テキストマイニングツール TTM (TinyTextMiner) の理念と使い方

松村真宏・三浦麻子・金明哲

テキストマイニングの登場

- テキストデータのような定性データは，大量のデータを分析することで安定した傾向が見いだせますが，人手で大量のテキストデータを分析することは現実的にはほとんど不可能でした
- テキストマイニングの登場によって，大量のデータを統一的な視点・基準から少ない労力で分析することが可能になりました
- テキストマイニングは，世の中で流行っている話題や，人々のニーズや不満を定量的に把握する手段として徐々に使われ始めています

テキストマイニングの主な基盤技術

- 自然言語処理
  - 形態素解析，構文解析
- 統計解析
  - 多変量解析，仮説検定
- データマイニング
  - 分類器，予測器

テキストマイニングの内側

- テキストマイニングの処理は,
  「前処理」と「後処理」に大別できます
- 前処理（テキストデータならではの処理）
  - 語の切り出しと集計
  - これが面倒…
- 後処理（多変量解析やデータマイニングと同じ処理）
  - 集計データの統計処理
  - 従来の手法が流用できます！

テキストデータの特徴 (1/4)

- 語彙や表現の揺れ（漢字・仮名・カタカナ・大文字と小文字・全角と半角・誤字・新語）が多い
  - 「内閣総理大臣」と「首相」
  - 「打ち合わせ」と「打合せ」
  - 「インタフェース」と「インタフェイス」
  - 「税金」と「血税」
  - 「スナナレ」「もしドラ」
  - 「ドコモ」と「DoCoMo」と「ｄｏｃｏｍｏ」

テキストデータの特徴 (2/4)

- 日本語は文法も曖昧
  - クロールで泳いでいる彼女を見た
  - 望遠鏡で泳いでいる彼女を見た
  - プールで泳いでいる彼女を見た
  - 先生とお酒を飲む
  - ビールとお酒を飲む

テキストデータの特徴 (3/4)

- 語の境界に曖昧性がある
  - 「そこではきものをぬげ」
    - そこで/はきもの/を/ぬげ
    - そこでは/きもの/を/ぬげ
- うなぎ文
  - 「ぼくはウナギだ」
- こんにゃく文
  - 「こんにゃくは太らない」

テキストデータの特徴 (4/4)

- 意味の文脈依存性
  - 「学校で遊ぶ」
    - このときの「学校」は場所としての学校
  - 「学校が談話を発表した」
    - このときの「学校」は法人的側面を表す

形態素解析

- 形態素解析は，自然言語で書かれた文章を語（形態素）に分割する処理のことです
- Chasen (奈良先端大)，MeCab (工藤拓氏)，JUMAN (京都大学) が公開しているオープンソースのソフトウェアが有名です
- 新聞記事を対象とした場合の精度は99%以上ですが，話し言葉を対象とすると精度は下がります（それでも十分実用的な精度です）
- 常に新しい言葉が生まれているので，未知語（辞書に載っていない語）問題はなかなかやっかいです

形態素解析の実行例

- 「親譲りの無鉄砲で子供の時から損ばかりしている。」を MeCab にかけた結果です

```
親譲り	名詞, 一般, *, *, *, *, 親譲り, オヤユズリ, オヤユズリ
の	助詞, 連体化, *, *, *, *, の, ノ, ノ
無鉄砲	名詞, 一般, *, *, *, *, 無鉄砲, ムテッポウ, ムテッポー
で	助詞, 格助詞, 一般, *, *, *, で, デ, デ
子供	名詞, 一般, *, *, *, *, 子供, コドモ, コドモ
の	助詞, 連体化, *, *, *, *, の, ノ, ノ
時	名詞, 非自立, 副詞可能, *, *, *, 時, トキ, トキ
から	助詞, 格助詞, 一般, *, *, *, から, カラ, カラ
損	名詞, 一般, *, *, *, *, 損, ソン, ソン
ばかり	助詞, 副助詞, *, *, *, *, ばかり, バカリ, バカリ
し	動詞, 自立, *, *, サ変・スル, 連用形, する, シ, シ
て	助詞, 接続助詞, *, *, *, *, て, テ, テ
いる	動詞, 非自立, *, *, 一段, 基本形, いる, イル, イル
。	記号, 句点, *, *, *, *, 。, 。, 。
```

機能語と内容語

- 語は，助詞や助動詞といった「機能語」と，名詞，形容詞，動詞，副詞といった「内容語」に大別できます
- 機能語は，それ単体では意味を持たない語なので，文章の内容を理解する際の助けにはなりません
- 内容語は，名称，性質，動作，状況など，文章の内容の一部を表しているので，内容を理解する際の助けになります．しかし，名詞と結びつかないと意味が特定できない場合が多いです
- したがって，名詞は必須で，分析の目的に応じて形容詞，副詞，動詞を用いることが多いです

未知語について

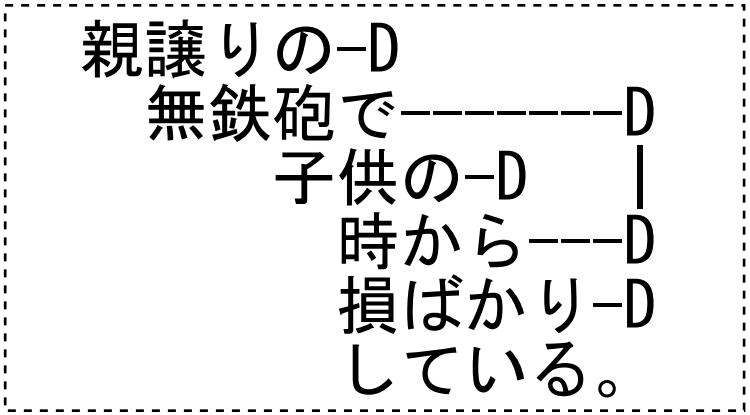
- 形態素解析器の辞書に登録されていない語は「未知語」として出力されます
- 未知語の品詞推定は研究レベルでは実装されていますが，まだ実用レベルには達していません
- 未知語は，単なるゴミであることも多いのですが，世の中の流行を反映した「新しい語」（例えば「婚活」や「H1N1」など）が含まれていることもあるので油断なりません
- なので，取り敢えず未知語は分析対象に加えて，不便があれば臨機応変に対応することが多いです

構文解析

- 構文解析は，語と語の係り受け関係を分析する処理のことです
- CaboCha (工藤拓氏)，KNP (京都大学) が公開しているオープンソースのソフトウェアが有名です
- 新聞記事を対象とした場合でも精度は80%くらいですが，確からしい結果だけを利用すれば十分使えます
- 特定の語と関係する語（例えば，「美味しい」の係り受け先など）を見たいときなど，用途を限定した場合にも十分使えます

構文解析の実行結果

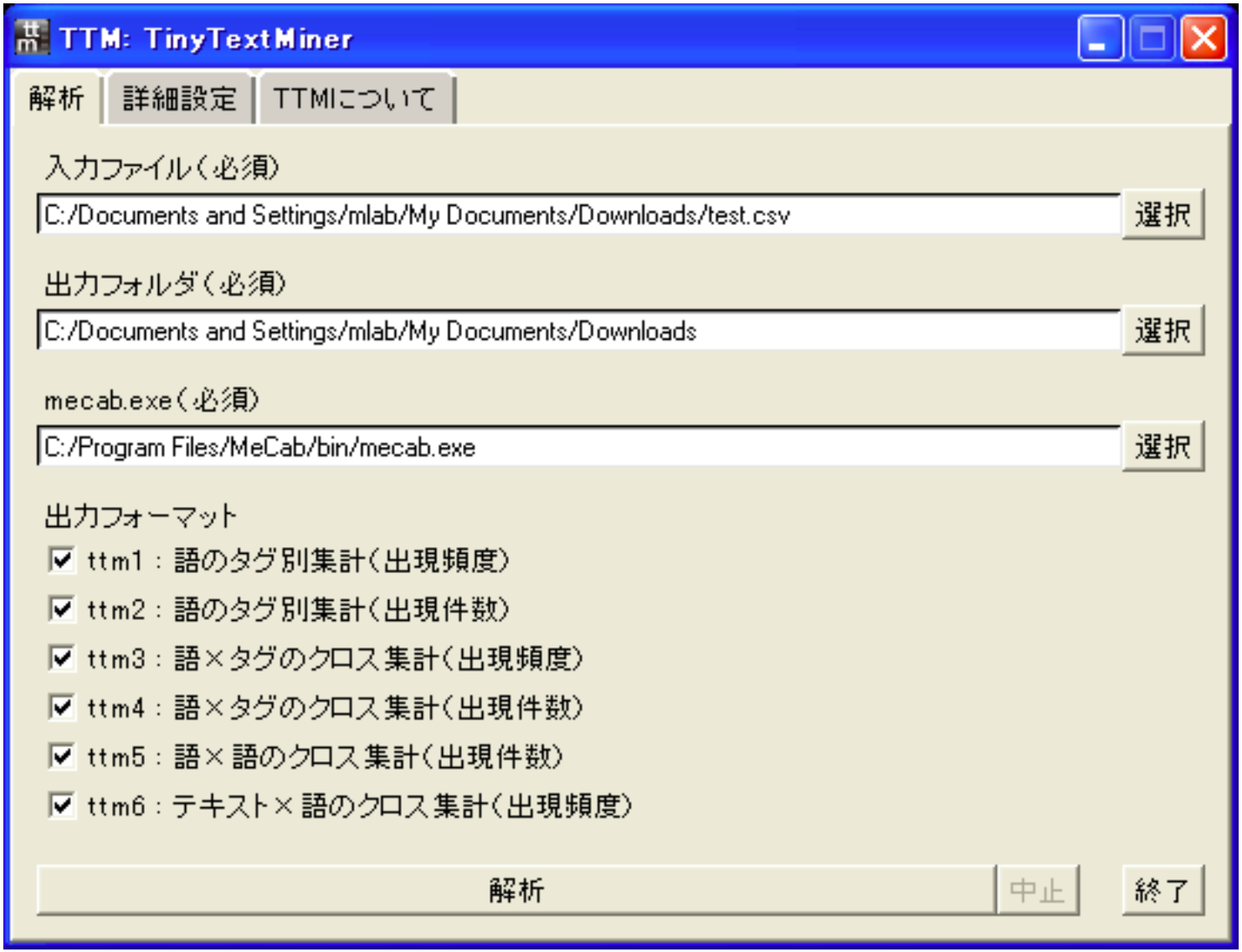
- 「親譲りの無鉄砲で子供の時から損ばかりしている。」を CaboCha にかけた結果です



