APORTACIÓN DEL ANÁLISIS CANÓNICO DE CORRESPONDENCIAS AL ANÁLISIS TEXTUAL

Belchin Adriyanov Kostov

Tutor del proyecto
Mónica Bécue Bertaut

INTRODUCCIÓN

Este proyecto final de carrera forma parte de un proyecto más amplio llamado "Aportación de los métodos de la estadística textual a la búsqueda de información ET-BI". Dicho proyecto se realiza en colaboración con el IDT (Instituto de Derecho y Tecnología de Universidad Autónoma de Barcelona) y Wolters Kluwer España (editora de bases de datos jurídicos "la Ley"), bajo la dirección de la Profesora Mónica Bécue-Bertaut del Departamento de Estadística e Investigación Operativa de la Universidad Politécnica de Cataluña.

La creación de bases de datos jurídicos y jurisprudenciales, que contienen principalmente las sentencias emitidas por los tribunales, ha conducido a crear herramientas múltiples para facilitar su interrogación. Actualmente, se dispone de buscadores que permiten contestar a las consultas de los usuarios a partir de la presencia de determinadas palabras o secuencias de palabras (siguiendo un modelo parecido a las consultas efectuadas desde Google, por ejemplo). El proyecto ET-BI tiene como objetivo estudiar la aportación del análisis de correspondencias y de extensiones como el análisis factorial múltiple para las tablas de contingencias y el análisis canónico de correspondencias (ACC) como herramientas para organizar los corpus previamente a su interrogación.

Este proyecto final de carrera está dedicado a introducir el último método citado (análisis canónico de correspondencias) como herramienta de la Estadística Textual. El trabajo realizado se expone en esta memoria.

El capítulo 1 resume el análisis de correspondencias de tablas léxicas. En el capítulo 2, se propone tener en cuenta información "cerrada" complementaria, mediante el análisis canónico de correspondencias cuyos principios se exponen. El siguiente capítulo, capítulo 3, está dedicado a presentar los packages *R* utilizados, así como las extensiones programadas para utilizar el ACC sobre datos textuales. Finalmente, el capítulo 4 presenta la aplicación del ACC a datos extraídos de una encuesta real.

CAPÍTULO 1

ANÁLISIS CLÁSICO DE RESPUESTAS ABIERTAS

En este capítulo, se presenta una estrategia clásica de análisis estadístico de respuestas abiertas. Dicha estrategia, propuesta por Lebart (Lebart et al., 2000), parte del recuento de las ocurrencias de las diferentes palabras en el conjunto de las respuestas analizadas. Dicho recuento conduce a construir la tabla Respuestas Individuales×Palabras a la cual se puede aplicar métodos estadísticos multidimensionales como el análisis de correspondencias y los métodos de clasificación o métodos más propios del dominio textual como la selección de las palabras características y la extracción de las respuestas modales. Presentamos dichos métodos en este capítulo. En la sección 1.1 se presenta el método del análisis de correspondencias y su aplicación a respuestas abiertas; en la sección 1.2, se exponen los métodos de clasificación. La sección 1.3 habla de la calidad de partición y, finalmente, en la sección 1.4 se explican los problemas específicos en el análisis directo de respuestas abiertas.

1.1. Análisis de correspondencias aplicado a un corpus de respuestas abiertas

1.1.1. Análisis de Correspondencias

Se puede atribuir el método de análisis de correspondencias tal como se emplea hoy a J.P Benzécri y Brigitte Escofier (ver Escofier & Pagès, 1990, para más información). En Lebart et al (2000), se ofrece una exposición del método orientada al análisis textual.

El análisis de correspondencias simple (AC) permite describir la relación entre dos variables categóricas. Las unidades estadísticas (o individuos) de una muestra están descritas por los valores tomados en dos variables categóricas. El AC representa en un espacio de pequeña dimensión las asociaciones y repulsiones entre las categorías de las dos variables. Así, este método permite estudiar e interpretar, por un lado, las similitudes entre categorías de una misma variable y, por otro, las relaciones entre las categorías de ambas variables.

El análisis de correspondencias se utiliza también para estudiar tablas de frecuencia. En este caso las unidades estadísticas, o individuos (en filas) de una muestra están descritas por la frecuencia de una serie de eventos (en columnas).

1.1.2. Análisis de correspondencias: herramienta de comparación de perfiles léxicos

Se considera una serie de documentos descritos por la frecuencia de las diversas palabras en cada uno de ellos. El análisis de correspondencias se puede aplicar con provecho a la tabla documentos $\times$ palabras.

El AC proporciona una descripción de las relaciones entre palabras y documentos mediante la comparación de los perfiles-columna por una parte, y de los perfiles-fila por otra. Opera a partir de la definición de una distancia entre perfiles-columna y entre perfiles-fila. El principio seguido para definir la distancia -llamada distancia de chi-dos - es el de la equivalencia distribucional. Por distribución de una palabra, se entiende el conjunto de todos los contextos posibles. El AC sintetiza las características distribucionales de las palabras.

Dicha síntesis conduce a una representación simultánea de las proximidades entre perfiles-textos por una parte, y perfiles-palabras por otra, es decir, una representación esquemática de la información contenida en la tabla de frecuencias. Para ello, el método busca la mejor representación de las palabras y de los documentos en un espacio de dimensión reducido, pero conservando lo mejor posible las distancias, es decir, la mayor parte de la información contenida en la tabla.

La representación visual obtenida permite efectuar una comparación de los perfiles-palabras (distribución de las palabras en los distintos documentos) por una parte, y de los perfiles-documento (frecuencias relativas con la cual cada documento utiliza cada una de las palabras), por otra parte.

Se puede utilizar una representación en un espacio de dimensión mayor que dos, estudiando de forma sucesiva varios planos. De hecho, uno de los resultados proporcionados por el propio método es una medición de la validez de la representación obtenida según la dimensión conservada. Para dicha medición, se utiliza la varianza (o inercia) de la tabla original, y se calcula el porcentaje de varianza conservada por cada eje.

1.2. Estrategia de análisis combinando AC y clasificación

Una estrategia clásica y provechosa para tratar con los datos de las encuestas es combinar dos métodos complementarios que son el análisis de correspondencias y la clasificación.

La clasificación permite agrupar los individuos en los ejes principales a partir de sus coordenadas y, por lo tanto, se resumen los resultados vinculados por los ejes. La eliminación de los últimos ejes ayuda a filtrar las fluctuaciones aleatorias que podrían

enmascarar las características importantes. El uso previo de los métodos de ejes principales como análisis de correspondencias es importante para la clasificación. Estos métodos proporcionan una protección eficaz frente a la inestabilidad de los métodos de clasificación respecto a la selección de muestras (pequeños cambios en los individuos podrían trasmitir grandes diferencias en el resultado de partición).

Las clases se pueden representar sobre el mismo gráfico resultante del análisis de correspondencias. Esto permite analizar conjuntamente los resultados de los dos métodos desde un punto de vista provechoso y fácil de interpretar. Las clases se pueden describir fácilmente mediante las características de los individuos pertenecientes a las mismas.

1.3. Métodos de Clasificación

El objetivo de una clasificación es reagrupar las unidades u observaciones en clases homogéneas. Para hacer esto, se calculan las distancias entre las observaciones y se agrupan en clases en función de sus proximidades, determinadas según la distancia escogida, cuya elección determina los resultados. Escoger una distancia es escoger un punto de vista.

Los métodos de clasificación son muy numerosos. Se pueden dividir, principalmente, en métodos de clasificación jerárquica y de clasificación por partición directa. A continuación, se presentan cuatro métodos de clasificación: un método de clasificación jerárquica y tres métodos de partición directa: *k-MEANS, PAM* y *CLARA*. Los dos primeros métodos están disponibles en *SPAD* y los otros, así como los primeros, en el paquete *R* llamado *CLUSTER*. Una descripción más completa se puede encontrar en Lebart et al. (2000) y Hastie et al. (2001).

1.3.1. Métodos jerárquicos

Los algoritmos de clasificación jerárquica pueden ser de dos tipos: ascendente y descendente. En los primeros se parte de alto número de pequeñas clases que son gradualmente unidas en un número menor de clases mayores (normalmente se empieza considerando cada individuo como una clase). Por el contrario, en los algoritmos de clasificación jerárquica descendente se parte de un pequeño número de clases numerosas que van dividiendo en un mayor número de clases más reducidas. El algoritmo de clasificación jerárquica ascendente, el que se usa habitualmente, es el siguiente:

- **E** = conjunto de objetos a clasificar
- Calcular la matriz de distancias de **E** en **D**
- Encontrar los dos elementos más próximos (*a,b*) en **D**
- Formar *h*= *a* agregado con *b*
- Actualizar **E** = **E** – {*a,b*}+{*h*}
- Actualizar la matriz de distancias de **E** en **D**

En la clasificación jerárquica hay varios métodos para calcular las distancias entre las observaciones: salto mínimo, diámetro, distancia media e índice de Ward. El último es muy usado cuando la clasificación opera a partir de los ejes factoriales; en efecto, descompone también la inercia lo que permite una interpretación conjunto con los ejes factoriales.

1.3.2. Métodos por partición directa

Los algoritmos de clasificación por partición directa se pueden resumir así:

- Se extraen al azar unos individuos que jugarán el papel de centros provisionales de las clases.
- Se asigna cada uno de los individuos al centro provisional más próximo. Se construye así una partición del conjunto de los individuos.
- Se calculan nuevos centros provisionales que son ahora los “centroides” (centros de gravedad, por ejemplo) de las clases que se acaban de obtener, y se reitera el proceso.

En lo que sigue, se privilegia la clasificación empleada en complemento de un análisis factorial, es decir, tomando como variables los ejes factoriales. Por lo tanto, sólo se considera la distancia euclídea clásica aunque lo expuesto se pueda fácilmente extender a otras distancias.

1.3.2.1. Método k-Means

El algoritmo conocido con el nombre de *k-means* es un algoritmo de partición directa. El número de clases a obtener se debe fijarse a priori. El algoritmo es el siguiente:

1. Tomar k “centroides” iniciales (al azar o por elección determinista)
2. Asignar cada individuo a la clase del centroide más cercano
3. Calcular el nuevo centroide de las clases
4. Repetir los pasos 2 y 3 hasta que no se mejore la función escogida. En muchos casos, se escoge la siguiente función objetivo

$$Min\left(\frac{InerciaIntra}{InerciaTotal}\right)$$

1.3.2.2. PAM (Partition Around Medoids)

La idea de partida de este método es considerar que cada clase está representada por una observación o *k-medoid,* considerada la más apropiada para representar la clase. Se decide a priori el número *k* de clases a formar. Las *k* clases se construyen asignando cada observación al *medoid* más cercano. Se busca mejorar la partición inicial mediante iteraciones en las cuales se va intercambiando una observación normal y un *medoid*. Si se mejora la calidad de la partición, la observación no representativa pasa a ser un *medoid*. El criterio utilizado es la minimización de la suma de distancias entre observaciones y *medoids.*

Se puede resumir el algoritmo de la siguiente manera:

- Escoger *k* representantes $M_1, M_2,\ldots M_k$
- Escoger al azar un representante M_r y otra observación (no representante) O_j
- Calcular la calidad de la nueva partición si se intercambian M_r y O_j
- Intercambiar M_r y O_j si la calidad es superior
- Volver al primer paso hasta que se estabilice la partición

El algoritmo *PAM* es más robusto que *k-means*, pero es de mayor complejidad algorítmica. Por esta razón cuando el tamaño de muestra es grande, no se opera directamente sobre la totalidad de la muestra, si no que se segmenta dicha muestra en varias submuestras. Se obtiene así el algoritmo *CLARA* presentado en la siguiente sección.

1.3.2.3. CLARA (Clustering Large Applications)

Se efectúa una búsqueda local de los representantes a partir de varias muestras del conjunto de datos (muestra total). Cada vez, se aplica el algoritmo de *PAM* y al final se conserva la mejor muestra.

Los pasos que sigue el algoritmo son los siguientes:

I. Extraer una muestra de tamaño *s*
II. Dividir en *k* grupos la muestra aplicando el algoritmo de PAM
III. Calcular la calidad de la partición a partir de la suma de disimilaridades intra grupos. El objetivo es obtener la suma mínima.
IV. Si la calidad es mejor que la de las particiones anteriores, memorizar la partición
V. Repetir los pasos 1-4 tantas veces como número de muestras extraídas
VI. Asignar cada observación de la muestra total al *medoid* más próximo

1.4. Calidad de la partición

1.4.1 Enfoque SILHOUETTE

"Ancho de la silhouette" es un indicador numérico de bondad de clasificación. Se considera A la clase a la cual pertenece el individuo x_i. C indica cualquier clase a la cual no pertenece este individuo. Se definen:

$$a(x_i) = \frac{1}{|A|-1} \sum_{\substack{i \in A \\ i \neq i'}} D(x_i, x_{i'}) \qquad D(x_i, C) = \frac{1}{|C|} \sum_{x_{i'} \in C} D(x_i, x_{i'})$$

Se nota

$$b(x_i) = \min\ D(x_i, C)$$

Se calcula para cada individuo:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \qquad s(x_i) \in [-1, +1]$$

Si x_i es la única observación de su grupo, entonces $s(x_i) = 0$. $s(x_i)$ se interpreta de la siguiente manera:

- +1, o próximo a +1, x_i bien clasificado
- 0, o próximo a 0, observación entre dos grupos
- -1, o próximo a -1, x_i mal clasificado

De aquí se calcula la índice de calidad de una partición en *k* clases

$$Q(k) = \frac{1}{n} \sum_{r=1}^{k} n_r \bar{s}_r$$

siendo $\bar{s}_r$ la media de $s(x_i)$ para todos los individuos *i* de la clase *r*.

1.4.2 Descripción de las clases

La descripción de las clases consiste en encontrar las modalidades más, y menos, características de cada grupo. Para encontrar estas modalidades, se comparan las frecuencias absolutas de la modalidad en la clase y en la muestra total. Se emplea el modelo hipergeométrico para determinar si la diferencia es significativa.

El modelo hipergeométrico compara la frecuencia observada f_{ij} de la modalidad *i* dentro de la clase *j* con la frecuencia esperada en caso de una selección aleatoria de las ocurrencias (sin reposición), lo que constituye la hipótesis nula a contrastar.

Se definen la frecuencia total del conjunto de datos (f), la frecuencia de la clase *j* ($f_{.j}$) y la frecuencia de la modalidad *i* ($f_{i.}$). Si $f_{ij} / f_{.j}$ es mayor que $f_{i.} / f$ se calcula la probabilidad dada por la formula (1) y si es menor, se la calcula por la formula (2).

Formula 1 $\longrightarrow$ $$P_{ij} = \sum_{x=f_{ij}}^{f_{.j}} \frac{\binom{f_{i.}}{x}\binom{f - f_{i.}}{f_{.j} - x}}{\binom{f}{f_{.j}}}$$

Formula 2 $\longrightarrow$ $$P_{ij} = \sum_{x=1}^{f_{ij}} \frac{\binom{f_{i.}}{x}\binom{f - f_{i.}}{f_{.j} - x}}{\binom{f}{f_{.j}}}$$

La hipótesis a contrastar es la siguiente:

Hipótesis en el caso 1 ($f_{ij} / f_{.j} > f_{i.} / f$)

$$H_0 : \frac{f_{ij}}{f_{.j}} \leq \frac{f_{i.}}{f}$$

$$H_1 : \frac{f_{ij}}{f_{.j}} > \frac{f_{i.}}{f}$$

Hipótesis en el caso 2 ($f_{ij} / f_{.j} < f_{i.} / f$)

$$H_0 : \frac{f_{ij}}{f_{.j}} \geq \frac{f_{i.}}{f}$$

$$H_1 : \frac{f_{ij}}{f_{.j}} < \frac{f_{i.}}{f}$$

Se observa que dichas pruebas son unilaterales. Por lo tanto, el valor riesgo será 5%, dicho de otra manera, las modalidades que tengan una probabilidad del test menor a 5% se consideraran como modalidades características de sus correspondientes clases.

Para facilitar la lectura de los resultados de la prueba, se traduce la probabilidad asociada a la comparación en *valor-test*. Dicho valor-test se puede leer como una realización de la variable de Laplace-Gauss centrada y reducida. La probabilidad de 0,05 es igual a 1,645 o -1,645 en términos de valor-test.

1.5. Problemas específicos en el análisis directo de respuestas abiertas

La estrategia presentada en este capítulo presenta varias dificultades cuando se aplica a respuestas abiertas de encuestas, habitualmente y relativamente, cortas. En este caso, las respuestas se distinguen más por la presencia o ausencia de formas que por verdaderas variaciones entre perfiles de frecuencia.

