2016

Tutorial de R-Text Mining Solution

Prof. Dr. José Pino-Díaz
Universidad de Málaga, Andalucía Tech, Escuela Técnica Superior de Ingeniería Industrial, Campus de Teatinos s/n, 29071 Málaga, España
18/07/2016

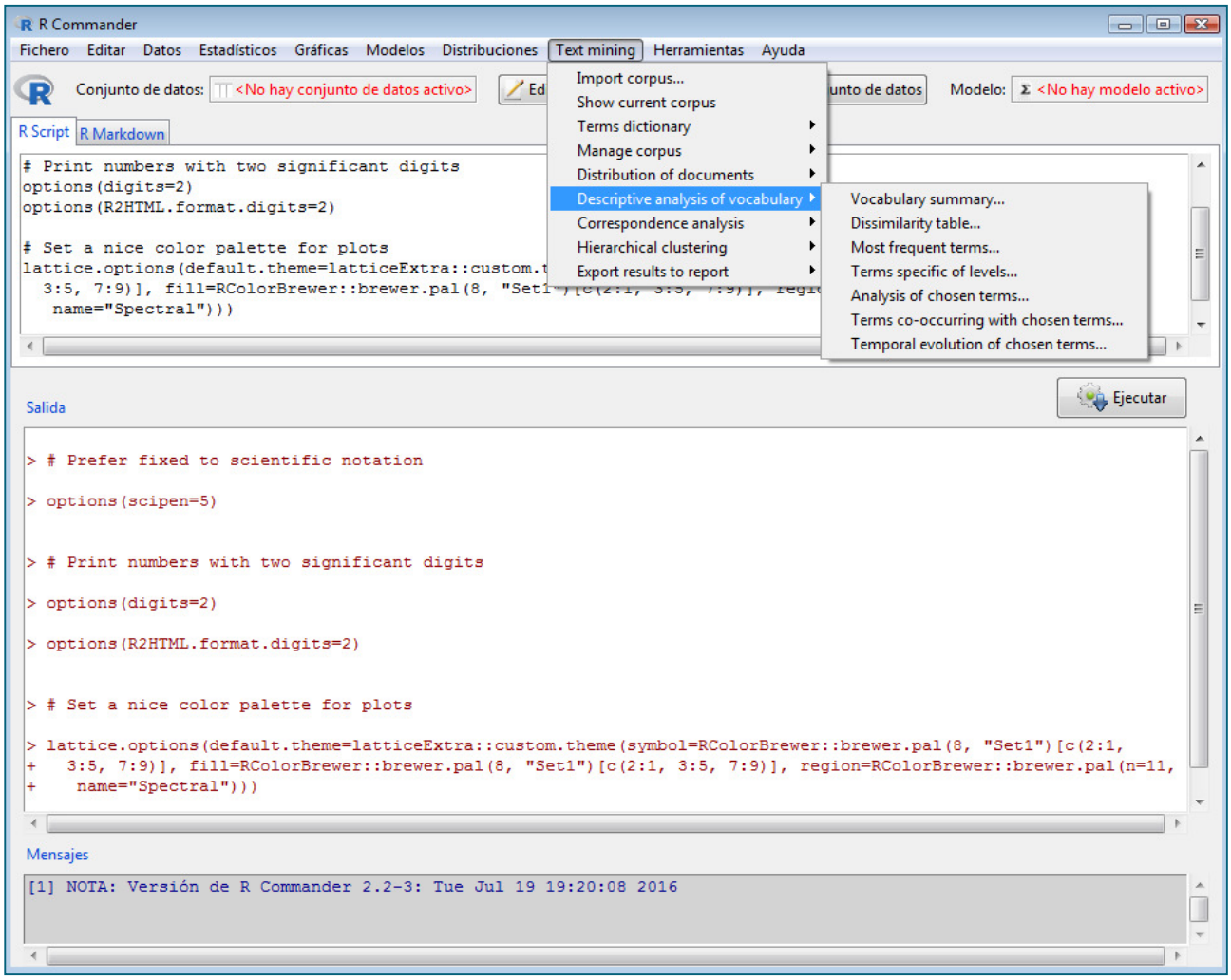
Contenido

Introducción

R.TeMiS (R Text MIning Solution) (Bouchet-Valat & Bastin, 2013) es un paquete de R (RcmdrPlugin.temis) (Bouchet-Valat, 2016), concebido como plugin de R Commander, que permite analizar, manipular y crear corpus de textos (Garnier, 2014).

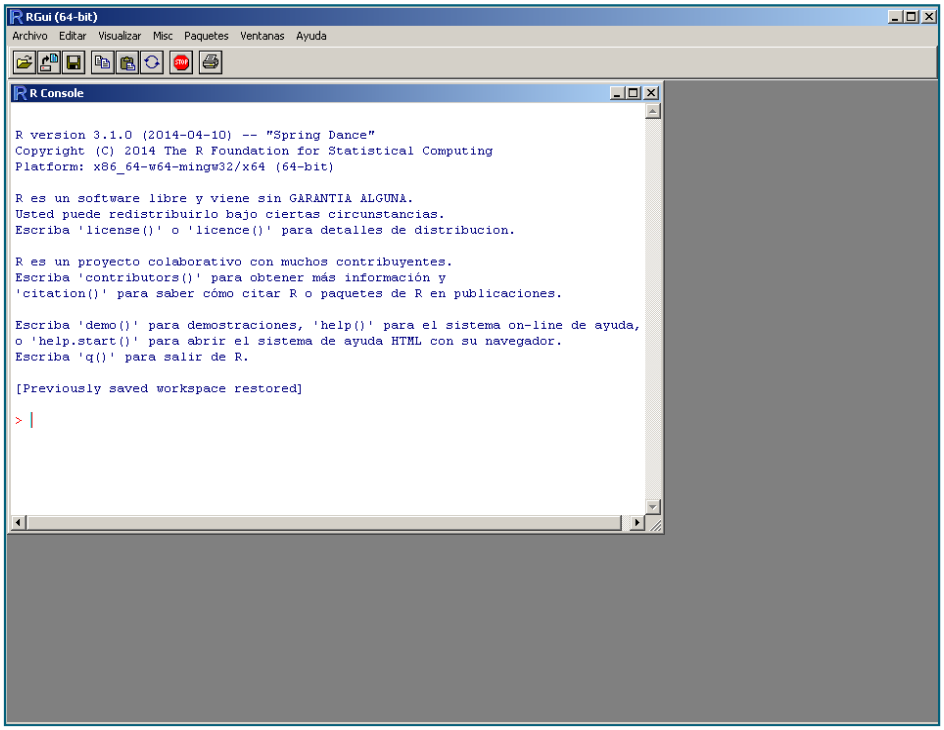
La arquitectura estadística de RTemis corre a cargo del paquete tm desarrollado por Ingo Feinerer (Feinerer, 2008 ; 2011 ; Feinerer, Hornik y Meyer, 2008). R.TeMiS se ha completado con otros paquetes clásicos de R, como el paquete para la representación de los análisis factoriales de correspondencias de Nenadic y Greenacre (2007). También se han desarrollado paquetes específicos para facilitar el uso de R.TeMiS en los estudios de prensa, por ejemplo para la gestión de los corpus de artículos de prensa de la base de datos Factiva.

R.TeMiS se presenta como un plugin de R Commander, desarrollado por Fox (2005), lo cual facilita su utilización para los no usuarios de R.

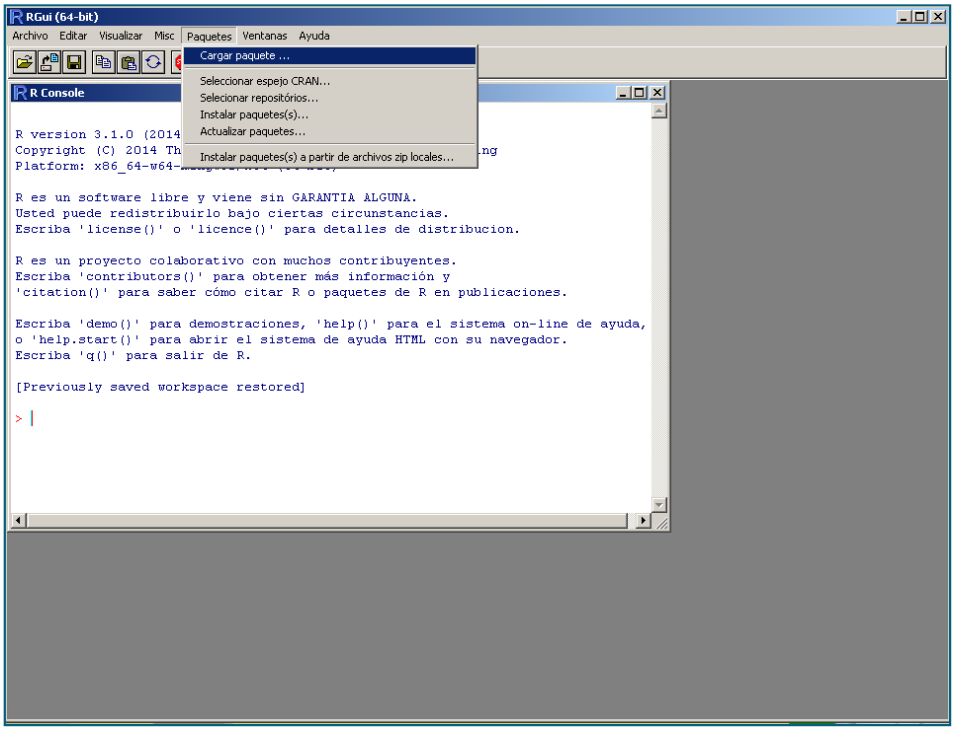


Comenzando a trabajar con R.TeMiS

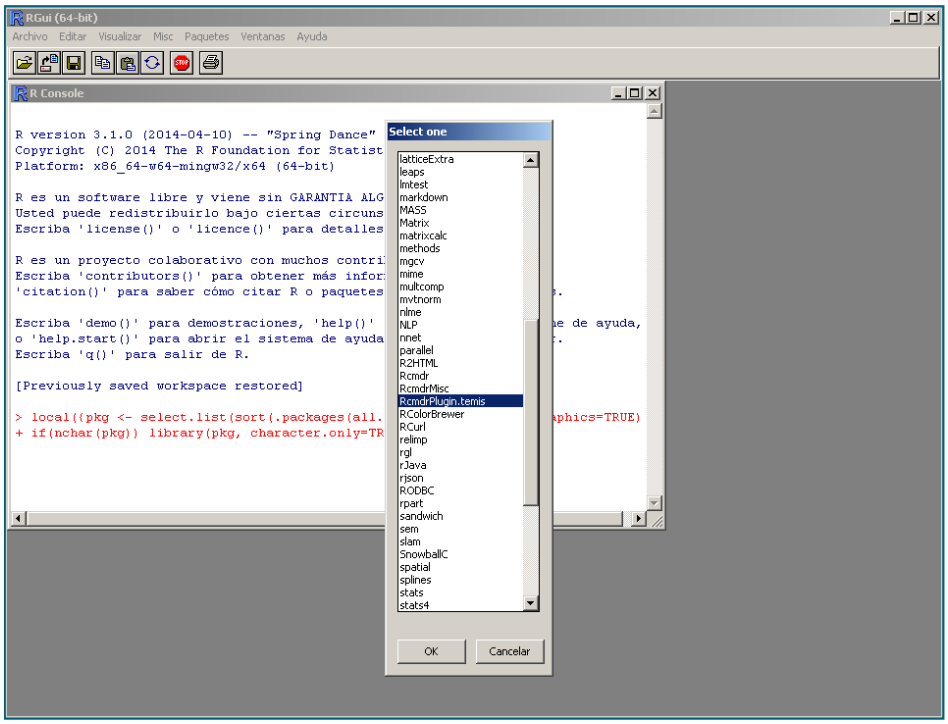
1º) Cargamos **R** en nuestro equipo. Obtenemos la siguiente pantalla una vez picamos dos veces en el logo de acceso a **R**.



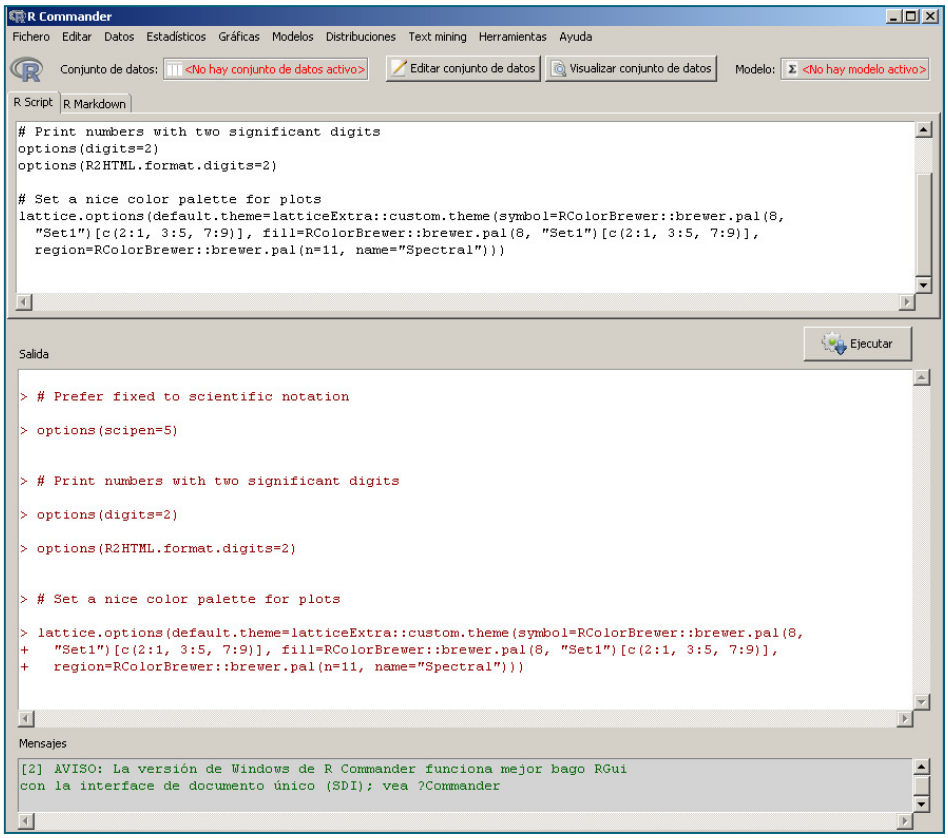
2º) A continuación nos situamos con el puntero del ratón en la opción **Paquetes** del menú de **R**. Se nos abre un menú desplegable y picamos en cargar paquete:



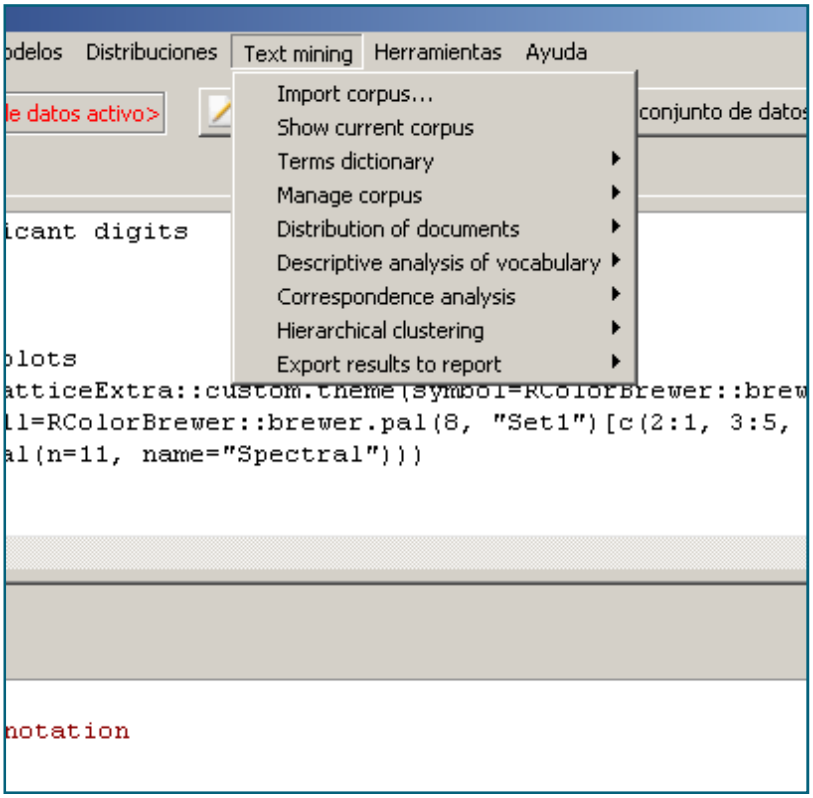
3º) Seleccionamos en la ventana de paquetes, **RcmdrPlugin.temis**:



4º) En un instante aparecerá la interfaz de **R.Commander** en la pantalla del equipo:



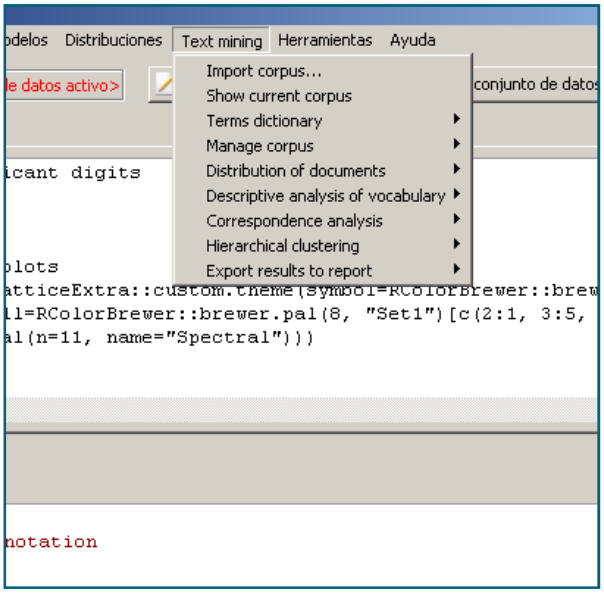
5º) Observamos que en el menú de la interfaz de R.Commander aparece la opción **Text mining**. Si picamos sobre el botón **Text mining** se nos despliega el siguiente menú:



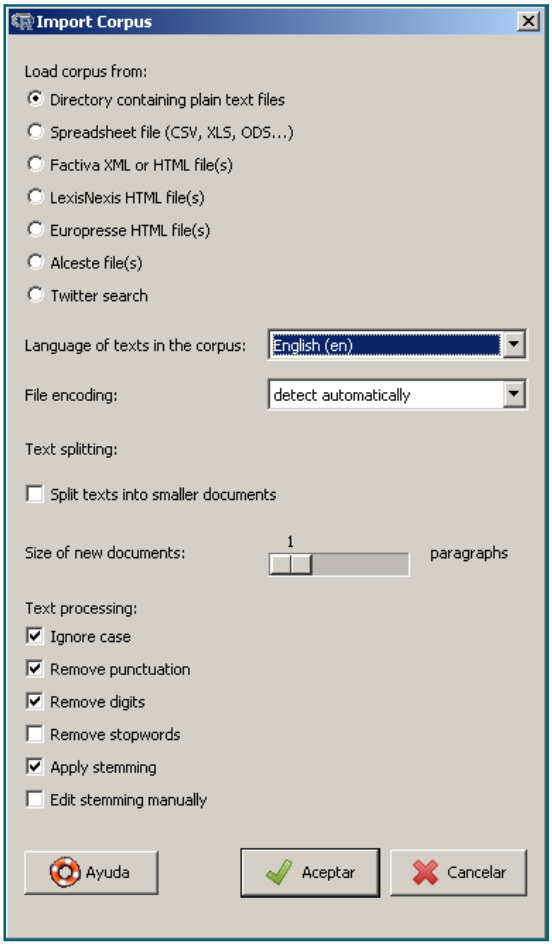
6º) En las siguientes secciónes del curso se explicarán cada una de las opciones del menú desplegable de **Text mining**.

Importar Corpus

1º) La primera opción que nos aparece en el menú ***Text mining*** es la de **Importar corpus**.



Picando sobre ella nos aparece la siguiente ventana de opciones de procesamiento del texto:



a) Se ha de seleccionar el tipo de archivo que contiene el documento o documentos que se van a analizar. Cuatro tipos de corpus pueden ser importados por R.TeMiS:

- archivos de texto plano (en el formato .txt)
- archivos tipo tabla [en formatos .csv, .xls (dependiendo de la versión de Excel, 32 o 64 bits, no siempre son reconocidos por el programa) u .ods. Las líneas corresponden a individuos (i.e., personas encuestadas mediante cuestionario) o a datos de elementos (i.e., artículos científicos, exposiciones, etc.) y las columnas a las variables descriptivas (una de ellas es la variable que contiene el texto que vamos a analizar)
- archivos en .html o .xml exportados de las bases de prensa Factiva, LexisNexis o Europresse.
- Resultados de investigaciones sobre corpus extraídos de twitter

b) Se selecciona el idioma en el que están escritos los documentos.

c) Se selecciona la codificación de los documentos de texto (UNICODE, UTF 8, Windows 1250, ISO8859-x, etc.). Si se desconoce la codificación del texto, dejar la opción *"detect automatically"*.

d) Seleccionar o no, según se desee, cortar el documento o documentos en otros más pequeños.

➔ Al importar un corpus el usuario puede decidir cortarlo en unidades más pequeñas. La unidad mínima a seleccionar es el párrafo. Si se elige esta opción cada párrafo será considerado como un documento, esto puede permitir mejorar la calidad del análisis de los textos, en particular de la clasificación jerárquica ascendente.