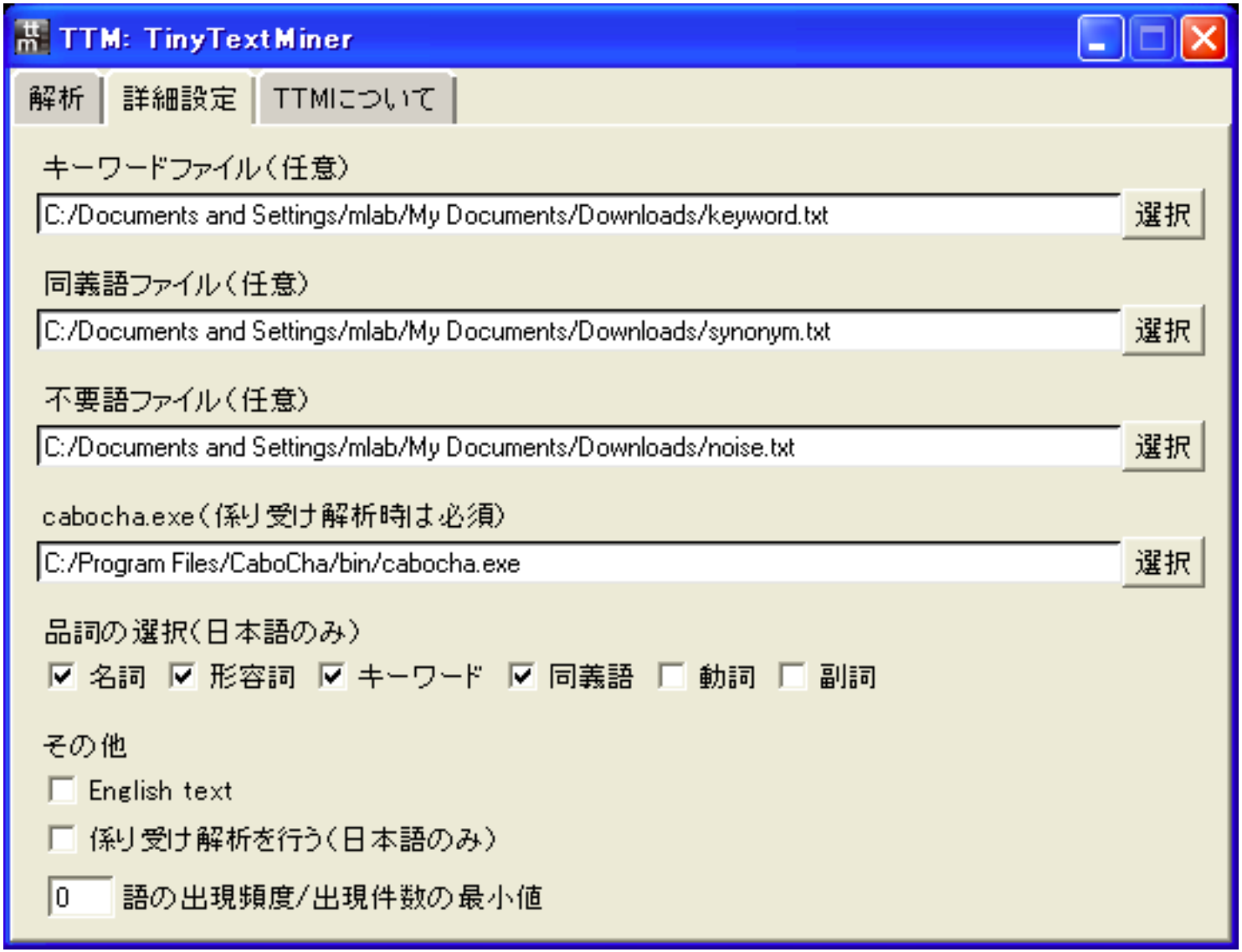
の紹介

- 「形態素解析は分かった，構文解析も分かった，それで，どうすればいいの？」という皆さんの心の声に答えるために を作りました
- はテキストデータを形態素解析器，構文解析器にかけて，その分析結果を読み込んで集計し，CSVファイルを出力するフリーウェアです
- はテキストマイニングの前処理に特化していますので，ここまでしか行いません．後処理は，みなさんの使い慣れたソフトウェアに読み込ませて，好きなように分析してもらいたいと思っています

のスクリーンショット (1/2)

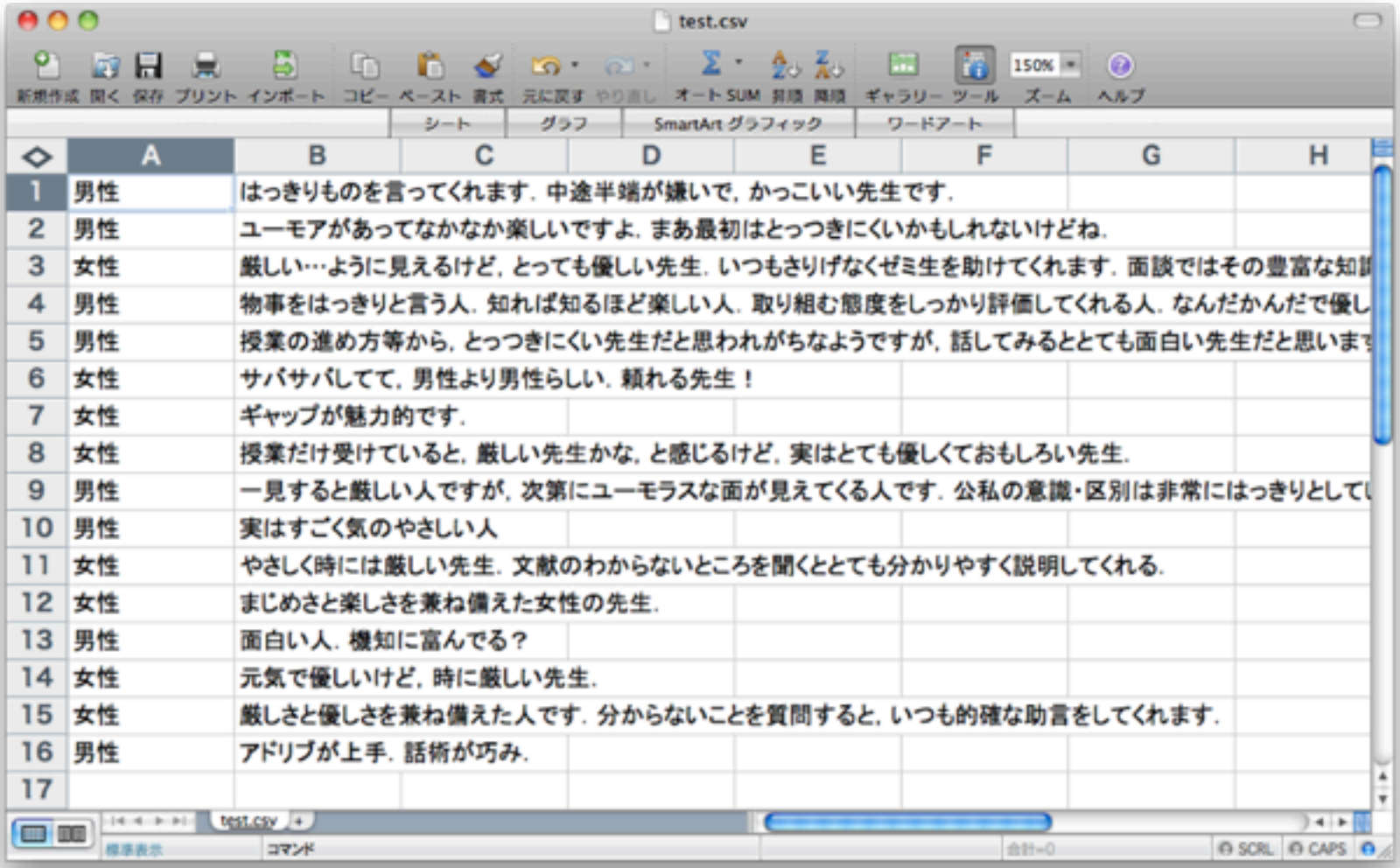


のスクリーンショット (2/2)



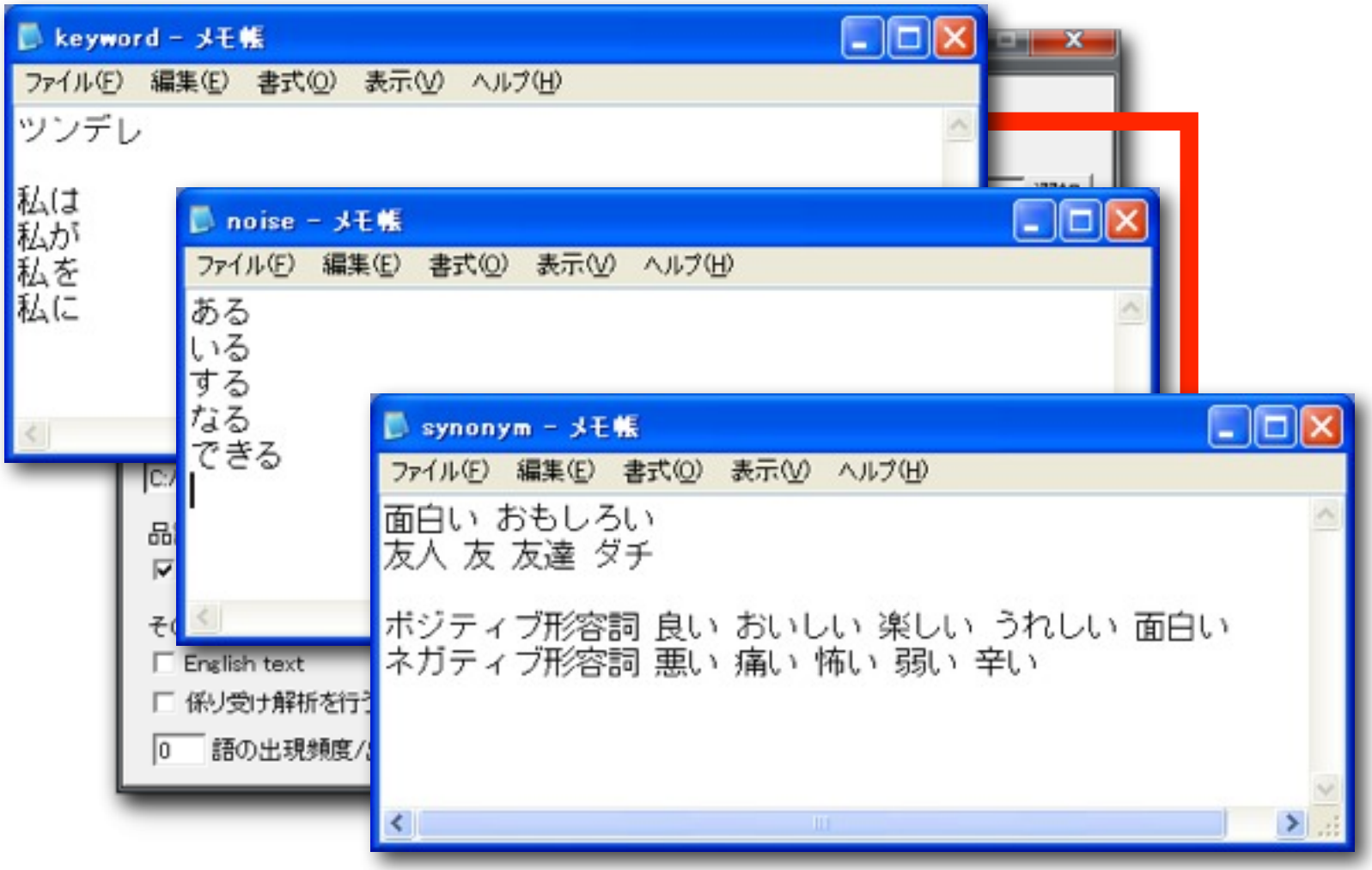
ttm の特徴 (1/4)