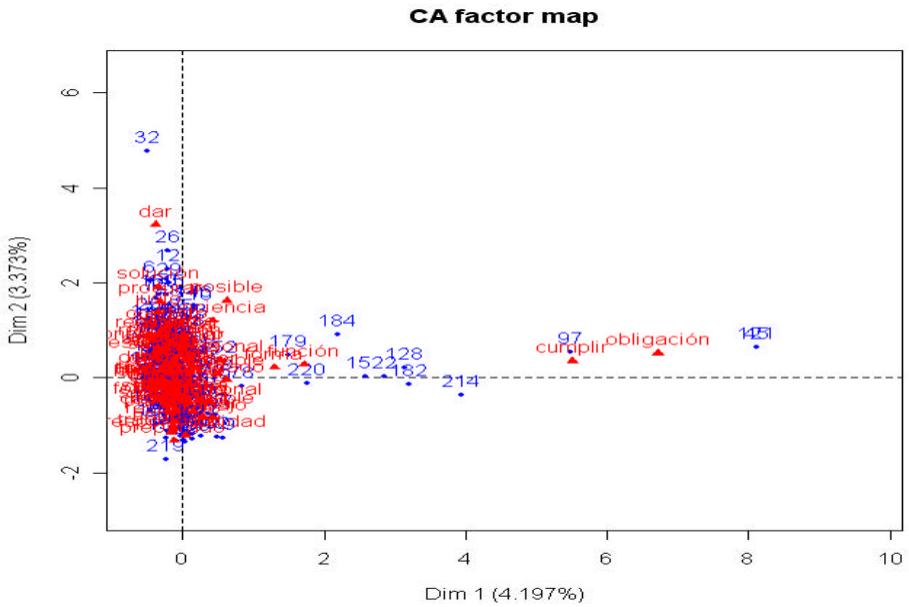

Figura 1.1. Proyecciones de los individuos y de las palabras sobre el subespacio de dimensiones 1 y 2 del AC

Como ejemplo, reproducimos en la figura 1.1 el primer plano factorial obtenido en el análisis de correspondencias de la tabla Respuestas×Palabras correspondiente a la encuesta "Jueces-jóvenes 2002". Se tratan estos datos en el capítulo 4. De momento, sólo se muestra este plano factorial como situación-tipo clásica en este tipo de análisis.

La razón por lo cual se obtiene este tipo de resultados es que la nube de puntos-individuo (y la nube de puntos-palabra) son nubes de puntos casi esféricas, sin direcciones de dispersión privilegiadas. Así, los resultados proporcionados por el análisis de correspondencias son pobres e difíciles de interpretar.

Se propone en el siguiente capítulo utilizar el análisis canónico de correspondencias, es decir, analizar la variabilidad del vocabulario pero en función de diversas características de individuos, conocidas a partir de las respuestas cerradas. Dichas características pueden ser categóricas o continuas. A continuación, se pueden clasificar a los individuos a partir de sus coordenadas sobre los ejes factoriales correspondientes a este método.

CAPÍTULO 2

ANÁLISIS CANÓNICO DE CORRESPONDENCIAS

2.1. Introducción

El análisis canónico de correspondencias (ACC) es un método desarrollado por Cajo J.F. Ter Braak (1986) e implementado inicialmente en el programa CANOCO por el mismo autor. Dicho método analiza la relación entre una tabla Individuos×Eventos–las casillas de la tabla contiene la frecuencia de una serie de eventos que forman un "todo" que se debe estudiar conjuntamente– y una tabla de variables, cuantitativas o cualitativas, que se consideran explicativas de las frecuencias observadas. Es un método usual en el campo de la ecología; en este caso la tabla de frecuencias es una tabla de abundancia de diferentes especies (o una tabla ausencia/presencia) en diferentes sitios ecológicos mientras que la otra tabla describe dichos sitios por sus características ambientales. Como las especies son atraídas por condiciones favorables, las variables ambientales son consideradas como explicativas. Así, los dos conjuntos de variables no juegan un papel simétrico. El ACC consigue introducir las variables explicativas dentro del análisis dándoles un papel activo y explicar los datos relacionándolos con ellas.

Esta metodología se puede aplicar a otros tipos de datos, como las respuestas de encuesta incluyendo respuestas cerradas y abiertas. El primer conjunto de preguntas conduce a crear una tabla o varias tablas Individuos×Variables; las segundas a construir una o varias tablas léxicas Individuos×Palabras. Determinadas características recogidas por las preguntas cerradas tienen una influencia sobre la frecuencia de las palabras. El problema planteado por la variabilidad del lenguaje tal como lo inducen dichas características es similar a la variabilidad de las especies según las condiciones ecológicas.

En este capítulo, se exponen los principios del ACC y su implementación en dos paquetes de *R*: *Vegan* y *ADE4*.

Primero se presenta la estructura de datos considerada en el ACC (Sección 2). Después se recuerdan los principios básicos (Sección 3). Un pequeño ejemplo expuesto en la sección 4 permitirá seguir paso a paso dos implementaciones, respectivamente en *Vegan* y *ADE4* (sección 5). Finalmente, se comparan las dos implementaciones del ACC y, también, el ACC y el AC.

2.2. Estructura de datos analizados por el ACC

La estructura de datos analizada por el ACC se presenta en la figura 2.1. Se tiene, por una banda, una matriz Individuos×Variables (en ecología, Sitios × Variables ambientales; las variables ambientales pueden ser tanto variables cualitativas como cuantitativas.). Dicha matriz, notada *X* describe las características de los individuos. En ecología, los sitios (que corresponden, por ejemplo, a los lugares a donde están conectadas pequeñas trampas para los animales o a puntos concretos del área bajo estudio) vienen descritos por sus características ambientales (tipo de terreno, climatología, etc.). Se tiene, por otra banda, una matriz, notada *Y*, que indica la presencia o frecuencia de determinados eventos para cada individuo. En ecología se trata de la frecuencia de diferentes especies (fauna o flora) en cada uno de los sitios.

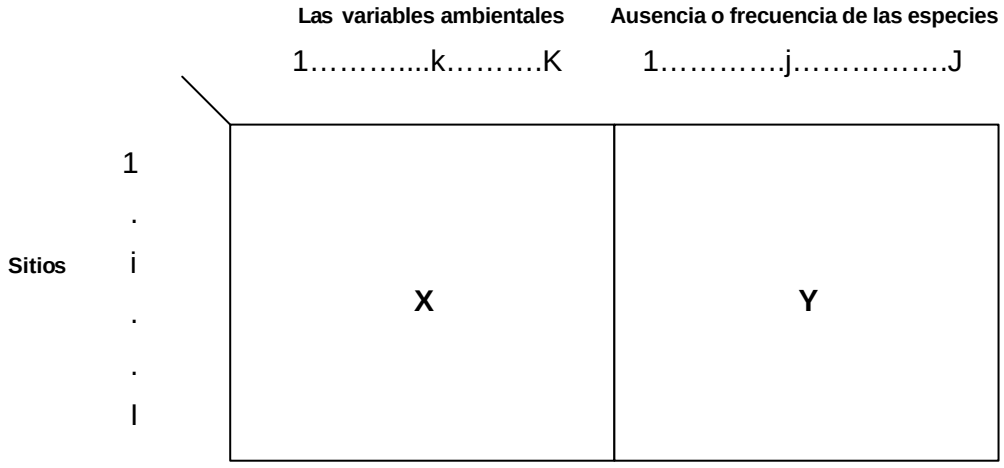

Figura 2.1. La tabla yuxtapuesta de las variables ambientales y las especies

En el análisis de respuestas abiertas, los individuos son las personas que han contestado a la encuesta. Dichos individuos están descritos por un conjunto de variables que, en el marco del estudio, se pueden considerar *explicativas* de las respuestas abiertas. La tabla de frecuencia *Y* corresponde a la tabla Individuos×Palabras, construidas a partir de las respuestas abiertas.

2.3. Principios del ACC

El objetivo del ACC es analizar la tabla de frecuencia, pero teniendo en cuenta las variables ambientales; es un análisis donde la matriz *X* (las variables ambientales) interviene en los cálculos del análisis de los datos de la matriz *Y*, forzando que los ejes de máxima dispersión sean combinaciones lineales de las variables de *X*.

El análisis canónico de correspondencias combina dos conceptos diferentes para realizar el análisis: ordenación, es decir, búsqueda de ejes de máxima dispersión, y regresión. Como

los otros métodos de ordenación, el análisis canónico de correspondencias produce ejes ortogonales sobre los cuales se pueden proyectar los datos. También está relacionado con el análisis de regresión múltiple, método que sirve para modelar una variable respuesta usando un grupo de variables explicativas. La regresión múltiple interviene de tal forma que los ejes de dispersión sean combinaciones lineales de las variables de *X*.

Se trata de un análisis proyectado y, evidentemente, se observa una disminución de la variancia total explicada. La inercia total, o la variancia total explicada, se divide en dos partes: la inercia del subespacio de proyección, que es el espacio de las variables ambientales, y la inercia del subespacio ortogonal al espacio de proyección, no relacionado con estas variables. El análisis de correspondencias simple suele producir ejes que pueden no estar muy correlacionados con las variables explicativas. Esto es debido a que el AC intenta explicar la máxima inercia posible y puede haber más inercia en el subespacio no correlacionado con las variables que la que hay en el subespacio correlacionado.

2.4. Ejemplo

Los datos del ejemplo corresponden a la distribución de 3 especies de arañas capturadas en los tramos de una duna holandesa (una duna es una acumulación de arena, en los desiertos o el litoral, generada por el viento). Se recoge la frecuencia de cada especie en cada una de los 5 tramos de esta área y además se anotan una serie de variables que describen los rasgos del tramo correspondiente. Las variables explicativas están divididas en una escala de 0 a 9, donde 0 es la ausencia y 9, la concentración máxima. Las variables son las siguientes:

Concentración de Tierra (CA): Porcentaje de la cantidad de tierra seca

Musgo Cubierta (MC): Porcentaje de la capa de musgo cubierta (los musgos son briófitas y son plantas no vasculares)

Reflejo de la Luz (RL): Reflejo de la superficie del suelo con cielo sin nubes

	X			**Y**		
	CA	MC	RL	Aulo albi	Troc terr	Alop cune
Sitio 1	6	5	6	4	9	2
Sitio 2	8	1	5	4	9	2
Sitio 3	9	1	7	4	9	6
Sitio 4	6	5	8	3	8	4
Sitio 5	5	7	8	2	7	3

Tabla 2.1. El conjunto de datos. X es la matriz de sitios × variables; Y es la matriz de sitios × especies

El ejemplo presentado corresponde a una parte reducida de un estudio completo de *Ter Braak,* quien usó este estudio entre otros para ilustrar su método.

2.5. Análisis canónico de correspondencias: implementaciones en R

Actualmente en *R* hay dos paquetes que tienen implementados el método del ACC: *Vegan* y *ADE4*. Usan escalas y algoritmos diferentes pero permiten llegar a los mismos resultados. *Vegan* aplica el algoritmo de *Legendre & Legendre* (1998). En cambio, *ADE4* interpreta el ACC desde una perspectiva diferente. Lo considera como un análisis de componentes principales propio. Fueron Chessel y Lebreton quienes presentaron esta interpretación del ACC (Thioulouse et al., 2004).

2.5.1. Algoritmo propuesto por Legendre & Legendre

Dicho algoritmo está implementado en *Vegan* (Oksanen et al., 2008). Se sigue a continuación este algoritmo paso a paso:

1- Primer paso de este algoritmo consiste en calcular las matrices *P* y $\underset{CR}{X}$. P es la matriz de los pesos y $\underset{CR}{X}$ es la matriz de *X*, centrada y reducida. A la hora de centrar y reducir, se aplican los pesos de las filas.

$$N = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \qquad p_{ij} = y_{ij} / N \qquad P = \begin{pmatrix} 0.0526 & 0.1184 & 0.0263 \\ 0.0526 & 0.1184 & 0.0263 \\ 0.0526 & 0.1184 & 0.0789 \\ 0.0395 & 0.1053 & 0.0526 \\ 0.0263 & 0.0921 & 0.0395 \end{pmatrix}$$

$$\sum_{i=1}^{I} p_{i.} x_{ik} = \bar{x}_k \qquad \sum_{i=1}^{I} p_{i.} \left(x_{ik} - \bar{x}_k\right)^2 = V(x_k) \qquad \underset{CR\,ik}{X} = \frac{x_{ik} - \bar{x}_k}{\sqrt{V(x_k)}} \qquad \underset{CR}{X} = \begin{pmatrix} -0.6618 & 0.6218 & -0.6730 \\ 0.6794 & -1.0659 & -1.5548 \\ 1.3501 & -1.0659 & 0.2089 \\ -0.6618 & 0.6218 & 1.0907 \\ -1.3324 & 1.4656 & 1.0907 \end{pmatrix}$$

2- El ACC trabaja con las distancias chi-cuadrado. Por lo tanto, en vez de trabajar con la matriz original *Y*, trabaja con la matriz $\overline{Q}$ en los cálculos. $\overline{Q}$ tiene como término general:

$$\bar{q}_{ij} = \frac{p_{ij} - p_{i.} p_{.j}}{\sqrt{p_{i.} p_{.j}}} \qquad \overline{Q} = \begin{pmatrix} 0.0403 & 0.0283 & -0.0848 \\ 0.0403 & 0.0283 & -0.0848 \\ -0.0139 & -0.0530 & 0.0973 \\ -0.0222 & -0.0115 & 0.0403 \\ -0.0479 & 0.0164 & 0.0221 \end{pmatrix}$$

3- Calcular los coeficientes de la regresión múltiple ponderada, donde $\overline{Q}$ contiene las variables respuestas (variables explicadas) y la matriz $\underset{CR}{X}$, las variables explicativas.

$$D_I = \sum_{i\in I}\sum_{j=1}^{J} p_{ij} \qquad D_I^{1/2} = \begin{pmatrix} 0.44 & 0 & 0 & 0 & 0 \\ 0 & 0.44 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.44 & 0 \\ 0 & 0 & 0 & 0 & 0.39 \end{pmatrix} \qquad B = \left\lfloor \underset{CR}{X}' D_I \underset{CR}{X} \right\rfloor^{-1} \underset{CR}{X}' D_I^{1/2} \overline{Q} \qquad B = \begin{pmatrix} 0.0032 & -0.1069 & 0.1649 \\ 0.0144 & -0.0457 & 0.0575 \\ -0.0799 & -0.0508 & 0.1599 \end{pmatrix}$$

4- Calcular las predicciones de $\overline{Q}$.

$$\hat{Y} = D_I^{1/2} \underset{CR}{X} B \qquad \hat{Y} = \begin{pmatrix} 0.0270 & 0.0340 & -0.0804 \\ 0.0493 & 0.0245 & -0.0879 \\ -0.0139 & -0.0531 & 0.0974 \\ -0.0357 & -0.0058 & 0.0448 \\ -0.0279 & 0.0079 & 0.0154 \end{pmatrix}$$

5- Hacer la descomposición en valores y vectores propios.

$$S_{\hat{Y}'\hat{Y}} = \hat{Y}'\hat{Y} \qquad EIGEN\left(S_{\hat{Y}'\hat{Y}}\right)$$

Valores propios

$$l_1 = 0.03372 \qquad l_2 = 0.00227$$

Vectores propios

$$U = \begin{pmatrix} -0.3403 & 0.8127 \\ -0.3410 & -0.5754 \\ 0.8763 & 0.0916 \end{pmatrix}$$

6- Calcular las coordenadas para hacer la representación gráfica. Las de las especies se están recogidas en las filas de la matriz $\hat{F}$. Las de los sitios, en el subespacio *Y*, están recogidas en las filas de $\hat{V}$ y las de los sitios, en el subespacio *X*, en las filas de *Z*.

$$D_J = \sum_{j\in J}\sum_{i=1}^{I} p_{ij} \qquad D_J^{-1/2} = \begin{pmatrix} 2.1144 & 0 & 0 \\ 0 & 1.3452 & 0 \\ 0 & 0 & 2.1144 \end{pmatrix} \qquad \hat{F} = D_J^{-1/2} U \Lambda^{1/2} = \begin{pmatrix} -0.1321 & 0.0819 \\ -0.0842 & -0.0369 \\ 0.3402 & 0.0092 \end{pmatrix}$$

$$D_I^{-1/2} = 1 \Big/ D_I^{1/2} \qquad \hat{V} = D_I^{-1/2} \overline{Q} U \Lambda^{-1/2} = \begin{pmatrix} -1.198 & 0.413 \\ -1.198 & 0.413 \\ 1.178 & 1.182 \\ 0.574 & -0.365 \\ 0.412 & -2.448 \end{pmatrix} \qquad Z = D_I^{-1/2} \hat{Y} U \Lambda^{-1/2} = \begin{pmatrix} -1.118 & -0.237 \\ -1.251 & 0.847 \\ 1.178 & 1.182 \\ 0.654 & -1.016 \\ 0.278 & -1.363 \end{pmatrix}$$

7- Las variables de *X* se proyectan a partir de sus correlaciones con los ejes de ordenación.

$$CORR\left(D_I^{1/2} \underset{CR}{X}, V\right) = \begin{pmatrix} 0.2305 & 0.9633 \\ -0.0415 & -0.9627 \\ 0.7839 & -0.6207 \end{pmatrix}$$

2.5.1.1. Resultados gráficos

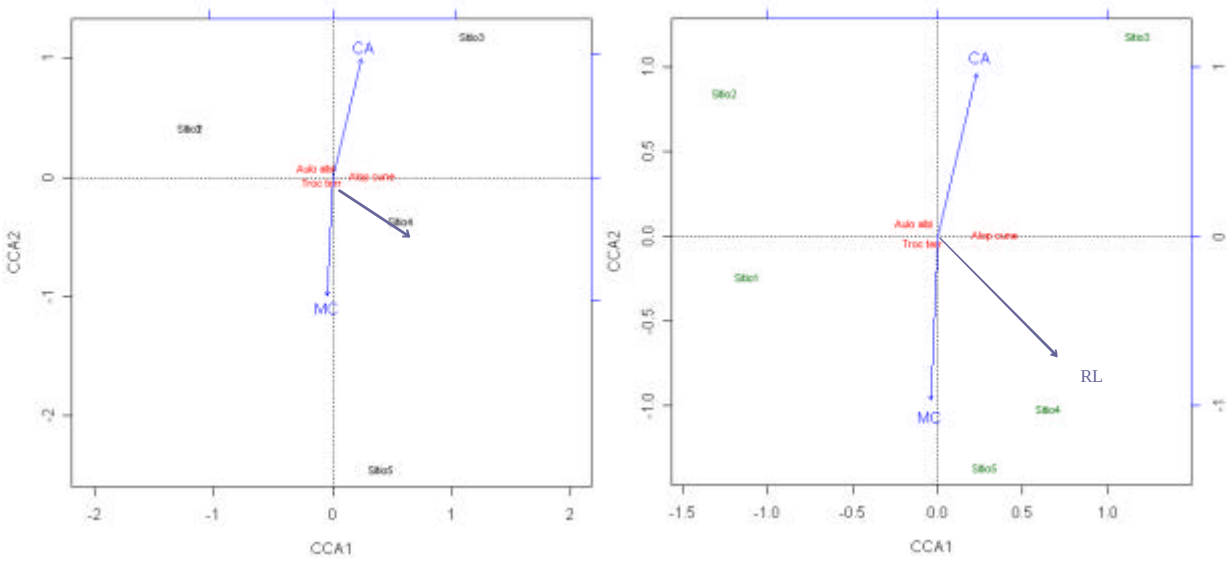

Figura 2.2. Las proyecciones en el subespacio Y y en el subespacio X (Vegan)

En la figura 2.2, se observa que las coordenadas de las especies y de las variables ambientales son las mismas en los dos gráficos. En cambio, las coordenadas de los sitios varían.