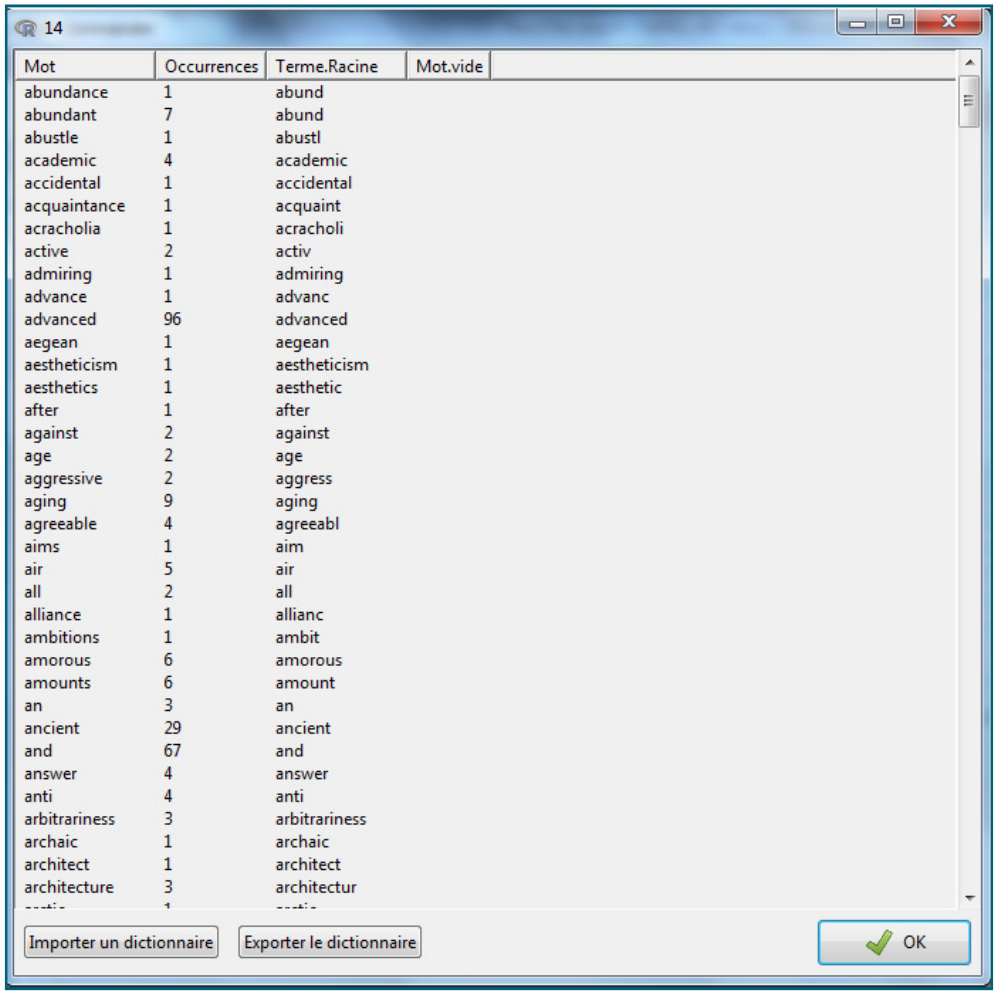
➔ La elección de cortar el documento en párrafos para el análisis del texto pretende tener en cuenta los formatos de escritura mediática de manera menos arbitraria que con el corte de segmentos de longitud uniforme (Jenny, 1999).

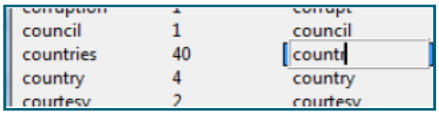
➔ Los archivos de texto tabulados (CSV, XLS), resultado de cuestionarios, encuestas o estudios similares, son cortados por defecto en tantos documentos como líneas.

e) A continuación se seleccionan las acciones con las que se quiere preparar el texto para el análisis (nivel de procesamiento léxico del corpus):

- pasar los términos a minúsculas (para que no haya diferencias entre el mismo término si aparece tanto escrito con una mayúscula como si no);
- eliminar los signos de puntuacion;

- eliminar los números;
- eliminar las palabras vacías (stopwords);
- extraer las raíces de los términos (para agrupar bajo la misma raíz todos los términos derivados);
- editar manualmente la lematización (modificar la lematización propuesta por defecto; esto se lleva a cabo por el paquete Snowball utilizando el algoritmo de Porter); la edición manual de la lematización del corpus permite reagrupar bajo una misma raíz todos los términos derivados de ésta.





  - (i.e., se modifica la raíz *contr* para que agrupe los términos *country* y *countries*)

f) Una vez fijado el nivel de procesamiento se pica en el botón aceptar.

g) Por último, en la ventana del explorador de *windows* que aparece, seleccionar la carpeta que contiene los documentos de texto plano (i.e., si se

ha seleccionado *directory containing plain text files*) o el archivo tabulado (CSV, XLS, etc.) a analizar; y picar en aceptar.

h) Carga del corpus en la memoria:

1.- Si es seleccionada en el explorador del ordenador una carpeta que contiene archivos de texto plano (TXT), el programa comienza el análisis y nos aparece en el interfaz de R.TeMiS un primer resultado:

```
> corpus
<<VCorpus (documents: 4, metadata (corpus/indexed): 3/0)>>

> dtm
<<DocumentTermMatrix (documents: 4, terms: 1001)>>
Non-/sparse entries: 1275/2729
Sparsity           : 68%
Maximal term length: 16
```

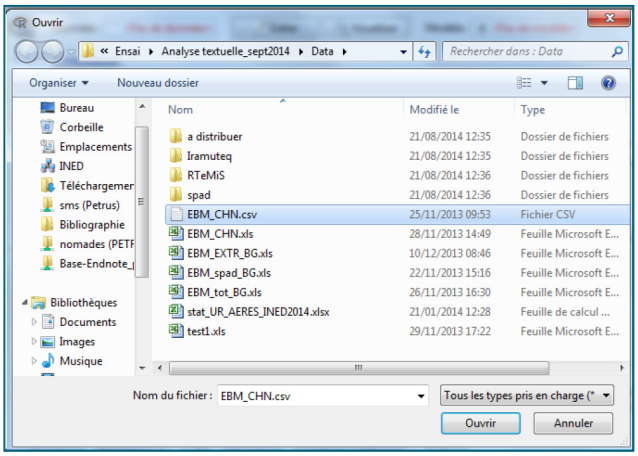
Ejemplo de primer resultado obtenido por R.TeMis de un corpus de cuatro documentos de texto plano contenidos en una carpeta.

R.TeMiS construye una matriz [documentos (filas) x términos (columnas)], de manera que en las celdas se anota el número de frecuencias de cada término en cada documento.

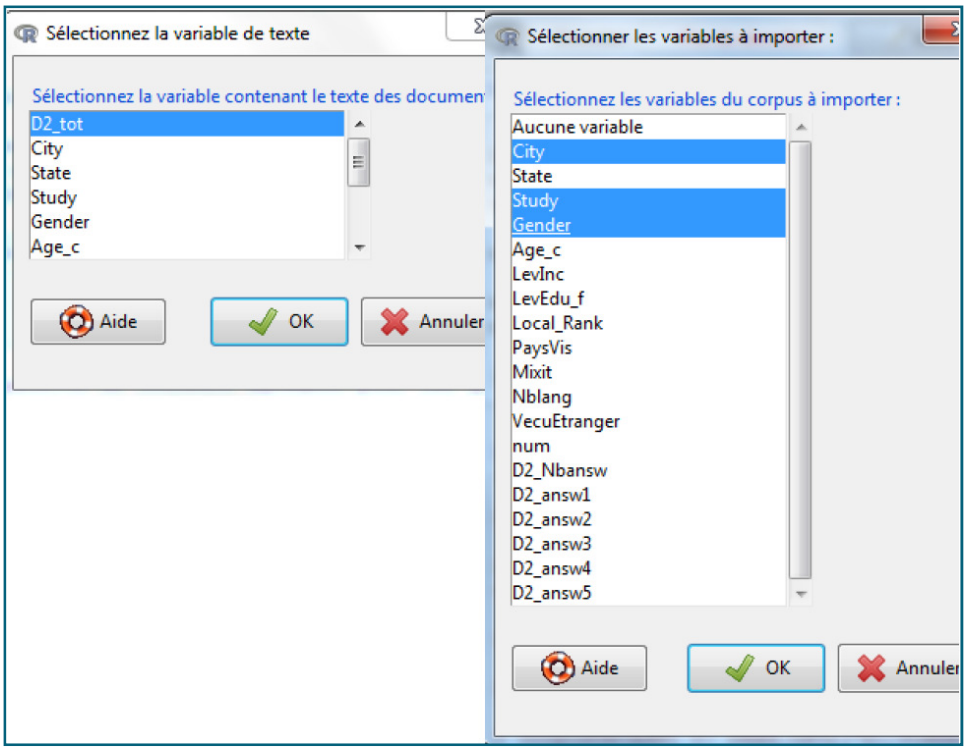
El parámetro de esparsidad se obtiene dividiendo el total celdas vacías por el total de celdas de la matriz (i.e., en el ejemplo: 1275 *nonzero elements* y 2729 *zero elements*, 68% *sparsity*). La esparsidad de corpus con documentos de léxicos muy similares es más baja que la de corpus con documentos de léxicos muy diferentes.

Otro dato que nos aporta este primer resultado es la longitud máxima de los términos (i.e., en el ejemplo, 16 letras).

2.- Si en el explorador del ordenador se ha seleccionado para su análisis un archivo tipo tabla (CSV o XLS):



- aparece una ventana con el listado de las variables de la tabla (que figuran en el encabezamiento de las columnas); se debe seleccionar entonces la variable que contiene el texto que se quiere analizar, y a continuación, una vez seleccionada la variable de texto.
- aparece una segunda ventana con el listado de las variables, menos la seleccionada para analizar el texto; se deben seleccionar las demás variables de la tabla que en el análisis se quiere relacionar con la del texto.



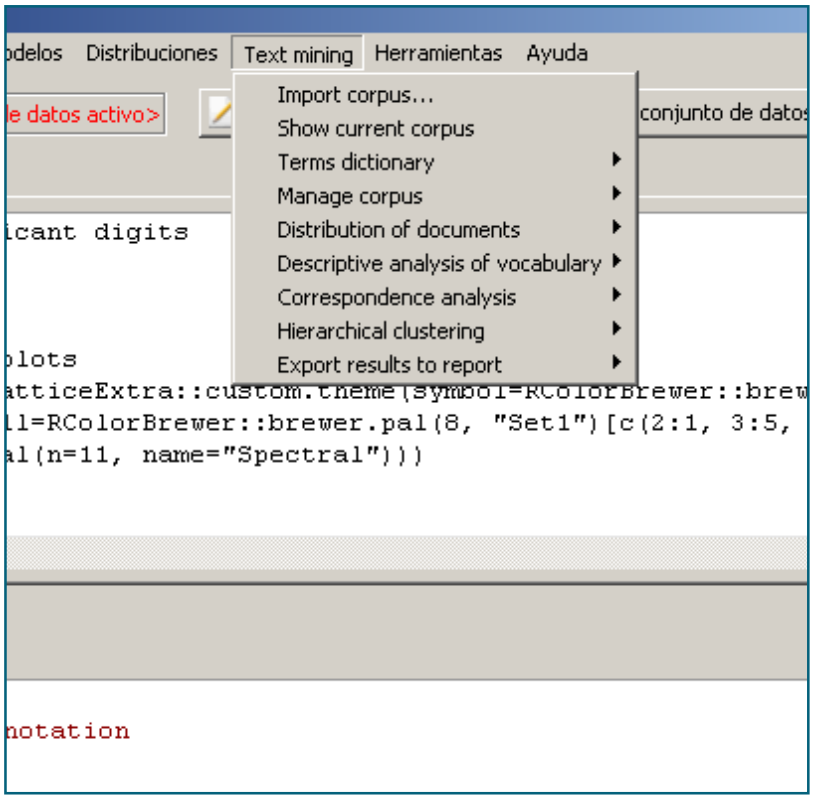
El programa comienza el análisis y nos aparece en el interfaz de R.TeMiS un primer resultado:

```
> dtm
<<DocumentTermMatrix (documents: 1140, terms: 974)>>
Non-/sparse entries: 5010/1105350
Sparsity           : 100%
Maximal term length: 15
```

En este ejemplo la matriz está formada por 1140 filas (documentos) y 974 columnas (términos). Contiene 5010 celdas con valores no nulos y 1105350 celdas vacías (valor cero) y el término más largo contiene 15 letras.

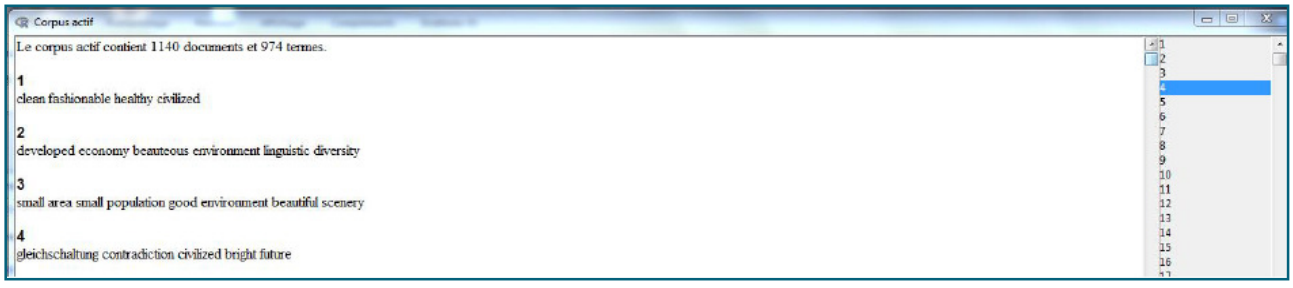
Visualización del corpus activo y de los diccionarios de términos

Continuamos explicando las opciones que aparecen en el menú ***Text mining*** de R.Temis.



Visualizar el corpus de documentos

Esta opción permite visualizar el corpus activo; los documentos aparecen con el número de serie que le asigna el programa y su texto.



Aquí nos aparecerán los diferentes documentos. Su tamaño (texto completo, párrafos o texto de la celda de la variable de texto elegida) dependerá de las opciones seleccionadas en la ventana de procesamiento del texto.

Visualizar el diccionario de términos

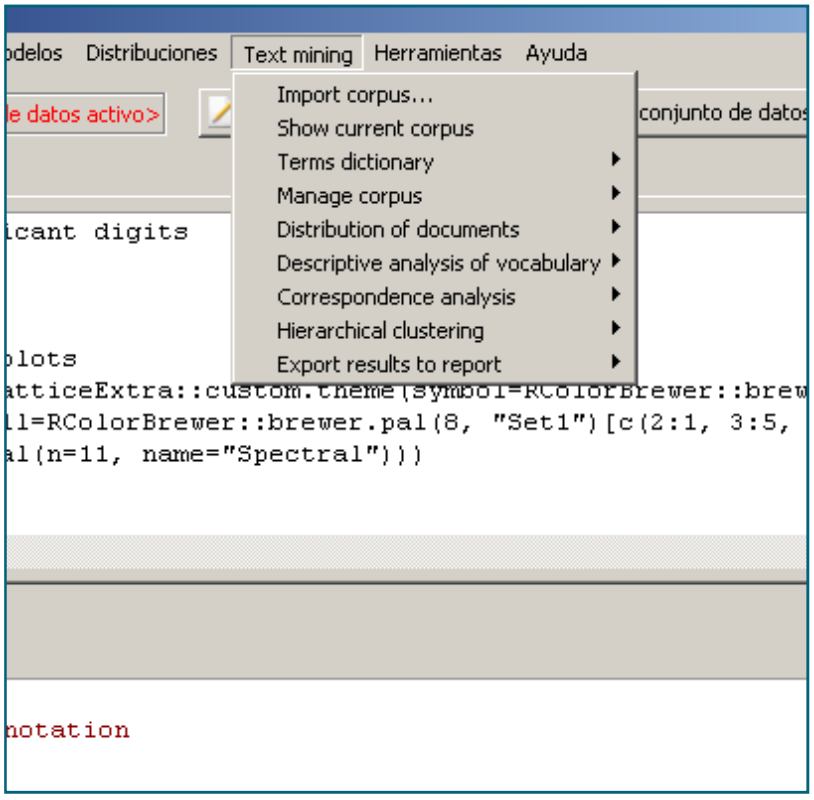
Esta opción permite visualizar el diccionario de los términos del corpus analizado, resultado del procesamiento del texto. Se obtienen dos listados, por orden alfabético y por nº de ocurrencias. En los listados aparecen para cada término: su número de ocurrencias, la raíz del término, el número de ocurrencias de la raíz y las palabras vacías suprimidas.

```
> attr(dict, "title") <- "Dictionnaire des termes par ordre alphabétique"

> dict
                Occurrences    Terme.Racine Occ. racine Mot vide Supprimé
abundance                 1           abund           8
abundant                  7           abund           8
abustle                   1          abustl           1
academic                  4          academ           4
accidental                1        accident           1
acquaintance              1        acquaint           1
acracholia                1      acracholia           1
active                    2           activ           2
admiring                  1           admir           1
advance                   1          advanc          97
advanced                 96          advanc          97
aegean                    1          aegean           1
aesthetics                1         aesthet           1
aestheticism              1       aesthetic           1
after                     1           after           1 Mot vide
against                   2         against           2 Mot vide
```

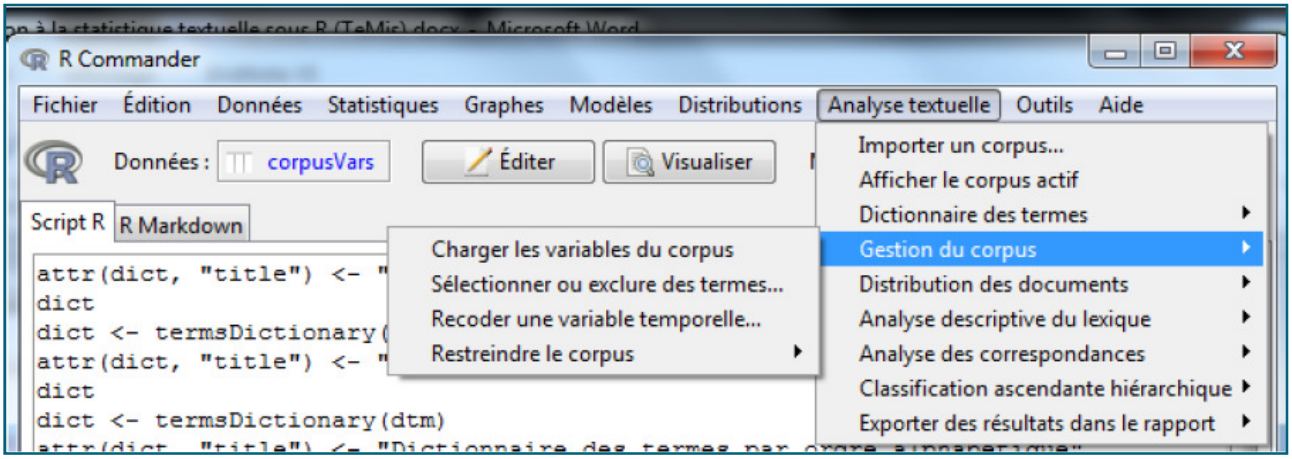
Para visualizar las palabras vacías en inglés, escribir y ejecutar el comando stopwords("en") en la ventana del script.

Gestión y distribución del corpus

Continuamos explicando las opciones que aparecen en el menú ***Text mining*** de R.Temis.

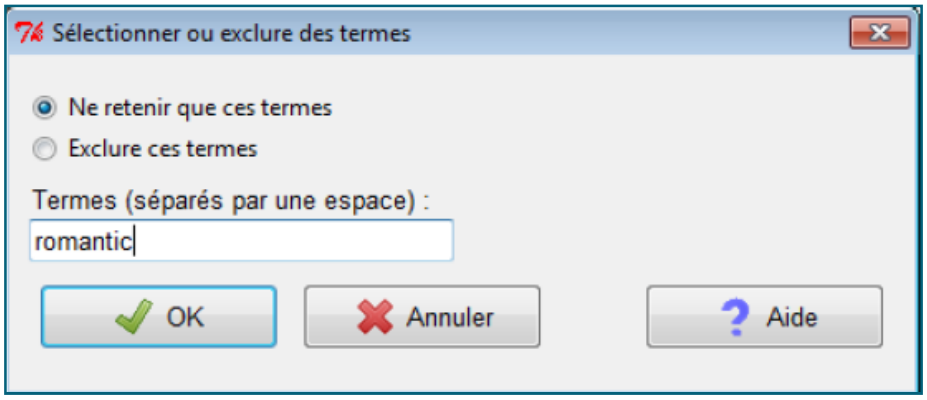


La opción ***Manage corpus*** contiene las siguientes acciones.



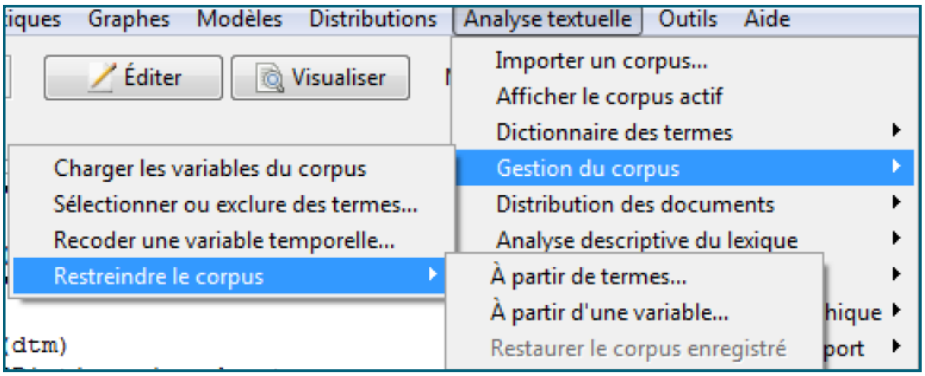
- Cargar las variables del corpus. Se deben cargar de nuevo las variables si se hubiesen editado-modificado con **Rcommander** o si se hubiesen cargado nuevas variables desde otro fichero.

- Seleccionar o excluir términos. Esta acción permite seleccionar o excluir uno o más términos del vocabulario.

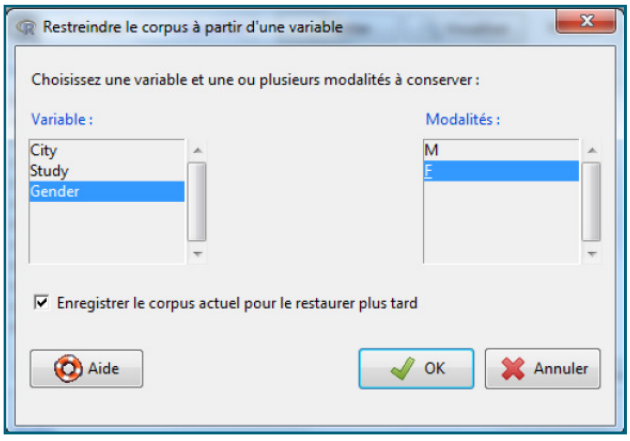


- Modificar (recodificar) los parámetros de una variable temporal.

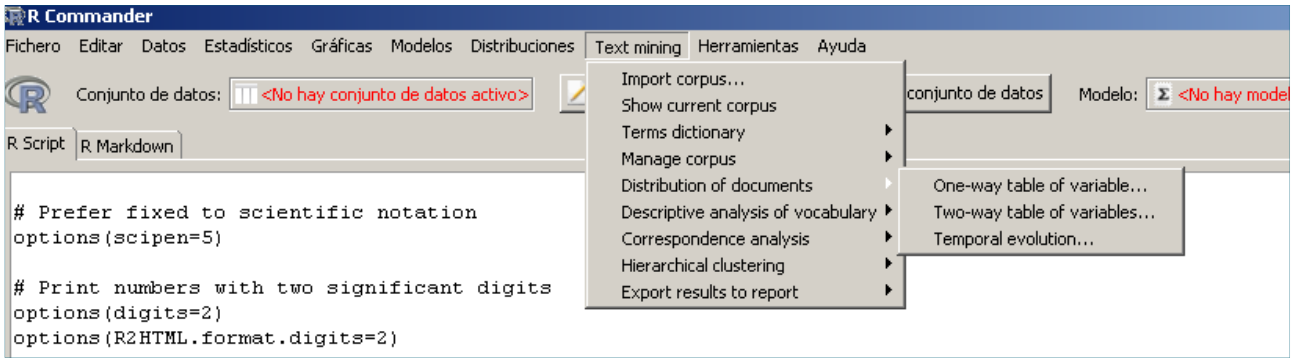
- Creación de subcorpus a partir de determinados términos o de una variable. Esto permite estudier el contexto de uso del término elegido.