- CSV 形式の「タグ付きテキスト」を読み込みます



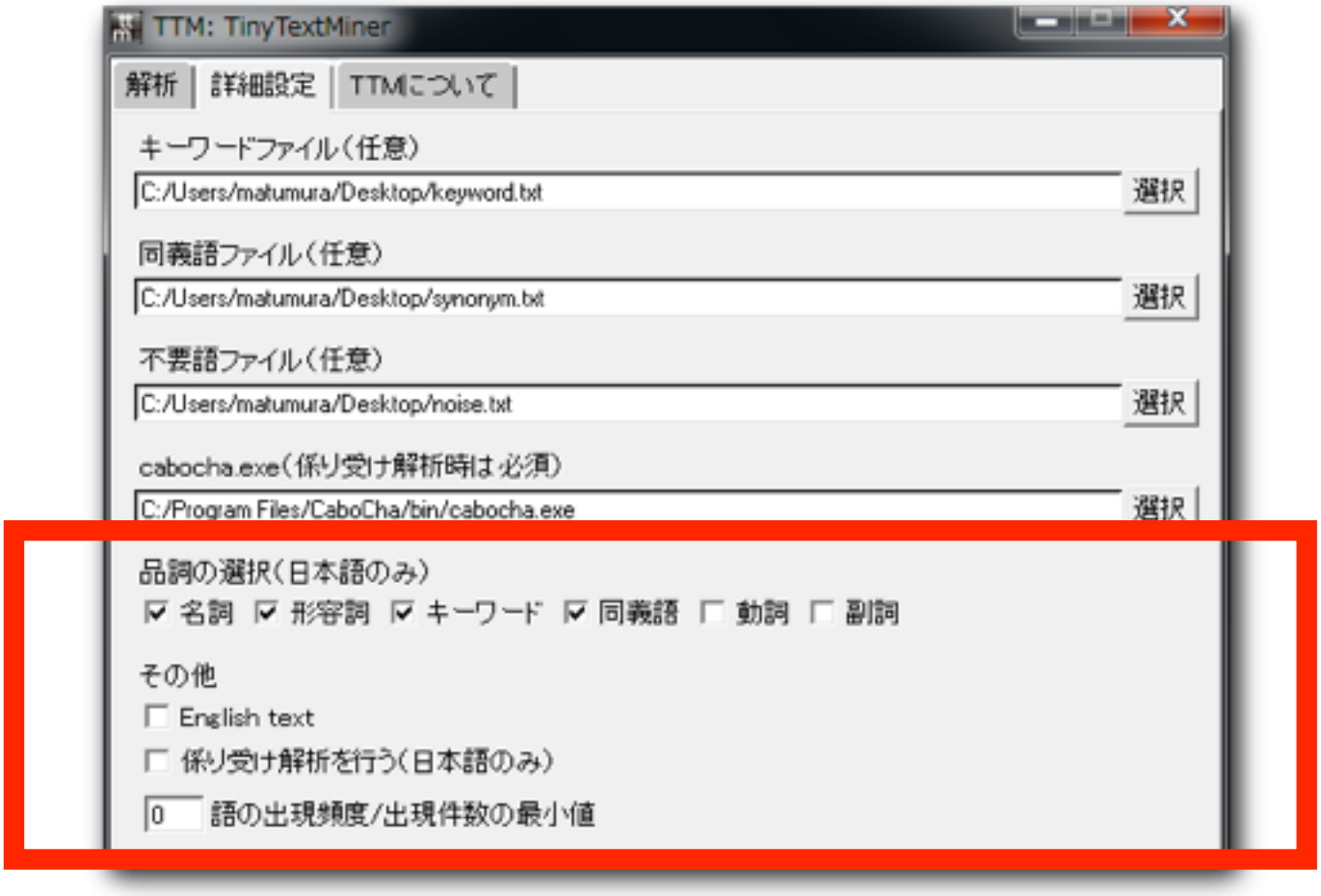
の特徴 (2/4)

- キーワード，同義語，不要語を指定できます



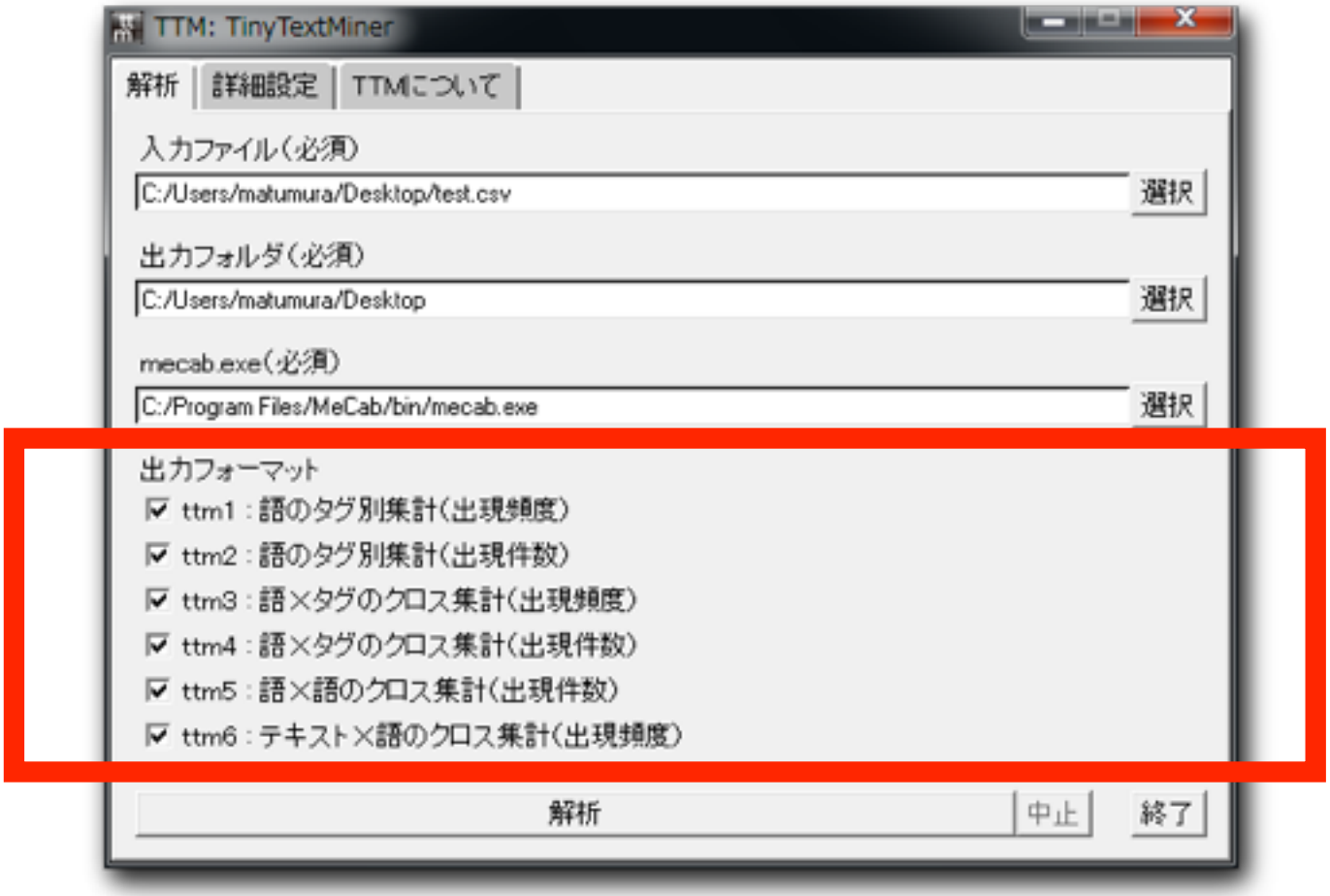
の特徴 (3/4)

- 品詞，係り受け解析，閾値，英文の設定もできます



の特徴 (4/4)

- 6種類の出力ファイルを提供します



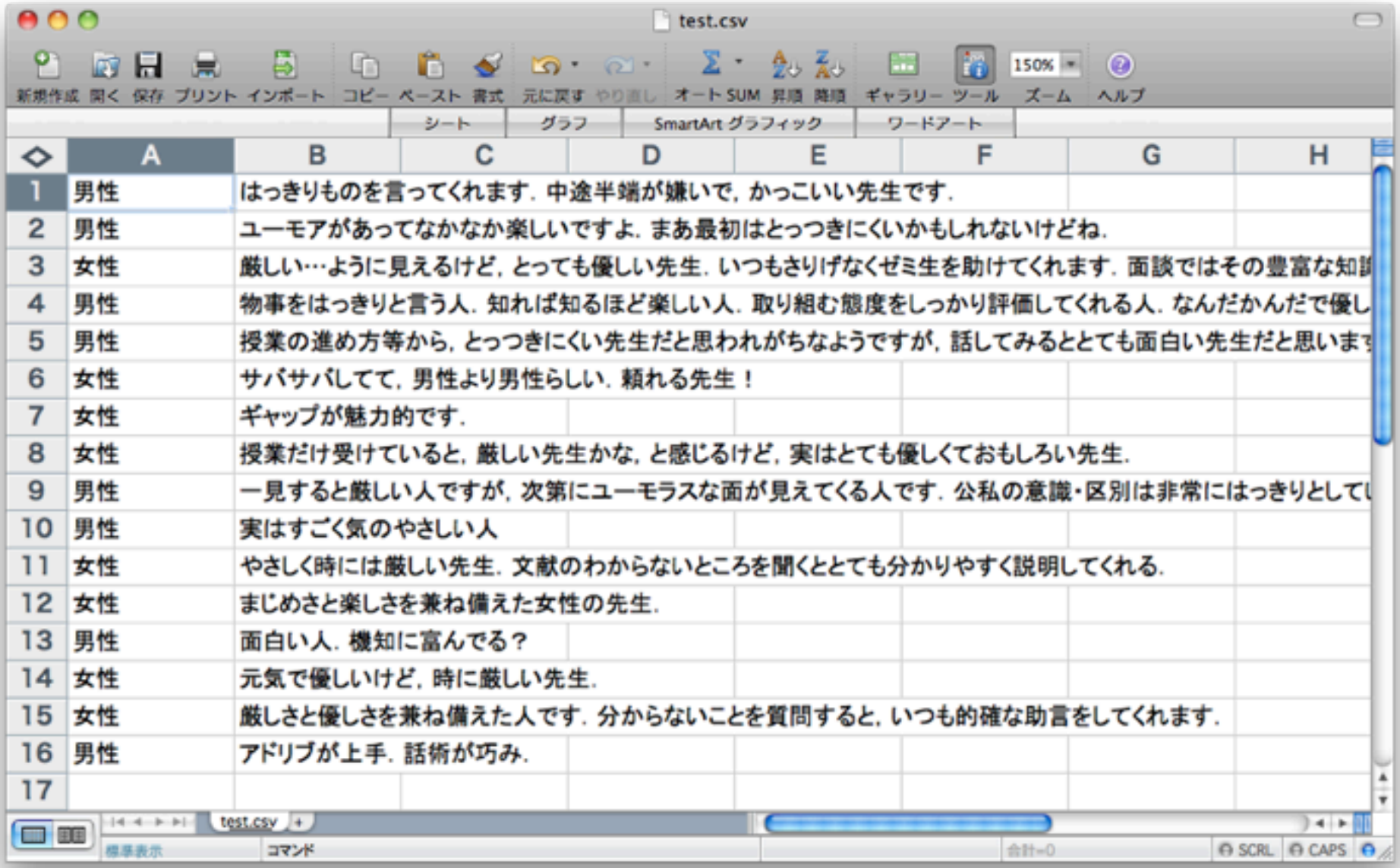
 のインストール

- Windows版とMac OSX版（10.5以降）があります
- http://mtmr.jp/ttm/ から ttm.exe をダウンロードするだけで  本体のインストールは終わりです
- 形態素解析を実行するためのソフトウェア MeCabを別途インストールする必要があります（構文解析を行うときは CaboCha もインストールします）．詳細は http://mtmr.jp/ttm/ をご覧ください
- Mac OSX版はOS内蔵のMeCabを使いますのでMeCabを別途インストールする必要はありません

サンプルデータ test.csv

- 三浦麻子先生のゼミに所属する16名（男女8名ずつ）の大学生が「三浦先生ってどんな人？」という質問に対して自由に記述した文章
  （http://mtmr.jp/ttm/test.csv からDLできます）