Las variables *"Concentración de Tierra" (CA)* y *"Musgo Cubierta" (MC)* tienen correlaciones altas respecto al segundo eje, de manera que, la parte positiva del eje indica alta porcentaje de tierra seca y la parte negativa, alta porcentaje de capa musgo cubierta. La tercera variable, *"Reflejo de la Luz" (RL)*, tiene correlaciones altas con cada uno de los dos ejes. Se asocia con la parte positiva del primero y con la parte negativa del segundo.

Las especies se posicionan muy cerca del centro de gravedad. La especie "alop cune" se asocia con la parte positiva del primer eje. Sabiendo que existe una relación entre *RL* y esta parte del eje, se puede confirmar que esta especie es más abundante en los sitios donde la luz del sol llega con más facilidad. En el segundo eje, las especies no destacan hacia ninguna dirección de crecimiento, ni de decrecimiento.

2.5.2. Algoritmo propuesto por Chessel y Lebreton

Dicho algoritmo está implementado en *ADE4*. Se sigue el algoritmo paso a paso sobre el ejemplo.

1. Calcular las matrices de los pesos de las filas y de las columnas.

$$N = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \qquad p_{ij} = y_{ij} / N \qquad D_I = \sum_{i\in I}\sum_{j=1}^{J} p_{ij} \qquad D_J = \sum_{j\in J}\sum_{i=1}^{I} p_{ij}$$

$$P = \begin{pmatrix} 0.0526 & 0.1184 & 0.0263 \\ 0.0526 & 0.1184 & 0.0263 \\ 0.0526 & 0.1184 & 0.0789 \\ 0.0395 & 0.1053 & 0.0526 \\ 0.0263 & 0.0921 & 0.0395 \end{pmatrix} \qquad D_I = \begin{pmatrix} 0.1974 & 0 & 0 & 0 & 0 \\ 0 & 0.1974 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0.1974 & 0 \\ 0 & 0 & 0 & 0 & 0.1579 \end{pmatrix} \qquad D_J = \begin{pmatrix} 0.2237 & 0 & 0 \\ 0 & 0.5526 & 0 \\ 0 & 0 & 0.2237 \end{pmatrix}$$

2. Calcular la matriz $\underset{CR}{X}$, la matriz *X* centrada y reducida.

$$\sum_{i=1}^{I} p_{i.} x_{ik} = \bar{x}_k \qquad \sum_{i=1}^{I} p_i \left(x_{ik} - \bar{x}_k\right)^2 = V(x_k) \qquad \underset{CR}{X}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{V(x_k)}}$$

$$\underset{CR}{X} = \begin{pmatrix} -0.6618 & 0.6218 & -0.6730 \\ 0.6794 & -1.0659 & -1.5548 \\ 1.3501 & -1.0659 & 0.2089 \\ -0.6618 & 0.6218 & 1.0907 \\ -1.3324 & 1.4656 & 1.0907 \end{pmatrix}$$

3. Calcular la matriz *DF*.

$$DF_{ij} = \frac{y_{ij}}{N \times p_{i.} p_{.j}} - 1 = \frac{p_{ij} - p_{i.} p_{.j}}{p_{i.} p_{.j}} \qquad DF = \begin{pmatrix} 0.1922 & 0.0857 & -0.4039 \\ 0.1922 & 0.0857 & -0.4039 \\ -0.0588 & -0.1429 & 0.4118 \\ -0.1059 & -0.0349 & 0.1922 \\ -0.2549 & 0.0556 & 0.1176 \end{pmatrix}$$

4. Calcular las predicciones de la matriz *DF*.

$$B = \left[\underset{CR}{X}^{'} D_I^{1/2} D_I^{1/2} \underset{CR}{X}\right]^{-1} \underset{CR}{X}^{'} D_I^{1/2} D_I^{1/2} DF = \left[\underset{CR}{X}^{'} D_I \underset{CR}{X}\right]^{-1} \underset{CR}{X}^{'} D_I DF \qquad \hat{DF} = \underset{CR}{X} B$$

$$B = \begin{pmatrix} 0.0067 & -0.1438 & 0.3486 \\ 0.0305 & -0.0615 & 0.1215 \\ -0.1690 & -0.0684 & 0.3380 \end{pmatrix} \qquad \hat{DF} = \begin{pmatrix} 0.1283 & 0.1029 & -0.3826 \\ 0.2347 & 0.0742 & -0.4181 \\ -0.0588 & -0.1429 & 0.4118 \\ -0.1697 & -0.0177 & 0.2134 \\ -0.1485 & 0.0268 & 0.0822 \end{pmatrix}$$

5. Calcular y diagonalizar la matriz *E* para obtener la descomposición en valores y vectores propios.

$$E = \sqrt{D_I} \times \hat{DF} \times \sqrt{D_J} \qquad E = \begin{pmatrix} -0.1391 & 0.2053 & -0.1414 \\ 0.1428 & -0.3520 & -0.3267 \\ 0.3193 & -0.3962 & 0.0494 \\ -0.1391 & 0.2053 & 0.2292 \\ -0.2504 & 0.4329 & 0.2050 \end{pmatrix}$$

$$EIGEN\left(E^{'} E\right)$$

Valores propios

$$l_1 = 0.03372 \qquad l_2 = 0{,}00227$$

Vectores propios

$$U = \begin{pmatrix} -0.3403 & 0.8127 \\ -0.3410 & -0.5754 \\ 0.8763 & 0.0916 \end{pmatrix}$$

6. Las especies se representan a partir de las filas de la matriz *Co* y los sitios de las filas de la matriz *L*. Esta representación se realiza sobre el subespacio creado como combinación lineal de las variables de *X*.

$$C_0 = \frac{1}{\sqrt{D_J}} \times U \times \Lambda^{1/2} \qquad C_0 = \begin{pmatrix} -0.1321 & 0.0818 \\ -0.0842 & -0.0368 \\ 0.3402 & 0.0092 \end{pmatrix}$$

$$L = DF \times \sqrt{D_J} \times U \qquad L = \begin{pmatrix} -0.2053 & -0.0113 \\ -0.2299 & 0.0404 \\ 0.2163 & 0.0563 \\ 0.1203 & -0.0484 \\ 0.0511 & -0.0650 \end{pmatrix}$$

7. Las variables se proyectan a partir de sus correlaciones con los ejes de ordenación.

$$COR = \left(D_I \times \underset{CR}{X}\right) \times L \times \Lambda^{-1/2} \qquad COR = \begin{pmatrix} 0.2317 & 0.9634 \\ -0.0430 & -0.9628 \\ 0.7831 & -0.6204 \end{pmatrix}$$

2.5.2.1. Resultados gráficos

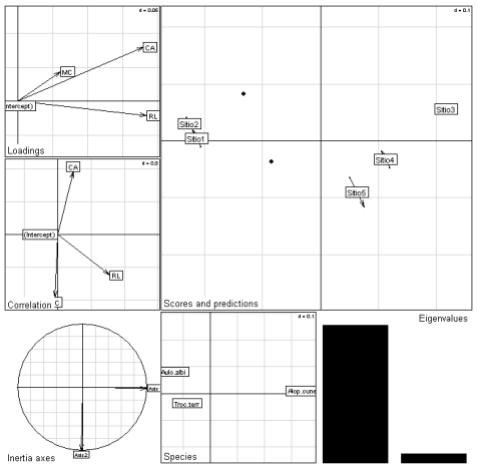

Figura 2.3. Las proyecciones en el subespacio X (ADE4)

Las variables ambientales (*CA, MC* y *RL*) tienen las mismas correlaciones que en el caso de *Vegan* y las especies, las mismas coordenadas. Por lo tanto, sus interpretaciones son las mismas que las de antes.

Las coordenadas de los sitios son las que cambian respecto a antes. Aquí es el primer eje el que hace una mejor separación entre los sitios. Recordando que este eje estaba asociado con la variable *RL*, los sitios que tienen un valor más alto de *RL* (los sitios 3, 4 y 5) se quedan en la parte positiva del primer eje. Esta parte corresponde a la dirección de crecimiento de la variable. De la misma manera, los sitios 1 y 2 están en la parte negativa del eje, la dirección de decrecimiento.

2.6. Comparación de las dos implementaciones del ACC

- *Vegan* trata y considera el ACC como un método que mezcla ordenación y regresión. *ADE4* añade a este punto de vista una interpretación diferente (ACP), que enriquece el método y es muy útil en el caso de que se quisiera trabajar con individuos suplementarios y variables suplementarias.

- *Vegan* permite proyectar los objetos tanto en el subespacio original de *Y* como el de *X*, combinación lineal de las variables ambientales. En cambio, *ADE4* solamente permite hacerlo en el de *X*.

- Con cada uno de los dos métodos se obtienen las mismas coordenadas para las especies y las variables ambientales. Lo único que varía son las coordenadas de los sitios. Los sitios proyectados en *Vegan* se ponderan por la raíz de los valores propios para proyectarlos en *ADE4* ($L = Z \times \Lambda^{1/2}$).

- *Vegan* también descompone la variabilidad del subespacio ortogonal a las variables ambientales. Esto permite estudiar el efecto de las terceras variables que no han sido incluidas dentro del análisis.

- *ADE4* permite trabajar con variables categóricas. En cambio, *Vegan* no lo permite hacer directamente, y para trabajar con variables categóricas requiere que se construya previamente la tabla disyuntiva completa.

- El que se interesa, cuando se aplica el ACC, es ver las relaciones entre las especies y las variables. La interpretación de los sitios tiene un papel secundario. Como los dos métodos nos dan las mismas interpretaciones para las especies y las variables, se pueden usar indistintamente siempre teniendo en cuenta las ventajas e inconvenientes de cada uno.

2.7. Comparación entre el AC y el ACC

La función *cca()* de *Vegan* y de *ADE4* piden dos parámetros de entrada, la matriz *Y* y la *X*, para hacer un análisis canónico de correspondencias. Si únicamente se entra la matriz *Y* de las frecuencias de las especies, entonces en este caso se hace un análisis de correspondencias simple.

Los valores propios que devuelve este análisis de correspondencias simple son:

$$l_1 = 0.033925112 \qquad l_2 = 0.003149192$$

Habría que recordar cuales eren los valores propios del ACC:

$$l_1 = 0.03372 \qquad l_2 = 0.00227$$

Se confirma que estos valores propios son más grandes que los del ACC como se esperaba (el AC intenta explicar máxima variabilidad posible y por lo tanto siempre encuentra ejes con unas inercias y valores propios más grandes que los del ACC), pero la diferencia en este caso es muy pequeña. Para tener más información se pueden calcular los ratios entre los valores propios.

$$r_1 = \frac{0.03372}{0.033925112} = 0.994 \qquad r_2 = \frac{0.00227}{0.003149192} = 0.721$$

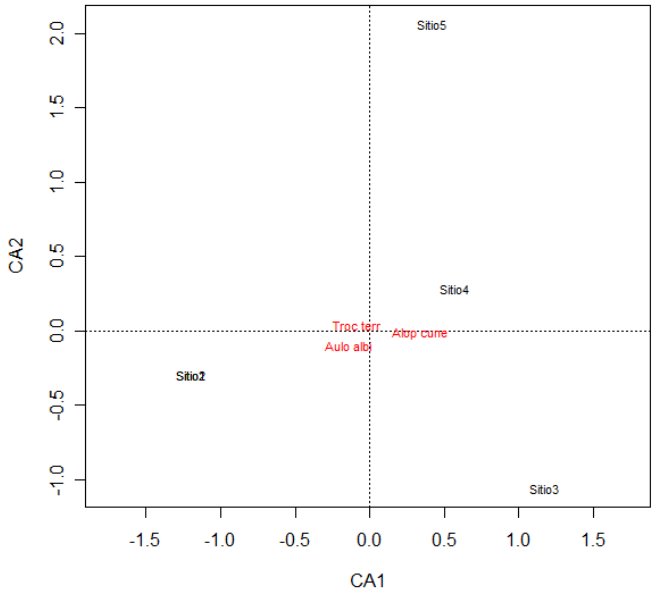

Figura 2.4. El análisis de correspondencias simple de las especies y los sitios

La figura 2.4, el gráfico del AC, tiene una forma bastante similar a los otros gráficos obtenidos mediante el ACC. Esto es debido a que hay pocos datos en este caso y los valores propios del AC y del ACC son muy similares y prácticamente explican la misma variabilidad de los datos.

CAPÍTULO 3

UTILIZACIÓN DEL SOFTWARE R

3.1. Introducción

R es un entorno de software libre para la estadística y computación gráfica. Se compila y ejecuta en una amplia variedad de plataformas de UNIX, Windows y MacOS.

Text Mining (Feinerer, 2007, Feinerer et al., 2007) es el paquete principal que ayuda a crear la tabla de frecuencias de documentos $\times$ palabras. A parte de ello, también se usan otros paquetes como *FactoMineR* (Husson et al., 2007) para hacer el AC, *Cluster* (Maechler, *2007)* para hacer la clasificación y dos paquetes más para realizar el ACC, *Vegan* (Oksanen et al., 2008) y *ADE4* (Chessel et al., 2008, Thioulouse et al., 2004).

Los algoritmos y las implementaciones de *Vegan* y *ADE4* se han explicado en el capítulo anterior. Las implementaciones de la función AC y las funciones de clasificación se pueden encontrar entre los anexos.

Este capítulo, está dedicado a explicar los pasos a seguir en *Text Mining* para obtener la matriz de documentos $\times$ palabras. Además de todo esto, se presentan funciones propias creadas para mejorar *Text Mining.*

3.2. Text Mining

Text Mining (*tm*) es un paquete de *R* que permite manejar los documentos de formato de texto. Con este paquete se pueden modificar estos textos añadiendo o eliminando palabras. También se puede crear una matriz donde cada fila corresponde a un documento y las columnas sean las palabras que aparecen en estos documentos. La creación de esta matriz es lo que tiene interés en este proyecto.

Se prueba *tm* sobre la encuesta "*jueces jóvenes 2002*". El conjunto de datos contiene 149 variables y 268 individuos. Entre otras cosas, se pregunta a los jueces "¿Qué es un buen juez?". Usando el paquete *tm* se crea la matriz de documentos $\times$ palabras correspondiente a dicha pregunta.

3.2.1. Las etapas para crear la matriz de documentos x palabras

Antes de crear dicha matriz, los documentos se tienen que pasar a un formato especial.

3.2.1.1. Paso 1

El primer paso consiste en crear una "colección de documentos de texto" y, de esta manera, ajuntar los varios documentos que hay. Para hacer esto se usa la siguiente función de *tm* con sus parámetros correspondientes.

> Función => **Corpus(object, readerControl=list(reader, language, load))**
>
> *object:* Tipo y nombre del archivo de fuente donde se encuentran los documentos. Puede ser de varios tipos como DirSource(…), CSVSource(…), GmaneSource(…), ReutersSource(…).
>
> *reader:* El formato de archivo de los documentos que forman el archivo de fuente. Puede ser de tipos como readPlain, readRCV1, readReut21578XML, readGmane, readNewsgroup, readPDF o readHTML. Si se quiere trabajar con los formatos readPlain (archivos .txt) o readPDF (archivos .pdf), estos documentos primero se tendrían que poner dentro de un directorio y trabajar con archivo de fuente tipo DirSource(…).
>
> *language:* Aunque existe este parámetro, actualmente la única opción es "inglés". El valor que tiene que recibir es "en_US".
>
> *load:* Es un booleano utilizado para indicar si la "colección de documentos de texto" creada debe guardarse inmediatamente en la memoria (TRUE) o solamente se guarde cuando haga falta (FALSE). La opción por defecto es "FALSE".

