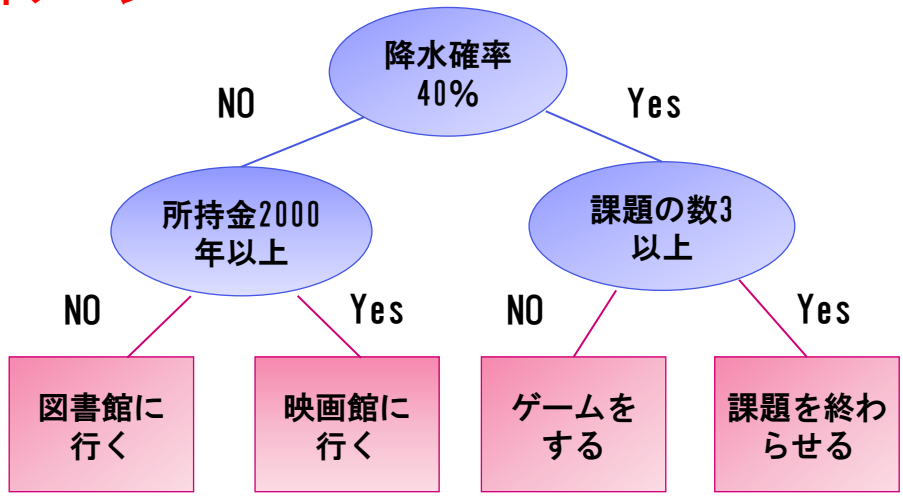
特徴：

- ノンパラメトリックな教師あり学習方法
- 解析対象のデータの分布を仮定しない
- 事前に与えられたデータから未知のデータを推定

74

CART (決定木)

イメージ：



75

CART (決定木)

長所：

- 可読性が高い
  木が生成されるイメージで結果を出力し，
  わかりやすい
- 説明変数・目的変数共に名義尺度・間隔尺度など様々
  データの対応
- 外れ値に対して頑健

短所：

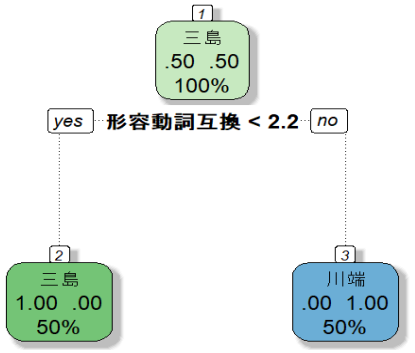
- 分類性能の高い手法ではない
- 過学習を起こしやすい
  パラメータの調整や枝の刈り込みを上手に行う必要

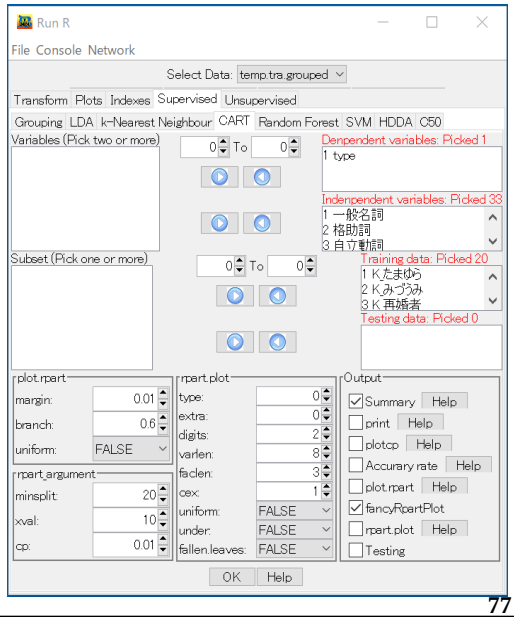
76

CART (決定木)