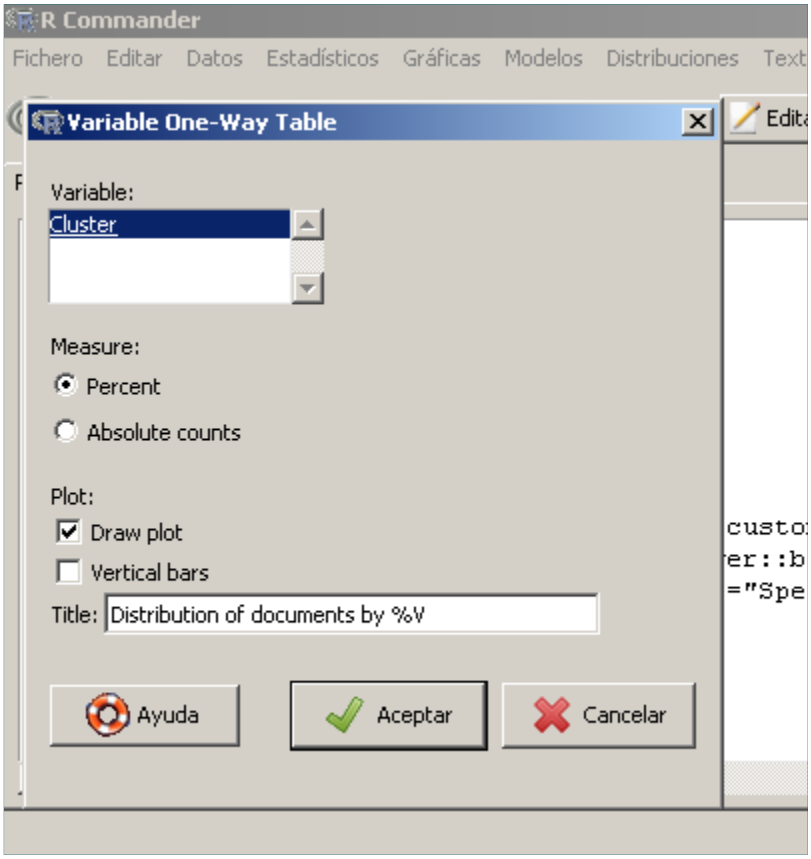
Creado un subcorpus, i.e. a partir de una variable cualitativa, como por ejemplo se ve en la imagen de abajo, "Género", la opción restaurar el corpus permite volver al corpus registrado.

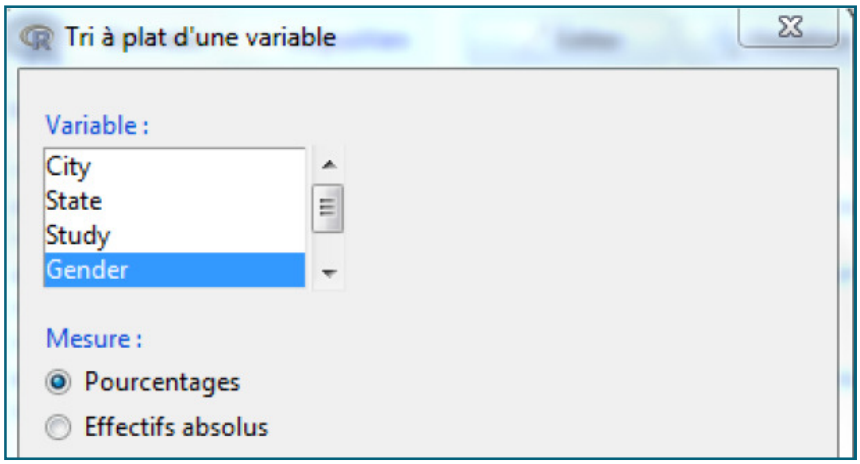


La opción **Distribución del corpus** permite tres acciones:



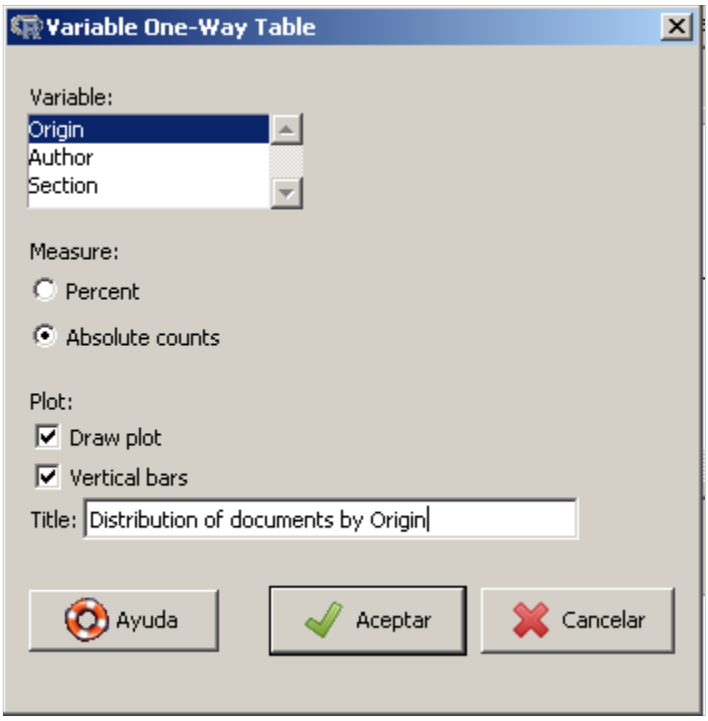
- Clasificar el corpus de documentos según una sola variable. Por ejemplo la variable cualitativa "Género" permite clasificar el corpus de respuestas (cada respuesta constituye un documento en los análisis de los corpus de cuestionarios, encuestas, etc.) según sean hombres (H), o mujeres (M), las dos modalidades de la variable. El dato se puede obtener en porcentages o en frecuencias absolutas (ver las tres imágenes siguientes).

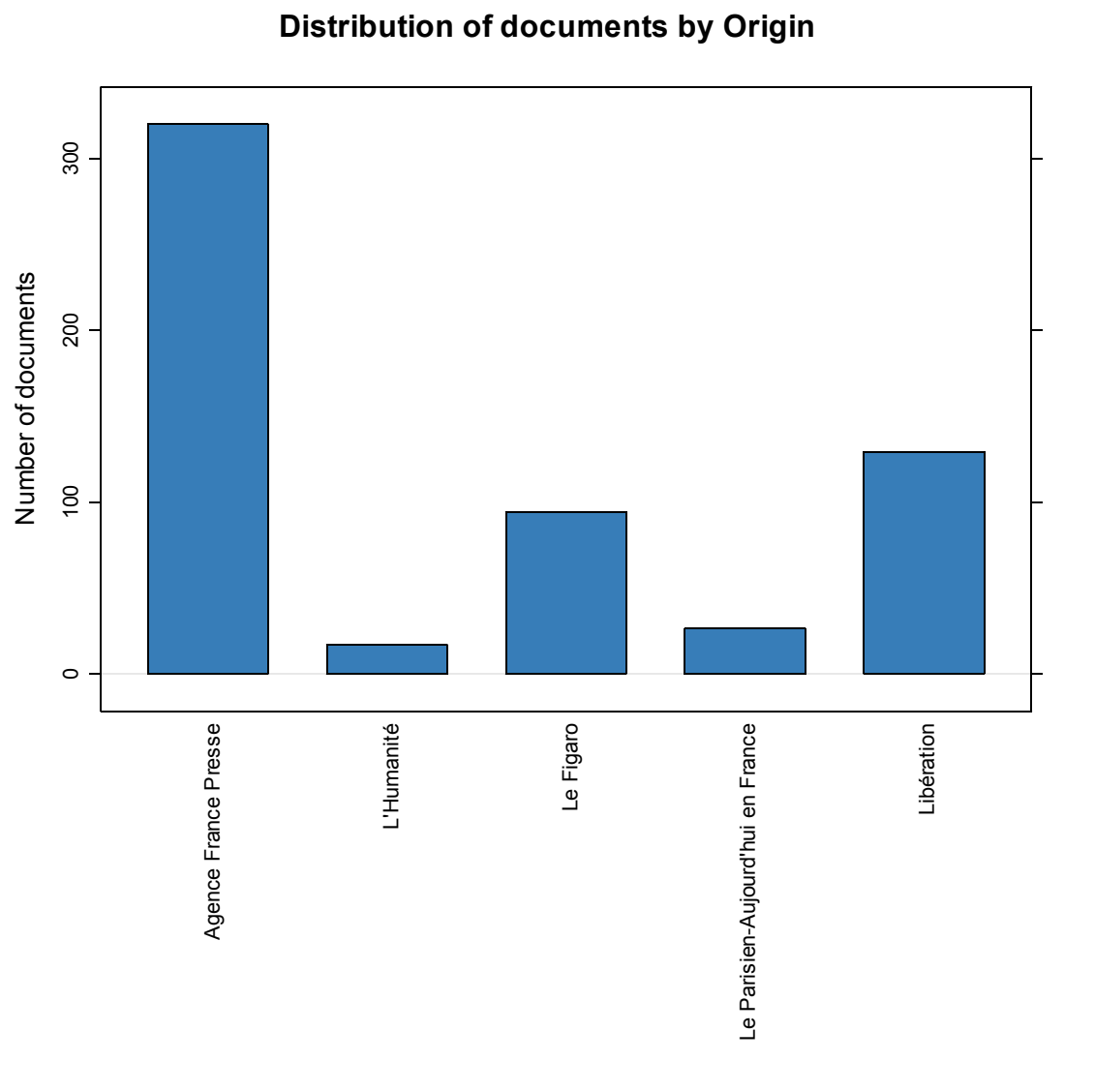





```
> varFreqs
Gender
  F   M Sum
 52  48 100
```

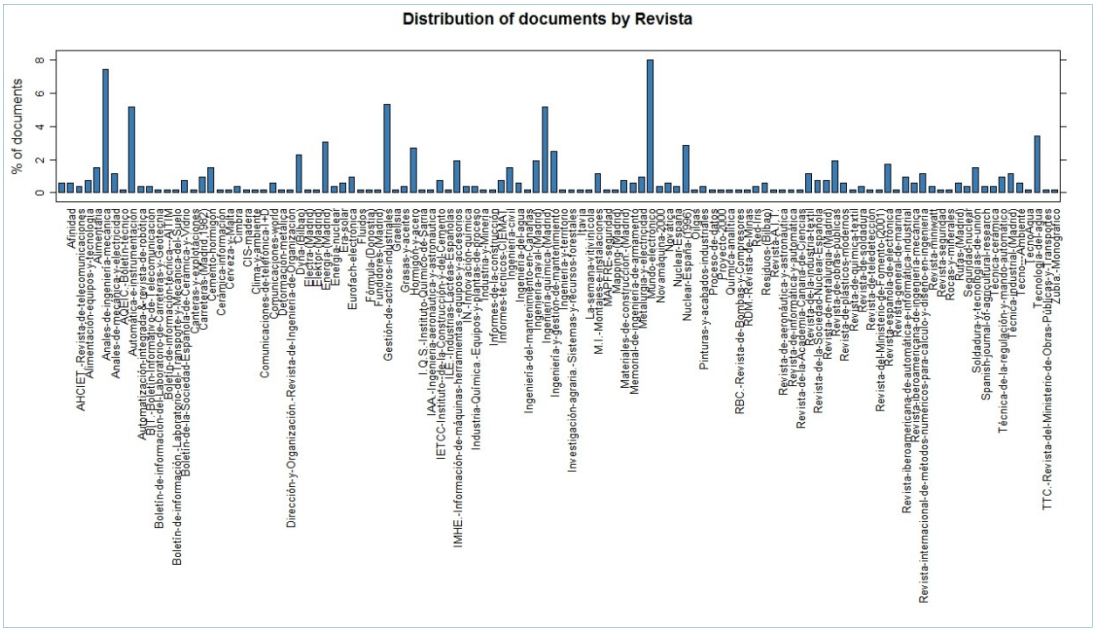
Ejemplo 1: Se analiza un corpus de Factiva (formato HTML) de 586 noticias de prensa sobre Julian Assange (escándalo Wikileaks), de diferentes medios (Agencia France Presse, L'humanité, Le Figaro, Le Parisien y Libération). Se quiere obtener el gráfico de distribución de noticias según el medio que la ha publicado (ver las tres imágenes siguientes).





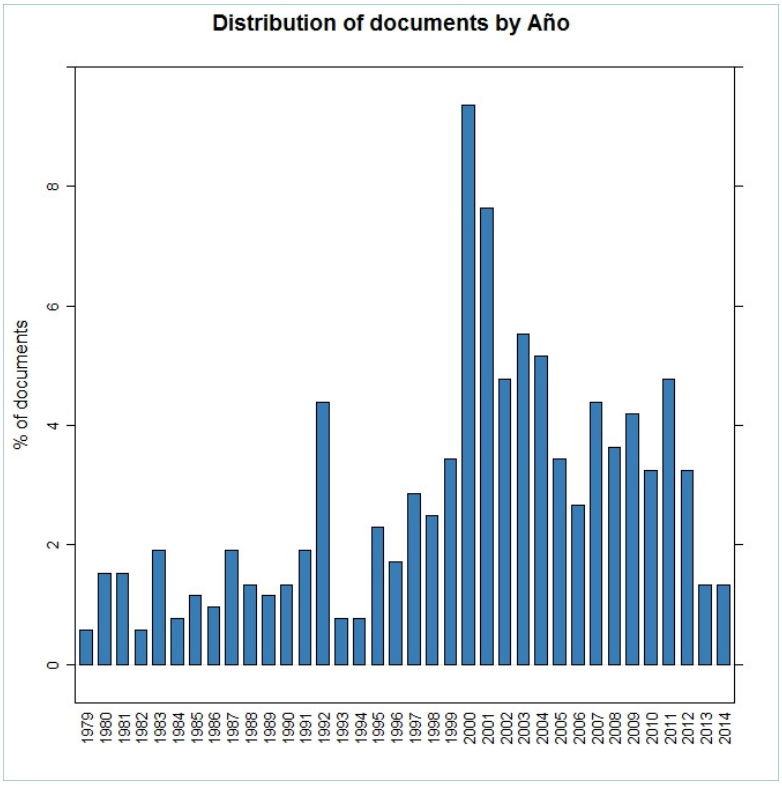

```
Origin
                Agence France Presse                            L'Humanité
                                 320                                    17
                           Le Figaro Le Parisien-Aujourd'hui en France
                                  94                                    26
                         Libération                                   Sum
                                 129                                   586
```

Ejemplo 2: Se analiza un corpus de artículos científicos obtenidos de una base datos (formato CSV). Los artículos han sido publicados en diferentes revistas científicas. Se quiere obtener el gráfico de distribución de artículos según la revista que lo ha publicado (ver la imagen siguiente).



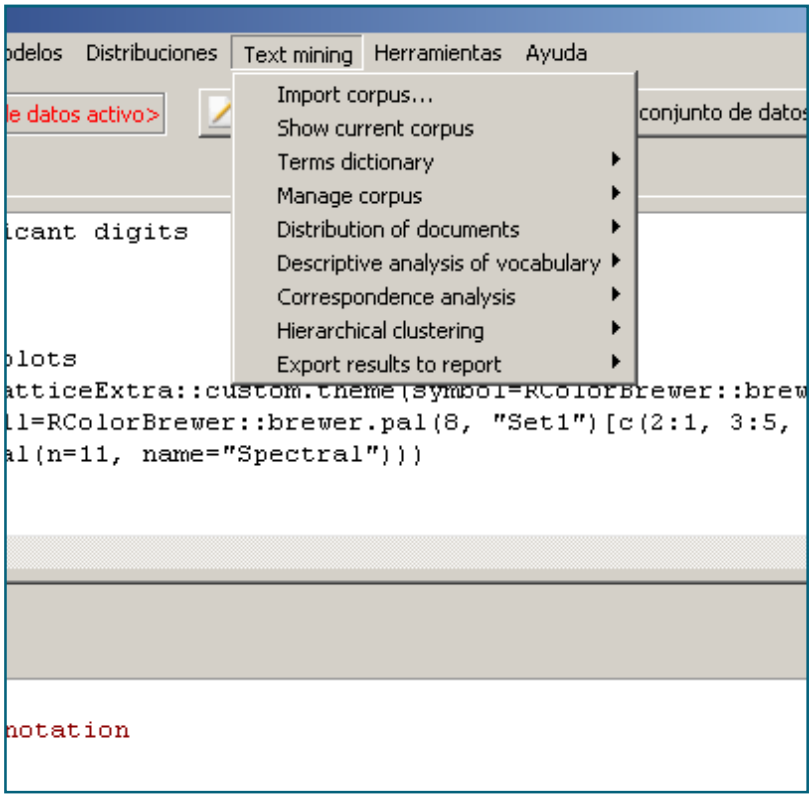
- Clasificar el corpus según dos variables.
- Conocer la evolución temporal del corpus, por ejemplo, según la fecha de publicación de cada documento.

Ejemplo 3: Gráfico de distribución de los artículos publicados por año.

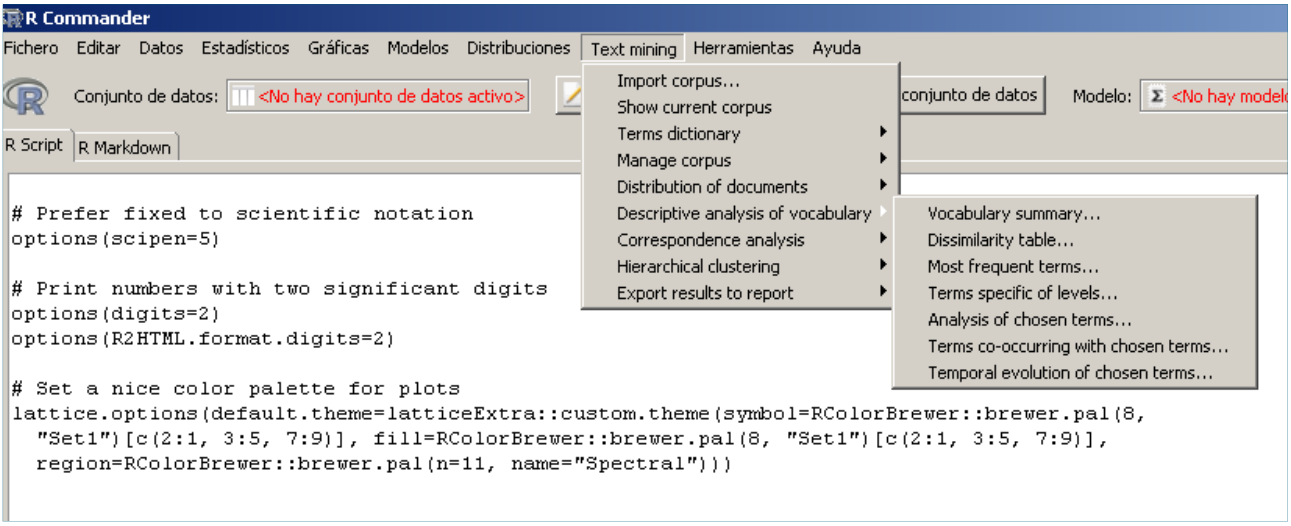


Análisis descriptivo del léxico

Continuamos explicando las opciones que aparecen en el menú ***Text mining*** de R.Temis.

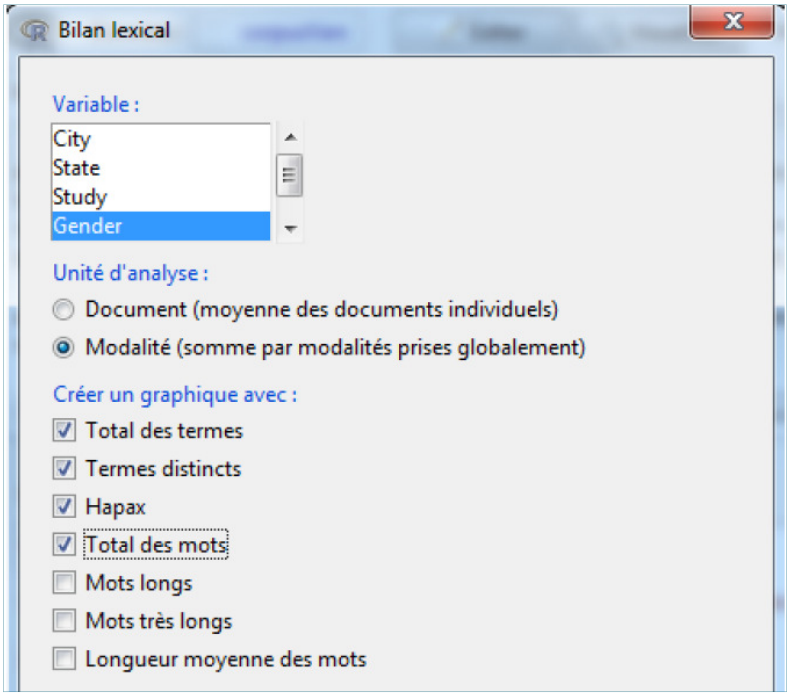


La opción ***Descriptive analysis of vocabulary*** contiene las siguientes opciones.



Resumen cuantitativo del vocabulario de términos.

Esta opción, una vez seleccionada la variable, permite obtener un resumen cuantitativo del número total de términos, el número de términos distintos, número de hapax (un hápax o *hápax legómenon* es una palabra que ha aparecido registrada solamente una vez en un corpus; es una palabra que sólo aparece una vez dentro de un contexto, ya sea en el registro escrito de un idioma entero, en las obras de un autor o dentro de un solo texto), etc., presentes en el corpus. Ejemplo:



Seleccionada la variable, se selecciona la unidad de análisis (que puede ser el documento o la modalidad de la variable.; por ejemplo, si la variable elegida es "Género", la selección de la modalidad de la variable realiza el análisis para "masculino", "femenino" y para el total).

```
> voc

Total par catégorie :                    F      M Total du corpus
  Nombre de termes                2604.0 2455.0          5059.0
  Nombre de termes distincts       694.0  782.0          1134.0
  Pourcentage de termes distincts   26.7   31.9            22.4
  Nombre de hapax                  408.0  522.0           672.0
  Pourcentage de hapax              15.7   21.3            13.3
  Nombre de mots                  2604.0 2455.0          5059.0
  Nombre de mots longs            1649.0 1502.0          3151.0
  Pourcentage de mots longs         63.3   61.2            62.3
  Nombre de mots très longs       403.0  404.0           807.0
  Pourcentage de mots très longs    15.5   16.5            16.0
  Longueur moyenne des mots          7.2    7.1             7.2
```

La ventana permite seleccionar los parámetros del gráfico asociado al resumen léxico cuantitativo.