test.csv の内容



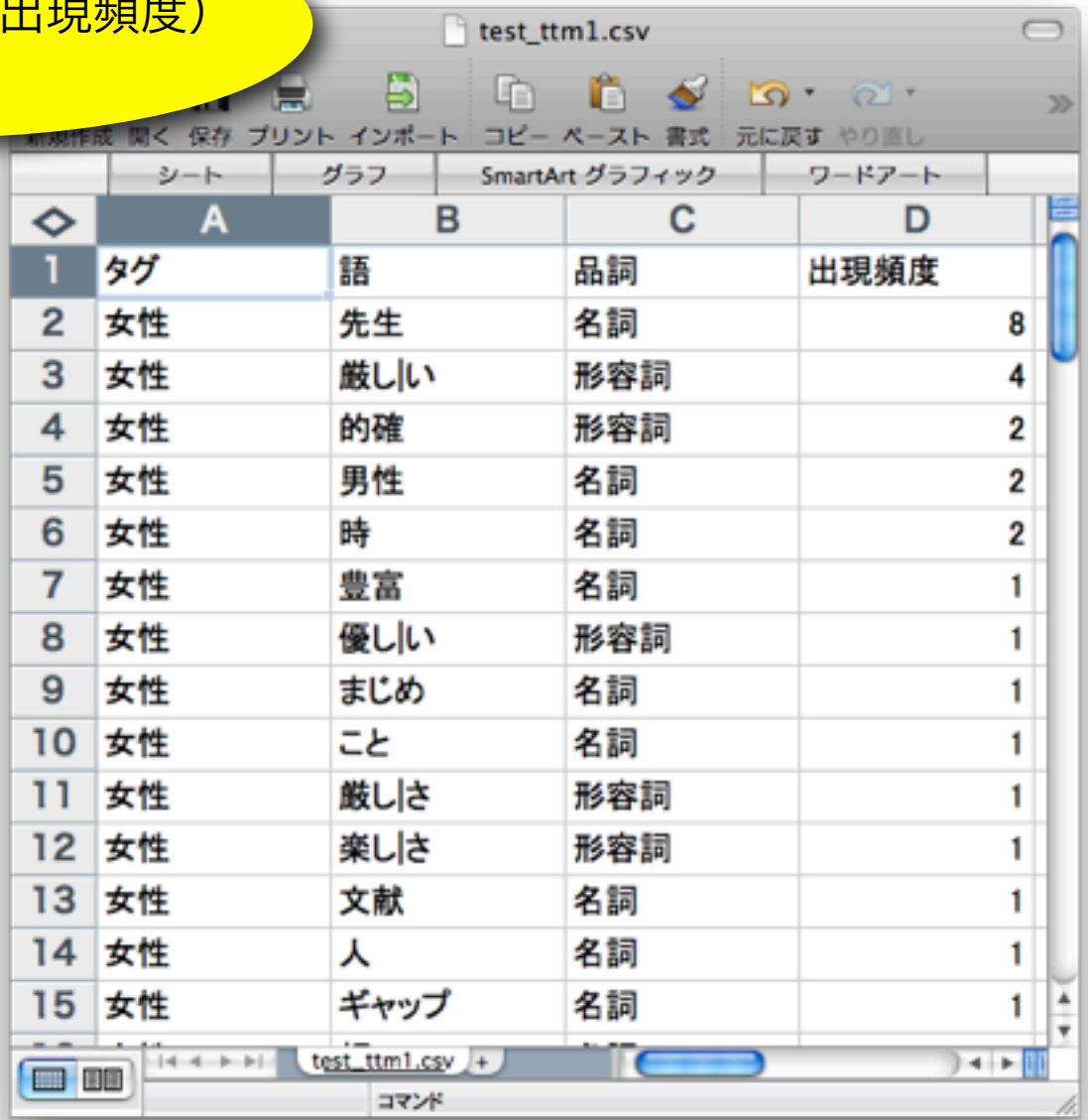
	A	B
1	男性	はっきりものを言ってくれます．中途半端が嫌いで，かっこいい先生です．
2	男性	ユーモアがあってなかなか楽しいですよ．まあ最初はとっつきにくいかもしれないけどね．
3	女性	厳しい…ように見えるけど，とっても優しい先生．いつもさりげなくゼミ生を助けてくれます．面談ではその豊富な知識
4	男性	物事をはっきりと言う人．知れば知るほど楽しい人．取り組む態度をしっかり評価してくれる人．なんだかんだで優し
5	男性	授業の進め方等から，とっつきにくい先生だと思われがちなようですが，話してみるととても面白い先生だと思いま
6	女性	サバサバしてて，男性より男性らしい．頼れる先生！
7	女性	ギャップが魅力的です．
8	女性	授業だけ受けていると，厳しい先生かな，と感じるけど，実はとても優しくておもしろい先生．
9	男性	一見すると厳しい人ですが，次第にユーモラスな面が見えてくる人です．公私の意識・区別は非常にはっきりとしてい
10	男性	実はすごく気のやさしい人
11	女性	やさしく時には厳しい先生．文献のわからないところを聞くととても分かりやすく説明してくれる．
12	女性	まじめさと楽しさを兼ね備えた女性の先生．
13	男性	面白い人．機知に富んでる？
14	女性	元気で優しいけど，時に厳しい先生．
15	女性	厳しさと優しさを兼ね備えた人です．分からないことを質問すると，いつも的確な助言をしてくれます．
16	男性	アドリブが上手．話術が巧み．
17		

分析条件

- 次の条件で  で test.csv を分析してみましょう
  - 出力する品詞は「名詞，形容詞」
  - オプションファイルは「設定せず」
  - 出現頻度/出現件数の最小値は「0」

test_ttm1.csv

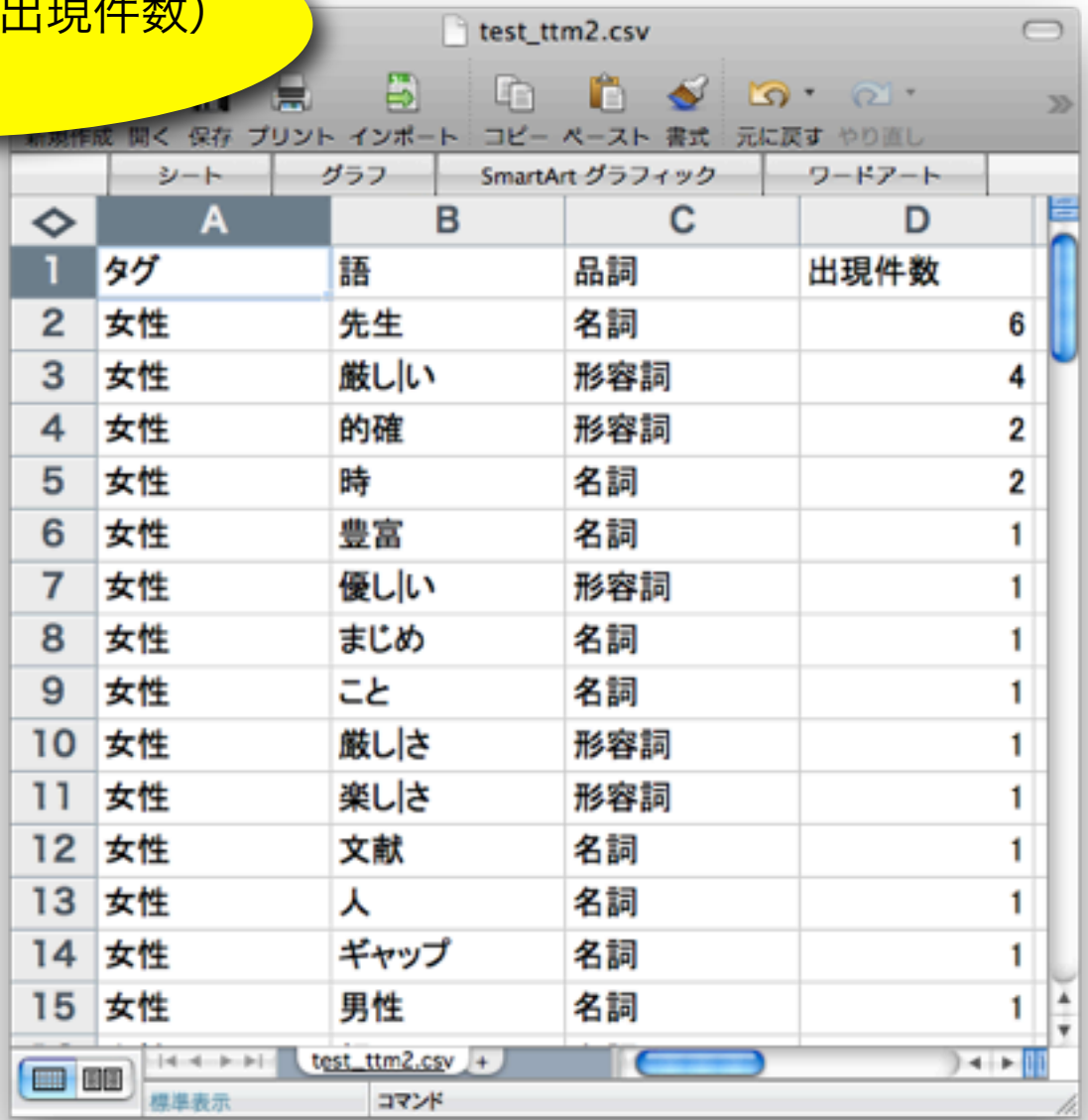
語のタグ別集計（出現頻度）



	A	B	C	D
1	タグ	語	品詞	出現頻度
2	女性	先生	名詞	8
3	女性	厳し|い	形容詞	4
4	女性	的確	形容詞	2
5	女性	男性	名詞	2
6	女性	時	名詞	2
7	女性	豊富	名詞	1
8	女性	優し|い	形容詞	1
9	女性	まじめ	名詞	1
10	女性	こと	名詞	1
11	女性	厳し|さ	形容詞	1
12	女性	楽し|さ	形容詞	1
13	女性	文献	名詞	1
14	女性	人	名詞	1
15	女性	ギャップ	名詞	1

test_ttm2.csv

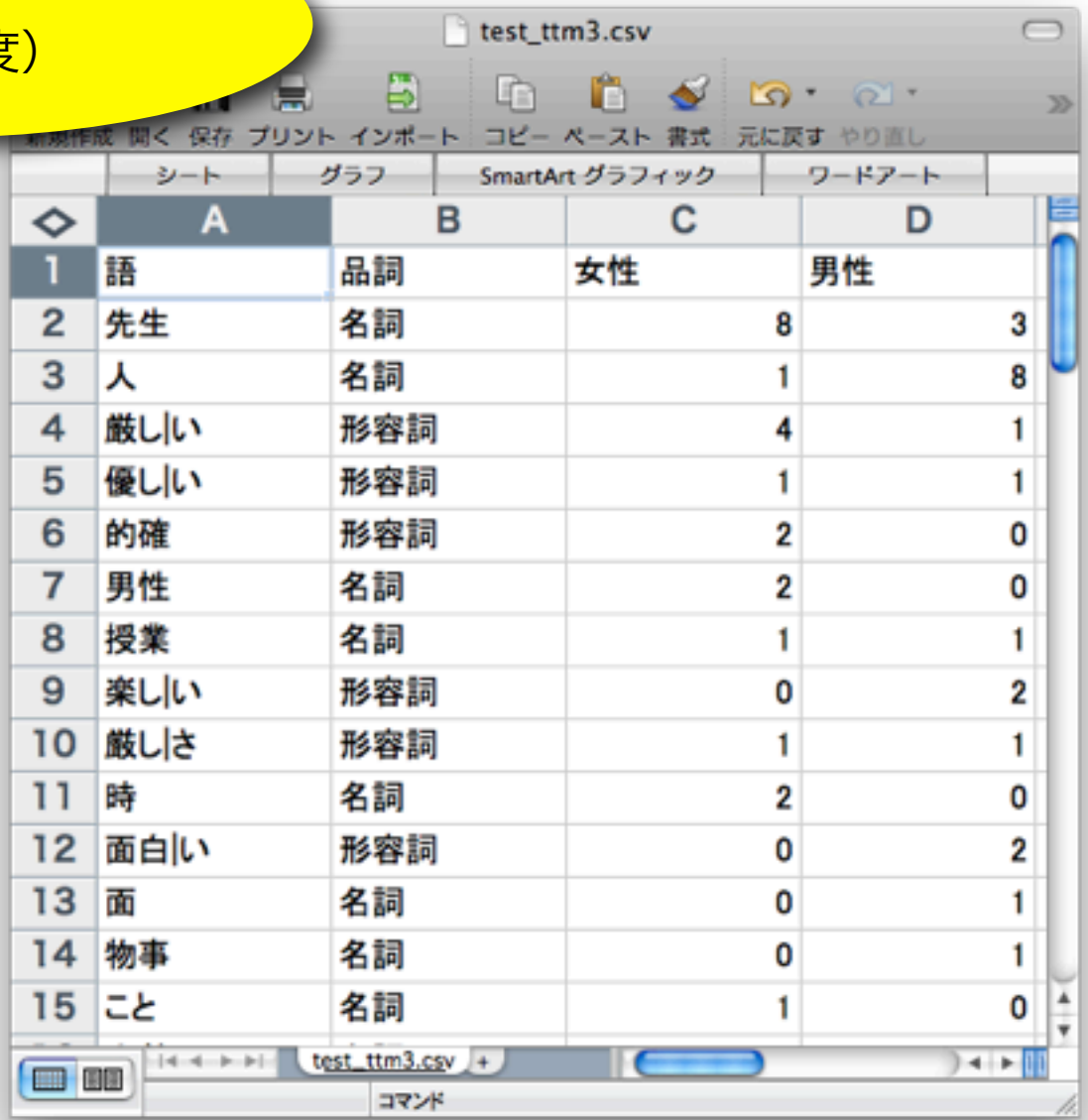
語のタグ別集計（出現件数）



	A	B	C	D
1	タグ	語	品詞	出現件数
2	女性	先生	名詞	6
3	女性	厳し|い	形容詞	4
4	女性	的確	形容詞	2
5	女性	時	名詞	2
6	女性	豊富	名詞	1
7	女性	優し|い	形容詞	1
8	女性	まじめ	名詞	1
9	女性	こと	名詞	1
10	女性	厳し|さ	形容詞	1
11	女性	楽し|さ	形容詞	1
12	女性	文献	名詞	1
13	女性	人	名詞	1
14	女性	ギャップ	名詞	1
15	女性	男性	名詞	1