Ejemplo:

Cada respuesta a la pregunta "¿Qué es un buen juez?" se considera como un documento. Estos documentos se guardan dentro del directorio "Que es un buen juez". Se ejecuta la siguiente función y el resultado se guarda en el objeto *tdc*.

```
tdc <= Corpus(DirSource("D:/Que es un buen juez"),
readerControl=list(reader=readPlain,language="en_US",load=TRUE))
```

3.2.1.2. Paso 2

Una vez creada la "colección de documentos de texto", se lematizan las palabras. Consiste en convertir los adjetivos de género femenino en masculino y las palabras plurales en singulares. Por otra banda, los verbos se convierten en infinitivo.

Hace falta usar dos funciones de *R* para poder lematizar las palabras manualmente. La primera es “createDictionary()” y sirve para visualizar las palabras de la matriz. Después se usa la función “replaceWords()” para crear las equivalencias (en el apartado siguiente, funciones secundarias, se puede ver como se usan estas dos funciones).

Para acabar, se eliminan las palabras que llevan poca información para la interpretación. En este caso solamente se conservan sustantivos, verbos, adjetivos y adverbios.

3.2.1.3. Paso 3

El tercer paso es crear la matriz de documentos × palabras.

> Función => **TermDocMatrix(object, control=list(tolower, removeNumbers, stemming, stopwords, dictionary, minDocFreq, minWordLength))**
>
> *object:* La “colección de documentos de texto” creada en el primer paso.
>
> *tolower:* Convertir los caracteres en minúsculas.
>
> *removeNumbers:* Booleano para indicar si se quieren borrar los caracteres numéricos de los documentos.
>
> *stemming:* Convertir las palabras en su “raíz”. Funciona únicamente para las palabras en inglés. La opción por defecto de este parámetro es “FALSE”.
>
> *stopwords:* Se refiere a las palabras de tipo conjunciones, artículos y preposiciones que no aportan mucha información a la interpretación y por lo tanto, se pueden eliminar. Las opciones disponibles que proporciona *R* son: danish, dutch, english, finnish, french, german, hungarian, italian, norwegian, portuguese, russian, spanish o swedish. Como se puede ver, está disponible la opción en castellano.
>
> *dictionary:* Crear un diccionario con las palabras indicadas. Para hacerlo, se tiene que entrar un vector que tendrá como elementos las palabras con las cuales se formará el propio diccionario.
>
> *minDocFreq:* Eliminar todas las palabras que tienen una frecuencia menor que este valor. Se compara con cada una de las frecuencias de la misma palabra en diferentes documentos (diferentes filas de la misma columna) y, si alguna de estas filas tiene una frecuencia igual o mayor que “minDocFreq”, entonces la palabra se conserva. Su valor por defecto es 1.

minWordLength: Eliminar las palabras que tienen una longitud menor que el valor de este parámetro. Por defecto vale 3. Es recomendable cambiar a 1 para no perder palabras de longitud 1 y 2.

Ejemplo:

Después de lematizar las palabras y eliminar las juzgadas poco informativas, se crea la matriz de documentos × palabras. La matriz resultante se guarda en el objeto *tdm*.

```
tdm <= TermDocMatrix(tdc,control=list(minWordLength=1))
```

3.2.2. Funciones secundarias

Existen unas funciones secundarias en *tm* que ayudan a manejar y visualizar la información mientras que se esté creando la "colección de documentos de texto" y la matriz de documentos × palabras. Las más importantes son las siguientes:

- **inspect**

Permite visualizar los elementos que pertenecen a una "colección de documentos de texto".

Función => inspect(object)
object: Corpus ("colección de documentos de texto").

Ejemplo: inspect(tdc)

- **createDictionary**

Crea un diccionario de todas las palabras que aparecen en el corpus.

Función => createDictionary(object)
object: TermDocMatrix (la matriz de documentos × palabras).

Ejemplo: createDictionary(tdm)

- **removePunctuation**

Elimina todos los signos de puntuación de un documento.

Función => removePunctuation(object)
object: PlainTextDocument (un sol documento de texto de los que forman el corpus).

Ejemplo: removePunctuation(tdc[[1]])

- **c**

Concatena varios documentos o corpus en uno.

Función => c(object)
object: Corpus o TextDocument.

Ejemplo: c(tdc,crude) (“Crude” es una base de datos que está integrada en el paquete *tm*)

- **appendElem**

Añade un documento a un corpus.

Función => appendElem(object, data, meta)
object: Corpus.
data: Documento de texto.
meta: Información de meta data. Por defecto es nulo.

Ejemplo: appendElem(tdc,crude[[1]])

- **replaceWords**

Reemplaza las palabras indicadas en el parámetro “words” con la del “by”.

Función => replaceWords(object, words, by)
object: PlainTextDocument.
words: Las palabras a reemplazar.
by: La palabra que reemplazará a las que están en “words”.

Ejemplo: replaceWords(tdc[[1]],c(“es”,“será”,“son”),“ser”)

- **searchFullText**

Devuelve un valor booleano indicando si la palabra buscada aparece o no en el documento.

Función => searchFullText(object, pattern)
object: PlainTextDocument.
pattern: La palabra de la cual quiere comprobarse su existencia en el documento.

Ejemplo: searchFullText(tdc[[1]],“juez”)

- **tmMap**

Aplica la función del parámetro “FUN” a cada uno de los elementos del objeto.

Función => tmMap(object, FUN)
object: Corpus.
FUN: La función que se quiere aplicar.

Ejemplo: tmMap(tdc,FUN=replaceWords,c(“es”,“será”,“son”),“ser”)

- **removeWords**

Elimina las palabras que pertenecen a uno de los “stopwords”.

Función => removeWords(object, stopwords)
object: PlainTextDocument.
stopwords: Se eliminan “stopwords” que pertenecen a uno de los siguientes idiomas: danish, dutch, english, finnish, french, german, hungarian, italian, norwegian, portuguese, russian, spanish o swedish.

Ejemplo: removeWords(tdc[[1]],stopwords(“spanish”))

- **tmFilter**

Hace un filtro aplicando una función sobre cada uno de los documentos que pertenecen al corpus.

Función => tmFilter(object, FUN, doclevel)
object: Corpus.
FUN: La función a aplicar.
doclevel: Parámetro booleano para indicar si se quiere guardar los cambios sobre el mismo objeto. Su valor por defecto es “FALSE”.

Ejemplo: tmFilter(tdc,FUN=searchFullText,“juez”,doclevel=TRUE)

3.2.3. Insuficiencias y modificaciones necesarias sobre el paquete tm

En el paquete *tm* se ha observado:

- La no ordenación alfabética de las palabras en la matriz. Esto dificulta la búsqueda de las palabras cuando, por ejemplo, se desea conocer su frecuencia.

- La no existencia de un vector con las frecuencias de las palabras.

- La inexistencia de un filtro para eliminar las palabras con poca frecuencia. El parámetro existente “minDocFreq” no trabaja con las frecuencias totales de las palabras, sino trabaja con las frecuencias parciales y, por lo tanto, no es un filtro muy útil.

- Que el vector de “stopwords” en castellano contiene diferentes formas verbales como *tener*, *haber* y otros verbos. Haciendo un filtro con este parámetro para eliminar las preposiciones, los artículos, los determinantes y las conjunciones, se eliminan también estos verbos. Por lo tanto, se ha descartado el uso de este vector de “stopwords” por no corresponder a las necesidades.

3.3. Funciones complementarias para crear la matriz de documentos x palabras

Para poder mejorar y complementar las insuficiencias y puntos débiles de *tm*, se crean dos funciones en *R*.

3.3.1. Selección por frecuencias

La primera función creada se llama “*SortTermDocMatrix*”. Tiene dos parámetros de entrada y 3 valores de salida.

Función => **SortTermDocMatrix(tdm, minfreq)**

tdm: La matriz resultante de la función “TermDocMatrix” de *tm* (Matriz de documentos × palabras).

minfreq: Eliminar las palabras que tienen una frecuencia global inferior a este. Si no se especifica ningún valor, se coge la opción por defecto, 2% de número total de documentos.

Valores que devuelve la función son:

Dataframe: La matriz de documentos × palabras donde las palabras se ordenan alfabéticamente, al mismo tiempo aplicando “minfreq”.

tf: El vector que indica la frecuencia total de cada palabra. Solamente tiene en cuenta las palabras que han sido seleccionadas con “minfreq”.

df: La matriz donde aparecen el número total de palabras usadas y el número total de palabras conservadas de cada documento.

Ejemplo:

Se aplica la función creada sobre *tdm*. Se eliminan las palabras que tienen una frecuencia menor a 2% de número total de individuos. El resultado se guarda en *docterms*.

```
docterms <= SortTermDocMatrix(tdm)
```

3.3.2. Filtrar palabras

Función "filter" es una alternativa a "stopwords" de *tm.* Solamente necesita como entrada la matriz de documentos $\times$ palabras y una lista de las palabras que se quieren eliminar de esta matriz. En este sentido la función es muy flexible. Se pueden crear y entrar varias listas de palabras para diferentes idiomas. También se pueden modificar. Existe una lista de palabras para castellano creada juntamente con esta función. La función es la siguiente:

Función => **filter(df, sw)**

df: La matriz de documentos $\times$ palabras.

sw: La lista de palabras que se quieren eliminar.

Ejemplo:

Se aplica la función sobre *docterms* y el resultado final se guarda en *filterdocterms.*

```
filterdocterms <= filter(docterms,castellano)
```

CAPÍTULO 4

APLICACIÓN: JUECES JÓVENES 2002

4.1. Encuesta "jueces jóvenes 2002"

La encuesta "*jueces jóvenes 2002*" es una encuesta hecha dentro de un proyecto nacional del Ministerio de Ciencia y Tecnología de España. Se realizó en el año 2002 con el objetivo de conocer mejor las dificultades actuales de los jueces jóvenes (Ayuso et al., 2005). Se tenía especial interés en saber el uso de las nuevas tecnologías para buscar información judicial.

Para formar la muestra, se escogió al azar 129 jueces con menos de 4 años de experiencia entre 352 jueces de la población. Para poder hacer la comparación, también se cogieron al azar 139 jueces señor entre 2352 jueces que habían en este grupo. La muestra total está compuesta por 268 jueces.

A partir de las respuestas se han construido 143 variables (5 variables cuantitativas, 134 variables cualitativas y 4 preguntas abiertas). Estas variables se pueden dividir en 3 bloques: uso de nuevas tecnologías, datos personales y variables de opinión.

Uso de nuevas tecnologías: Variables que hacen referencia a la frecuencia de uso de bases de datos, los portales, las bibliotecas y las publicaciones. También recopila información sobre la consulta de documentación de varios tipos como doctrina, estadística, judicial, sociológica, etc.

Datos personales: Variables como "Sexo", "Año de nacimiento", "Estado civil", "Número de hijos". Por otra banda, variables que hacen referencia a las relaciones con asociaciones y colaboración con las otras instituciones.

Opinión: Es el bloque más amplio cuanto a número de variables. Variables como opinión sobre la formación recibida ("Valoración formación facultad"), valoración del uso de nuevas tecnologías ("Valoración red telemática", "Calidad Centro Documentación Judicial"), valoración de la justicia ("La justicia es lenta"), impacto de varios temas (drogas, violencia, delitos, inmigración, etc.). En abierto, se les preguntaba que manifestaran su opinión personal mediante sus respuestas a las siguientes preguntas: "Qué es un buen juez ?" y "Como son los jueces actuales ?".

4.2. Creación de la matriz de individuos x palabras

Usando la metodología explicada en el tercer capítulo, se crea la matriz de individuos × palabras correspondiente a la pregunta abierta "Qué es un buen juez". Se ofrece en la tabla 4.1 algunos indicadores sobre el correspondiente corpus. Después de operar la lematización, eliminar las palabras que corresponden a *stopwords* y escoger un umbral de frecuencia igual a 6, se conservan 76 palabras distintas (tablas 4.2 y 4.3).

Longitud total	4187
Longitud conservada	1095
Palabras en total	959
Palabras conservadas	76

Tabla 4.1. Información sobre la matriz de individuos × palabras

La matriz Individuos×Palabras tiene 76 columnas-palabra. Se yuxtapone dicha tabla con la tabla Individuos×Variables creada a partir de las respuestas cerradas de "*jueces jóvenes 2002*". Se obtiene una base de datos con 268 individuos y 219 variables (143 columnas-variable + 76 columnas-palabra).

aplicar	cumplir	formación	justicia	persona	realidad	social
asunto	dar	función	justiciable	personal	realizar	sociedad
bueno	decidir	gran	justo	posible	resolución	solución
calidad	decisión	haber	juzgado	práctico	resolutivo	técnico
capacidad	dejar	hacer	ley	preparado	resolver	tener
caso	derecho	honesto	mantener	problema	responsabilidad	tiempo
ciudadano	día	humano	más	procurar	responsable	trabajador
común	escuchar	imparcial	mucho	profesional	saber	trabajar
conciencia	estar	intentar	muy	prudente	sentido	trabajo
conflicto	estudiar	juez	no	público	ser	tratar
conocimiento	forma	jurídico	obligación	razonable	servicio	

Tabla 4.2. Las palabras seleccionadas ordenadas alfabéticamente

persona	**101**	juez	**19**	más	**12**	estudiar	**8**	obligación	**6**
ser	**73**	capacidad	**18**	personal	**12**	imparcial	**8**	posible	**6**
sentido	**58**	ley	**18**	conocimiento	**11**	trabajar	**8**	práctico	**6**
tener	**45**	asunto	**17**	resolución	**11**	forma	**7**	preparado	**6**
común	**41**	escuchar	**17**	responsabilidad	**11**	honesto	**7**	procurar	**6**
trabajo	**38**	social	**17**	calidad	**10**	juzgado	**7**	prudente	**6**
bueno	**36**	jurídico	**16**	caso	**10**	mucho	**7**	público	**6**
no	**36**	problema	**16**	cumplir	**10**	resolutivo	**7**	razonable	**6**
saber	**32**	haber	**15**	justo	**10**	técnico	**7**	realizar	**6**
resolver	**27**	profesional	**14**	conflicto	**9**	conciencia	**6**	sociedad	**6**
trabajador	**27**	ciudadano	**13**	dar	**9**	decidir	**6**	solución	**6**
derecho	**26**	formación	**13**	decisión	**9**	dejar	**6**	tratar	**6**
responsable	**22**	intentar	**13**	estar	**9**	día	**6**		
aplicar	**21**	gran	**12**	muy	**9**	función	**6**		
justicia	**21**	hacer	**12**	servicio	**9**	justiciable	**6**		
realidad	**20**	humano	**12**	tiempo	**9**	mantener	**6**		

Tabla 4.3. Las palabras seleccionadas ordenadas por frecuencia

4.3. Aplicación del ACC sobre la encuesta

4.3.1. Los datos del análisis

La matriz *Y*, que indica las frecuencias de los eventos en el ACC, está formada por las frecuencias de las palabras que usan los jueces en sus respuestas a la pregunta "Qué es un buen juez?". Es la matriz de individuos $\times$ palabras. En total hay 268 jueces que participan a la encuesta pero 45 de ellos se eliminan de la matriz por no haber usado ninguna de las palabras seleccionadas y otros 24 por no haber respondido a una gran mayoría de las preguntas de la encuesta. En resumen, la matriz *Y* tiene 199 jueces y 76 palabras que son sus filas y columnas, respectivamente.

Las variables consideradas como variables explicativas son las nueve variables categóricas que hacen referencia al uso de las nuevas tecnologías en la consulta de datos jurídicos (tabla 4.4).

Consulta documentación jurisprudencia en publicaciones papel (NO/SÍ)
Consulta documentación jurisprudencia en bases de datos (NO/SÍ)
Consulta documentación doctrina (NO/SÍ)
Utiliza Internet (SÍ/NO)
Ayudaría red telemática a tomar decisiones (SÍ/NO/NC)
Frecuencia uso bases de datos de CGPJ (POCO/REGULAR/FRECUENTE)
Utiliza centro de documentación judicial de CGPJ (SÍ/NO)
Frecuencia uso portal web de CGPJ (NUNCA/REGULAR/FRECUENTE/NC)
Frecuencia uso publicaciones en papel de CGPJ (POCO/REGULAR/FRECUENTE)

Tabla 4.4. Las variables de la matriz X y las categorías

4.3.2. Análisis de los datos por el ACC

Las variables explicativas son en este caso, variables categóricas. Como se explica en el capitulo anterior, *Vegan* no permite trabajar directamente con variables categóricas y requiere que se construya previamente la tabla disyuntiva completa. Por este motivo, el ACC de estos datos se hacen en *ADE4*.

Aplicando la función ACC, se calculan las coordenadas de los jueces, las palabras y las categorías. Los gráficos de *ADE4* se sustituyen por gráficos de *SPAD* que ofrece una mejor visibilidad. Las diferentes nubes de puntos (individuos-juez, columnas-palabra y columnas-modalidad) se presentan en gráficos separados.

A continuación, se presentan los resultados de este análisis. Las reglas de interpretación son parecidas a las reglas de interpretación de un AC.