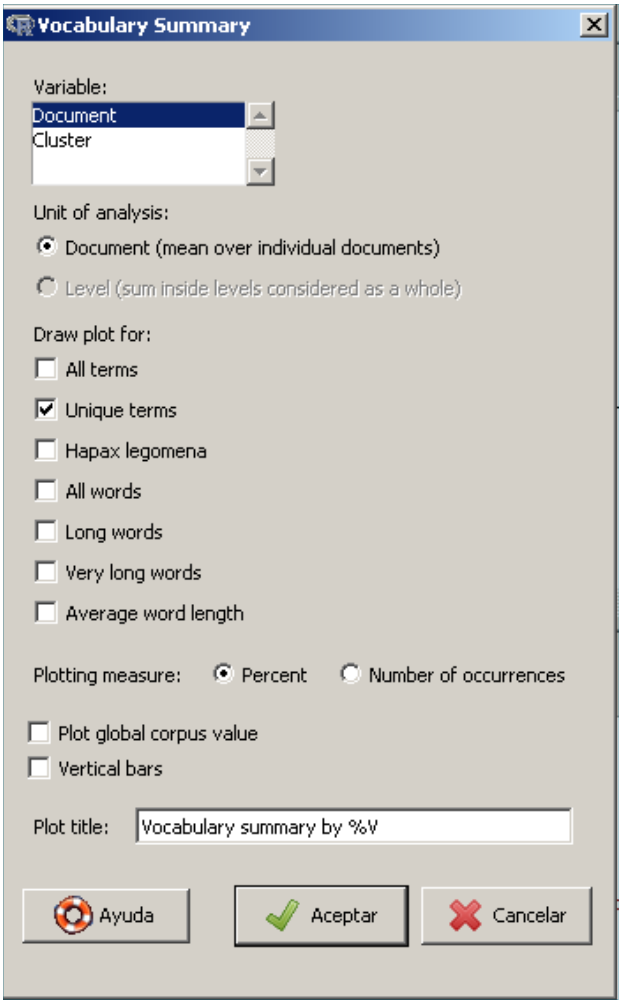
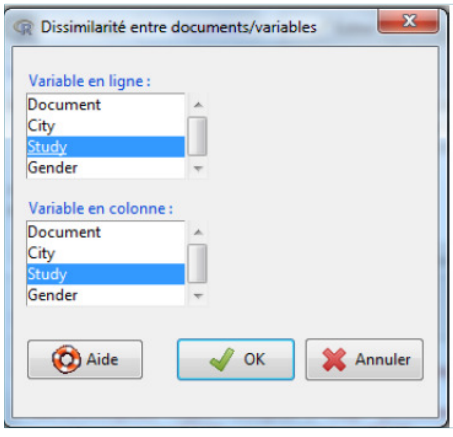


Tabla de disimilaridad.

La opción **Tabla de disimilaridad** nos permite conocer si los vocabularios de los documentos o de las modalidades de las variables son más o menos similares (el programa calcula la distancia chi cuadrado, de modo que un menor valor de la distancia implica una mayor similaridad). Ejemplo:




```
> diss
    ART BUS ENG HEA POL
BUS 1.8
ENG 1.9 1.9
HEA 1.8 1.7 1.8
POL 1.8 1.8 1.8 1.7
SHS 1.8 1.8 1.8 1.7 1.7
```

Le vocabulaire est plus proche entre les étudiants en sciences sociales (SHS), en santé (HEA) et sciences politiques (POL)

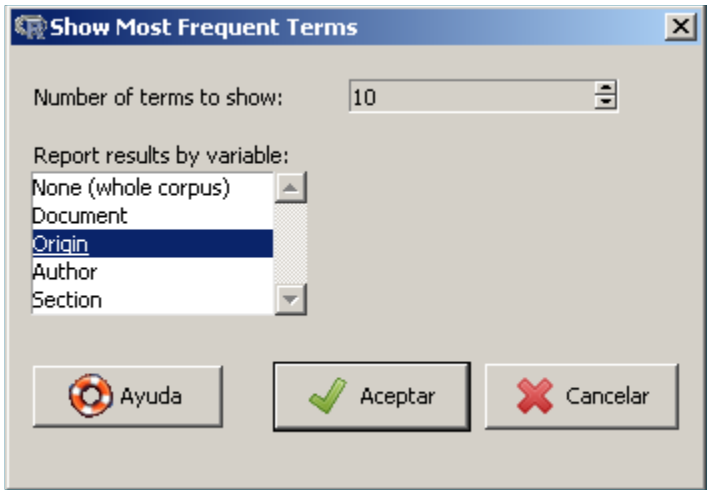
Ejemplo: De un corpus de 586 noticias de prensa sobre Julian Assange (escándalo Wikileaks) extraído de Factiva (formato HTML), procedente de diferentes medios (Agencia France Presse, L'humanité, Le Figaro, Le Parisien y Libération), se quiere obtener la tabla de disimilaridad de las noticias según lel medio donde se ha publicado (variable "Origin").

```
> diss
                                  Agence France Presse L'Humanité Le Figaro
L'Humanité                                        2.4
Le Figaro                                         1.1         2.3
Le Parisien-Aujourd'hui en France                 1.8         2.8       1.9
Libération                                        1.1         2.3       1.1
                                  Le Parisien-Aujourd'hui en France
L'Humanité
Le Figaro
Le Parisien-Aujourd'hui en France
Libération                                                     1.9
```

Se comprueba como los vocabularios empleados en los textos de las noticias más próximos son los de *Agence France Press* y *Le Figaro*, *Agence France Press* y *Libération,* y *Libération* y *Le Figaro.*

Términos más frecuentes.

Esta opción permite conocer cuáles son los términos más frecuentes en los documentos, en todo el corpus o según una variable seleccionada (en la imagen de abajo se ha seleccionado la variable "Origin" y un número de términos a visualizar de 10).



El resultado nos aparece en forma de tabla, donde:

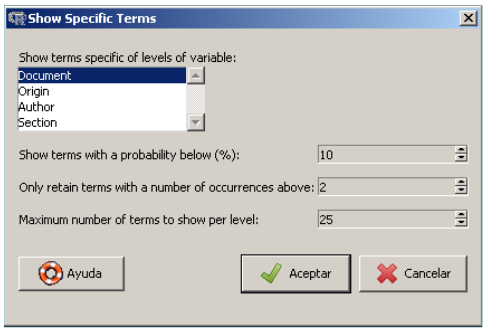
- **% Term/Level** = % ; Nº de ocurrencia del término en la modalidad sobre el total de las ocurrencias de todos los términos que aparecen en la categoría o modalidad de la variable seleccionada.

- **% Level/Term** = % ; Nº de ocurrencias del término en la modalidad de la variable elegida sobre el total de frecuencias de dicho término en el corpus.
- **Global %** = % ; Nº de ocurrencias del término en el conjunto del corpus sobre el total de ocurrencias de todos los términos del corpus.
- **Level** = Número de ocurrencias del término en la modalidad (en la imagen de abajo las modalidades son Agence france Presse y L'Humanité) de la variable seleccionada (en este caso la variable es "Origin" y las modalidades cada uno de los medios que han publicado las noticias).
- **Global** = Número de ocurrencias del término en todo el corpus.
- **t value** = Parámetro del valor del test (si este valor es positivo el término está sobrerrepresentado, si es negativo el término está subrepresentado)
- **Prob.** = Probabilidad de obtener el término en el conjunto del corpus.

```
> freqTerms
$`Agence France Presse`
          % Term/Level % Level/Term Global % Level Global t value  Prob.
assang            2.53           70     1.83  1708   2441     Inf 0.0000
wikileak          1.50           57     1.32  1009   1760    5.65 0.0000
julian            1.34           65     1.04   905   1385     Inf 0.0000
sued              0.95           81     0.59   640    792     Inf 0.0000
extradit          0.76           80     0.48   510    634     Inf 0.0000
fondateur         0.73           70     0.53   494    705     Inf 0.0000
suédois           0.73           73     0.50   492    672     Inf 0.0000
cour              0.68           79     0.44   461    581     Inf 0.0000
américain         0.81           55     0.74   545    992    2.70 0.0035
sit               0.70           52     0.68   470    903    0.83 0.2039

$`L'Humanité`
          % Term/Level % Level/Term Global % Level Global t value  Prob.
état              0.86         18.9     0.15    37    196     Inf 0.0000
journal           0.58          5.8     0.33    25    433    2.62 0.0044
plus              0.90          4.5     0.66    39    876    1.86 0.0312
docu              0.65          4.2     0.50    28    663    1.30 0.0970
américain         0.90          3.9     0.74    39    992    1.13 0.1286
fait              0.63          4.0     0.51    27    679    0.98 0.1643
être              0.56          3.7     0.48    24    641    0.63 0.2648
sit               0.72          3.4     0.68    31    903    0.26 0.3985
assang            0.67          1.2     1.83    29   2441   -6.51 0.0000
wikileak          0.97          2.4     1.32    42   1760   -2.04 0.0206
```

Términos específicos por modalidades de la variable.

Esta opción permite recuperar los términos específicos por modalidades de la variable. Se seleccionan los parámetros: mínima probabilidad de ocurrencia, mínimo nº de ocurrencias y el nº de términos específicos a obtener en el resultado.



En la imagen anterior se ha seleccionado que el resultado solo muestre los términos con una probabilidad por debajo del 10%, los términos con una ocurrencia superior a 2 y el número máximo de términos por modalidad (25, en la imagen anterior).

Tabla de términos específicos de la modalidad "Agence France Presse" (extraído de un análisis de artículos de prensa de la base de datos Factiva):

```
> specTerms
$`Agence France Presse`
            % Term/Level % Level/Term Global % Level Global t value Prob.
afp                 0.21           92    0.116   141    154     Inf     0
agress              0.40           76    0.267   272    356     Inf     0
ajout               0.19           85    0.112   126    149     Inf     0
appel               0.46           71    0.327   307    435     Inf     0
arrêt               0.59           77    0.387   399    516     Inf     0
assang              2.53           70    1.833  1708   2441     Inf     0
australien          0.50           77    0.327   337    436     Inf     0
britann             0.58           79    0.372   394    496     Inf     0
conditionnel        0.13           91    0.074    89     98     Inf     0
cour                0.68           79    0.436   461    581     Inf     0
décis               0.32           80    0.203   217    270     Inf     0
déclar              0.33           83    0.202   223    269     Inf     0
demand              0.46           71    0.331   311    441     Inf     0
dev                 0.45           78    0.292   302    389     Inf     0
dh                  0.22          100    0.112   149    149     Inf     0
```

Tabla de términos específicos de la modalidad "Libération" (extraído de un análisis de artículos de prensa de la base de datos Factiva):

```
$Libération
           % Term/Level % Level/Term Global % Level Global t value Prob.
assang            1.068         15.7     1.83   384   2441   -13.3     0
sued              0.225         10.2     0.59    81    792   -11.7     0
extradit          0.167          9.5     0.48    60    634   -11.0     0
mard              0.019          2.8     0.19     7    250   -10.2     0
britann           0.128          9.3     0.37    46    496    -9.8     0
londr             0.108          8.8     0.33    39    444    -9.6     0
dh                0.000          0.0     0.11     0    149    -9.4     0
australien        0.122         10.1     0.33    44    436    -8.7     0
cour              0.200         12.4     0.44    72    581    -8.5     0
julian            0.667         17.3     1.04   240   1385    -8.5     0
fondateur         0.270         13.8     0.53    97    705    -8.4     0
arrêt             0.170         11.8     0.39    61    516    -8.4     0
justic            0.186         12.2     0.41    67    547    -8.4     0
jeud              0.011          2.6     0.12     4    156    -8.1     0
mandat            0.108         10.2     0.29    39    382    -8.1     0

attr(,"title")
[1] "Specific terms by Origin"
```

Los términos son clasificados por su valor test (t value); si el valor test es positivo el término es sobrerrepresentado en la categoría, si es negativo, el término está bajorrepresentado.

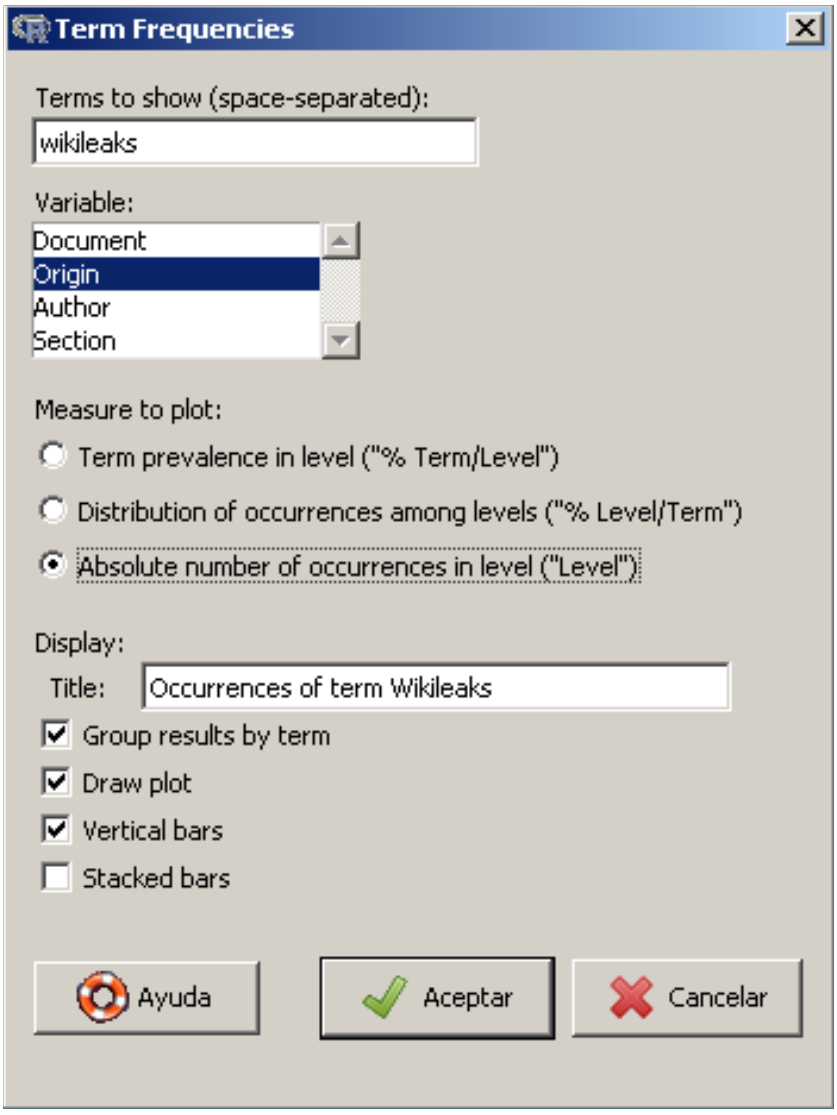
La palabra "assang" representa el 2,53% del conjunto de ocurrencias de las palabras citadas en la modalidad "Agence France Presse" y el 1,068% del conjunto de ocurrencias de las palabras citadas en n la modalidad "Libération"

El 77% de las ocurrencias de la palabra "australien" aparecen en las noticias de "Agence France Presse" y el 10,1% aparecen en las noticias de "Libération"

Análisis de términos concretos.

Se puede seleccionar un término (por ejemplo "wikileaks") y conocer:

a) Los resultados del témino para las distintas modalidades de la variable.

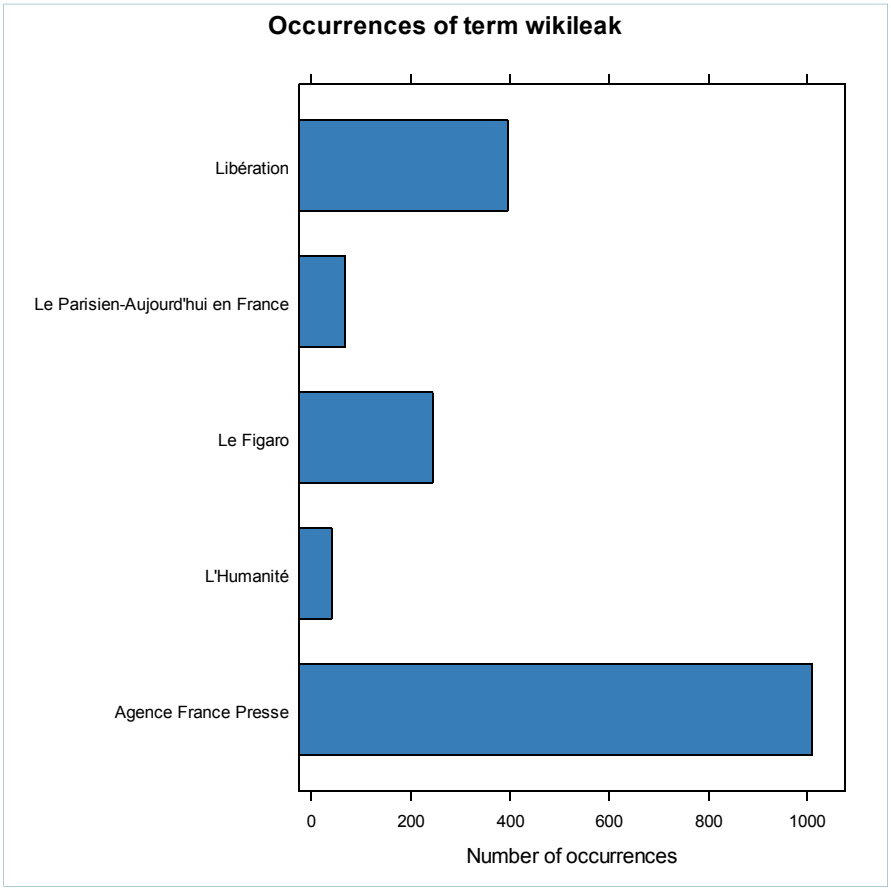


```
> termFreqs
, , wikileak

                                % Term/Level % Level/Term Global % Level Global t value  Prob.
Agence France Presse                    1.50         57.3      1.3  1009   1760     5.7 0.0000
L'Humanité                              0.97          2.4      1.3    42   1760    -2.0 0.0206
Le Figaro                               1.13         13.9      1.3   244   1760    -2.7 0.0035
Le Parisien-Aujourd'hui en France       1.77          3.9      1.3    69   1760     2.3 0.0097
Libération                              1.10         22.5      1.3   396   1760    -4.3 0.0000

attr(,"title")
[1] "Occurrences of term wikileak"
```

En al imagen superior se observa que el término "wikileak" ha sido citado 1760 veces en el conjunto de artículos de todas los medios. Representa el 1,50% del conjunto de ocurrencias de todas las palabras que aparecen en los artículos de "Agence France Presse". El 57,3% del total de ocurrencias de la palabra "wikileak" están en "Agence France Presse".

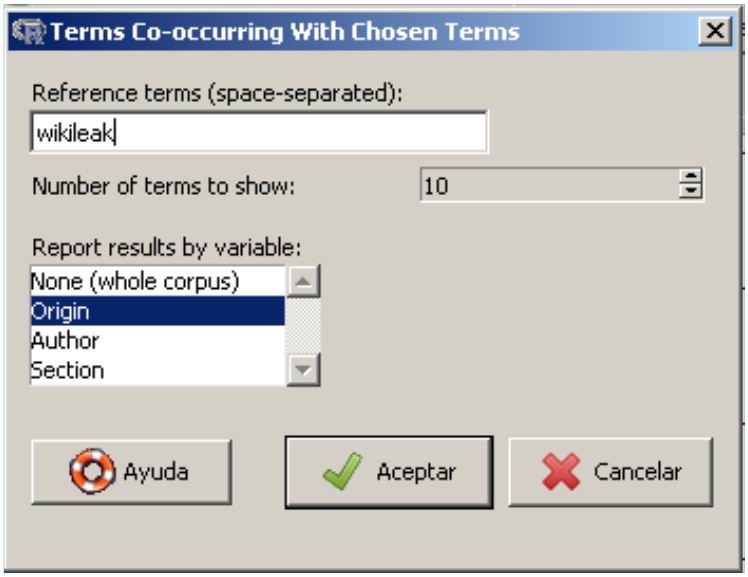
b) Obtener un gráfico para visualizarr:

- El porcentaje del término sobre el conjunto de términos citados en cada modalidad.
- El porcentaje de ocurrencias del término en cada modalidad.
- El número absoluto de ocurrencias del término por modalidad.



Términos que coocurren con otros concretos.

Esta opción permite realizar la búsqueda de los términos que coocurren con un término dado. Es posible incluir varios términos en la ventana para conocer cuáles otros coocurren con éstos, pero no se debe confundir, no se busca la coocurrencia entre ellos.



Se puede restringir a subcorpus de una variable especificada la búsqueda de los términos que coocurren.

Evolución temporal de términos concretos.

Esta opción permite conocer la evolución en el tiempo del número de ocurrencias de términos concretos.

Análisis de correspondencias aplicado a un corpus de documentos

Concepto de similaridad, disimilaridad, distancia y proximidad entre documentos.

Dos documentos son tanto más similares entre sí cuanto mayor es el número de términos que comparten entre ellos.

Si la similaridad la expresamos en una escala entre 0 a 1, dos documentos, A y B, son tanto más similares cuánto más se acerca su valor de similaridad a 1.

El concepto contrario a similaridad es disimilaridad. Se obtiene:

$$Disim = 1 - Sim$$

Al representar sobre un plano el conjunto de documentos de un corpus en función de su similaridad, de modo que cada documento se represente por un punto, obtenemos una nube de puntos; observaremos que hay puntos más próxmos y otros más alejados entre sí. La proximidad entre los puntos es indicativa de la similaridad entre los documentos; aparecen más próximos entre sí cuanto más similares son entre ellos. Hablamos entonces de distancia entre dos puntos como medida de la proximidad-similaridad entre esos dos puntos, de modo que cuánto mayor similitud entre dos documentos menor distancia en el plano hay entre los puntos que los representan. La distancia entre dos puntos del plano se puede expresar como:

$$Distancia = 1 - Sim$$

La estadística descriptiva multidimensional permite calcular las similaridades y distancias entre documentos de un corpus y obtener su representación gráfica.

Análisis de correspondencias.

El análisis de correspondencias o análisis factorial de correspondencias forma parte de la estadística descriptiva multidimensional.

Dado un conjunto de n elementos y m variables, el objetivo del análisis de correspondencias es representar las similitudes entre las categorías de las variables en función de los elementos.

Imaginemos ahora n puntos y cada punto con m coordenadas. El conjunto de todos los puntos situados según sus m coordenadas forma una "nube de puntos" en el espacio m-dimensional. El análisis de correspondencias se emplea para representar la "nube de puntos" en un plano de dos dimensiones

de manera que la nube proyectada sobre dicho plano se ajuste lo máximo posible a la nube real. La solución pasa por proyectar la "nube de puntos" real sobre unos ejes factoriales (el número de ejes factoriales o factores es fijado por el analista), de manera que se respete lo máximo posible la distancia entre los puntos en el espacio m-dimensional. La finalidad del análisis de correspondencias es obtener los principales ejes factoriales o factores que expliquen la mayor variabilidad entre las variables o entre las categorías de una variable analizadas (Navarro Gómez, 1983).