test_ttm3.csv

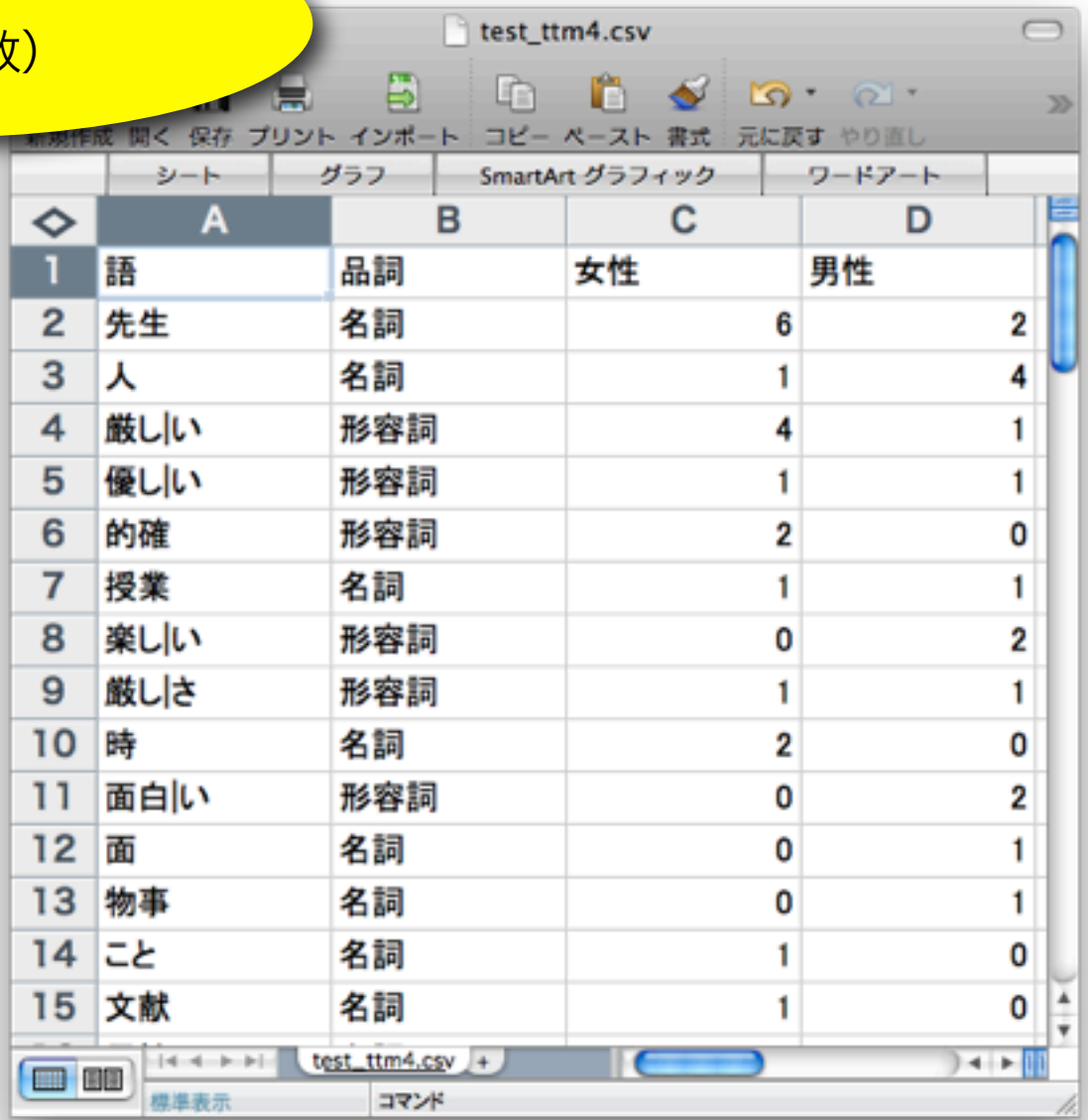
語×タグのクロス集計
（出現頻度）



	A	B	C	D
1	語	品詞	女性	男性
2	先生	名詞	8	3
3	人	名詞	1	8
4	厳し|い	形容詞	4	1
5	優し|い	形容詞	1	1
6	的確	形容詞	2	0
7	男性	名詞	2	0
8	授業	名詞	1	1
9	楽し|い	形容詞	0	2
10	厳し|さ	形容詞	1	1
11	時	名詞	2	0
12	面白|い	形容詞	0	2
13	面	名詞	0	1
14	物事	名詞	0	1
15	こと	名詞	1	0

test_ttm4.csv

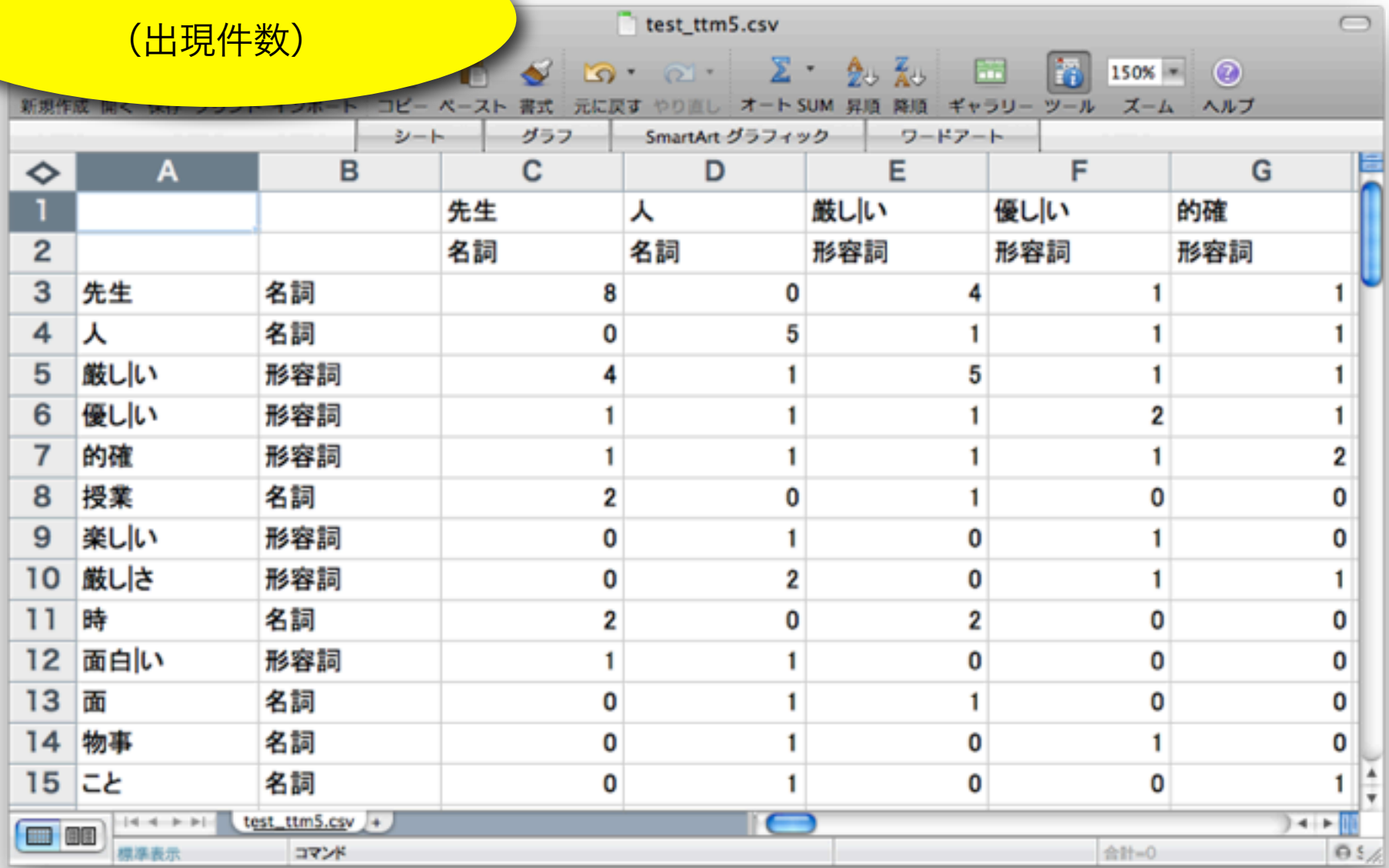
語×タグのクロス集計
（出現件数）



	A	B	C	D
1	語	品詞	女性	男性
2	先生	名詞	6	2
3	人	名詞	1	4
4	厳し|い	形容詞	4	1
5	優し|い	形容詞	1	1
6	的確	形容詞	2	0
7	授業	名詞	1	1
8	楽し|い	形容詞	0	2
9	厳し|さ	形容詞	1	1
10	時	名詞	2	0
11	面白|い	形容詞	0	2
12	面	名詞	0	1
13	物事	名詞	0	1
14	こと	名詞	1	0
15	文献	名詞	1	0