4.3.3. Valores propios

Eje	Valor propio	Porcentaje	Porcentaje Acumulada
1	0.149	15,32%	15,32%
2	0.108	11,08%	26,40%
3	0.099	10,14%	36,54%
4	0.089	9,19%	45,73%
5	0.083	8,48%	54,21%
6	0.080	8,23%	62,44%
7	0.069	7,12%	69,56%
8	0.063	6,51%	76,07%
9	0.052	5,32%	81,39%
10	0.047	4,88%	86,27%
11	0.039	4,04%	90,31%
12	0.037	3,79%	94,10%
13	0.032	3,28%	97,38%
14	0.026	2,62%	100,00%

Tabla 4.5. Valores propios, porcentajes y porcentajes acumulados

4.3.4. Columnas-palabra

Se tiene que mirar las contribuciones de las palabras que proporcionan una información complementaria a la representación gráfica. Casi todas las funciones que aplican un método de análisis multivariante dan esta información. En cambio, la función *cca()* del *ADE4* no lo hace. Por lo tanto, las contribuciones se calculan externamente creando una pequeña función en *R*. Se calculan las contribuciones de las palabras para identificar cuales son las más contributivas de cada eje.

EJE 1		EJE 2	
Parte positiva		Parte positiva	
Derecho	**5.68%**	Resolución	**8.24%**
Hacer	**4.44%**	Realizar	**5.80%**
Aplicar	**4.30%**	Conflicto	**5.68%**
Bueno	**3.77%**	Social	**3.64%**
Sociedad	**3.21%**	Procurar	**1.10%**
Humano	**3.11%**		
Resolutivo	**2.75%**		
Parte negativa		Parte negativa	
Asunto	**8.45%**	Gran	**4.99%**
Responsabilidad	**5.22%**	Derecho	**4.24%**
Estudiar	**3.71%**	Conocimiento	**3.92%**
Forma	**3.71%**	Capacidad	**3.46%**
Imparcial	**3.45%**	Personal	**2.92%**
Capacidad	**3.26%**	Resolutivo	**2.69%**
Problema	**3.06%**	Práctico	**2.64%**

Tabla 4.6. Las 23 palabras más contributivas

Las palabras más contributivas a la inercia del primer eje son: *aplicar, asunto, bueno, capacidad, derecho, estudiar, forma, hacer, humano, imparcial, problema, resolutivo, responsabilidad* y *sociedad*.

Estas palabras se agrupan en dos grupos (las que están en la parte positiva del eje y las que están en la parte negativa).

En la parte positiva quedan agrupadas las palabras *aplicar, bueno, derecho, hacer, humano, resolutivo, sociedad* (en expresiones como “aplicar el derecho”, “hacer bien su trabajo”, “tener buena formación”, “ser humano”, “ser resolutivo”) y, en la parte negativa; *asunto, capacidad, estudiar, forma, imparcial, problema, responsabilidad* (en expresiones como “estudiar el asunto”, “capacidad de trabajo”, “ser imparcial”, “resolver y solucionar problemas”, “sentido de responsabilidad”).

De la misma manera se definen las palabras más contributivas del segundo eje. La parte positiva contiene las palabras como *conflicto, procurar, realizar, resolución, social* (en expresiones como “resolver el conflicto”, “realizar su trabajo”, “motivar sus resoluciones”, “realidad social”) y, la parte negativa del eje; *capacidad, conocimiento, derecho, gran, personal, práctico, resolutivo* (en expresiones como “gran capacidad de trabajo”, “conocimientos jurídicos”, “aplicar el derecho”, “trato personal”, “sentido práctico”, “ser resolutivo).

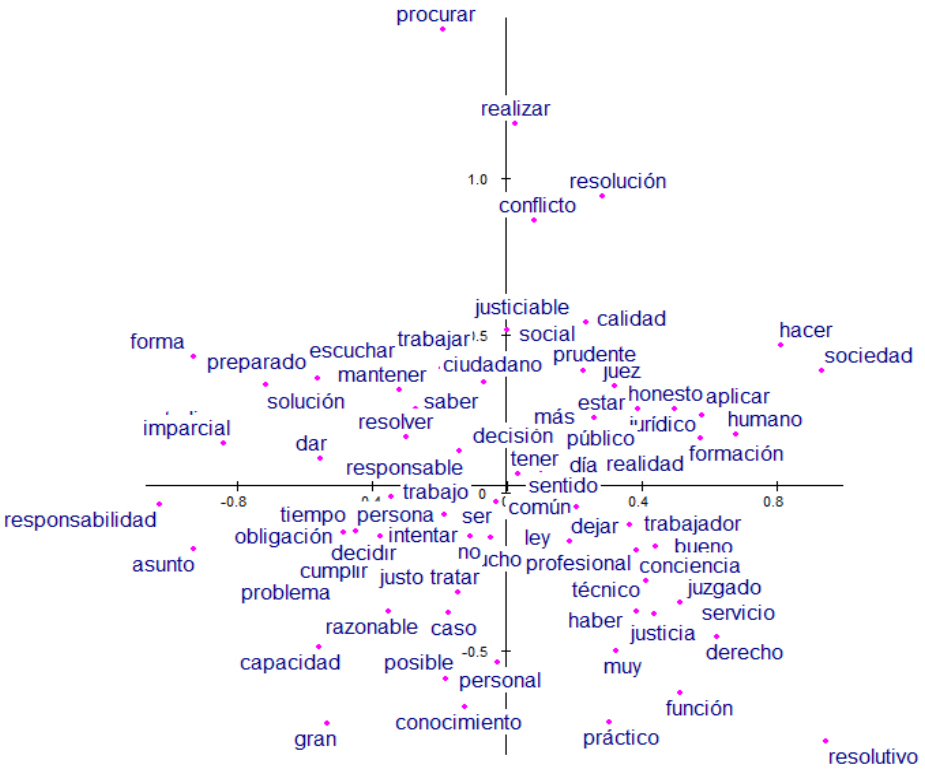

Figura 4.1. La representación gráfica de las palabras

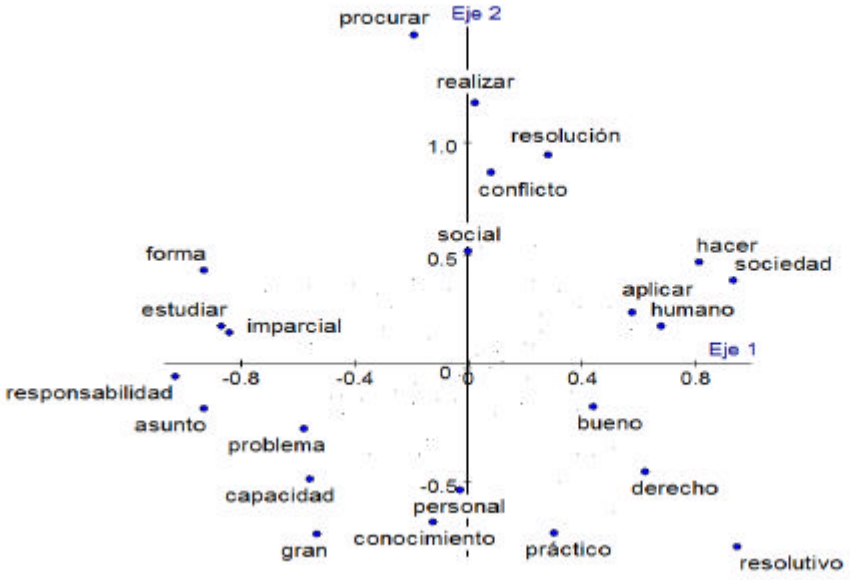

Figura 4.2. La representación gráfica de las palabras más contributivas

4.3.5. Filas-juez

Para los individuos también se calculan las contribuciones y se identifican los individuos que tienen una mayor contribución sobre los ejes. Lo interesante de este gráfico es ver cuales son los jueces que más se han diferenciado del resto y averiguar cual ha sido su vocabulario y palabras usadas para contestar a la pregunta.

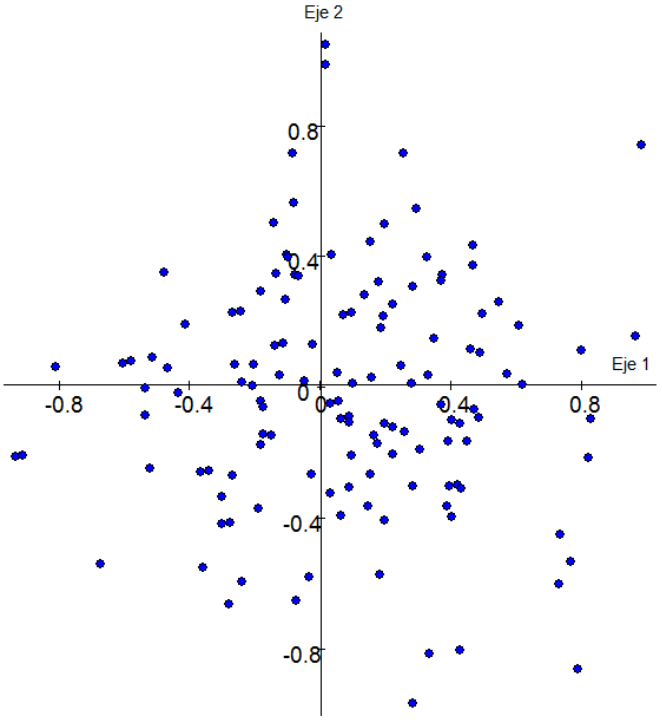

Figura 4.3. La representación gráfica de los individuos (jueces)

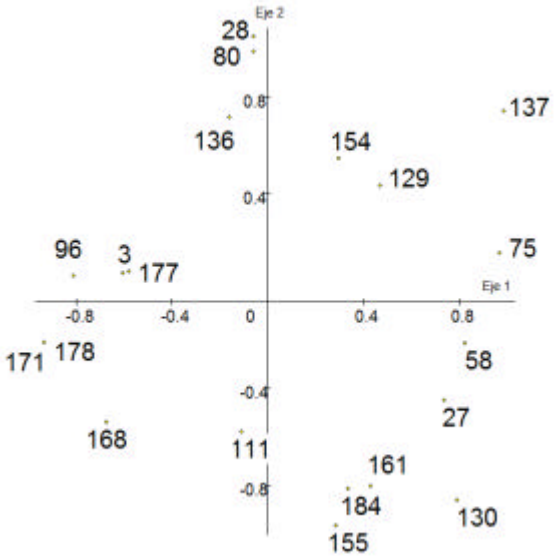

Figura 4.4. La representación gráfica de los jueces más contributivos

En la figura 4.4 se pueden observar cuales son estos individuos con una mayor contribución sobre los ejes. A continuación, tabla 4.7, se pueden ver algunas respuestas de estos individuos más contributivos. En la misma tabla, se observa que el individuo más contributivo del primer eje y del segundo es el mismo individuo. Este individuo ha sido el más contributivo porque ha usado algunas de las más contributivas como *capacidad, bueno, práctico* y *gran.*

Individuo	Eje	Respuesta
130	1 (+)	HA DE REUNIR MUCHAS CUALIDADES, POR EJEMPLO: UNA GRAN SENSATEZ (ES LO MÁS IMPORTANTE); INTELIGENCIA MEDIA COMO MÍNIMO, ORIENTADA AL SENTIDO PRÁCTICO, UNA IMPORTANTE CAPACIDAD DE DECISIÓN, UNA BUENA FORMACIÓN JURÍDICA, UN BUEN TALANTE PERSONAL, CAPACIDAD
137	1 (+)	EL QUE ES CAPAZ DE MIRAR CADA ASUNTO PERSONALIZADAMENTE Y VER EN EL MISMO NO UN NÚMERO SINO EL PROBLEMA DE UNAS PERSONAS Y DESPUÉS RESOLVERLO TÉCNICAMENTE CONFORME A LA LEY, ADECUANDO ÉSTA O DEDUCIENDO DE ÉSTA UNA NORMA PARA ESE PROBLEMA CONCRETO
168	1 (-)	UNA PERSONA RESPONSABLE, CON SENTIDO JURÍDICO Y SENTIDO COMÚN Y PERSONA DE SU TIEMPO
171	1 (-)	"ANTE TODO, UNA PERSONA JUSTA; INCLUSO POR ENCIMA DE LA LEY EN ALGUNAS OCASIONES"
28	2 (+)	CUMPLIDOR DE LAS CONDICIONES DEL ART.117 CONSTITUCIÓN ESPAÑOLA (EN ESPECIAL LA INDEPENDENCIA Y LA RESPONSABILIDAD), TAMBIÉN QUE SEA TRABAJADOR
80	2 (+)	UNA PERSONA CON SENTIDO COMÚN
130	2 (-)	"HA DE REUNIR MUCHAS CUALIDADES, POR EJEMPLO: UNA GRAN SENSATEZ (ES LO MÁS IMPORTANTE); INTELIGENCIA MEDIA COMO MÍNIMO, ORIENTADA AL SENTIDO PRÁCTICO, UNA IMPORTANTE CAPACIDAD DE DECISIÓN, UNA BUENA FORMACIÓN JURÍDICA, UN BUEN TALANTE PERSONAL, CAPACIDAD
111	2 (-)	QUIEN LAS DICTA DESDE LA PERSPECTIVA DEL JUSTICIABLE Y LE INTENTA EXPLICAR LAS RAZONES DE QUE SU PRETENSIÓN SEA ACOGIDA O NO

Tabla 4.7. Las respuestas de 4 individuos más contributivos de cada eje (2 parte positiva y 2 parte negativa)

4.3.6. Columnas-modalidad

El ACC facilita la información que corresponde a las coordenadas de las modalidades pero *ADE4* representa, de cada variable activa, todas las modalidades menos una porque hay una redundancia en el conjunto de las modalidades (la subnube de las modalidades de una misma variable está centrada). Aprovechando las facilidades ofrecidas por *SPAD*, se construye una gráfica en la cual figuran todas las modalidades, lo que facilita la interpretación.

Las variables categóricas no tienen contribuciones en ACC ya que afectan el análisis de una manera indirecta y no directa. No obstante, se puede estudiar si ocupan una posición significativa mediante un test clásico. El p-value, por comodidad de lectura, se traduce en "valor-test".

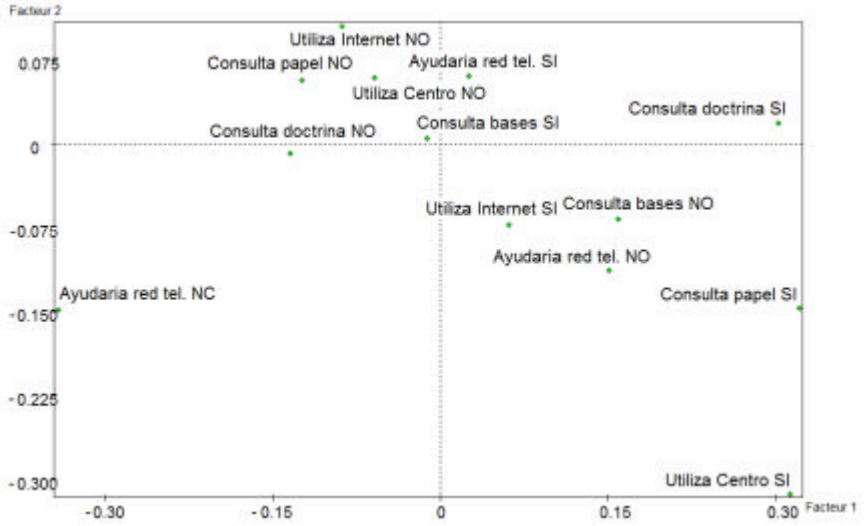

Figura 4.5. La representación gráfica de las modalidades (SI/NO)

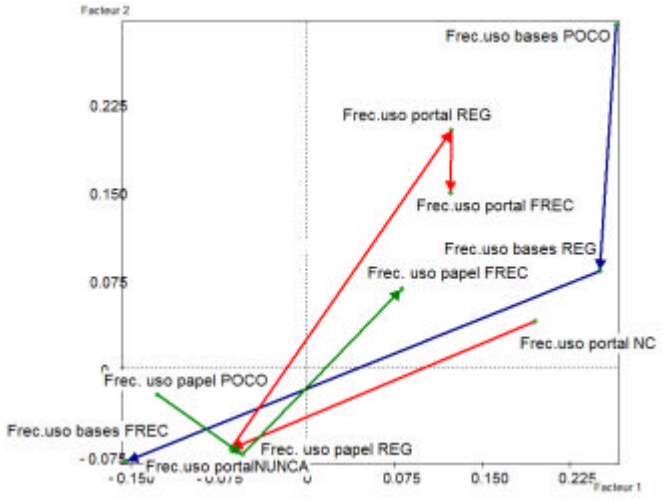

Figura 4.6. La representación gráfica de las modalidades de las variables de frecuencias

Leyendo conjuntamente las figuras 4.5 y 4.6, se observa que se identifican 4 perfiles de jueces, figura 4.7.

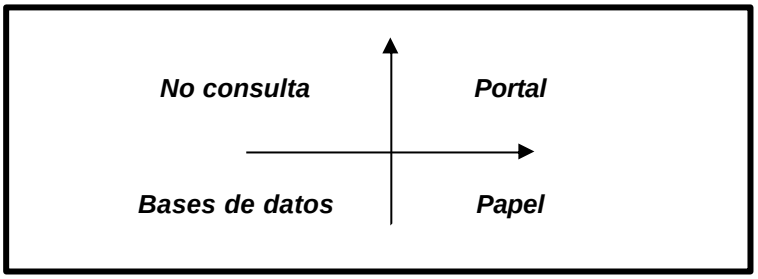

Figura 4.7. Los perfiles de los jueces según las variables activas del análisis

4.3.7. Columnas suplementarias

También es posible proyectar otras variables categóricas que no se han usado en el análisis como activas. Aquí se emplean la variables: edad recodificada, Edad×Sexo y Joven/mayor×Sexo (en este caso joven y mayor se refieren a los años de experiencia. Los jóvenes son los que llevan hasta 4 años y los mayores son los que llevan más de 4 años).