各パラメータはRパッケージの中の名前と同じであるので，各パッケージのサイトに確認

パラメータを変換することで，異なるプロットが描かれる





77

Random Forest

決定木を複数組み合わせ、各決定木の予測結果を多数決することによって結果を得る

アルゴリズム：

① ランダムにデータを抽出する
② 決定木を成長させる
③ ステップ①②を指定回繰り返す
④ 予測結果を多数決することによって分類ラベルを決定する

78

Random Forest

長所：

- 考慮するパラメータが少ない

  主なパラメータ：

  ① サンプリング数

  ② 決定木を成長させる際に使用する変数の数
- 重要度の高い変数を出力できる

短所：

- データが少ない場合は、分類精度が高くない

79

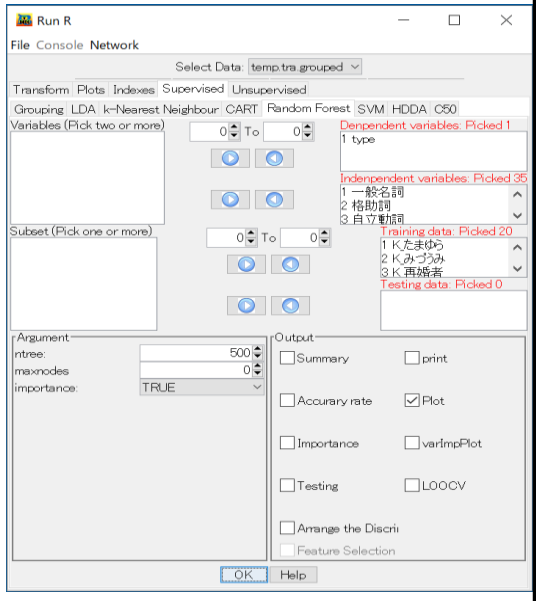
Random Forest

- Summary: 要約の出力
- Print: OOBでの予測結果
- Accurary rate:正解率
- Plot:木の数と誤り率との対応図
- Important:変数重要度の計算
- VarImpPlot:分類における変数の重要度のランキング
- Testing: 予測
- LOOCV：Leave-one-outを行う
- Arrange the Discrimination Maker:

Mean Decrease Accuracyによる変数重要度に基づき、変数を降順でソートし、結果をarrangeに入れる

- Feature Selection:

テキスト分類に用いるべき変数の数を決める



80

Random Forest

OOB：out-of-bag

Random Forestでは、**各決定木で異なるサンプルを使って学習する。**

学習データのうち、平均的に約1/3のデータは学習に使われない、それをOOBと呼ぶ

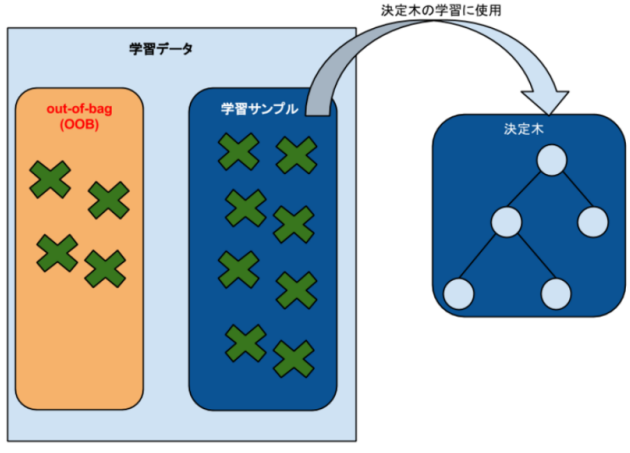
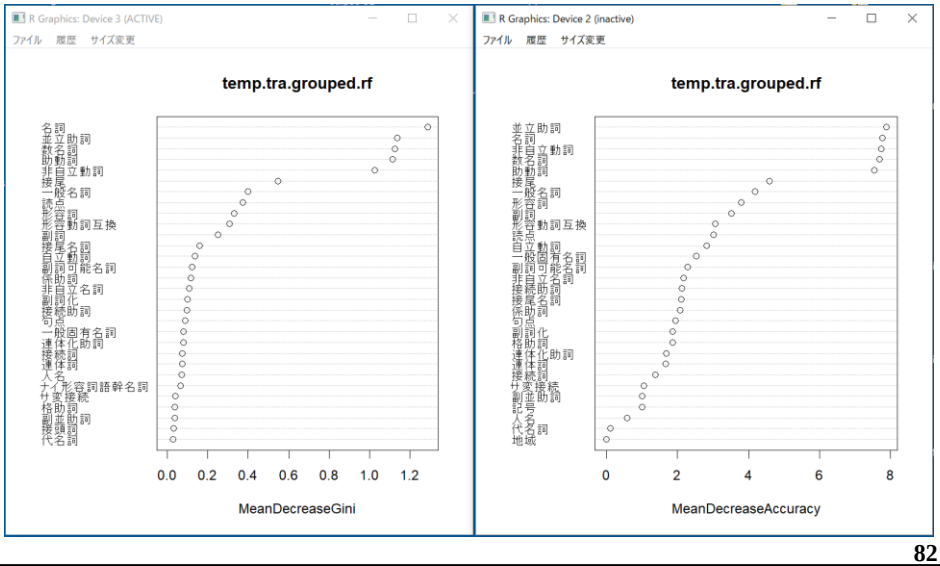


図　決定木の学習（各決定木でサンプリングと学習を繰り返す）

81

Random Forest

変数重要度の結果



82

ご清聴ありがとうございました