test_ttm5.csv

語×語のクロス集計
（出現件数）

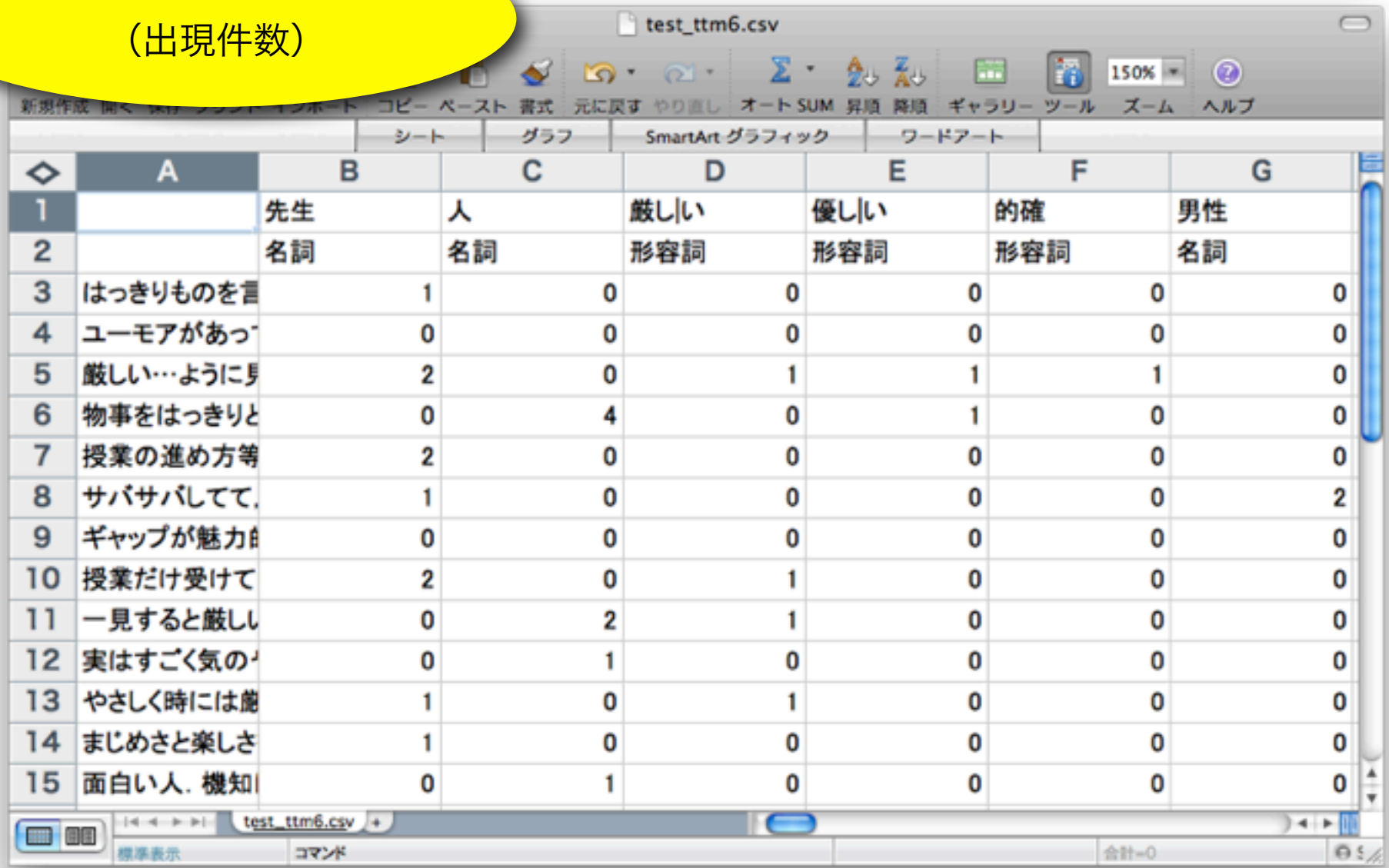


test_ttm5.csv

	A	B	C	D	E	F	G
1			先生	人	厳し|い	優し|い	的確
2			名詞	名詞	形容詞	形容詞	形容詞
3	先生	名詞	8	0	4	1	1
4	人	名詞	0	5	1	1	1
5	厳し|い	形容詞	4	1	5	1	1
6	優し|い	形容詞	1	1	1	2	1
7	的確	形容詞	1	1	1	1	2
8	授業	名詞	2	0	1	0	0
9	楽し|い	形容詞	0	1	0	1	0
10	厳し|さ	形容詞	0	2	0	1	1
11	時	名詞	2	0	2	0	0
12	面白|い	形容詞	1	1	0	0	0
13	面	名詞	0	1	1	0	0
14	物事	名詞	0	1	0	1	0
15	こと	名詞	0	1	0	0	1

test_ttm6.csv

テキスト×語のクロス集計
（出現件数）



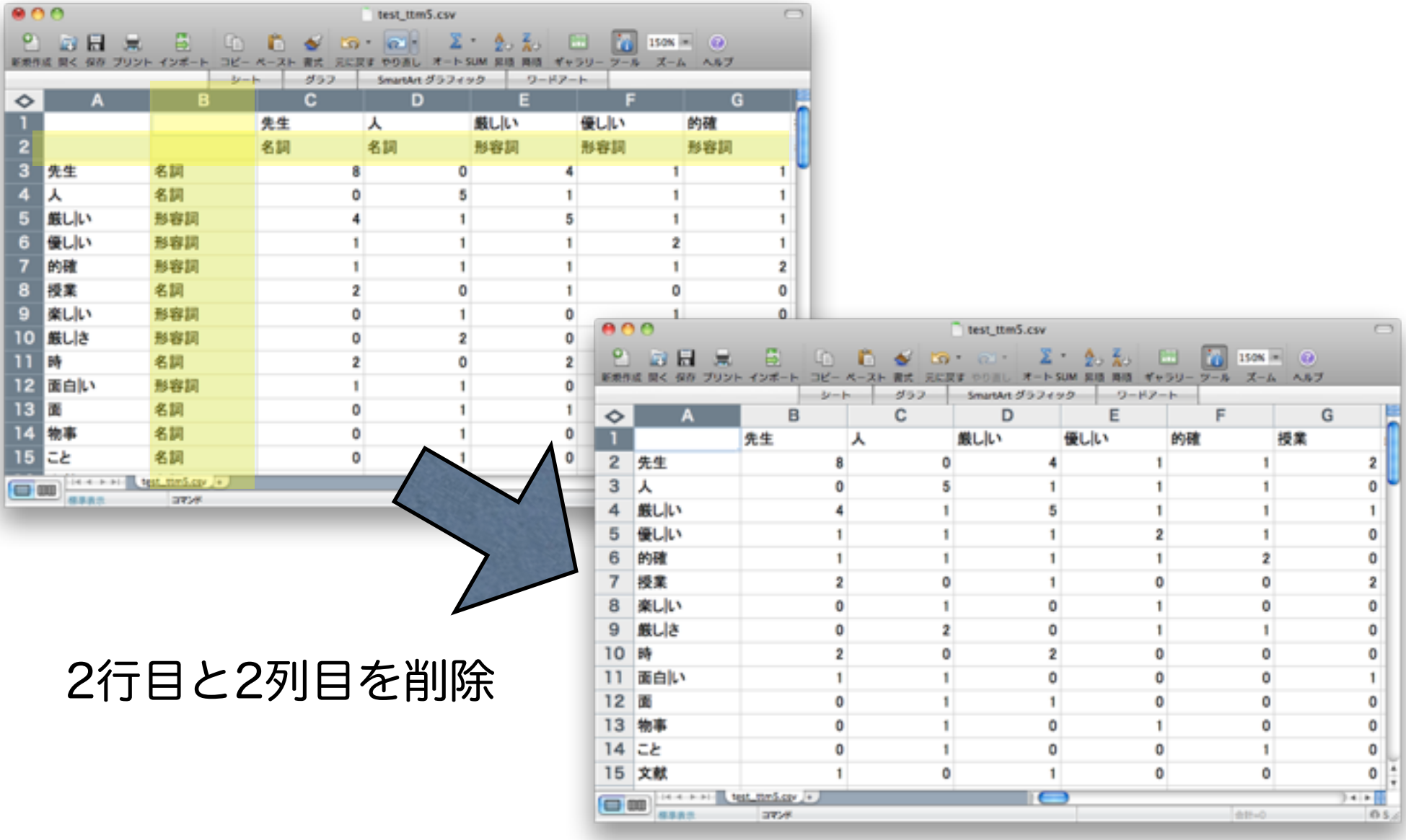
test_ttm6.csv

	A	B	C	D	E	F	G
1		先生	人	厳しい	優しい	的確	男性
2		名詞	名詞	形容詞	形容詞	形容詞	名詞
3	はっきりものを言	1	0	0	0	0	0
4	ユーモアがあっ	0	0	0	0	0	0
5	厳しい…ように見	2	0	1	1	1	0
6	物事をはっきりと	0	4	0	1	0	0
7	授業の進め方等	2	0	0	0	0	0
8	サバサバしてて、	1	0	0	0	0	2
9	ギャップが魅力的	0	0	0	0	0	0
10	授業だけ受けて	2	0	1	0	0	0
11	一見すると厳しい	0	2	1	0	0	0
12	実はすごく気の	0	1	0	0	0	0
13	やさしく時には厳	1	0	1	0	0	0
14	まじめさと楽しさ	1	0	0	0	0	0
15	面白い人．機知	0	1	0	0	0	0

を使ってみます

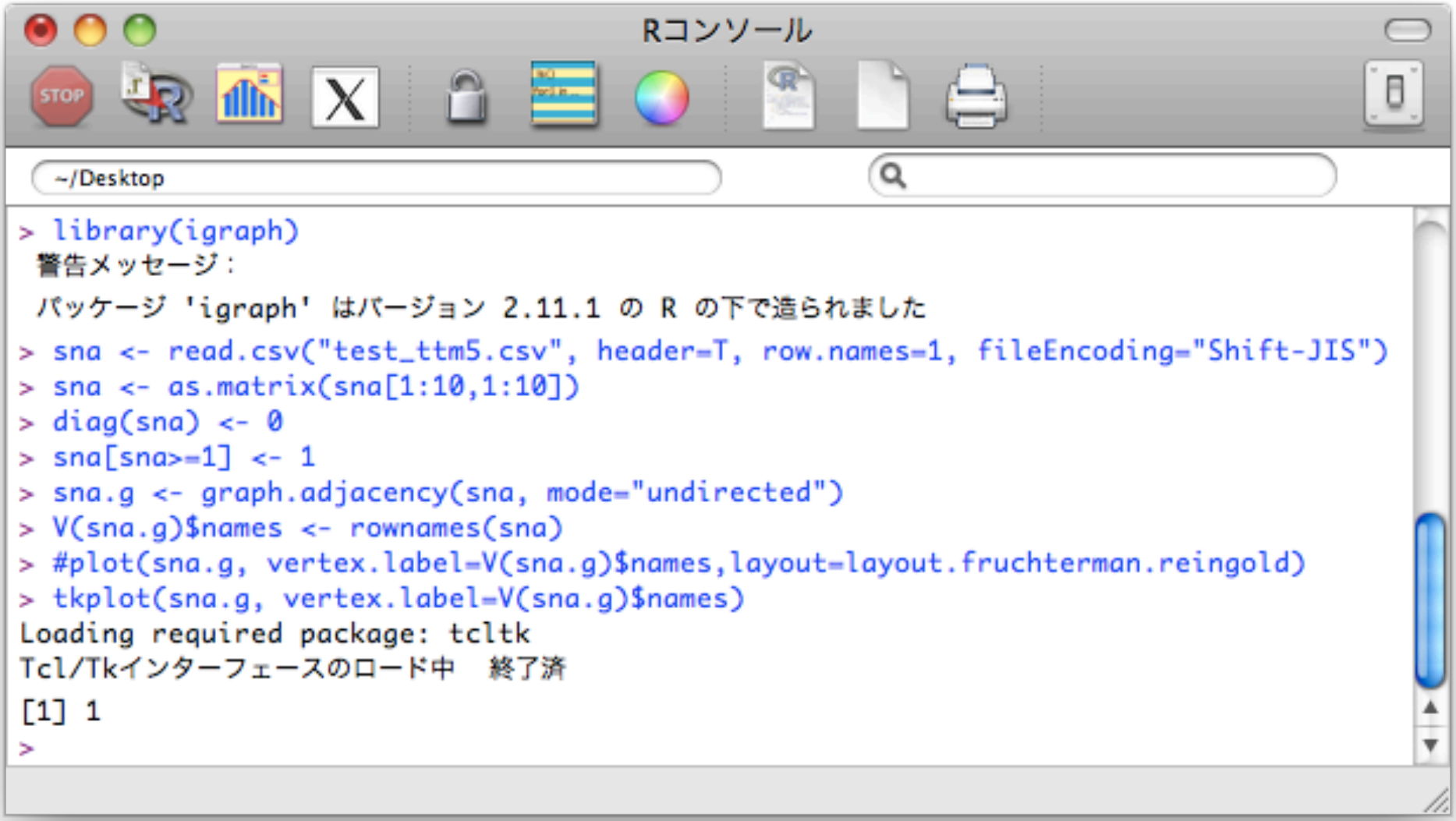
- 統計解析のフリーソフトウェアです
- いろんな人が便利な関数やパッケージを公開している(2009年6月17日現在，CRANには1849ものパッケージが登録されています）ので，コレ一つで大抵のことはできます
- 形態素解析や構文解析を行うパッケージもあります
- データマイニングのパッケージもあります
- たくさんの書籍が出版されているので，マニュアルも充実しています