La figura 4.7 muestra que las categorías que corresponden a los hombres y las que corresponden a las mujeres, se separan con claridad (eje 1). Se tendría que investigar la razón de la existencia de un vocabulario claramente diferenciado de hombres y mujeres.

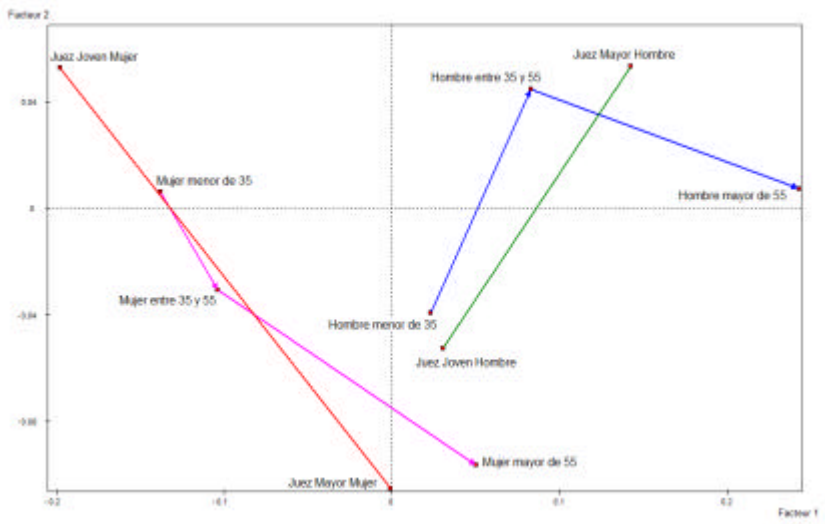

Figura 4.8. La representación gráfica de las modalidades de las variables SexoXEdad y SexoXJuez

4.4. Síntesis de los resultados

Se utilizan métodos de clasificación para segmentar el primer plano factorial en zonas homogéneas. Se usa un método de clasificación por partición directa (*CLARA*), y otro método de clasificación jerárquica, de manera que, la información que da un método complementa a la de otro.

Según el método de *CLARA*, figura 4.9, la mejor clasificación es de 4 clases (el valor más alto de “ancho de la silhouette”). La clasificación jerárquica ascendente también considera una clasificación en 4 clases como la mejor opción, figura 4.10.

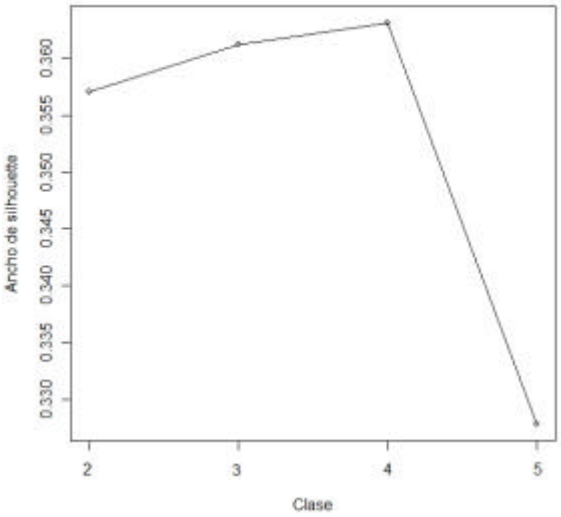

Figura 4.9. Valores de “ancho de la silhouette” de clasificación con CLARA

```
NUM. AINE  BENJ   EFF.    POIDS    INDICE   HISTOGRAMME DES INDICES DE NIVEAU
348   322   155     3     13.00   0.00007   *
349   342   317     7     49.00   0.00007   *
350   346   239    10     37.00   0.00008   *
351   319   336     6     31.00   0.00009   *
352   289   291     6     30.00   0.00009   *
353   312   326     9     41.00   0.00009   *
354   101   304     4     17.00   0.00009   *
355   278   344     9     47.00   0.00010   *
356   328    91     6     39.00   0.00011   *
357   109   285    10     49.00   0.00011   *
358   330   334     7     38.00   0.00012   *
359   337   305    14     82.00   0.00015   *
360   118   293    10     64.00   0.00016   *
361   136   325     4     22.00   0.00018   *
362   339   338     5     20.00   0.00023   *
363   351   288    10     64.00   0.00025   *
364   354   358    11     55.00   0.00025   *
365   343   352    10     59.00   0.00026   *
366   299   340     7     52.00   0.00029   *
367   324   335     8     31.00   0.00031   *
368   320   327     5     18.00   0.00053   *
369   349   321    12     67.00   0.00054   *
370   277   269     6     30.00   0.00062   *
371   353   357    19     90.00   0.00063   *
372   366   362    12     72.00   0.00066   *
373   298   347    21    125.00   0.00068   *
374   130   341     4     25.00   0.00071   *
375   364   345    21    108.00   0.00090   *
376   373   350    31    162.00   0.00092   *
377   367   360    18     95.00   0.00100   *
378   355   356    15     86.00   0.00122   *
379   374   348     7     38.00   0.00150   **
380   377   168    19    108.00   0.00160   **
381   311   361     7     37.00   0.00164   **
382   247   371    26    125.00   0.00172   **
383   375   359    35    190.00   0.00205   **
384   365   368    15     77.00   0.00208   **
385   370   363    16     94.00   0.00208   **
386   378   372    27    158.00   0.00299   ***
387   384   137    16     84.00   0.00319   ***
388   309   380    22    140.00   0.00402   ****
389   383   369    47    257.00   0.00424   ****
390   376   382    57    287.00   0.00537   *****
391   387   386    43    242.00   0.00755   *******
392   385   390    73    381.00   0.01231   **********
393   391   381    50    279.00   0.01426   ***********
394   389   379    54    295.00   0.01536   ************
395   388   392    95    521.00   0.02439   ********************
396   393   394   104    574.00   0.03929   *********************************
397   396   395   199   1095.00   0.09857   ****************************************************************************************
```

4 clases

Figura 4.10. Histograma de índices de nivel de la clasificación jerárquica

Las dos particiones, *CLARA* y jerárquica, son muy similares. A continuación, se usarán las clases formadas en *SPAD* para hacer la descripción de estas clases.

Las 4 clases que se forman, figura 4.11, se separan unas de las otras mutuamente, de manera que, cada clase ocupe uno de los cuadrantes del plano factorial sin que los individuos de una clase se mezclasen con los de otra (lo que se podría esperar, dando que se parte sólo de los dos primeros ejes).

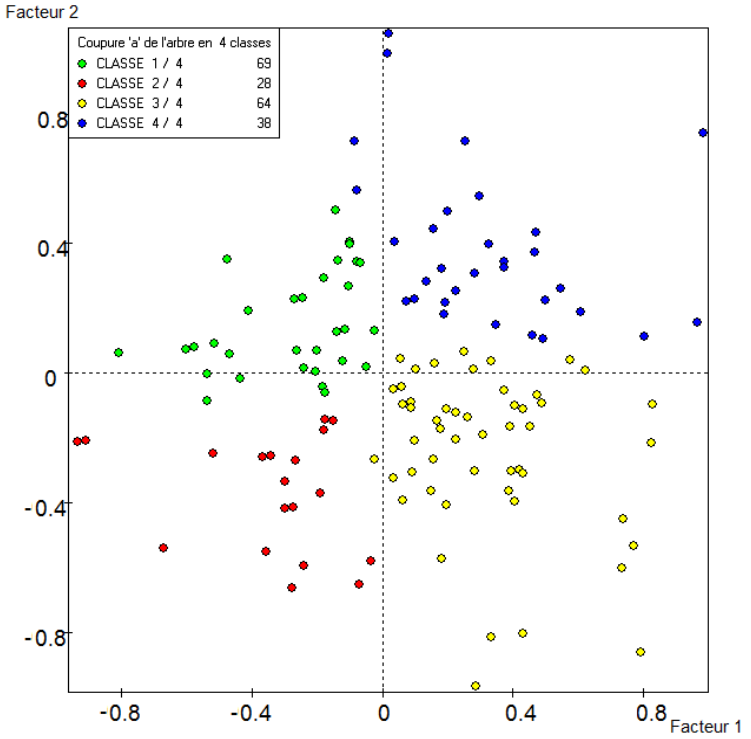

Figura 4.11. La clasificación de los individuos (jueces) en 4 clases

La tabla 4.8 resume los resultados obtenidos con el análisis.

La utilización de la clasificación y la descripción de las clases, muestra que la lectura del primer plano factorial se tiene que hacer según los 4 cuadrantes.

No se debe olvidar que el ACC estudia la variabilidad de las palabras. En este caso, el tamaño del corpus es limitado y los resultados son bastante frágiles.

De este estudio surge una interesante pregunta: el vocabulario de los hombres y de las mujeres difiere. Se tendría que investigar si corresponde a una concepción distinta del trabajo.

% int. - % global	CLASE 1
52.82 - 26.76	Juez Joven Mujer
92.94 - 72.05	No consulta documentos en papel
99.15 - 84.11	No utiliza centro de doc. judicial
84.46 - 62.28	Usa frecuentemente bases
89.27 - 69.13	No consulta doctrina
70.34 - 48.95	Formación en la escuela de jueces
41.53 - 24.20	Mujer menor de 35
50.00 - 31.60	No magistrado
57.63 - 40.73	No utiliza internet
55.65 - 44.38	Valoración buena de for. recibida
48.02 - 38.81	Expectativas cumplidas de carrera
97.18 - 92.97	Consulta bases
9.32 - 6.30	Valoración de vida baja
72.60 - 67.67	Red tel. ayudaría a tomar decisiones

Palabras características
estudiar, imparcial

Respuestas características

Persona que sepa escuchar, imparcial y uso presionable externamente resolviendo en conciencia lo que cree justo.

Una persona será, estudiosa, razonable, de principios, con un sentido de la ética muy pronunciado. Totalmente imparcial y objetivo, con plena libertad e independencia a la hora de resolver las cuestiones. Sepa escuchar.

% int. - % global	CLASE 4
30.91 - 9.04	Usa bases poco
60.45 - 30.87	Consulta doctrina
57.73 - 32.69	Juez mayor hombre
39.09 - 19.27	Usa regularmente el portal web
50.00 - 28.68	Usa regularmente bases
45.45 - 26.48	Hombre entre 35 y 55
70.91 - 51.05	No formación en la escuela de jueces
85.00 - 68.40	Magistrado
82.27 - 67.67	Red tel. ayudaría a tomar decisiones
14.55 - 7.67	Valoración muy buena de for. recibida
59.09 - 47.85	Usa frecuentemente doc. en papel
49.55 - 38.81	Expectativas cumplidas de carrera
12.27 - 7.03	No consulta bases
90.45 - 84.11	No utiliza centro de doc. judicial

Palabras carac terísticas
resolución, conflicto, procurar, realizar, hacer, sociedad

Respuestas características

Ser un buen ciudadano que realiza su labor de pacificar el conflicto haciendo justicia. Lo que conlleva a la serenidad, el reposo, el buen hacer diario, todo ello con decisión cada vez más rápida en una sociedad que nos está soluciones.

Persona que resuelve de forma equilibrada; prudencia y sentido común; procurando la paz social, resolución de los conflictos eficazmente.

% int. - % global	CLASE 2
97.73 - 62.28	Usa frecuentemente bases
98.86 - 69.13	No consulta doctrina
90.91 - 59.27	Utiliza Internet
88.07 - 72.05	No consulta papel
85.23 - 68.68	No usa nunca el portal web
42.05 - 31.60	No magistrado
49.43 - 38.81	Expectativas cumplidas de carrera
51.14 - 41.55	Valoración regular de for. recibida

Palabras características
capacidad, asunto, gran, conocimientos, problema

Respuestas características

Una persona con amplios conocimientos jurídicos, permanentemente actualizados, con ideas claras, capacidad de ver el problema, capacidad de trabajo, gran sentido común.

El que es capaz de mirar cada asunto personalizadamente y ver en el mismo no un número sino el problema de unas personas y después resolverlo técnicamente conforme a la ley, adecuando esta o deduciendo de esta una norma para ese problema concreto.

% int. - % global	CLASE 3
57.10 - 27.95	Consulta documentos en papel
33.62 - 15.89	No utiliza centro de doc. judicial
46.09 - 28.68	Usa regularmente bases
47.83 - 30.87	Consulta doctrina
32.46 - 19.00	Red tel. no ayudaría a tomar decisiones
60.58 - 45.48	Expectativas reg. cumplidas de carrera
82.03 - 68.40	Magistrado
30.14 - 19.36	Juez mayor mujer
24.93 - 16.16	Hombre menor de 35
30.14 - 21.19	Juez joven hombre
17.39 - 11.23	Hombre mayor de 55
50.72 - 41.55	Valoración regular de for. recibida
59.13 - 51.05	No formación en la escuela de jueces
75.07 - 68.68	No usa nunca el portal web

Palabras características
Derecho, resolutivo, trabajador, honesto

Respuestas características

El que no es orgulloso, el que es trabajador, el que es consciente que con su trabajo afecta a la vida de las personas y el que tiene mucho sentido común.

Un buen trabajador como cualquier oficio; un buen estudioso; persona con ganas de aprender; resolutivo; prudente.

Tabla 4.8. La descripción de las clases, las palabras y las respuestas características

4.5. La comparación entre ACC y AC simple

La diferencia entre aplicar el método del ACC y el método del AC simple, se queda reflejada en la comparación de los gráficos proporcionados por los dos métodos.

Como se explica en los anteriores capítulos, el método del AC puede no ser apropiado para tratar con respuestas abiertas porque las frecuencias son muy variadas y la tabla léxica se puede dividir en subtablas disjuntas.

En estos casos, se usa el ACC en lugar del AC simple. Después, se puede utilizar un método de clasificación.

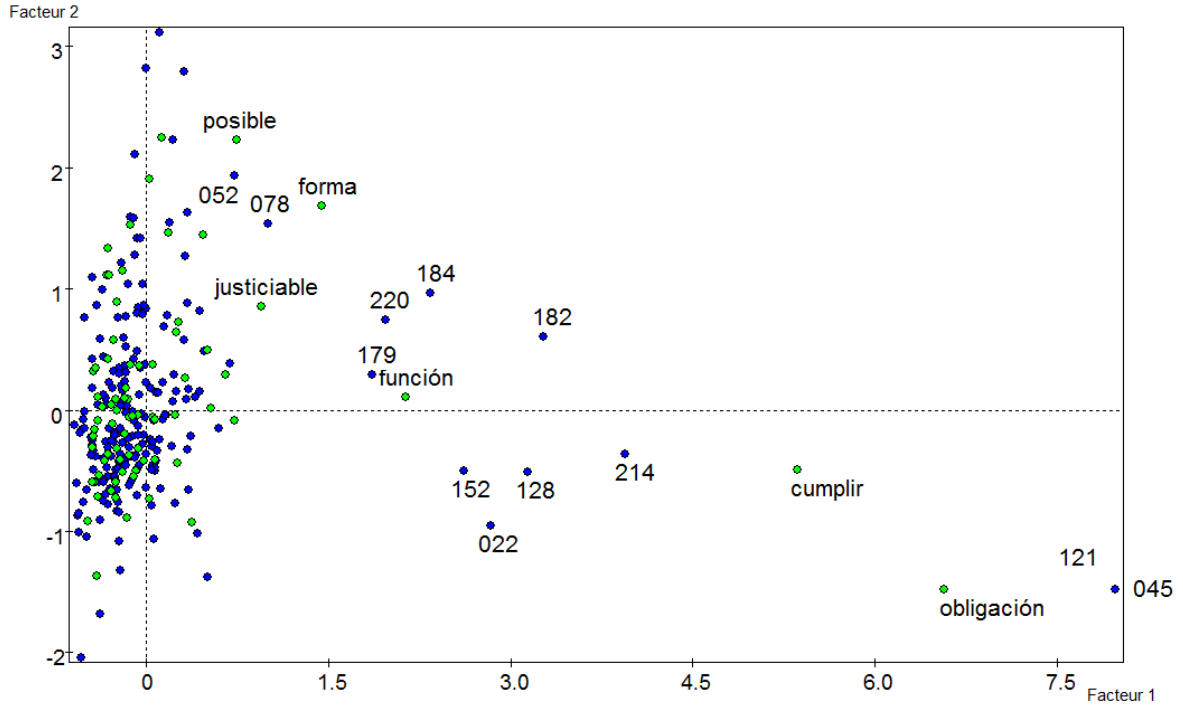

Figura 4.12. AC simple de las frecuencias de las palabras y los individuos

CONCLUSIONES

En conclusión de este proyecto de fin de carrera, conviene hacer comentarios relativos a la aplicación, la metodología y la importancia del trabajo requerido por el proyecto de fin de carrera.