Cada eje factorial o factor contribuye en un porcentaje determinado a explicar la variación entre categorías de una variable (la medida de la variabilidad se denomina "inercia" y se representa en %). El cálculo de las tasas de inercia permite evaluar la calidad global del ajuste. La tasa de inercia asociada a cada eje factorial indica la parte de la inercia total de la nube proyectada sobre este eje.

En el diagrama de correspondencias el eje horizontal (x) representa al primer factor y el eje vertical (y) representa el segundo factor. Cada factor explica, como se ha comentado, un porcentaje de variación de la nube de puntos. Los puntos del diagrama son el resultado de tomar las coordenadas de los mismos sobre el primer y segundo factor. El eje factorial o factor que explica más variabilidad en una categoría de una variable está más relacionado con ella.

Análisis de correspondencias con RTemis.

Se parte de matrices no cuadradas [por ejemplo, la matriz documentos (filas) - términos (columnas). En la matriz documentos-términos los valores de cada fila corresponden a un documento diferente y los valores de cada columna corresponden a un término diferente; en la intersección fila-columna figuran los valores de las ocurrencias del término en el documento (al número de frecuencias de aparición de un término en un documento le denominamos ocurrencia del término).

El análisis factorial de correspondencias se realiza sobre la tabla lexical completa, matriz documentos-términos (*full document-term matrix*), agregando o no variables (*aggregate document-term matrix by variables*)

Las matrices de ocurrencias se normalizan (se convierten los valores absolutos de las ocurrencias en valores entre 0 y 1) y se calculan las distancias "chi-cuadrado". Se fijan los factores a calcular y se realiza el análisis. Independientemente de los factores que se fijen para realizar el análisis, en el diagrama de correspondencias se visualizan los dos factores con mayor tasa de inercia, el factor de tasa de inercia más alta será la dimensión 1 y el siguiente factor en tasa de inercia será la dimensión 2.

Según el análisis realizado obtendremos que cada término, documento o categoría de variable quedará situado en el diagrama de correspondencias mediante dos coordenadas (x, y), la coordenada (x) es el valor que toma el término para el factor 1, o dimensión 1, la coordenada (y) es el valor que toma el término para el factor 2, o dimensión 2.

Como resultado del análisis de correspondencias en RTemis se obtienen dos salidas: el informe del análisis y el diagrama de correspondencias.

En el **informe del análisis de correspondencias** figura:

- el resumen del corpus analizado:

  Correspondence analysis of 795 documents, 85 terms and 25 supplementary variables.

- el resumen de los factores:

  Axes summary:

Axis	*1*	*2*	*3*	*4*	*5*
Inertia (%)	*6.3*	*3.6*	*3.3*	*3.1*	*2.7*
Cumulated inertia (%)	*6.3*	*9.9*	*13.2*	*16.2*	*19.0*

  (La tasa de inercia asociada a cada eje factorial indica la parte de la inercia total de la nube proyectada sobre ese eje. Cada eje factorial o factor contribuye en un porcentaje determinado a explicar la variación entre categorías)

- los términos y documentos más contributivos en la parte negativa o positiva de cada uno de los dos factores o dimensiones:

  Position: Indica la coordenada del término en el factor o dimensión.

  Contribution: Indica la inercia del término o documento en %. Es decir, el porcentaje en el que dicho término contribuye a explicar la variación entre entre categorías.

  Quality: Indica la calidad de un término o documento en %. Es una medida de la calidad de la representación de dicho término o documento en el plano de dos dimensiones.

  Most contributive terms on positive side of axis 1:

	Position	*Contribution (%)*	*Quality (%)*
barroco	*1.89*	*8.6*	*50.3*
siglo_XVII	*1.69*	*7.7*	*46.6*
siglo_XVI	*1.72*	*5.9*	*35.1*
siglo_XVIII	*1.62*	*5.6*	*34.6*
manierismo	*2.18*	*5.3*	*37.0*
romanticismo	*1.78*	*4.4*	*31.2*

neoclasicismo	*2.14*	*3.8*	*32.0*
siglo_XIX	*0.94*	*3.3*	*20.7*
artistas	*1.80*	*3.2*	*15.1*
renacimiento	*1.72*	*3.1*	*15.7*

...

Most contributive documents on positive side of axis 1:

	Position	*Contribution (%)*	*Quality (%)*
542	*2.2*	*2.18*	*47.5*
593	*1.8*	*1.72*	*43.3*
562	*2.0*	*1.68*	*45.7*
571	*1.7*	*1.60*	*42.9*
607	*1.9*	*1.56*	*28.0*
608	*1.9*	*1.56*	*28.0*

...

542

Pinturas Exposición_temática Barroco Manierismo Neoclasicismo Renacimiento Romanticismo Siglo_XIX Siglo_XV Siglo_XVI Siglo_XVII Siglo_XVIII Artistas Grandes_Maestros_/_Old_Masters Museo_de_El_Prado Revisión_colecciones

593

Esculturas Obras_en_papel Pinturas Platería Exposición_itinerante Exposición_temática Barroco Gótico Impresionismo Manierismo Neoclasicismo Renacimiento Rococó Romanticismo Antigüedad_romana Siglo_XIX Siglo_XV Siglo_XVI Siglo_XVII Siglo_XVIII Museo_de_El_Prado Obras_artísticas_de_pequeño_formato Revisión_colecciones

562

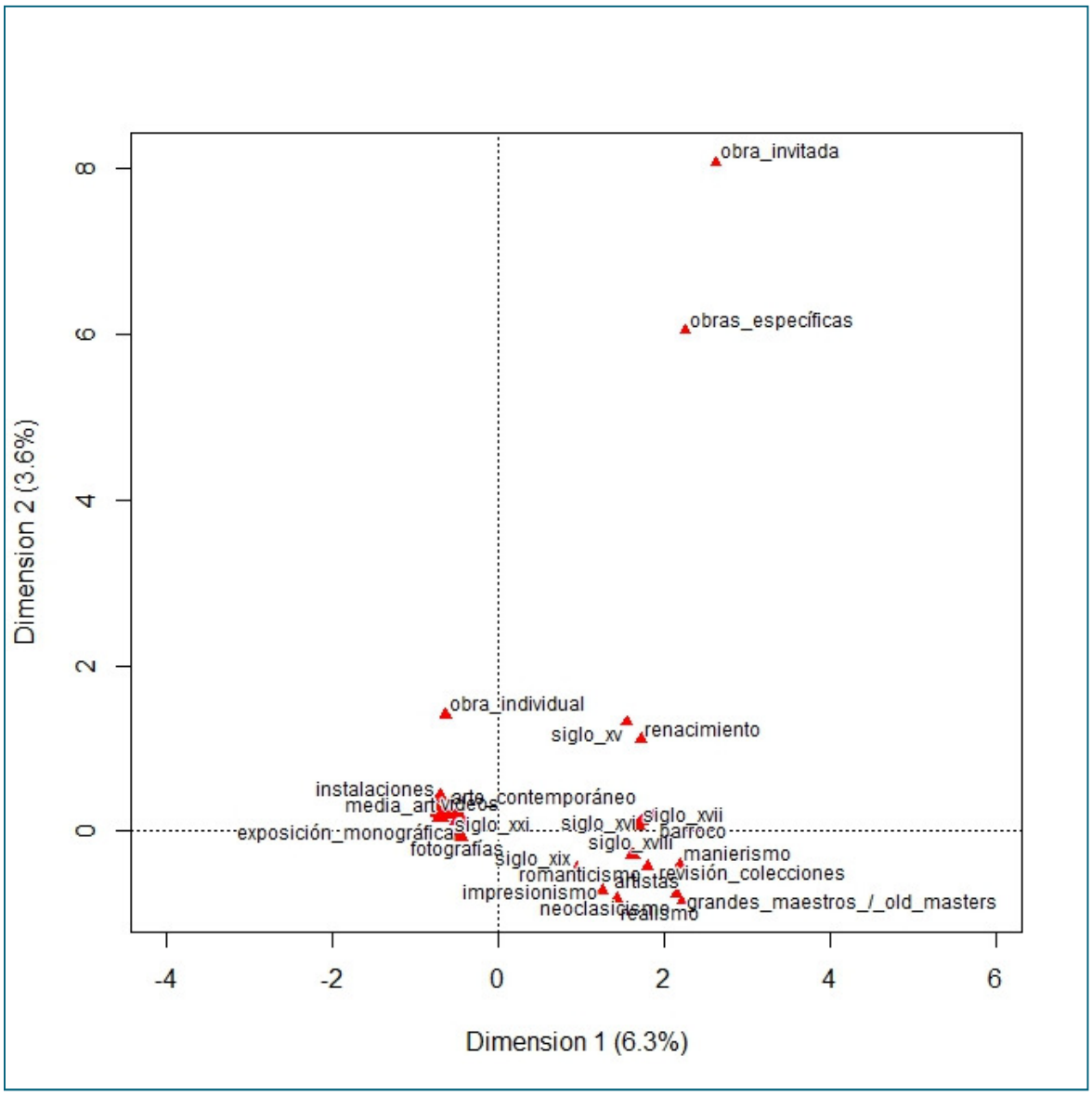
Pinturas Exposición_temática Barroco Impresionismo Manierismo Neoclasicismo Realismo Romanticismo Siglo_XIX Siglo_XVI Siglo_XVII Siglo_XVIII Artistas Evolución_de_la_pintura_española Grandes_Maestros_/_Old_Masters Pintura_española

571

Artes_visuales Bocetos Esculturas Pinturas Relieves Exposición_temática Barroco Impresionismo Manierismo Neoclasicismo Realismo Renacimiento Rococó Romanticismo Siglo_XIX Siglo_XV Siglo_XVI Siglo_XVII Siglo_XVIII Museo_de_El_Prado Obras_artísticas_de_pequeño_formato Revisión_colecciones

...

Interpretación del diagrama de correspondencias



En el diagrama de correspondencias la dimensión 1 representa el primer factor y la dimensión 2 representa el segundo factor. Cada factor explica un porcentaje de variación de la nube de puntos.

El eje factorial o factor que explica más variabilidad en una categoría de una variable está más relacionado con ella.

En el diagrama del ejemplo los factores 1 y 2 tienen una inercia (%) de 6.3 y 3.6 respectivamente. Estos valores son muy bajos por tanto se hace muy difícil interpretar el significado de los mismos. Más fácil es encontrar las similitudes entre los términos según su proximidad en el plano.

Procedimiento de análisis de correspondencias con RTemis.

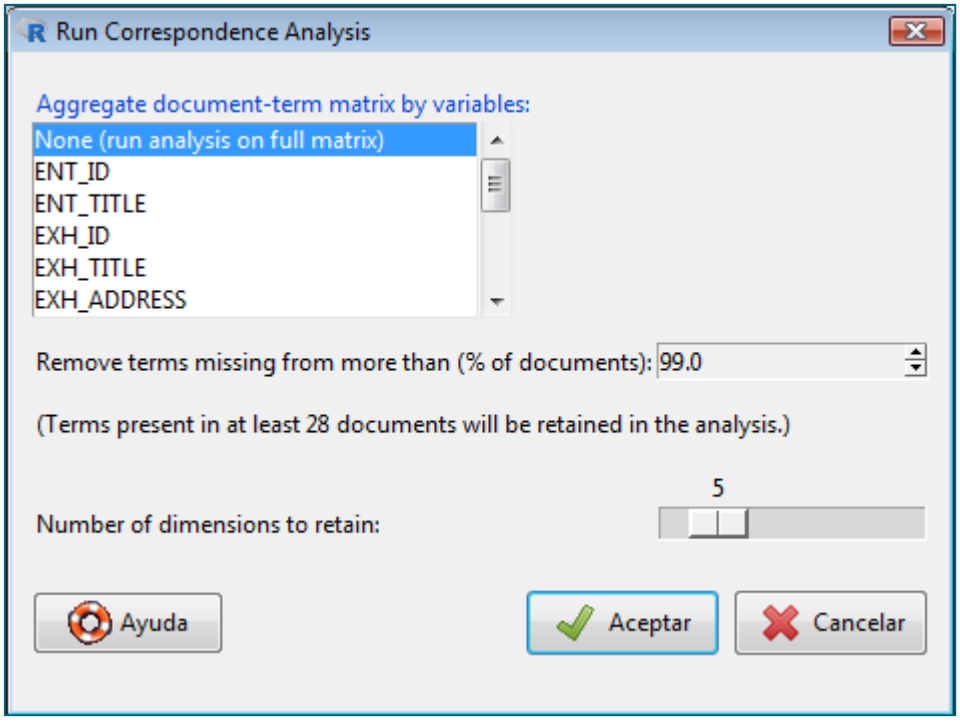
Hay dos formas de realizar el análisis factorial de correspondencias (AFC) con RTemis:

- AFC de la tabla lexical completa sin agregar ninguna variable (*full document-term matrix*)
- AFC de la tabla lexical completa agregando variables (*full document-term matrix by variables*)

I) AFC de la matriz documentos-términos sin agregar ninguna variable

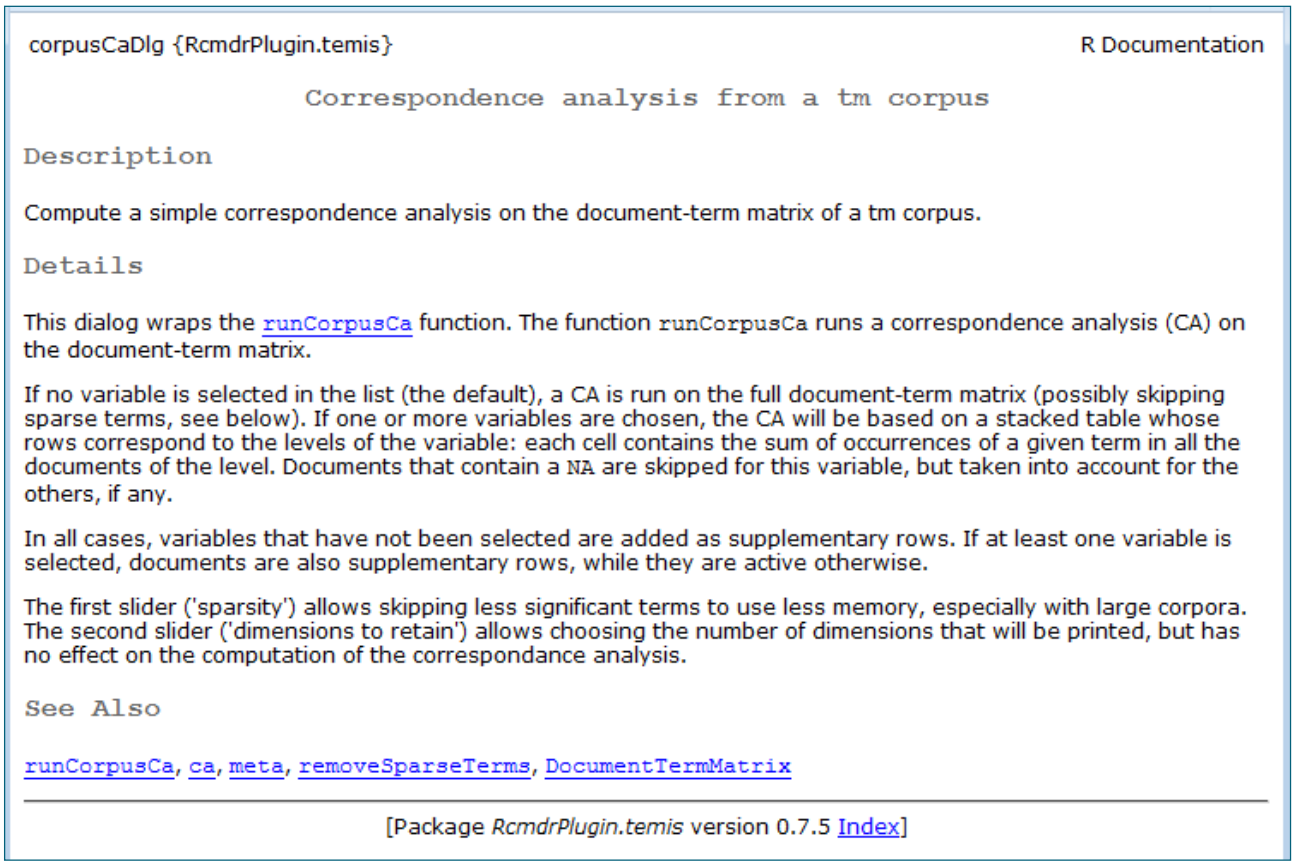
Este análisis de correspondencias permite representar gráficamente los términos coocurrentes en un plano de dos dimensiones (dos ejes factoriales o factores).

El informe del AC (análisis de la posición, contribución y calidad de los términos), el diagrama de correspondencias y otras ayudas, como el contexto de utilización o las coocurrencias de los términos, permiten localizar los temas.



El analista puede eliminar los términos no presentes en más de un porcentaje (%) de documentos a determinar.

También puede fijar el número de ejes factoriales o factores (dimensiones) a estudiar.
El usuario también dispone de una ayuda sobre el análisis a realizar.



corpusCaDlg {RcmdrPlugin.temis} R Documentation

Correspondence analysis from a tm corpus

Description

Compute a simple correspondence analysis on the document-term matrix of a tm corpus.

Details

This dialog wraps the runCorpusCa function. The function runCorpusCa runs a correspondence analysis (CA) on the document-term matrix.

If no variable is selected in the list (the default), a CA is run on the full document-term matrix (possibly skipping sparse terms, see below). If one or more variables are chosen, the CA will be based on a stacked table whose rows correspond to the levels of the variable: each cell contains the sum of occurrences of a given term in all the documents of the level. Documents that contain a NA are skipped for this variable, but taken into account for the others, if any.

In all cases, variables that have not been selected are added as supplementary rows. If at least one variable is selected, documents are also supplementary rows, while they are active otherwise.

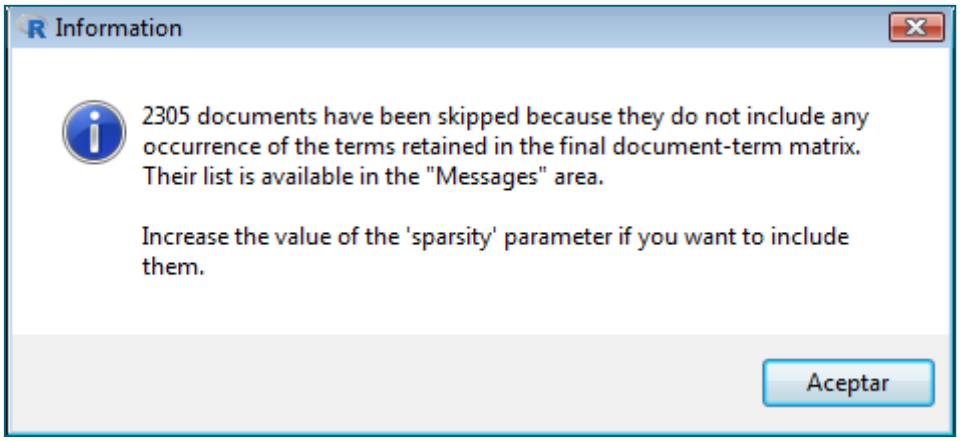
The first slider ('sparsity') allows skipping less significant terms to use less memory, especially with large corpora. The second slider ('dimensions to retain') allows choosing the number of dimensions that will be printed, but has no effect on the computation of the correspondance analysis.

See Also

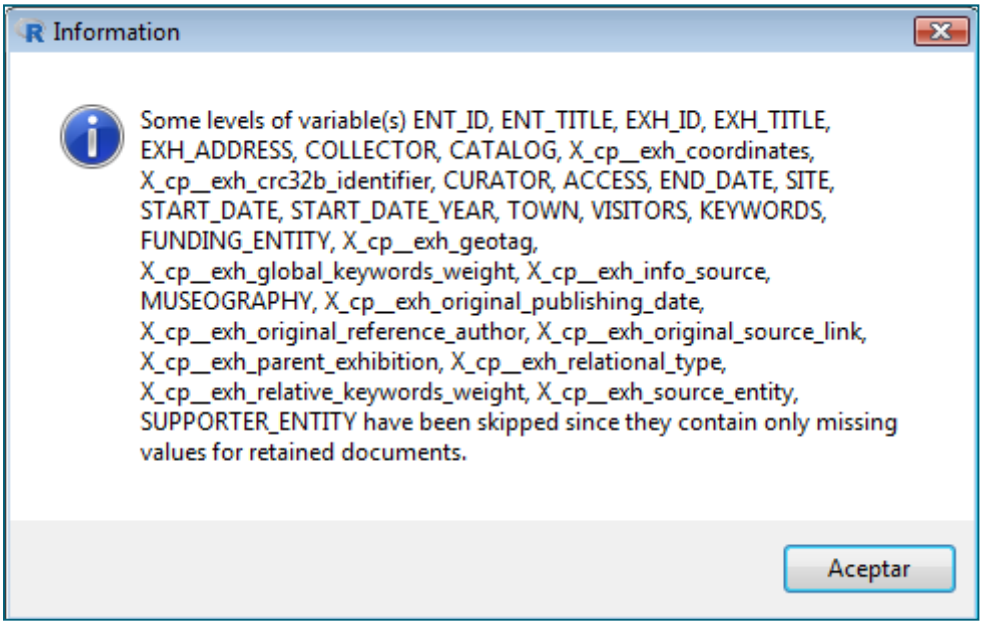
runCorpusCa, ca, meta, removeSparseTerms, DocumentTermMatrix

[Package *RcmdrPlugin.temis* version 0.7.5 Index]

Una vez fijados los parámetros del análisis se acepta y nos aparecen dos cuadros de diálogo:

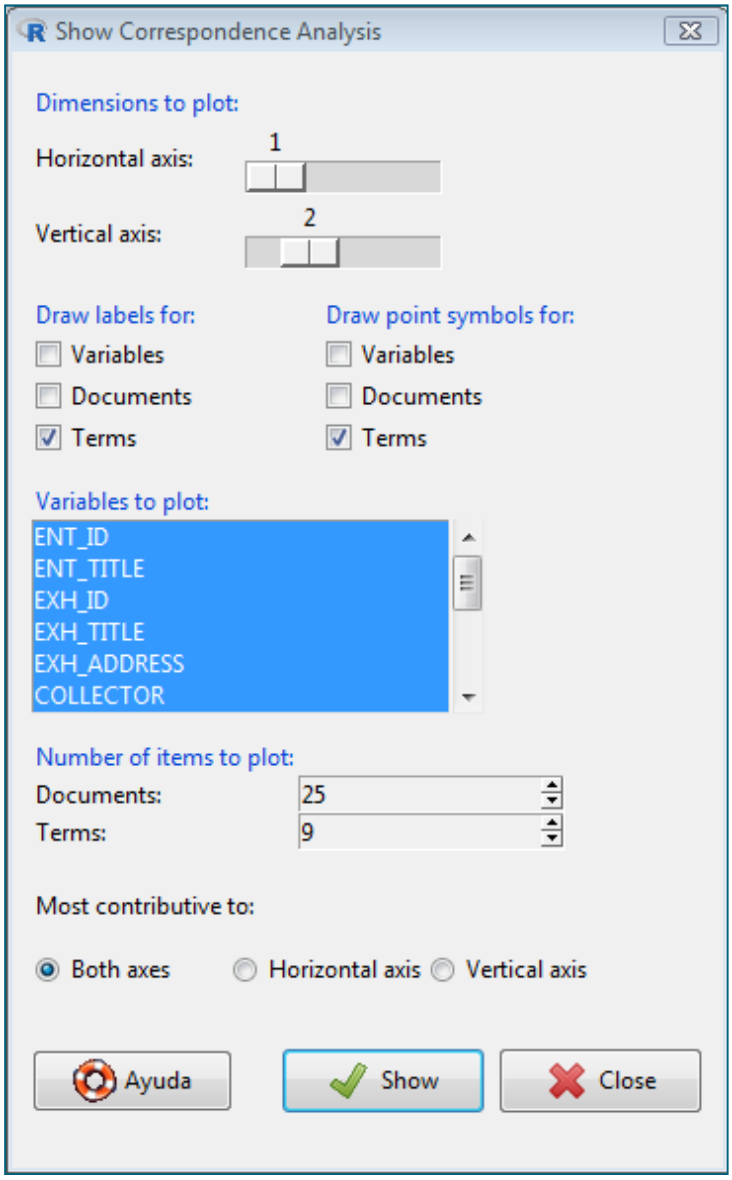


Este cuadro nos informa del número de documentos eliminado del análisis porque no incluyen ningún término de los retenidos. Disminuyendo el valor del porcentaje de documentos fijado en el cuadro inicial eliminaremos menos términos y retendremos más documentos.