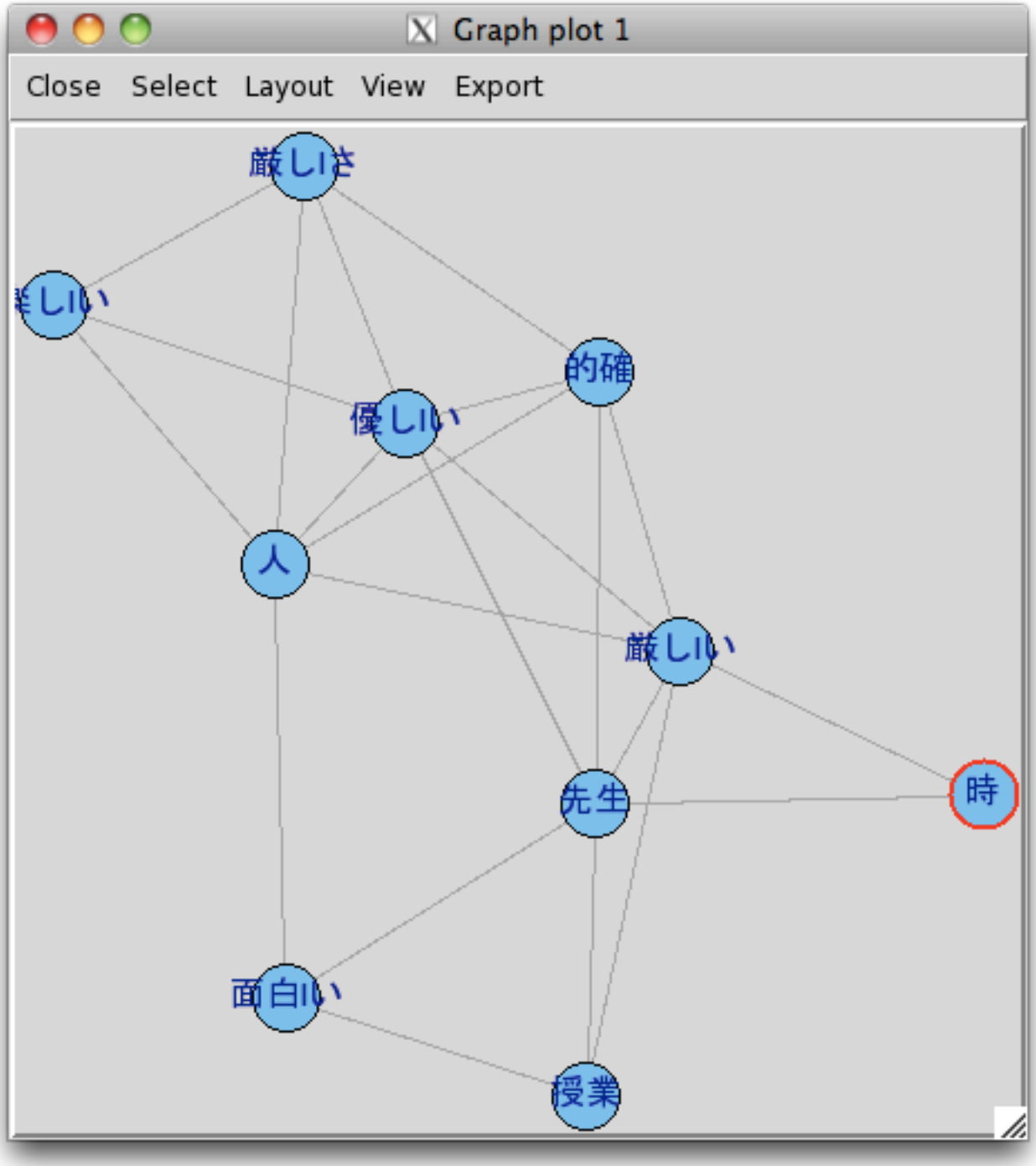
test_ttm5.csvを編集



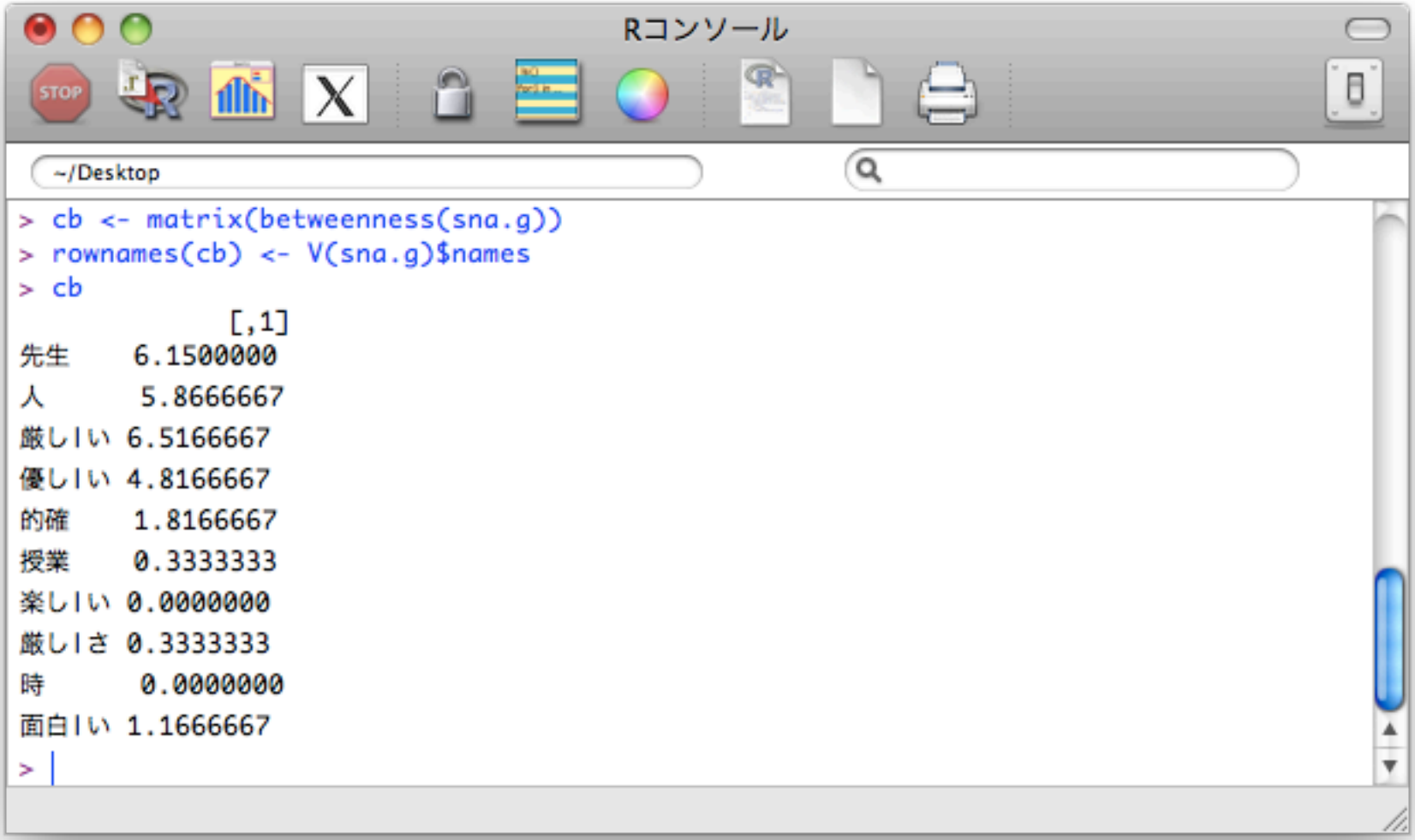
2行目と2列目を削除

共起グラフの描画


Rコンソール
~/Desktop
> library(igraph)
警告メッセージ:
パッケージ 'igraph' はバージョン 2.11.1 の R の下で造られました
> sna <- read.csv("test_ttm5.csv", header=T, row.names=1, fileEncoding="Shift-JIS")
> sna <- as.matrix(sna[1:10,1:10])
> diag(sna) <- 0
> sna[sna>=1] <- 1
> sna.g <- graph.adjacency(sna, mode="undirected")
> V(sna.g)$names <- rownames(sna)
> #plot(sna.g, vertex.label=V(sna.g)$names,layout=layout.fruchterman.reingold)
> tkplot(sna.g, vertex.label=V(sna.g)$names)
Loading required package: tcltk
Tcl/Tkインターフェースのロード中　終了済
[1] 1
>