En lo que concierne la aplicación, se debe mencionar que:

- El número de jueces entrevistados es relativamente pequeño (268 en total de bs cuales 69 se eliminaron en el análisis final) considerando que se analizan preguntas abiertas. En este caso, se requiere disponer de un corpus de al menos 5000 ocurrencias. En nuestro caso, el corpus tiene una longitud de 4187 cuando se consideran los 268 jueves. A parte de esto, el hecho que todos los encuestados compartan una misma "cultura", la cultura de juez, y el hecho que la pregunta "*Qué es un buen juez*?" es una pregunta que no permite interpretaciones muy variadas, conducen a unos resultados de estructura marcada. Por lo tanto, no se observan agrupaciones de las palabras muy claras y la interpretación no deja de ser frágil.

- Después de terminar este proyecto, se repetirá la aplicación del método sobre otro conjunto de datos para conocer mejor el proceso y su aplicabilidad. No se debe olvidar que no hay experiencia previa de la aplicación de este método en el análisis textual.

En lo que concierne la metodología, se puede decir que:

- Se ha experimentado la posibilidad de usar el análisis canónico de correspondencias en las tablas de frecuencias de los datos textuales como un método efectivo.

- La introducción de las variables consideradas explicativas como activas dentro del análisis ha enriquecido muchísimo la interpretación permitiendo hacer observaciones desde nuevos puntos de vista.

- Se ha comprobado que se puede utilizar un método de clasificación en complemento del ACC, de lo cual no había experiencia previa, según lo que sabemos.

- Se ha cumplido el objetivo del proyecto de aplicar el análisis canónico de correspondencias a datos textuales y valorar su aportación a este tipo de análisis.

Cuanto al propio proyecto, quiero mencionar que

❖ Ha significado para mí aprender nuevas técnicas y métodos de análisis multivariante como el ACC y los métodos de clasificación *PAM* y *CLARA*.

❖ Aprender el funcionamiento del paquete estadístico *R* y, dentro de *R*, de los paquetes *tm*, *Vegan*, *FactoMineR*, *ADE4* y *Cluster*. Para iniciarme en el software *R,* he seguido el curso sobre *R* ofrecido para los estudiantes del master y doctorado durante el primer cuatrimestre de este curso.

❖ Tuve que crear rutinas propias en *R*, en particular para completar el paquete *tm* que, por ser un paquete nuevo, tiene muchas insuficiencias. Así, se ha tenido que hacer manualmente ciertos procesos y crear pequeñas funciones.

❖ Al principio del proyecto se había hecho un planning detallado. Dicho planning se ha respetado con bastante exactitud. En este sentido se ha cumplido el objetivo de terminar el proyecto en junio y presentarlo antes de las evaluaciones del fin de curso.

BIBLIOGRAFÍA

- Ayuso, M.; Álvarez-Esteban, R.; Bécue-Bertaut, M.;. Statistical study of judicial practices. *Lectura notes in computer science,* 2005, vol. 3369, p. 25-35.
- Chessel, D.; Dufour, A.; Dray, S. *Analysis of ecological data: exploratory and euclideans method in environmental sciences* [en línia]. Versión 1.4-8. Lyon: University Claude Bernard Lyon 1, 2008 [Consulta: 20 de abril de 2008]. Disponible a: <http://cran.r-project.org/>.
- Escofier, B.; Pagès, J. *Analyses factorielles simples et multiples*. Paris: Dunod, 1990. (Traducción en castellano, publicado por la Universidad del País Vasco).
- Feinerer, I. *Text minig package* [en línia]. Versión 0.2-3.7. Viena: Vienna University of Economics and Business Administration, 2007 [Consulta: 10 de enero de 2008]. Disponible a: <http://cran.r-project.org/>.
- Feinerer, I. *Introduction to the tm package* [en línia]: *text minig in R.* CRAN, 2007 [Consulta: 13 de enero de 2008]. Disponible a: <http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>.
- Feinerer, I.; Hornik, K. *Text mining of supreme administrative court jurisdictions* [en línia]. Viena: ePub, 2007 [Consulta: 20 de enero de 2008]. Disponible a: <http://epub.wu-wien.ac.at/dyn/virlib/wp/mediate/epub-wu-01_bad.pdf?ID=epub-wu-01_bad>.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: data mining, inference and prediction.* Springer-Verlag, 2001.
- Husson, F.; Josse, J.; Le, S.; Mazet, J. *Factor analysis and data mining with R* [en línia]. Versión 1.07. Rennes: Agrocampus Rennes, 2007 [Consulta: 12 de febrero de 2008] . Disponible a: <http://cran.r-project.org/>.
- Lebart L., Salem A., Bécue. *Análisis estadístico de textos*. Lleida: Milenio, 2000.
- Legendre, P.; Legendre, L. *Numerical Ecology.* 2ª ed. Ámsterdam: Elsevier Science, 1998.
- Maechler, M. The cluster package [en línia]. Versión 1.11.9. Zurich: ETH Zurich, 2007 [Consulta: 23 de febrero de 2008]. Disponible a: <http://cran.r-project.org/>.
- Oksanen, J.; Kindt, R.; Legendre, P.; O'Hara, B.; Simpson, G.; Henry, M.; Stevens, H. *Community Ecology Package* [en línia]. Versión 1.11-4. University of Helsinki, 2008 [Consulta: 18 de abril de 2008]. Disponible a: <http://cran.r-project.org/>.
- Ter Braak, C. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology,* 1986, vol. 67, p. 1167-1179.

ANEXOS

ANEXO A: LAS PALABRAS LEMATIZADAS

ACTÚA + ACTUAR => **ACTUAR**

ADECUADA + ADECUADO => **ADECUADO**

ADMINISTRA + ADMINISTRAR => **ADMINISTRAR**

AFRONTA + AFRONTAR => **AFRONTAR**

AJENAS + AJENOS => **AJENO**

ALGUNA + ALGUNAS => **ALGUNO**

ALTA + ALTOS + ALTO => **ALTO**

AMPLIO + AMPLIOS => **AMPLIO**

ANALIZA + ANALIZAR => **ANALIZAR**

APLICA + APLICANDO + APLICAR + APLIQUE => **APLICAR**

APRUEBA + APROBADO => **APROBAR**

ASUME + ASUMIR => **ASUMIR**

ASUNTO + ASUNTOS => **ASUNTO**

ATENDIENDO + ATIENDE => **ATENDER**

BUEN + BUENA + BUENAS => **BUENO**

BUSCA + BUSCANDO + BUSCAR => **BUSCAR**

CASO + CASOS => **CASO**

CERCANA + CERCANO => **CERCANO**

CIUDADANO + CIUDADANOS => **CIUDADANO**

CLARA + CLARO + CLARAS => **CLARO**

COMPLETO + COMPLETA => **COMPLETO**

COMPRENDA + COMPRENDE + COMPRENDER => **COMPRENDER**

COMPRENSIVA + COMPRENSIVO => **COMPRENSIVO**

CONCILIADOR + CONCILIADORA => **CONCILIADOR**

CONDICIÓN + CONDICIONES => **CONDICIÓN**

CONFLICTO + CONFLICTOS => **CONFLICTO**

CONOCE + CONOZCA => **CONOCER**

CONOCIMIENTO + CONOCIMIENTOS => **CONOCIMIENTO**

CONSECUENCIA + CONSECUENCIAS => **CONSECUENCIA**

CONSIGO + CONSIGUE => **CONSEGUIR**

CONTINUA + CONTINUADA => **CONTINUA**

CONVICCIÓN + CONVICCIONES => **CONVICCIÓN**

CREA + CREE + CREERSE => **CREER**

CRITERIO + CRITERIOS => **CRITERIO**

CUAL + CUÁLES => **CUAL**

CUESTIÓN + CUESTIONES => **CUESTIÓN**

CUMPLE + CUMPLEN + CUMPLIENDO + CUMPLIR => **CUMPLIR**

CUYAS + CUYO => **CUYO**

DA + DANDO + DAR => **DAR**

DEBE + DEBERÍAN => **DEBER**

DECIDE + DECIDIR => **DECIDIR**

DICEN + DECIR => **DECIR**

DECISIÓN + DECISIONES => **DECISIÓN**

DEDICÁNDOSE + DEDICA => **DEDICAR**

DEJA + DEJARSE + DEJE => **DEJAR**

DERECHOS + DERECHO => **DERECHO**

EJERCE + EJERCER + EJERZA => **EJERCER**

EQUILIBRADA + EQUILIBRADO => **EQUILIBRADO**

ESCUCHA + ESCUCHAR => **ESCUCHAR**

ESTÁ + ESTAR => **ESTAR**

ESTUDIA + ESTUDIADO + ESTUDIANDO + ESTUDIAR => **ESTUDIAR**

ESTUDIOSO + ESTUDIOSA => **ESTUDIOSO**

EXIGE + EXIGIENDO => **EXIGIR**

FRÍO + FRÍA => **FRÍO**

FUNCIONARIO + FUNCIONARIOS => **FUNCIONARIO**

GUSTA + GUSTE => **GUSTAR**

HA + HABER + HAY + HAYA => **HABER**

HACE + HACER + HACERSE + HACIENDO + HAGA => **HACER**

HONESTA + HONESTO => **HONESTO**

HONRADA + HONRADO => **HONRADO**

HUMANA + HUMANAS => **HUMANO**

IMPARTE + IMPARTIR => **IMPARTIR**

IMPUTADO + IMPUTADOS => **IMPUTADO**

ÍNTEGRA + ÍNTEGRO => **ÍNTEGRO**

INTEGRADA + INTEGRADO => **INTEGRADO**

INTENTA + INTENTANDO + INTENTAR + INTENTE => **INTENTAR**

IR + VA => **IR**

JUDICIAL + JUDICIALES => **JUDICIAL**

JUECES + JUEZ => **JUEZ**

JURÍDICA + JURÍDICAS + JURÍDICO + JURÍDICOS => **JURÍDICO**

JUSTA + JUSTAS => **JUSTA**

JUSTICIAN + JUSTICIANDO => **JUSTICIAR**

LEY + LEYES => **LEY**

LIBERTAD + LIBERTADES => **LIBERTAD**

MANTENER + MANTENGA + MANTIENE => **MANTENER**

MÁXIMA + MÁXIMO => **MÁXIMO**

MEDIA + MEDIO => **MEDIO**

MISMA + MISMO => **MISMO**

MOTIVA + MOTIVAR + MOTIVE => **MOTIVAR**

MUCHA + MUCHAS + MUCHO + MUCHOS => **MUCHO**

NECESIDAD + NECESIDADES => **NECESIDAD**

OBLIGACIÓN + OBLIGACIONES => **OBLIGACIÓN**

PASANDO + PASARSE => **PASAR**

PERMITIDO + PERMITA => **PERMITIR**

PERSONAL + PERSONALES => **PERSONAL**

PERSONA + PERSONAS => **PERSONA**

PIENSA + PIENSE => **PENSAR**

PLAZO + PLAZOS => **PLAZO**

PODER + PODERES => **PODER**

POLÍTICO + POLÍTICOS => **POLÍTICO**

POSEE + POSEER => **POSEER**

POSIBLE + POSIBLES => **POSIBLE**

PRÁCTICA + PRÁCTICO => **PRÁCTICO**

PREPARADA + PREPARADO => **PREPARADO**

PRESIÓN + PRESIONES => **PRESIÓN**

PRINCIPIO + PRINCIPIOS => **PRINCIPIO**

PROBLEMA + PROBLEMAS => **PROBLEMA**

PROCEDIMIENTO + PROCEDIMIENTOS => **PROCEDIMIENTO**

PROCURA + PROCURANDO + PROCURAR => **PROCURAR**

PROFESIONAL + PROFESIONALES => **PROFESIONAL**

PROPIAS + PROPIOS + PROPIO => **PROPIO**

PUBLICA + PÚBLICO => **PÚBLICO**

PUEDA + PUEDE + PUEDEN => **PODER(V)**

RÁPIDA + RÁPIDO => **RÁPIDO**

RAZÓN + RAZONES => **RAZÓN**

REALICE + REALIZA + REALIZAR => **REALIZAR**

REALISTA + REALISTAS => **REALISTA**

RECIBE + RECIBIR => **RECIBIR**

RECTO + RECTA => **RECTO**

REFLEXIONADO + REFLEXIONE => **REFLEXIONAR**

RESOLUCIÓN + RESOLUCIONES => **RESOLUCIÓN**

RESOLUTIVA + RESOLUTIVO => **RESOLUTIVO**

RESOLVIENDO + RESOLVER + RESOLVERLO + RESUELTO + RESUELVA + RESUELVE => **RESOLVER**

RESPETA + RESPETE => **RESPETAR**

RESPONDE + RESPONDER => **RESPONDER**

REÚNA + REÚNE + REUNIR => **REUNIR**

RODEA + RODEAN => **RODEAR**

SABE + SABER + SABES + SABIENDO + SEPA => **SABER**

ES + SEA + SER + SERÁ + SERLO + SIENDO + SON => **SER**

SENCILLA + SENCILLO => **SENCILLO**

SENTIR + SENTIRSE => **SENTIR**

SIRVA + SIRVE => **SERVIR**

SITUACIÓN + SITUACIONES => **SITUACIÓN**

SOCIAL + SOCIALES => **SOCIAL**

SOLUCIÓN + SOLUCIONES => **SOLUCIÓN**

SOLVENTANDO + SOLVENTAR => **SOLVENTAR**

SOSEGADA + SOSEGADO => **SOSEGADO**

SUFICIENTE + SUFICIENTES **=> SUFICIENTE**

TÉCNICAS + TÉCNICO => **TÉCNICO**

TENER + TENGA + TENIENDO + TIENE + TIENEN => **TENER**

TEÓRICA + TEÓRICO => **TEÓRICO**

TOMANDO + TOMARSE => **TOMAR**

TRABAJADOR + TRABAJADORA => **TRABAJADOR**

TRABAJA + TRABAJAR => **TRABAJAR**

TRATAR + TRATE + TRATA => **TRATAR**

VALORACIÓN + VALORACIONES => **VALORACIÓN**

VIVA + VIVE => **VIVIR**

ANEXO B: LAS PALABRAS ELIMINADAS

Las palabras de la lista que está a continuación, se han eliminado de la matriz documentos × palabras por no ser de las clases de palabras escogidas (sustantivos, verbos, adjetivos y adverbios). Las palabras eliminadas son:

a	**al**	**alguien**	**aquel**	**cada**	**como**	**con**	**de**	**del**	**demás**	**e**
el	**en**	**la**	**las**	**le**	**lo**	**los**	**o**	**para**	**por**	**que**
se	**sin**	**sobre**	**su**	**sus**	**todo**	**un**	**una**	**uno**	**y**	

ANEXO C: LA FUNCIÓN CA DE R

Función => **CA(X,ncp,row.sup,col.sup,graph,axes,row.w)**

X: Tabla de contingencia de las variables categóricas.

ncp: Número que indica cuantas de las primeras dimensiones tendrán reflejados sus resultados como salida. Por defecto se cogen 5 primeras dimensiones y como máximo puede ser el menor número de categorías de las dos variables.

row.sup: Las filas, categorías de la variable en fila, suplementarias de la tabla de contingencia. Las filas suplementarias no contribuyen en la creación de los subespacios.

col.sup: Las columnas, categorías de la variable en columna, suplementarias. De la misma manera que las filas suplementarias, las columnas suplementarias no contribuyen en la creación de los subespacios.

graph: Valor booleano para indicar si se quiere disponer o no el gráfico de este análisis. La opción por defecto es “TRUE”.

axes: Vector que tiene como elementos las dos dimensiones con las cuales se forma el subespacio del gráfico. La opción por defecto es “c(1,2)”, las dimensiones de mayor información.

row.w: Los pesos de las filas. Por defecto se da un peso uniforme a todas las filas pero en el caso de considerar más importante unas filas que las otras, se podrían modificar estos pesos.

Los parámetros de salida son:

eig: Los valores propios.

col: Matriz con toda la información que hace referencia a las categorías en columnas (coordenadas, cosinus al cuadrado y contribuciones).

row: Matriz con toda la información que hace referencia a las categorías en filas (coordenadas, cosinus al cuadrado y contribuciones).

col.sup: Matriz con toda la información que hace referencia a las categorías suplementarias en columnas (coordenadas y cosinus al cuadrado).

row.sup: Matriz con toda la información que hace referencia a las categorías suplementarias en filas (coordenadas y cosinus al cuadrado).

call: Tabla de contingencia y las proporciones marginales de las categorías.

ANEXO D: LA FUNCIÓN CLARA DE R

Función => **clara(X,k,metric,stand,samples,sampsize,trace,medoids.x,keep.data,rngR)**

X: Las coordenadas de los ejes factoriales.

k: Numero de grupos. Tiene que ser un valor entero entre 0 i *n*, donde *n* es número de observaciones.

metric: Mesura de disimilaridad. Opciones disponibles son “euclidean” y “manhattan”.

stand: Booleano que indica si los valores en *X* se estandarizan o no antes de calcular las disimilaridades.

samples: Número de muestras a sacar de la muestra total. Por defecto se sacan 5 muestras.

sampsize: Número de observaciones de las muestras extraídas. Por defecto se calcula el mínimo entre n, número de observaciones, y 40 + 2*k donde k es número de grupos.

trace: Número de resultados parciales del algoritmo.

medoids.x: Booleano para decidir si se quiere tener los valores de *medoids*.

keep.data: Booleano que indica si la muestra total aparezca o no entre los resultados finales.

rngR: Booleano que indica si se quiere usar la función generadora de números aleatorios en lugar del valor primitivo de la función *CLARA*. El uso de este parámetro hará que *CLARA* devuelva cada vez un resultado diferente respecto a las clasificaciones de los grupos.