Este cuadro nos informa que algunas categorías de las variables han sido eliminadas ya que no contienen valores válidos para los documentos retenidos.

A continuación, una vez aceptados los cuadros de diálogo, nos aparece la ventana donde fijar los parámetros de visualización del diagrama de correspondencias.



Se fijan los elementos a visualizar en el diagrama de correspondencias:

- Los factores o dimensiones en los ejes (x, y).
- Las categorías de las variables.
- Las etiquetas y símbolos de los términos, documentos y variables.
- El número de ítems (documentos o términos) más contributivos a los dos ejes (factores), al eje horizontal o al eje vertical.

Informe del análisis de correspondencias:

Correspondence Analysis

Correspondence analysis of 446 documents, 9 terms and 31 supplementary variables.

Axes summary:

	Axis				
	1	2	3	4	5
Inertia (%)	17	17	17	16	13
Cumulated inertia (%)	17	34	50	66	79

Most contributive terms on negative side of axis 1:

	Position	Contribution (%)	Quality (%)
55293:_[sin_determinar]	-0.38	4.55	6.66
21610:_francisco_de_goya_y_lucientes	-0.38	1.78	3.27
6664:_pablo_ruiz_picasso	-0.38	1.75	2.08
16189:_peter_paul_rubens	-0.38	1.10	1.94
21598:_francisco_de_zurbarán	-0.38	0.93	2.18
21618:_bartolomé_esteban_murillo	-0.38	0.93	2.49
22131:_antonio_saura	-0.38	0.82	1.04
25924:_antoni_tàpies_i_puig	-0.38	0.80	0.96

Most contributive documents on negative side of axis 1:

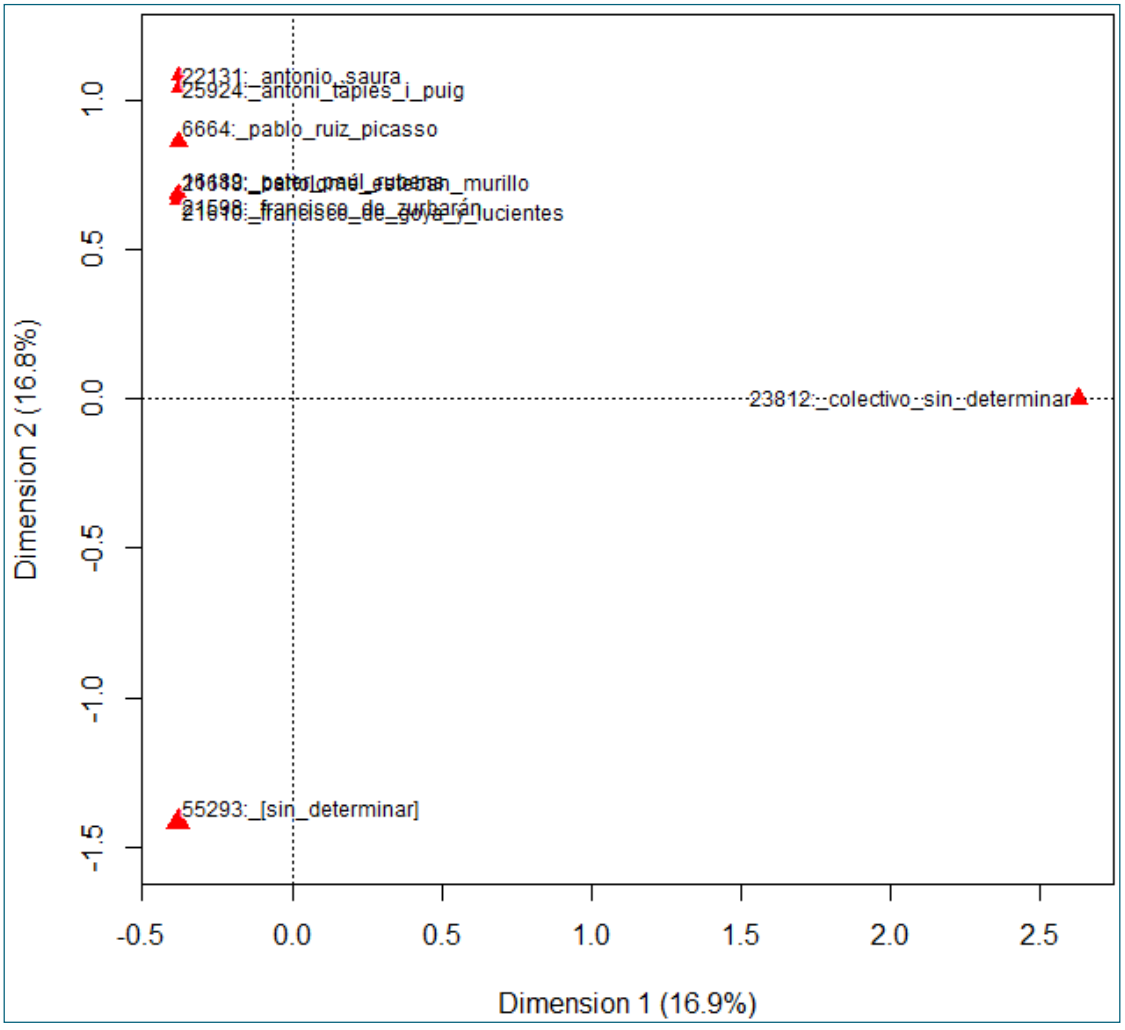
None among the 25 most contributive documents.

Most contributive terms on positive side of axis 1:

	Position	Contribution (%)	Quality (%)
23812:_colectivo_sin_determinar	2.6	87	100

Axes summary
Axis 1 - Negative Side:
Axis 1 - Positive Side:
14
25
168
171
173
174
453
457
505
556
657
682
705
707
711
712
716
774
1019
1065
1067
1071
1086
1087
1091
Axis 1 - Variables

Diagrama de correspondencias.



La ventana dispone de ayuda sobre los parámetros a fijar.

showCorpusCaDlg {RcmdrPlugin.temis} R Documentation

Show a correspondence analysis from a tm corpus

Description

Displays a correspondence analysis previously computed from a tm corpus.

Details

This dialog allows plotting and showing most contributive terms and documents from a previously computed correspondence analysis (see corpusCaDlg). It allows plotting any dimensions of the CA together, showing either documents, terms, or variables set on the corpus using the Text mining->Manage corpus->Set corpus variables menu.

Compared with most correpondence analyses, CAs of a corpus tend to have many points to show Thus, the dialog provides two sliders ("Number of items to plot") allowing to show only a subset of terms, documents, the most contributive to the chosen dimension. These items are the most useful to interpret the axes.

The text window shows the active items most contributive to the chosen axis, together with their position, their contribution to the inertia of the axis ("Contribution"), and the contribution of the axis to their inertia ("Quality of Representation"). (For supplementary variables or documents, depending on the parameters chosen for the CA, absolute contributions are not reported as they do not exist by definition.) The part of total inertia represented by each axis is shown first, but the rest of the window only deals with the selected axis (horizontal or vertical).

The 'Draw point symbols for' checkboxes allow representing documents, terms and variables masses (corresponding to the size of the symbols) and relative contributions (corresponding to the color intensities). See the contrib argument to plotCorpusCa for details.

See Also

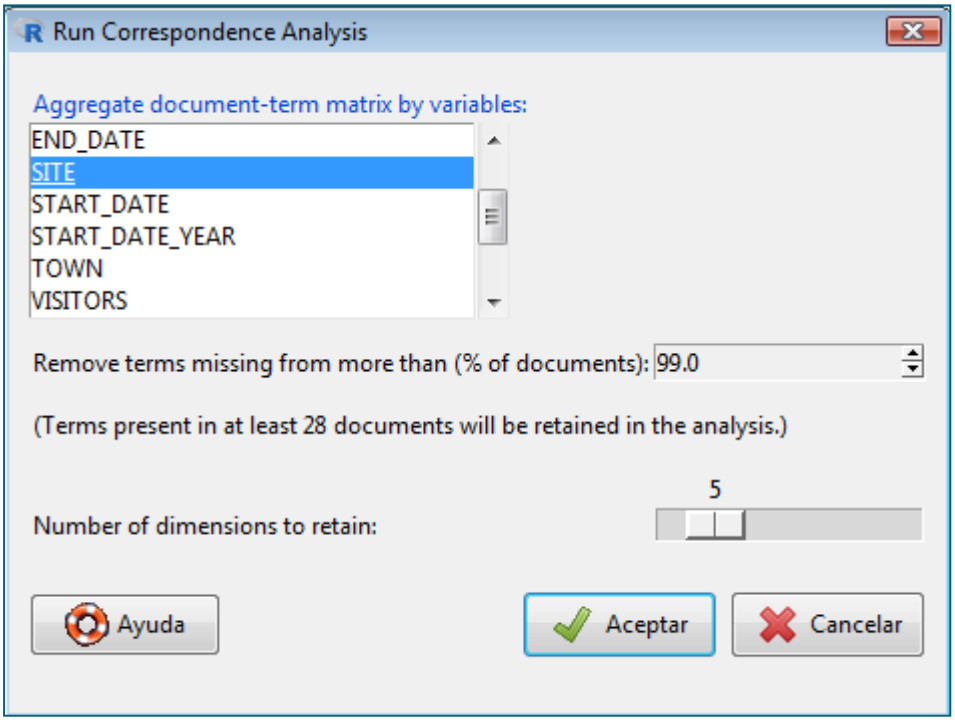
corpusCaDlg, plotCorpusCa, runCorpusCa, ca

[Package *RcmdrPlugin.temis* version 0.7.5 Index]

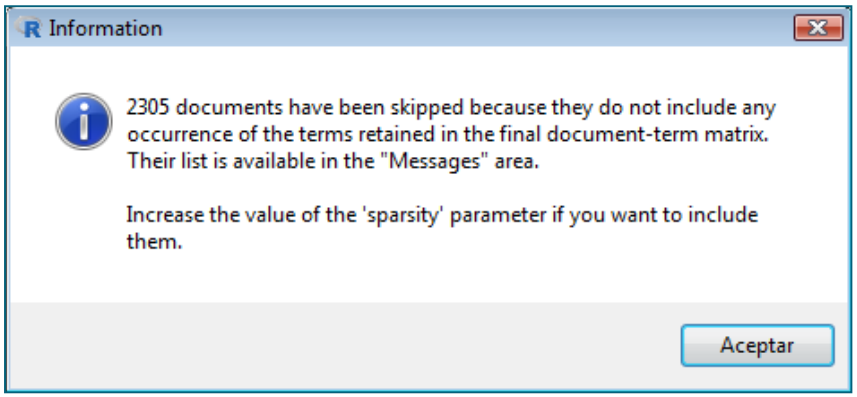
II) AFC de la tabla lexical completa agregando variables (*full document-term matrix by variables*)

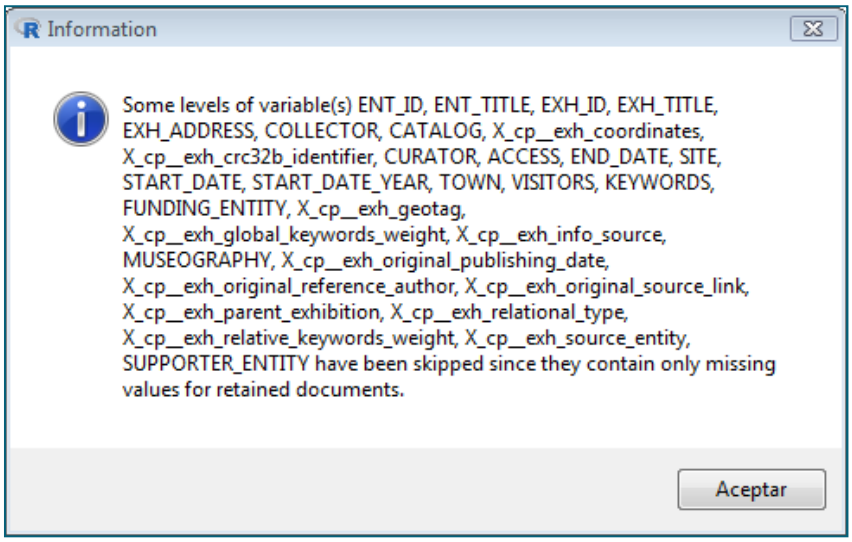
Este análisis de correspondencias permite cruzar el conjunto de términos y las categorías de las variables selecionadas; permite representar gráficamente en un plano de dos dimensiones los términos en función de las categorías de las variables elegidas.

El informe del AC (análisis de la posición, contribución y calidad de los términos) (Minguillón-Campos & Pino-Díaz, 2016), el diagrama de correspondencias y otras ayudas, como el contexto de utilización o las coocurrencias de los términos, permiten localizar los temas.

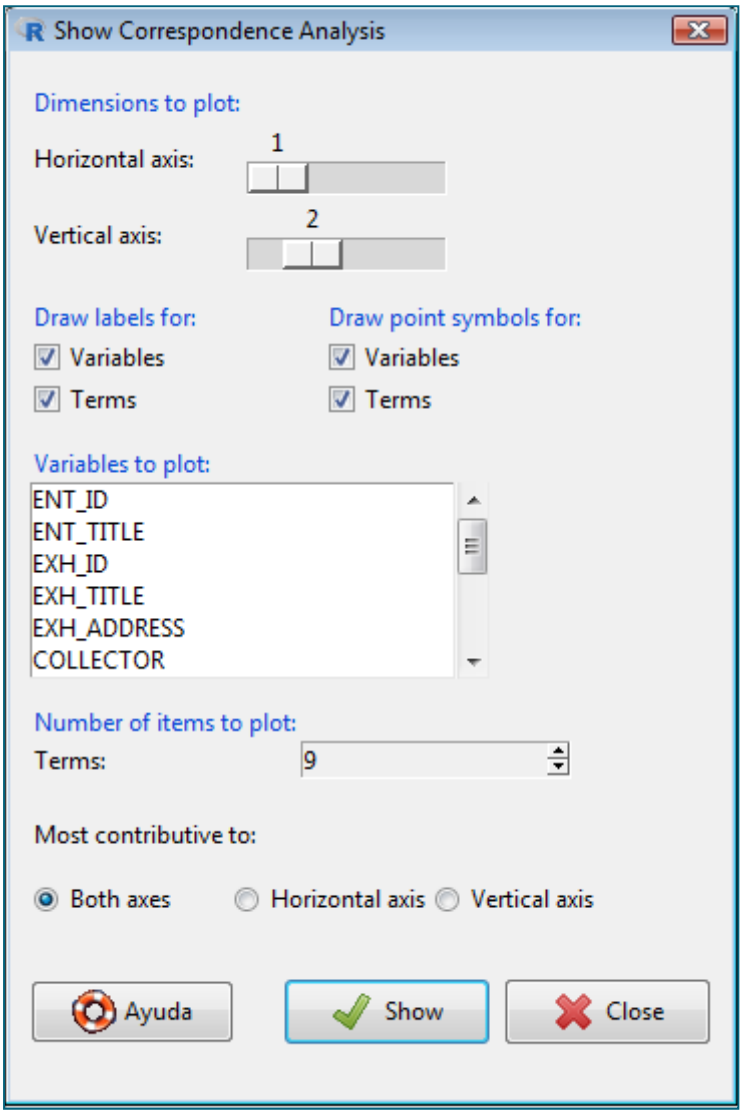


Una vez fijados los parámetros del análisis se acepta y nos aparecen dos cuadros de diálogo:





A continuación, una vez aceptados los cuadros de diálogo, nos aparece la ventana donde fijar los parámetros de visualización del diagrama de correspondencias.



Se fijan los elementos a visualizar en el diagrama de correspondencias:

- Los factores o dimensiones en los ejes (x, y).
- Las categorías de las variables.
- Las etiquetas y símbolos de los términos, documentos y variables.
- El número de ítems (documentos o términos) más contributivos a los dos ejes (factores), al eje horizontal o al eje vertical.

Informe del análisis de correspondencias:

Correspondence Analysis

Correspondence analysis of 1 active variable(s) (aggregating 446 documents), 9 terms and 30 supplementary variable(s).

Axes summary:

	Axis				
	1	2	3	4	5
Inertia (%)	24	23	19	15	7.3
Cumulated inertia (%)	24	47	66	81	88.4

Most contributive terms on negative side of axis 1:

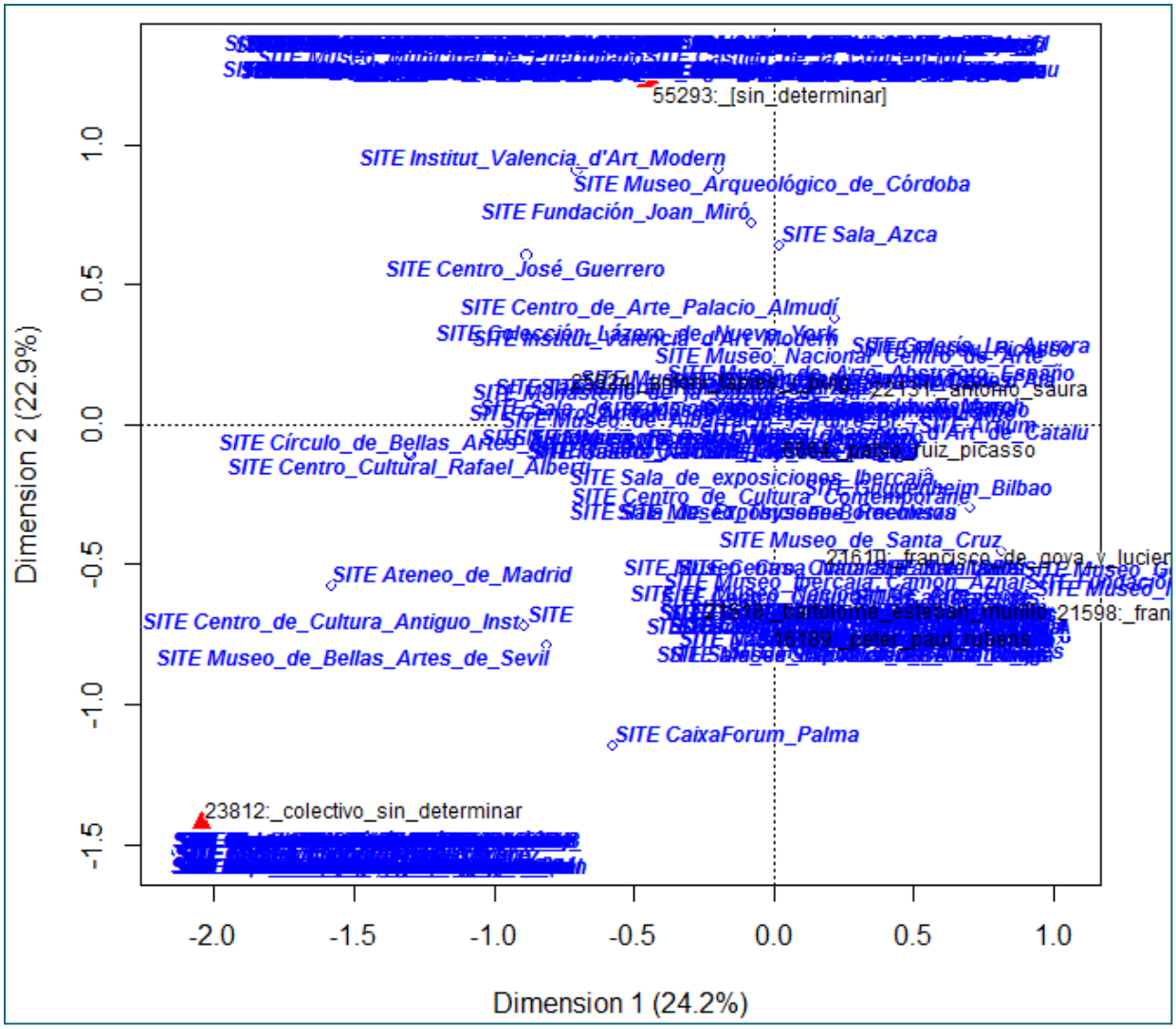
	Position	Contribution (%)	Quality (%)
23812:_colectivo_sin_determinar	-2.05	57.7	67
55293:_[sin_determinar]	-0.45	6.8	11

Active levels on negative side of axis 1:

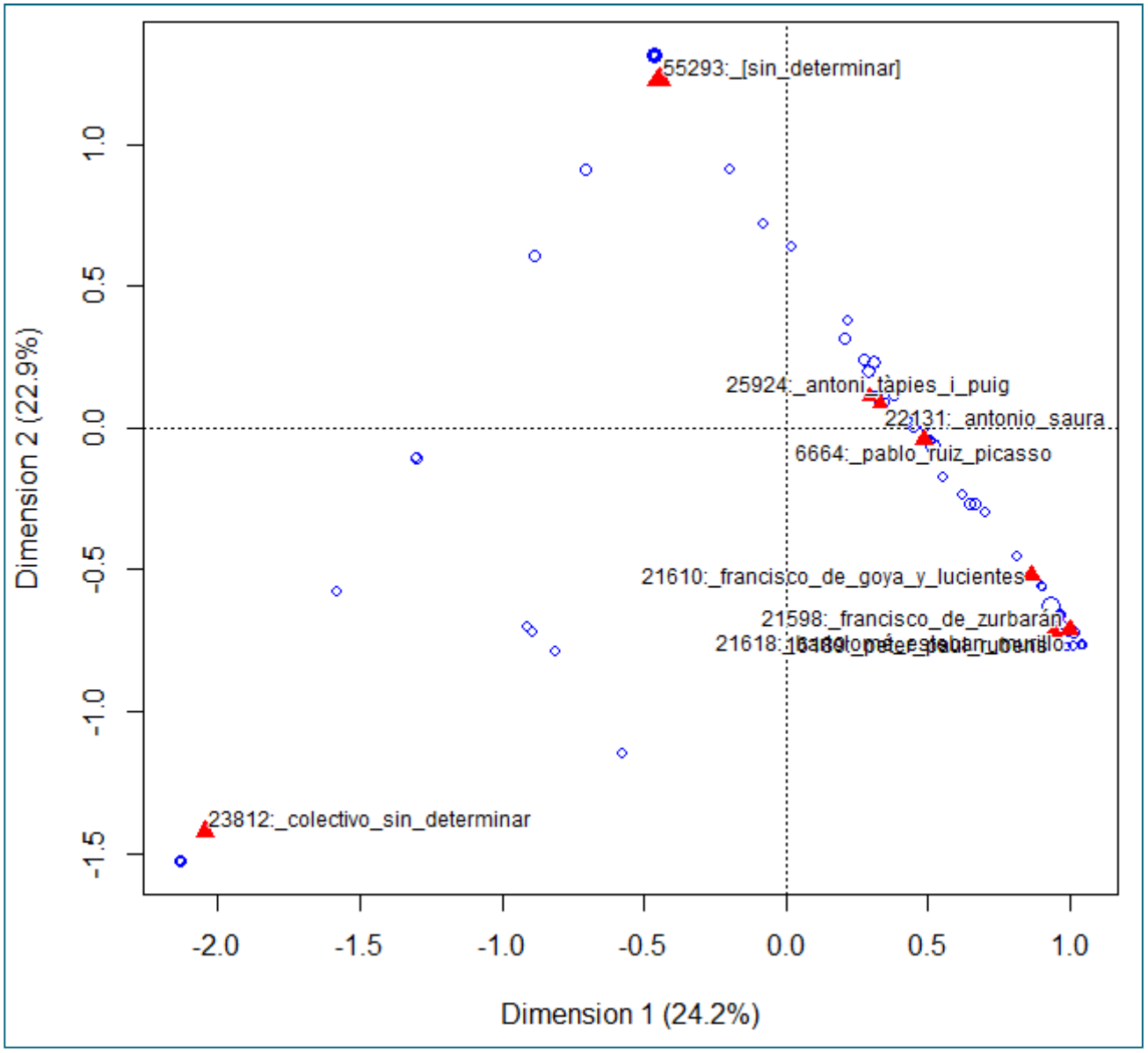
	Position	Contribution (%)	Quality (%)
SITE Centro_de_cultura_Antiguo_Inst	-2.13	3.7476	66.03
SITE Palacio_de_Villena_(Valladolid	-2.13	2.8107	66.03
SITE Museo_Arqueolóxico_do_Castro_d	-2.13	1.8738	66.03
SITE Ateneo_de_Madrid	-1.58	1.5373	86.97
SITE Círculo_de_Bellas_Artes_de_Val	-1.30	1.3915	95.48
SITE Museo_Nacional_de_Artes_Decora	-0.47	0.9857	9.96
SITE Biblioteca_municipal_de_Pasaró	-2.13	0.9369	66.03
SITE Casa_Bardín	-2.13	0.9369	66.03

Axes summary
Axis 1 - Negative Side:
14
25
31
168
171
173
174
453
457
505
556
657
682
705
707
711
712
716
774
1019
1065
1067
1071
1086

Diagrama de correspondencias:



Si volvemos a realizar el AC y no marcamos *draws labels for variables* obtenemos el siguiente diagrama de correspondencias:



Clasificación ascendente jerárquica aplicada a un corpus de documentos

Concepto de similaridad y distancia entre documentos.

Dos documentos son tanto más similares entre sí cuanto mayor es el número de términos que comparten entre ellos. Igualmente dos grupos o clases de documentos son tanto más similares entre sí cuanto mayor es el número de términos que comparten entre ellas. A mayor similitud entre documentos o clases menor distancia entre los mismos.

La clasificación de documentos consiste en agrupar documentos o clases similares.

Concepto de clasificación jerárquica.

La clasificación jerárquica (*hierarchical clustering*) se emplea para constituir grupos de elementos similares (clases o *clusters*). Los datos a partir de los cuales se realiza la clasificación son los valores binarios (0/1), cuantitativos o cualitativos representados de una matriz del tipo elementos (filas) / variables (columnas). En el caso de variables cualitativas se aconseja efectuar previamente un análisis de correspondencias (AC).

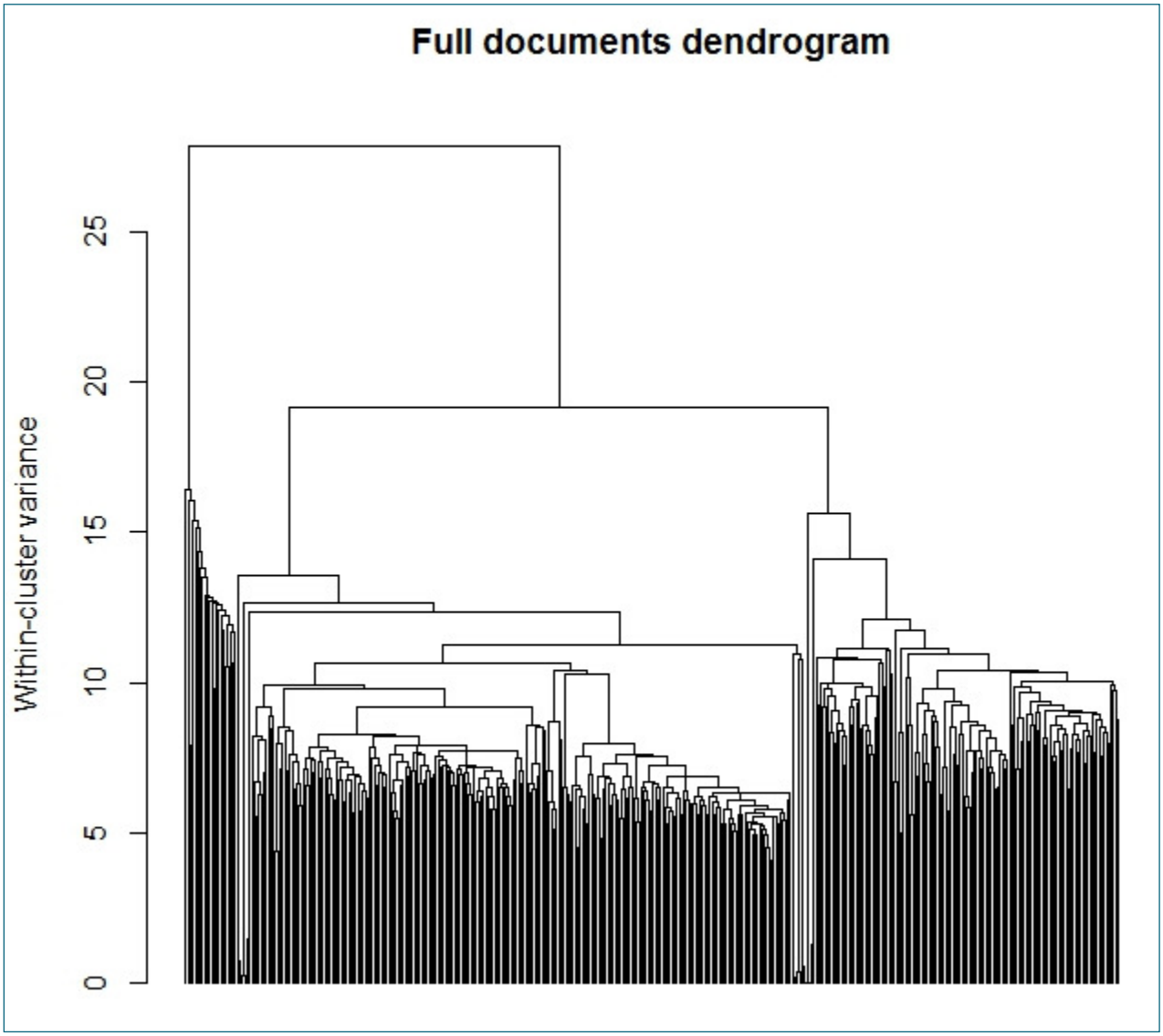
La clasificación jerárquica tiene por objetivo agrupar clases o *clusters* para formar una nueva, o bien separar alguna ya existente para dar origen a otras dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeracióon o división, se minimice alguna distancia o bien se maximice alguna medida de similitud entre los *clusters* (Gutiérrez, González, Torres, & Gallardo, 1994).

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como elementos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los elementos están englobados en un mismo conglomerado o cluster.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como elementos hay.

La representación gráfica de la clasificación jerárquica es el dendrograma.



Clasificación ascendente jerárquica.

Sea ***n*** el conjunto de elementos de la muestra y sea ***K*** el nivel de aglomeración; el número de clases o grupos del nivel de aglomeración ***K = 0***, será ***n*** grupos (se inicia la agrupación de elementos con tantos grupos como elementos hay).

En el siguiente nivel se agruparían aquellos dos elementos que tengan la mayor similitud (o menor distancia), resultando así ***(n – 1)*** grupos; a continuación, y siguiendo con la misma estrategia, se agruparían en el nivel posterior aquellos dos elementos (o *clusters* ya formados) con menor distancia o mayor similitud; de esta forma, en el nivel ***K = L*** tendremos ***(n – L)*** grupos formados.

Si se continúa agrupando de esta forma, se llega al nivel ***K = (n – 1***) en el que sólo hay 1 grupo, ***[n - (n - 1) = 1]***, formado por todos los individuos de la muestra.

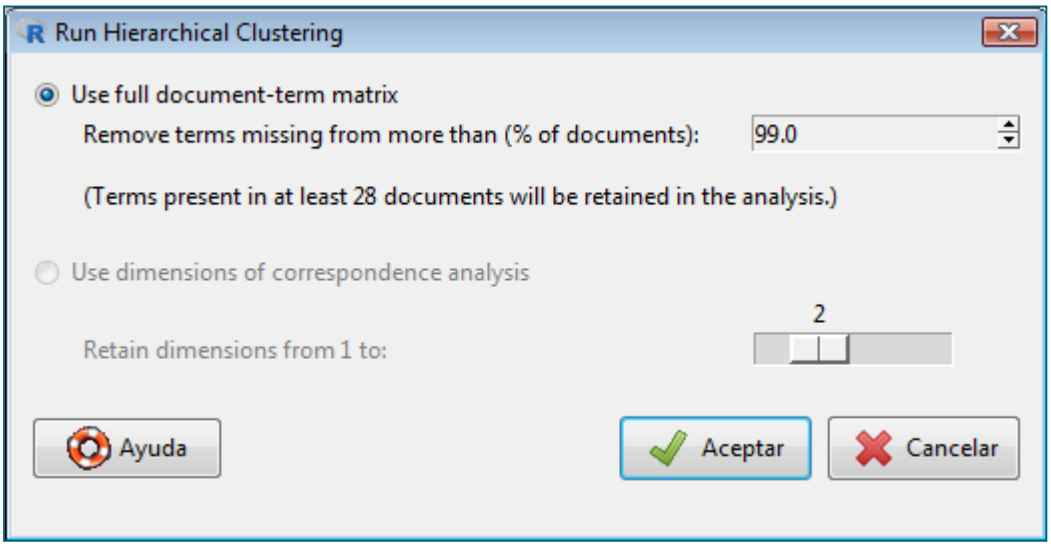
Clasificación ascendente jerárquica con RTemis.

La clasificación ascendente jerárquica en Rtemis se puede realizar sobre la matriz de documentos-términos (*full document-term matrix*) o tabla lexical completa (TLC), o sobre la tabla lexical completa agregando variables (TLA) (*aggregate document-term matrix by variables*). Es también posible realizar la clasificación ascendente jerárquica empleando las dimensiones o factores del análisis de correspondencias.

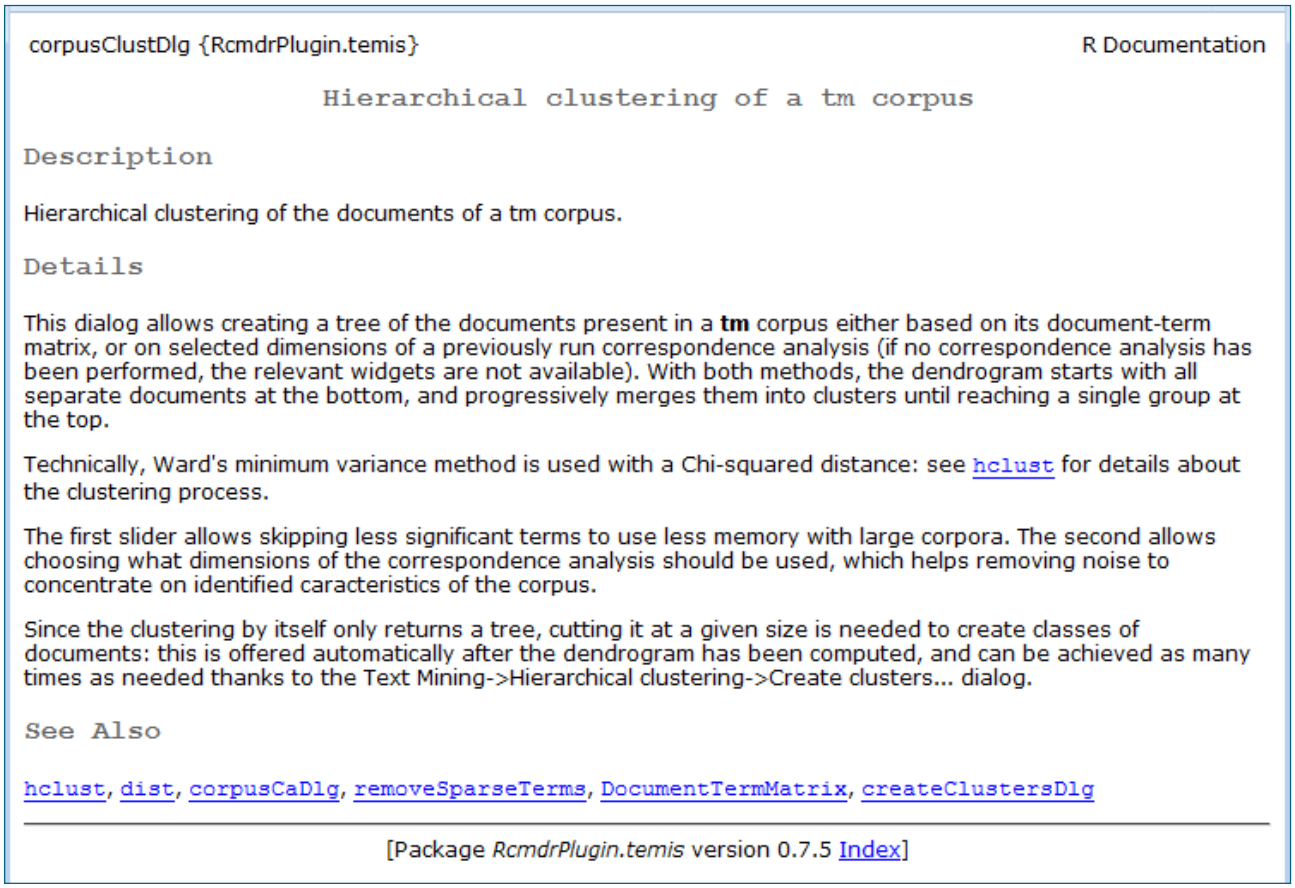
Iniciado la clasificación ascendente jerárquica aparece un cuadro de diálogo donde seleccionar:

a) usar la matriz de documentos-términos (eliminando los términos no presentes en más de un porcentaje (%) de documentos a fijar en el cuadro).

b) usar las dimensiones (factores) del análisis de correspondencias.



El cuadro de diálogo dispone de la opción de Ayuda en la que se explica el método de *Hierarchical clustering* empleado en RTemis:



corpusClustDlg {RcmdrPlugin.temis} R Documentation

Hierarchical clustering of a tm corpus

Description

Hierarchical clustering of the documents of a tm corpus.

Details

This dialog allows creating a tree of the documents present in a **tm** corpus either based on its document-term matrix, or on selected dimensions of a previously run correspondence analysis (if no correspondence analysis has been performed, the relevant widgets are not available). With both methods, the dendrogram starts with all separate documents at the bottom, and progressively merges them into clusters until reaching a single group at the top.

Technically, Ward's minimum variance method is used with a Chi-squared distance: see hclust for details about the clustering process.

The first slider allows skipping less significant terms to use less memory with large corpora. The second allows choosing what dimensions of the correspondence analysis should be used, which helps removing noise to concentrate on identified caracteristics of the corpus.

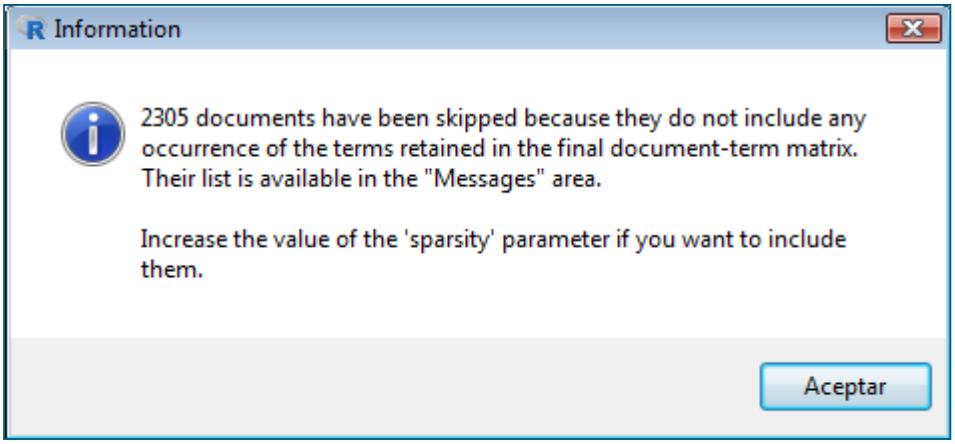
Since the clustering by itself only returns a tree, cutting it at a given size is needed to create classes of documents: this is offered automatically after the dendrogram has been computed, and can be achieved as many times as needed thanks to the Text Mining->Hierarchical clustering->Create clusters... dialog.

See Also

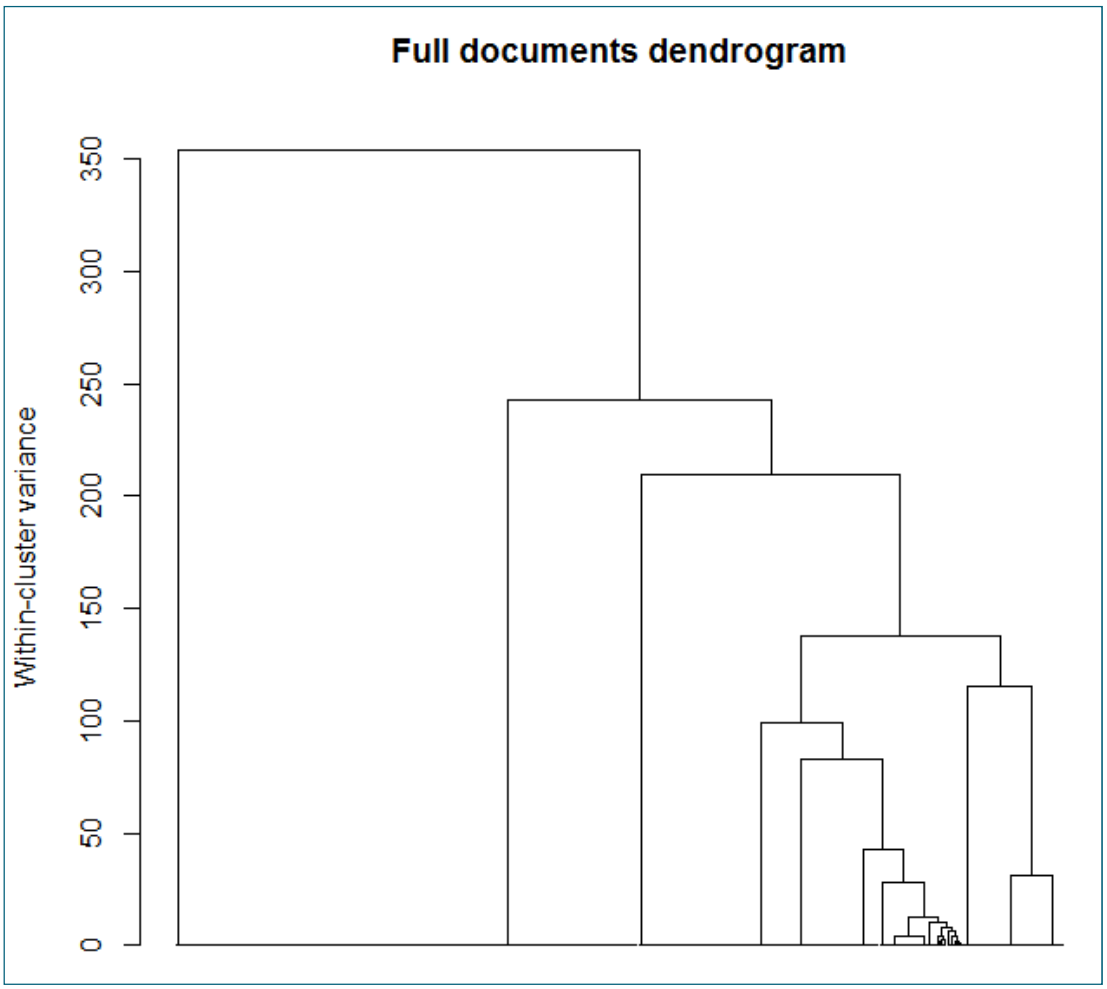
hclust, dist, corpusCaDlg, removeSparseTerms, DocumentTermMatrix, createClustersDlg

[Package *RcmdrPlugin.temis* version 0.7.5 Index]

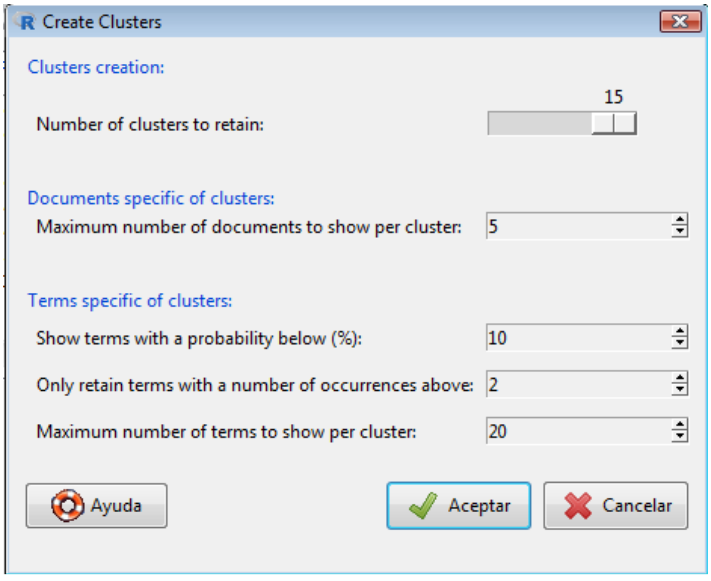
Una vez elegido el conjunto de documentos sobre los que hacer la clasificación jerárquica se acepta y nos aparece una ventana informativa:



Se acepta y aparecen dos ventanas:
1ª) el dendrograma completo antes de crear los *clusters* en el cuadro *Create clusters* (siguiente ventana)



2ª) el cuadro de diálogo *Create clusters*:



Se fijan los siguientes parámetros:

a) El número de *clusters* o clases a retener en el análisis.

b) El número máximo de documentos por clase a obtener en el informe del *Hierarchical clustering*.