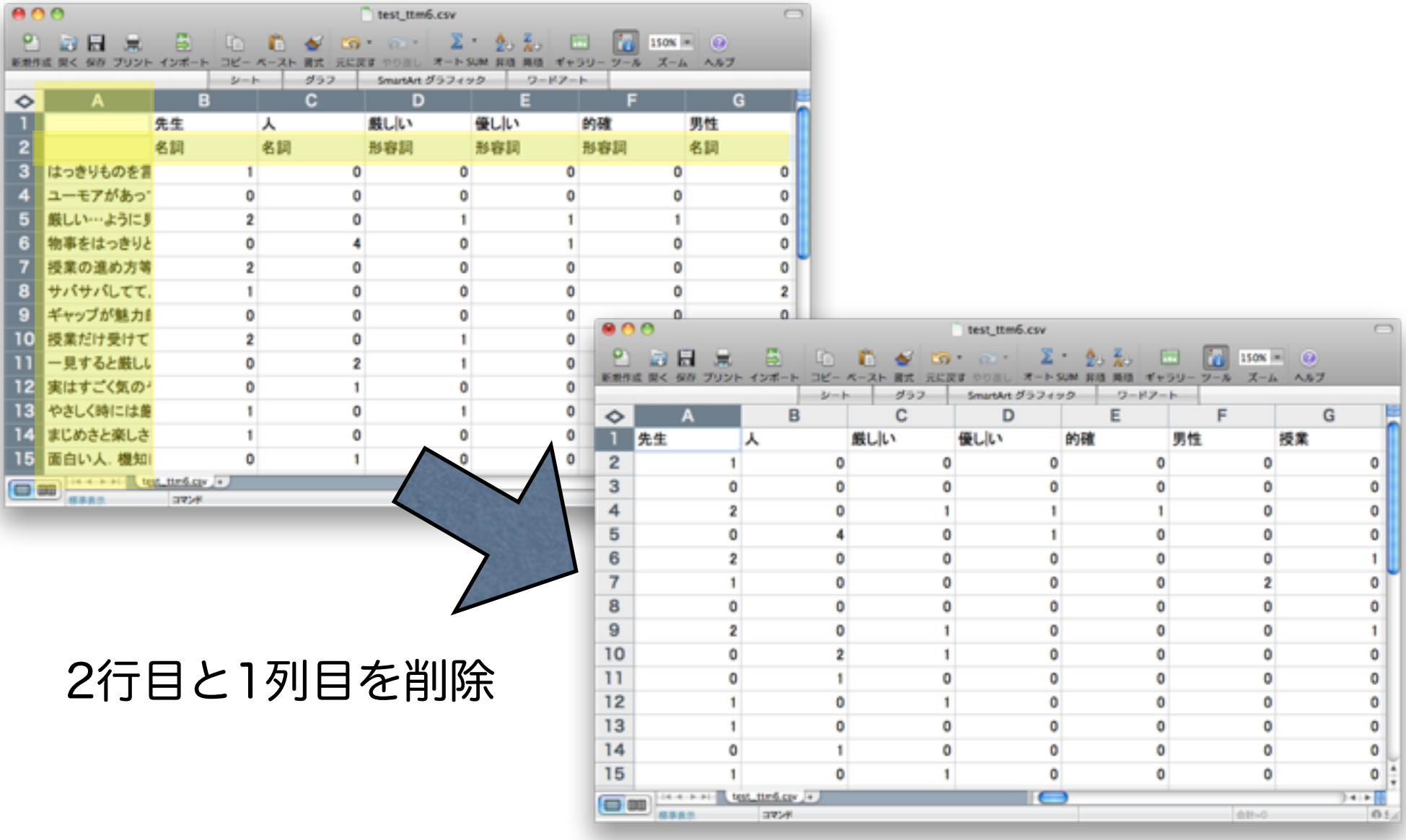

Graph plot 1
Close
Select
Layout
View
Export
厳しい
的確
優しい
人
厳しい
先生
時
面白い
授業

キーワード

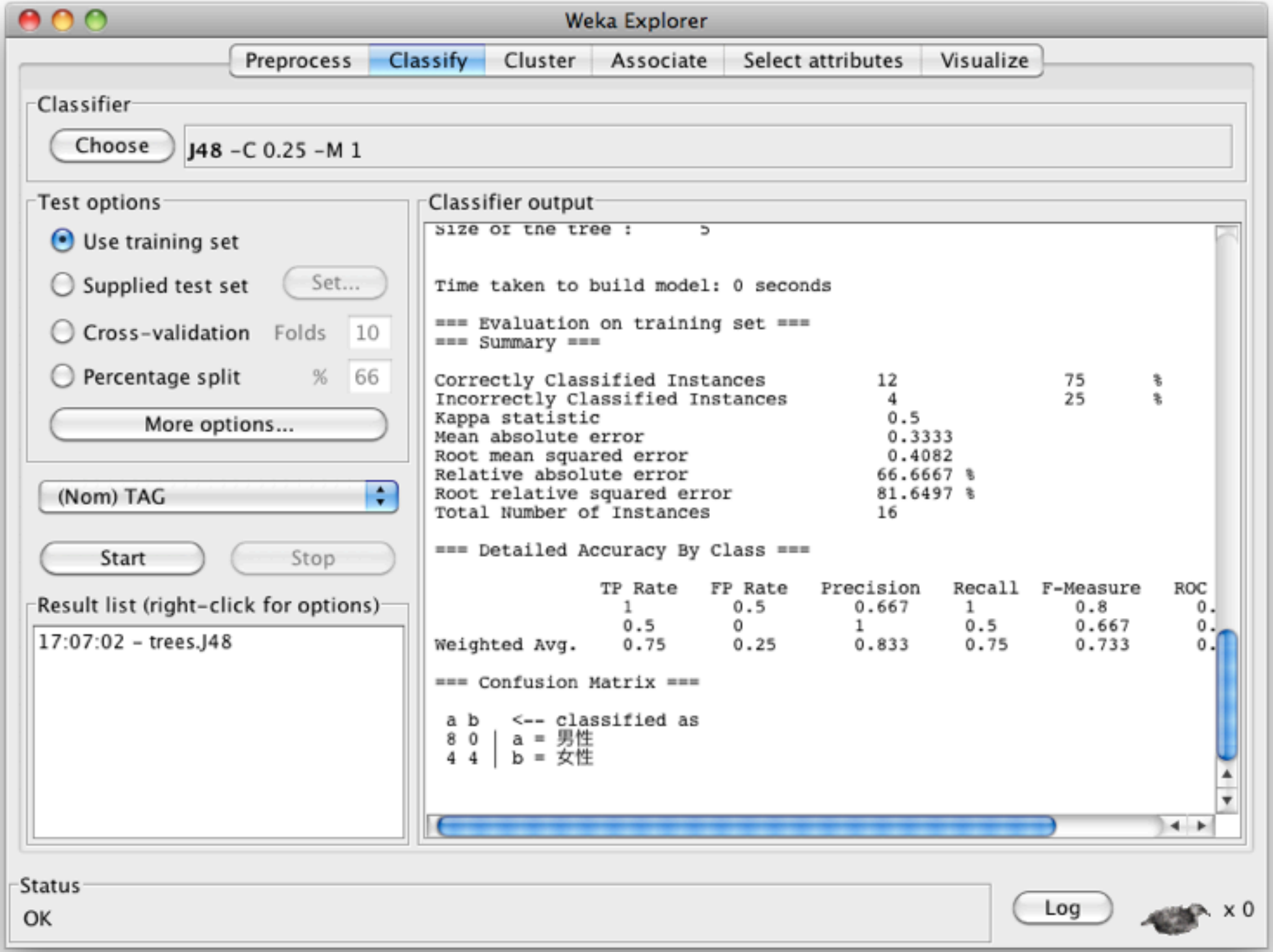

Rコンソール
~/Desktop
> cb <- matrix(betweenness(sna.g))
> rownames(cb) <- V(sna.g)$names
> cb
[,1]
先生 6.1500000
人 5.8666667
厳しい 6.5166667
優しい 4.8166667
的確 1.8166667
授業 0.3333333
楽しい 0.0000000
厳しさ 0.3333333
時 0.0000000
面白い 1.1666667
>

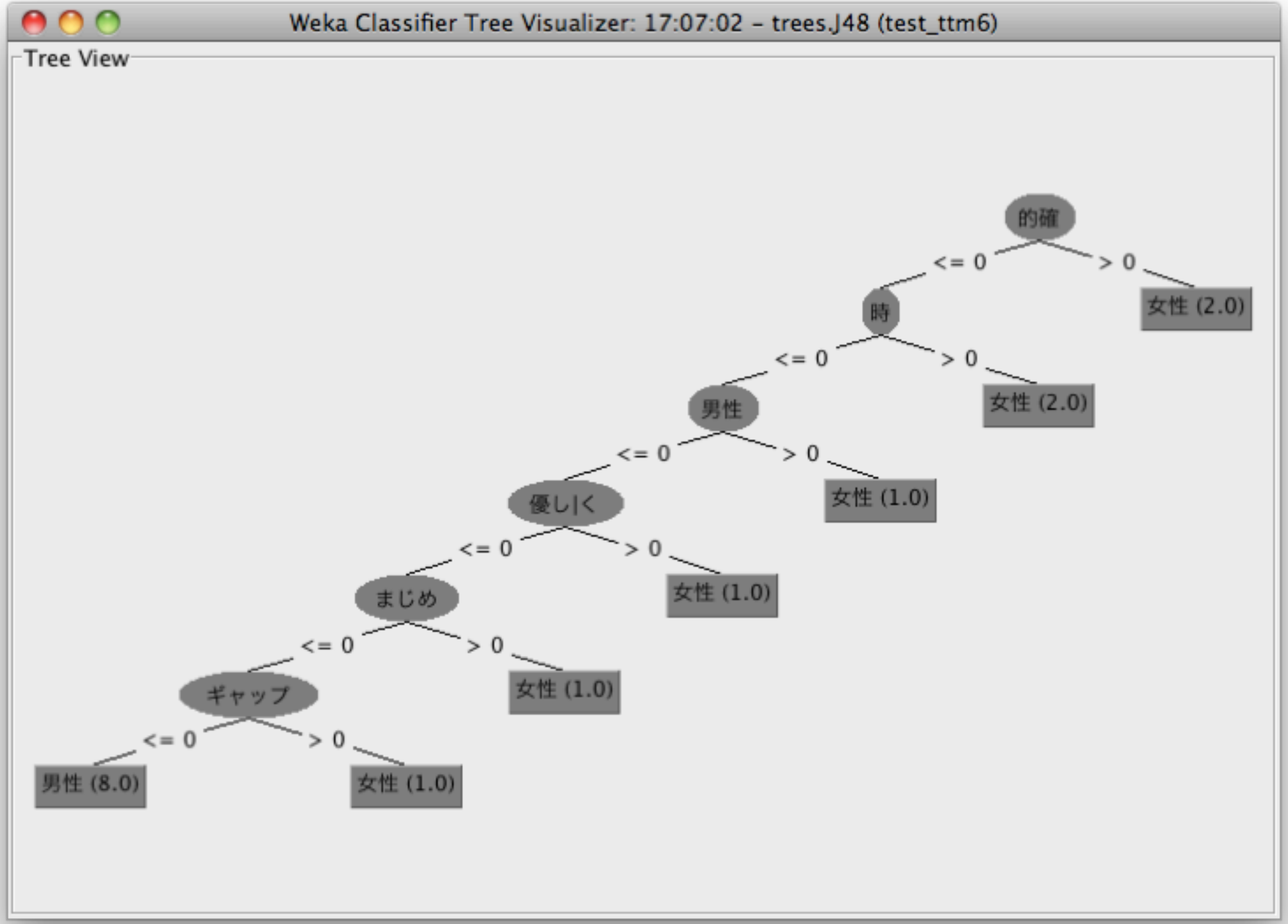
- データマイニングのフリーソフトウェアです
- 代表的なデータマイニングのアルゴリズムはほとんど使えます
- GUIなので操作も簡単
- 最近はKNIMEやRapidMinerといったデータマイニングのフリーソフトウェアもあります

test_ttm6.csvを編集



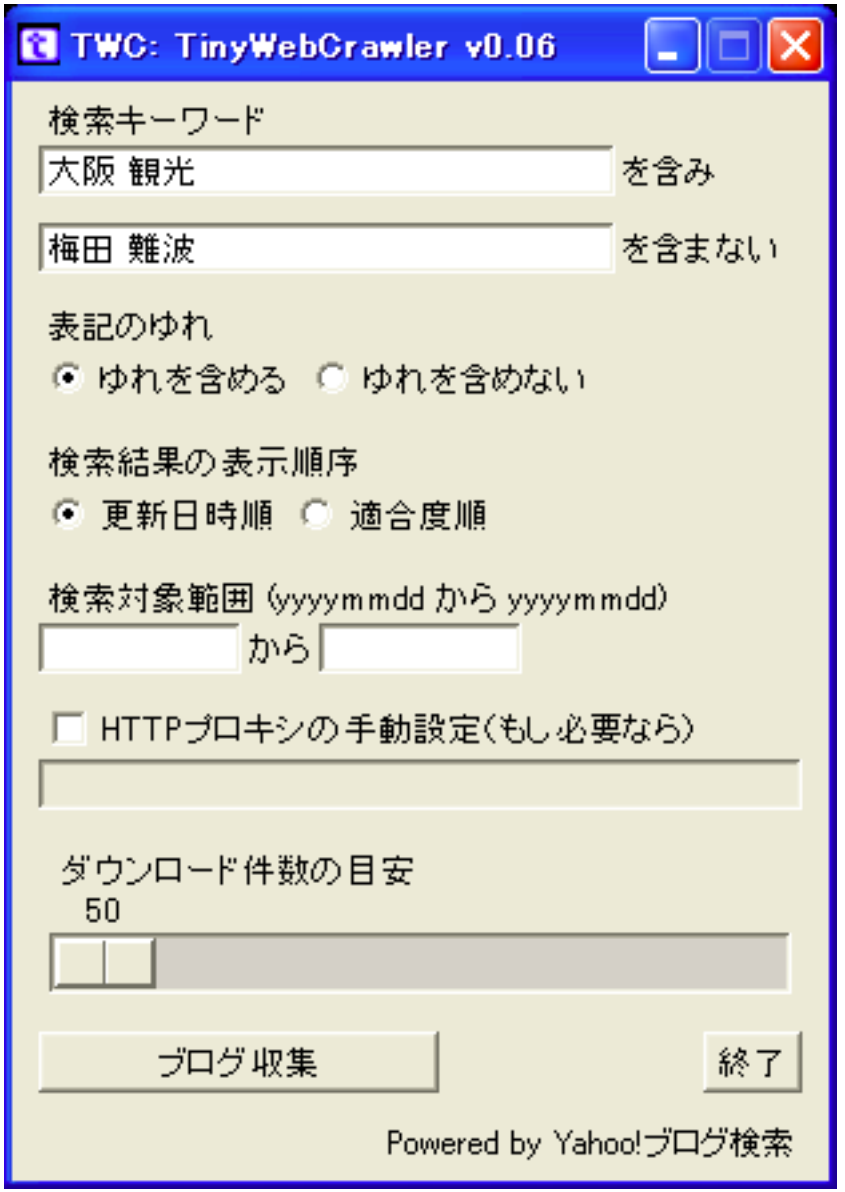
2行目と1列目を削除


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) TAG
Start
Stop
Result list (right-click for options)
17:07:02 - trees.J48
Classifier output
Size of the tree : 5
Time taken to build model: 0 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances 12 75 %
Incorrectly Classified Instances 4 25 %
Kappa statistic 0.5
Mean absolute error 0.3333
Root mean squared error 0.4082
Relative absolute error 66.6667 %
Root relative squared error 81.6497 %
Total Number of Instances 16
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC
1 0.5 0.667 1 0.8 0.
0.5 0 1 0.5 0.667 0.
Weighted Avg. 0.75 0.25 0.833 0.75 0.733 0.
=== Confusion Matrix ===
a b <-- classified as
8 0 | a = 男性
4 4 | b = 女性
Status
OK
Log
x 0


Weka Classifier Tree Visualizer: 17:07:02 - trees.J48 (test_ttm6)
Tree View
的確
<= 0
> 0
女性 (2.0)
時
<= 0
> 0
女性 (2.0)
男性
<= 0
> 0
女性 (1.0)
優しく
<= 0
> 0
女性 (1.0)
まじめ
<= 0
> 0
女性 (1.0)
ギャップ
<= 0
> 0
男性 (8.0)
女性 (1.0)

おまけ：テキストデータの収集ソフト

TTC (TinyTweetCrawler)
http://mtmr.jp/ttc/



TWC (TinyWebCrawler)
http://mtmr.jp/twc/

まとめ

- テキストマイニング
  ＝〈前処理〉＋〈多変量解析 or データマイニング〉
- テキストマイニングは語や表現の揺れが大きいので，
  それを如何に吸収するかが重要になります
- 前処理さえ済めば，あとはRやWekaといった各自の
  得意な土俵に持ち込んで勝負すればいいのです
- テキストマイニング恐るるに足らず！

宣伝

「人文・社会科学のためのテキストマイニング」
松村真宏・三浦麻子著，誠信書房（2009) 2,520円

目次