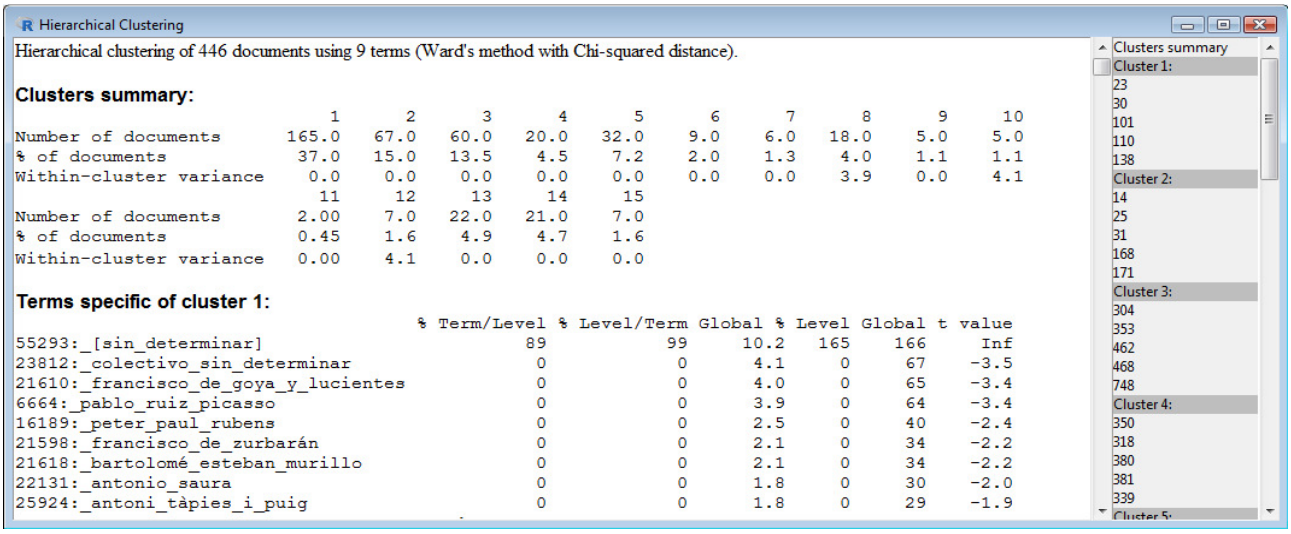
c) Retener términos con una probabilidad por debajo (%) del valor fijado.

d) Retener términos con una ocurrencia por encima del valor fijado.

e) El número máximo de términos a obtener en el informe.

Fijados los parámetros se acepta y se obtiene:

a) el informe de la clasificación jerárquica ascendente:



En el informe nos aparece en primer lugar datos sobre el análisis efectuado:

Hierarchical clustering of 446 documents using 9 terms (Ward's method with Chi-squared distance).

Y a continuación el resumen del análisis:

Clusters summary:

	1	2	3	4	5	6	7	8	9	10
Number of documents	165.0	67.0	60.0	20.0	32.0	9.0	6.0	18.0	5.0	5.0
% of documents	37.0	15.0	13.5	4.5	7.2	2.0	1.3	4.0	1.1	1.1
Within-cluster variance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	0.0	4.1
	11	12	13	14	15					
Number of documents	2.00	7.0	22.0	21.0	7.0					
% of documents	0.45	1.6	4.9	4.7	1.6					
Within-cluster variance	0.00	4.1	0.0	0.0	0.0					

En el sumario, para cada *cluster* o clase figura su número de documentos, el porcentaje de documentos sobre el total analizado y la varianza dentro de la clase.

El informe de cada *cluster* comprende el listado de términos específicos. Para cada término figuran los siguientes datos:

- **% Term/Level** = (%) Nº de ocurrencia del término en la clase sobre el total de las ocurrencias de todos los términos que aparecen en la clase.
- **% Level/Term** = (%) Nº de ocurrencias del término en clase sobre el total de frecuencias del término en el corpus.
- **Global %** = (%) Nº de ocurrencias del término en el conjunto del corpus sobre el total de ocurrencias de todos los términos del corpus.
- **Level** = Número de ocurrencias del término en la clase.
- **Global** = Número de ocurrencias del término en todo el corpus.
- **t value** = Parámetro del valor del test (si este valor es positivo el término está sobrerrepresentado, si es negativo el término está subrepresentado)
- **Prob.** = Probabilidad de obtener el término en el conjunto del corpus.

Terms specific of cluster 11:

	% Term/Level	% Level/Term	Global %	Level	Global
24053:_josé_camarón_y_boronat	5.3	50.0	0.12	1	2
24324:_juan_martínez_montañés	5.3	50.0	0.12	1	2
23579:_alonso_berruguete	5.3	33.3	0.18	1	3
23588:_francisco_herrera_el_viejo	5.3	33.3	0.18	1	3
24318:_guido_reni	5.3	33.3	0.18	1	3
24320:_giovanni_francesco_barbieri,_guercino	5.3	33.3	0.18	1	3
24046:_josé_antolínez	5.3	25.0	0.25	1	4
21598:_francisco_de_zurbarán	10.5	5.9	2.09	2	34
16189:_peter_paul_rubens	10.5	5.0	2.46	2	40
21823:_juan_de_valdés_leal	5.3	14.3	0.43	1	7
24051:_mariano_salvador_maella	5.3	12.5	0.49	1	8

	t value	Prob.
24053:_josé_camarón_y_boronat	2.0	0.023
24324:_juan_martínez_montañés	2.0	0.023
23579:_alonso_berruguete	1.8	0.035
23588:_francisco_herrera_el_viejo	1.8	0.035
24318:_guido_reni	1.8	0.035
24320:_giovanni_francesco_barbieri,_guercino	1.8	0.035
24046:_josé_antolínez	1.7	0.046
21598:_francisco_de_zurbarán	1.6	0.058
16189:_peter_paul_rubens	1.4	0.078
21823:_juan_de_valdés_leal	1.4	0.079
24051:_mariano_salvador_maella	1.3	0.090

Documents specific of cluster 11:

	Chi2 dist. to centroid
659	152.5
306	188.3

659

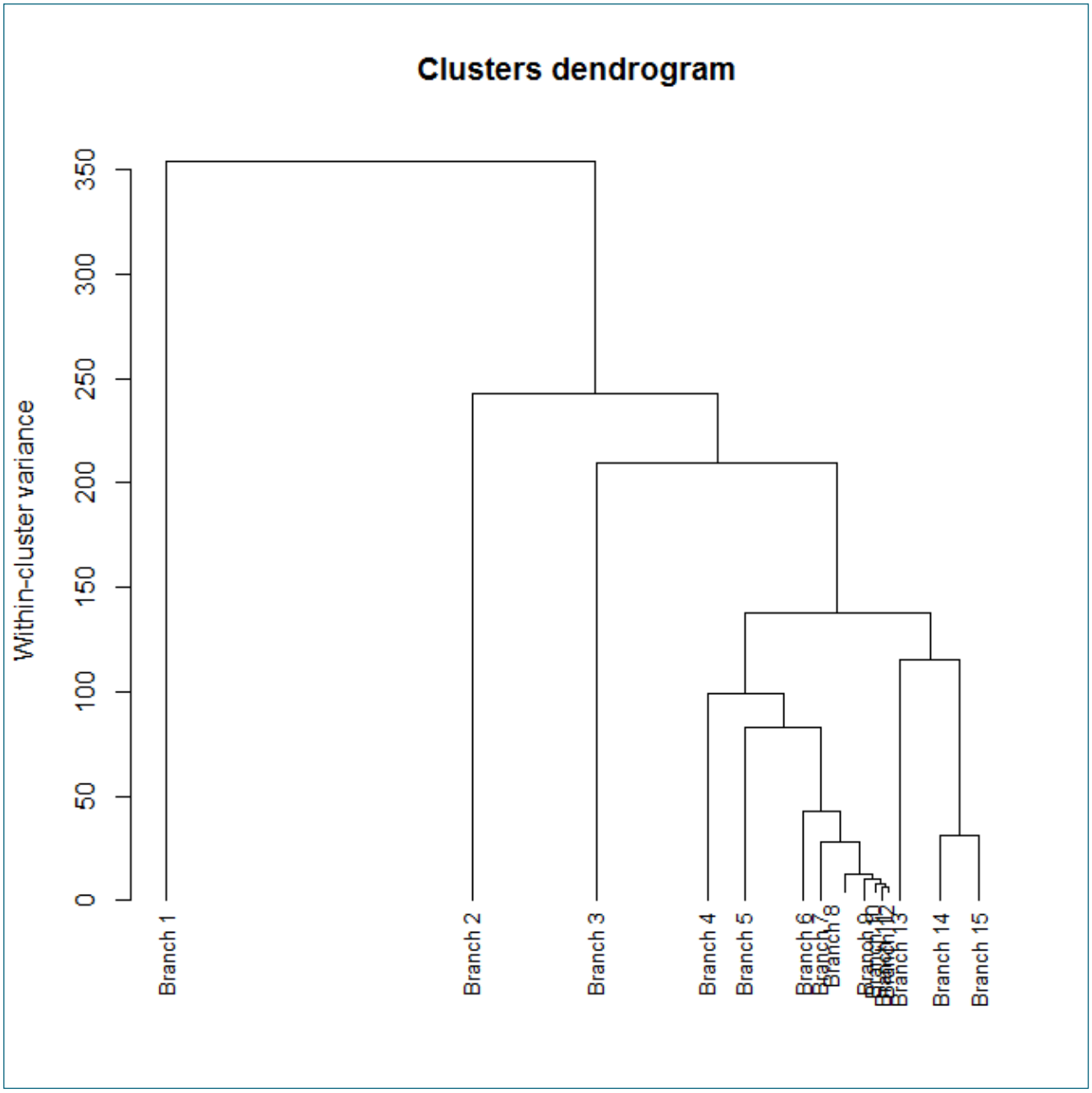
16189:_Peter_Paul_Rubens 21598:_Francisco_de_Zurbarán 21823:_Juan_de_Valdés_Leal 23579:_Alonso_Berruguete 23588:_Francisco_Herrera_el_Viejo 24046:_José_Antolínez 24051:_Mariano_Salvador_Maella 24053:_José_Camarón_y_Boronat 24114:_Juan_Bautista_Romero 2434

306

16189:_Peter_Paul_Rubens 16195:_Claudio_de_Lorena 21598:_Francisco_de_Zurbarán 24317:_Gregorio_Hernández 24318:_Guido_Reni 24320:_Giovanni_Francesco_Barbieri,_Guercino 24323:_Gregorio_Fernández 24324:_Juan_Martínez_Montañés 24325:_Michelangelo_Merisi_da_C

Además, el informe del *cluster* se completa con los documentos que lo componen (con indicación de la distancia de cada documento al centroide del *cluster*).

b) el dendrograma:



La clasificación jerárquica ascendente agrupa a los documentos en *clusters* según su similaridad, esto se visualiza en el dendrograma.

Los documentos similares, con términos comunes, son clasificados en el mismo *cluster.* Dependiendo de los documentos que lo forman, el cluster resultante tiene una varianza interna determinada; cuanto más similares sean los documentos menor será la varianza interna.

Los *clusters* se van uniendo de abajo a arriba (el *cluster* resultante agrupa a los documentos de los *clusters* que se unen), de manera que se observa como

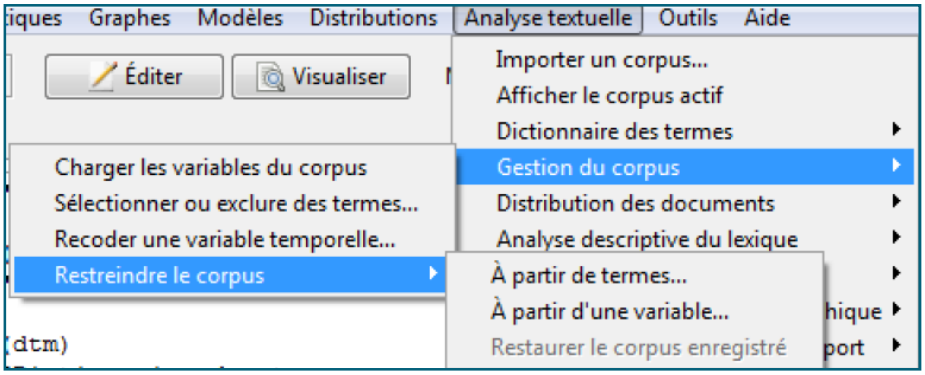
a modo de ramas se unen en otras ramas hasta que quedan dos que se unen arriba en una sola (este tronco común es el corpus completo).

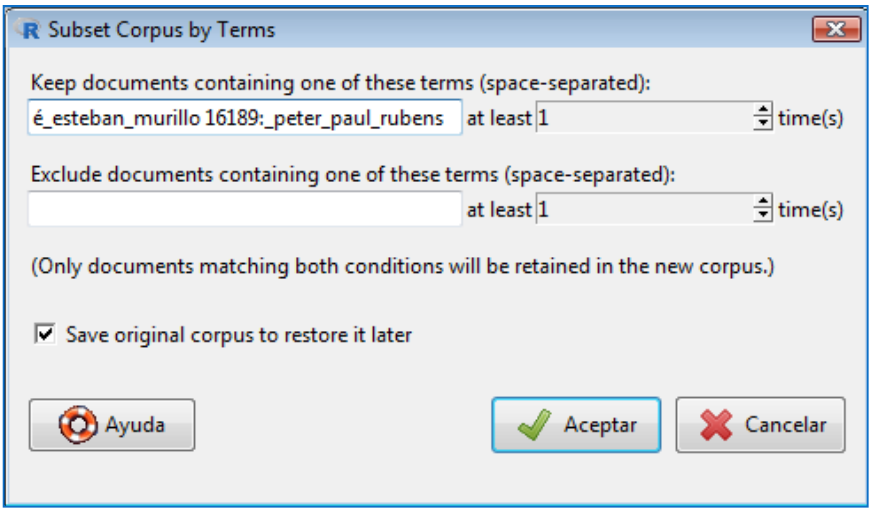
Dos *clusters* son más similares entre sí cuanto más cerca de la base del dendrograma se unen en una sola rama (más cerca de la base supone una más baja varianza dentro del cluster que se forma por la unión de estos dos); conforme ascendemos en el dendrograma la varianza interna en los *clusters* aumenta.

Ayuda a la interpretación del Análisis de Correspondencias y de la Clasificación ascendente jerárquica

Como ayuda a la interpretación de la proximidad entre dos términos en el diagrama de correspondencias se puede analizar el contexto de utilización de ambos términos. Para ello se crea un subcorpus con los documentos que contienen esos términos y se visualiza el conjunto de documentos que lo forman.

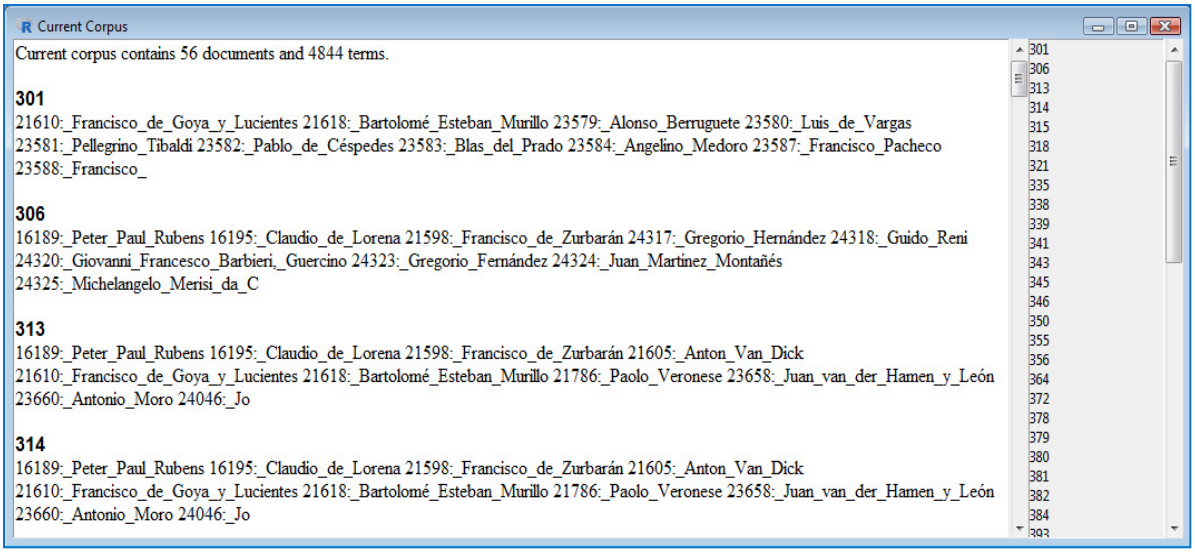
a) Creación de subcorpus a partir de determinados términos o de una variable. Esto permite estudiar el contexto de uso del término elegido.



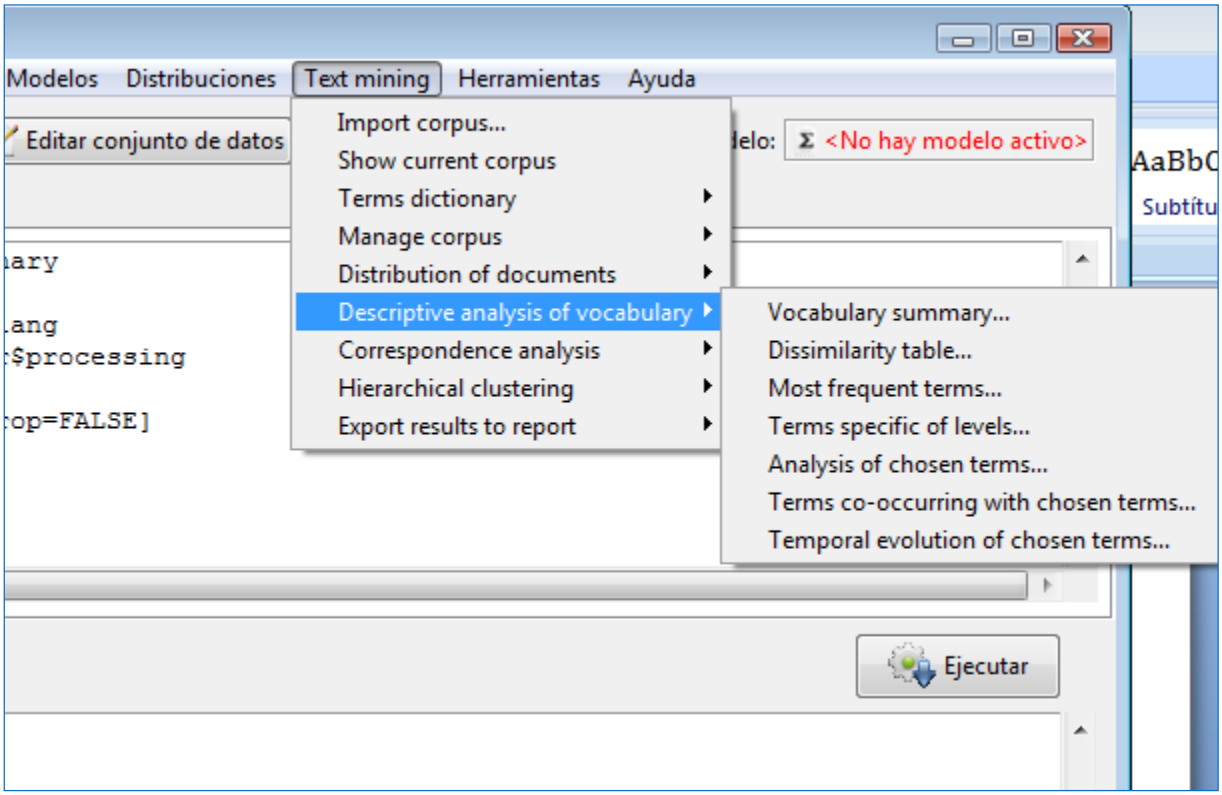



```
> corpus
<<VCorpus>>
Metadata:  corpus specific: 3, document level (indexed): 31
Content:  documents: 56

> dtm
<<DocumentTermMatrix (documents: 56, terms: 4844)>>
Non-/sparse entries: 462/270802
Sparsity           : 100%
Maximal term length: 78
Weighting          : term frequency (tf)
```

b) Visualización del subcorpus



También se pueden interpretar más finamente las proximidades gráficas entre los términos realizándo el análisis descriptivo del subcorpus.



Bibliografía

Bouchet-Valat, M. (2016). *Package RcmdrPlugin.temis.* Recuperado el 25 de Julio de 2016, de https://cran.r-project.org/web/packages/RcmdrPlugin.temis/RcmdrPlugin.temis.pdf

Bouchet-Valat, M., & Bastin, G. (2013). RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R. *The R Journal* , 188-196.

Garnier, B. (2014). *R.TeMiS. Une approche intégrée et libre de l'analyse de données textuelles.* Recuperado el 19 de julio de 2016, de rtemis.hypotheses.org

Gutiérrez, R., González, A., Torres, F., & Gallardo, J. (1994). *Métodos jerárquicos de análisis cluster.* Recuperado el 24 de Julio de 2016, de Técnicas de análisis de datos multivariable. Tratamiento computacional: http://www.ugr.es/~gallardo/pdf/cluster-3.pdf

Minguillón-Campos, J., & Pino-Díaz, J. (2016, Abril 1). *Aplicación de la técnica de Regresión Lineal Simple a la relación Contribution – Quality en el análisis de correspondencias en data mining con R.TeMiS [R Text Mining Solution].* Recuperado el 25 de Julio de 2016, de RIUMA, Repositorio Institucional de la Universidad de Málaga: http://riuma.uma.es/xmlui/handle/10630/11102

Navarro Gómez, M. L. (1983). Aspectos teóricos y una aplicación práctica del análisis factorial de correspondencias. *Estadística española* , 33-59.