Los resultados de salida:

sample: Las observaciones de la mejor muestra extraída, la cual se usa en el algoritmo para hacer la partición final.

medoids: Matriz de los valores de *medoids*. Es nulo si *medoids.x=FALSE*.

i.med: Índices de *medoids*.

clustering: El grupo correspondiente de cada observación.

clusinfo: Matriz que tiene coma filas los grupos y como columnas varias indicadores sobre estos grupos.

diss: La matriz de las disimilaridades entre las observaciones.

silinfo: Devuelve “ancho de la silhouette” de las observaciones, de la media de los grupos y la media total.

ANEXO E : DESCRIPCIÓN DE LAS CLASES : INFORMACIÓN GENERADA POR SPAD

DESCRIPTION DE PARTITION(S)
DESCRIPTION DE LA Coupure 'a' de l'arbre en 4 classes
CARACTERISATION DES CLASSES PAR LES MODALITES
CARACTERISATION DES CLASSES PAR LES CONTINUES
CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES
DE Coupure 'a' de l'arbre en 4 classes
CLASSE 1 / 4

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				32.33	CLASSE 1 / 4		aa1a	354
13.12	0.000	63.82	52.82	26.76	Juez Joven Mujer	JMSEX	JovM	293
11.44	0.000	41.70	92.94	72.05	Consulta papel NO	Consulta.doc..jurispapel	Mod1	789
10.95	0.000	38.11	99.15	84.11	Utiliza Centro NO	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod2	921
10.83	0.000	43.84	84.46	62.28	Frec.uso bases FREC	Frecuencia uso bases CGPJ 2	AF03	682
10.49	0.000	41.74	89.27	69.13	Consulta doctrina NO	Consulta.documentación.doctrina	Mod1	757
9.82	0.000	46.46	70.34	48.95	SI FORMACION EJB	Formación.EJB.del.CGPJ	SI	536
8.96	0.000	55.47	41.53	24.20	Mujer menor de 35	EDSEX	Ed1M	265
8.86	0.000	51.16	50.00	31.60	NO MAGISTRADO	Magistrado	NO	346
7.77	0.000	45.74	57.63	40.73	Utiliza Internet NO	Utiliza.Internet	Mod2	446
5.12	0.000	40.53	55.65	44.38	VAL. FOR. BUENA	Valoración.formación.facultad	BUEN	486
4.28	0.000	39.82	49.72	40.37	menor de 35	EDACLA	Eda1	442
4.24	0.000	40.00	48.02	38.81	EXP. CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	CUM	425
3.89	0.000	33.79	97.18	92.97	Consulta bases SI	Doc.juris..base.datos	Mod2	1018
3.84	0.000	54.17	11.02	6.58	VAL. OPOS. MUY BUENA	Valoración.oposición	MBUE	72
2.65	0.004	47.83	9.32	6.30	VAL. VIDA BAJA	Valoración.de.calidad.de.vida	BAJ	69
2.36	0.009	34.68	72.60	67.67	Ayudaria red tel. SI	Ayudaría.red.telemática.a.tomar.decisiones	Mod1	741
-2.39	0.009	29.00	46.61	51.96	VAL. OPOS. BUENA	Valoración.oposición	BUEN	569
-2.40	0.008	16.98	2.54	4.84	VAL. OPOS. NEGATIVA	Valoración.oposición	NEG	53
-3.14	0.001	23.71	15.54	21.19	Juez Joven Hombre	JMSEX	JovH	232
-3.16	0.001	24.83	20.34	26.48	Hombre entre 35 y 55	EDSEX	Ed2H	290
-3.58	0.000	7.50	0.85	3.65	VAL. VIDA MUY BAJA	Valoración.de.calidad.de.vida	MBAJ	40
-3.71	0.000	26.51	37.29	45.48	EXP. REG. CUMPLIDAR	La.carrera.judicial.responde.a.sus.expectativas	RCUM	498
-3.87	0.000	21.23	12.71	19.36	Juez Mayor Mujer	JMSEX	MayM	212
-3.89	0.000	12.99	2.82	7.03	Consulta bases NO	Doc.juris..base.datos	Mod1	77
-3.91	0.000	17.07	5.93	11.23	Hombre mayor de 55	EDSEX	Ed3H	123
-4.02	0.000	20.67	12.15	19.00	Ayudaria red tel. NO	Ayudaría.red.telemática.a.tomar.decisiones	Mod2	208
-4.11	0.000	18.00	7.63	13.70	más 55	EDACLA	Eda3	150
-4.19	0.000	14.14	3.95	9.04	Frec.uso bases POCO	Frecuencia uso bases CGPJ 2	AF01	99
-4.84	0.000	24.18	31.07	41.55	VAL. FOR. REG	Valoración.formación.facultad	REG	455
-5.10	0.000	16.38	8.19	16.16	Hombre menor de 35	EDSEX	Ed1H	177
-6.83	0.000	18.72	18.93	32.69	Juez Mayor Hombre	JMSEX	MayH	358
-7.77	0.000	23.11	42.37	59.27	Utiliza Internet SI	Utiliza.Internet	Mod1	649
-8.86	0.000	23.63	50.00	68.40	SI MAGISTRADO	Magistrado	SI	749
-9.03	0.000	13.06	11.58	28.68	Frec.uso bases REG	Frecuencia uso bases CGPJ 2	AF02	314
-9.82	0.000	18.78	29.66	51.05	NO FOMACION EJB	Formación.EJB.del.CGPJ	NO	559
-10.49	0.000	11.24	10.73	30.87	Consulta doctrina SI	Consulta.documentación.doctrina	Mod2	338
-10.95	0.000	1.72	0.85	15.89	Utiliza Centro SI	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod1	174
-11.44	0.000	8.17	7.06	27.95	Consulta papel SI	Consulta.doc..jurispapel	Cons	306

CLASSE 2 / 4

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				16.07	CLASSE 2 / 4		aa2a	176
12.10	0.000	25.22	97.73	62.28	Frec.uso bases FREC	Frecuencia uso bases CGPJ 2	AF03	682
10.99	0.000	22.99	98.86	69.13	Consulta doctrina NO	Consulta.documentación.doctrina	Mod1	757
10.02	0.000	24.65	90.91	59.27	Utiliza Internet SI	Utiliza.Internet	Mod1	649
8.25	0.000	42.47	35.23	13.33	Ayudaria red tel. NC	Ayudaría.red.telemática.a.tomar.decisiones	C103	146
6.67	0.000	25.71	61.36	38.36	Frec. uso papel REG	Frecuencia uso publicaciones papel CGPJ 2	AK02	420
5.42	0.000	19.65	88.07	72.05	Consulta papel NO	Consulta.doc..jurispapel	Mod1	789
5.36	0.000	19.95	85.23	68.68	Frec.uso portalNUNCA	Frecuencia uso portal web CGPJ 2	AI01	752
4.68	0.000	50.00	10.23	3.29	EXP. NC	La.carrera.judicial.responde.a.sus.expectativas	NC	36
3.16	0.001	29.21	14.77	8.13	VAL. OPOS. NC	Valoración.oposición	NC	89
3.11	0.001	21.39	42.05	31.60	NO MAGISTRADO	Magistrado	NO	346
3.04	0.001	20.47	49.43	38.81	EXP. CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	CUM	425
2.86	0.002	21.55	36.36	27.12	VAL. OPOS. REG	Valoración.oposición	REG	297
2.72	0.003	19.78	51.14	41.55	VAL. FOR. REG	Valoración.formación.facultad	REG	455
-2.40	0.008	0.00	0.00	2.47	Mujer mayor de 55	EDSEX	Ed3M	27
-2.58	0.005	12.85	36.36	45.48	EXP. REG. CUMPLIDAR	La.carrera.judicial.responde.a.sus.expectativas	RCUM	498
-2.60	0.005	10.10	11.93	19.00	Ayudaria red tel. NO	Ayudaría.red.telemática.a.tomar.decisiones	Mod2	208
-2.65	0.004	0.00	0.00	2.83	VAL. FOR. MUY NEG	Valoración.formación.facultad	MNEG	31
-2.66	0.004	5.41	2.27	6.76	EXP. NO CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	NCUM	74
-2.96	0.002	8.00	6.82	13.70	más 55	EDACLA	Eda3	150
-3.11	0.001	13.62	57.95	68.40	SI MAGISTRADO	Magistrado	SI	749
-3.16	0.001	0.00	0.00	3.65	VAL. VIDA MUY BAJA	Valoración.de.calidad.de.vida	MBAJ	40
-4.41	0.000	12.55	52.84	67.67	Ayudaria red tel. SI	Ayudaría.red.telemática.a.tomar.decisiones	Mod1	741
-4.50	0.000	0.00	0.00	6.30	VAL. VIDA BAJA	Valoración.de.calidad.de.vida	BAJ	69
-4.62	0.000	0.00	0.00	6.58	VAL. OPOS. MUY BUENA	Valoración.oposición	MBUE	72
-4.63	0.000	6.16	7.39	19.27	Frec.uso portal REG	Frecuencia uso portal web CGPJ 2	AI02	211
-5.42	0.000	6.86	11.93	27.95	Consulta papel SI	Consulta.doc..jurispapel	Cons	306
-5.60	0.000	0.00	0.00	9.04	Frec.uso bases POCO	Frecuencia uso bases CGPJ 2	AF01	99
-7.92	0.000	7.06	21.02	47.85	Frec. uso papel FREC	Frecuencia uso publicaciones papel CGPJ 2	AK03	524
-9.79	0.000	1.27	2.27	28.68	Frec.uso bases REG	Frecuencia uso bases CGPJ 2	AF02	314
-10.02	0.000	3.59	9.09	40.73	Utiliza Internet NO	Utiliza.Internet	Mod2	446
-10.99	0.000	0.59	1.14	30.87	Consulta doctrina SI	Consulta.documentación.doctrina	Mod2	338

CLASSE 3 / 4

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				31.51	CLASSE 3 / 4		aa3a	345
14.22	0.000	64.38	57.10	27.95	Consulta papel SI	Consulta.doc..jurispapel	Cons	306
10.40	0.000	66.67	33.62	15.89	Utiliza Centro SI	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod1	174
8.41	0.000	50.64	46.09	28.68	Frec.uso bases REG	Frecuencia uso bases CGPJ 2	AF02	314
8.04	0.000	48.82	47.83	30.87	Consulta doctrina SI	Consulta.documentación.doctrina	Mod2	338
7.40	0.000	53.85	32.46	19.00	Ayudaria red tel. NO	Ayudaría.red.telemática.a.tomar.decisiones	Mod2	208
6.75	0.000	41.97	60.58	45.48	EXP. REG. CUMPLIDAR	La.carrera.judicial.responde.a.sus.expectativas	RCUM	498
6.71	0.000	37.78	82.03	68.40	SI MAGISTRADO	Magistrado	SI	749
6.08	0.000	54.00	23.48	13.70	más 55	EDACLA	Eda3	150
5.89	0.000	49.06	30.14	19.36	Juez Mayor Mujer	JMSEX	MayM	212
5.27	0.000	69.57	9.28	4.20	EXP. DEFRAUDADO	La.carrera.judicial.responde.a.sus.expectativas	DEF	46
5.12	0.000	48.59	24.93	16.16	Hombre menor de 35	EDSEX	Ed1H	177
4.82	0.000	77.78	6.09	2.47	Mujer mayor de 55	EDSEX	Ed3M	27
4.75	0.000	44.83	30.14	21.19	Juez Joven Hombre	JMSEX	JovH	232
4.15	0.000	48.78	17.39	11.23	Hombre mayor de 55	EDSEX	Ed3H	123
4.10	0.000	38.46	50.72	41.55	VAL. FOR. REG	Valoración.formación.facultad	REG	455
3.80	0.000	52.70	11.30	6.76	EXP. NO CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	NCUM	74
3.57	0.000	36.49	59.13	51.05	NO FOMACION EJB	Formación.EJB.del.CGPJ	NO	559
3.06	0.001	34.44	75.07	68.68	Frec.uso portalNUNCA	Frecuencia uso portal web CGPJ 2	AI01	752
2.90	0.002	35.50	58.55	51.96	VAL. OPOS. BUENA	Valoración.oposición	BUEN	569
2.55	0.005	46.38	9.28	6.30	VAL. VIDA BAJA	Valoración.de.calidad.de.vida	BAJ	69
-2.71	0.003	27.16	38.26	44.38	VAL. FOR. BUENA	Valoración.formación.facultad	BUEN	486
-2.71	0.003	20.00	6.38	10.05	VAL. VIDA NC	Valoración.de.calidad.de.vida	NC	110
-3.17	0.001	28.34	60.87	67.67	Ayudaria red tel. SI	Ayudaría.red.telemática.a.tomar.decisiones	Mod1	741
-3.24	0.001	17.17	4.93	9.04	Frec.uso bases POCO	Frecuencia uso bases CGPJ 2	AF01	99
-3.43	0.000	23.57	20.29	27.12	VAL. OPOS. REG	Valoración.oposición	REG	297
-3.57	0.000	26.31	40.87	48.95	SI FORMACION EJB	Formación.EJB.del.CGPJ	SI	536
-3.64	0.000	25.81	37.10	45.30	entre 35 y 55	EDACLA	Eda2	496
-3.72	0.000	20.85	12.75	19.27	Frec.uso portal REG	Frecuencia uso portal web CGPJ 2	AI02	211
-3.89	0.000	22.41	18.84	26.48	Hombre entre 35 y 55	EDSEX	Ed2H	290
-4.53	0.000	15.75	6.67	13.33	Ayudaria red tel. NC	Ayudaría.red.telemática.a.tomar.decisiones	C103	146
-5.17	0.000	18.87	14.49	24.20	Mujer menor de 35	EDSEX	Ed1M	265
-6.05	0.000	24.78	48.99	62.28	Frec.uso bases FREC	Frecuencia uso bases CGPJ 2	AF03	682
-6.71	0.000	17.92	17.97	31.60	NO MAGISTRADO	Magistrado	NO	346
-8.04	0.000	23.78	52.17	69.13	Consulta doctrina NO	Consulta.documentación.doctrina	Mod1	757
-9.72	0.000	10.24	8.70	26.76	Juez Joven Mujer	JMSEX	JovM	293
-10.30	0.000	13.88	17.10	38.81	EXP. CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	CUM	425
-10.40	0.000	24.86	66.38	84.11	Utiliza Centro NO	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod2	921
-14.22	0.000	18.76	42.90	72.05	Consulta papel NO	Consulta.doc..jurispapel	Mod1	789

CLASSE 4 / 4

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				20.09	CLASSE 4 / 4		aa4a	220
11.06	0.000	68.69	30.91	9.04	Frec.uso bases POCO	Frecuencia uso bases CGPJ 2	AF01	99
10.19	0.000	39.35	60.45	30.87	Consulta doctrina SI	Consulta.documentación.doctrina	Mod2	338
8.54	0.000	35.47	57.73	32.69	Juez Mayor Hombre	JMSEX	MayH	358
7.75	0.000	40.76	39.09	19.27	Frec.uso portal REG	Frecuencia uso portal web CGPJ 2	AI02	211
7.48	0.000	35.03	50.00	28.68	Frec.uso bases REG	Frecuencia uso bases CGPJ 2	AF02	314
6.79	0.000	34.48	45.45	26.48	Hombre entre 35 y 55	EDSEX	Ed2H	290
6.60	0.000	27.91	70.91	51.05	NO FOMACION EJB	Formación.EJB.del.CGPJ	NO	559
6.14	0.000	24.97	85.00	68.40	SI MAGISTRADO	Magistrado	SI	749
5.30	0.000	24.43	82.27	67.67	Ayudaria red tel. SI	Ayudaría.red.telemática.a.tomar.decisiones	Mod1	741
5.12	0.000	27.02	60.91	45.30	entre 35 y 55	EDACLA	Eda2	496
3.87	0.000	38.10	14.55	7.67	VAL. FOR. MUY BUENA	Valoración.formación.facultad	MBUE	84
3.66	0.000	24.81	59.09	47.85	Frec. uso papel FREC	Frecuencia uso publicaciones papel CGPJ 2	AK03	524
3.55	0.000	25.65	49.55	38.81	EXP. CUMPLIDAS	La.carrera.judicial.responde.a.sus.expectativas	CUM	425
3.49	0.000	45.00	8.18	3.65	VAL. VIDA MUY BAJA	Valoración.de.calidad.de.vida	MBAJ	40
3.08	0.001	35.06	12.27	7.03	Consulta bases NO	Doc.juris..base.datos	Mod1	77
2.90	0.002	21.61	90.45	84.11	Utiliza Centro NO	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod2	921
2.55	0.005	38.46	6.82	3.56	VAL. FOR. NEG	Valoración.formación.facultad	NEG	39
-2.70	0.003	11.92	8.18	13.79	Frec. uso papel POCO	Frecuencia uso publicaciones papel CGPJ 2	AK01	151
-2.77	0.003	2.78	0.45	3.29	EXP. NC	La.carrera.judicial.responde.a.sus.expectativas	NC	36
-2.85	0.002	0.00	0.00	2.47	Mujer mayor de 55	EDSEX	Ed3M	27
-2.90	0.002	12.07	9.55	15.89	Utiliza Centro SI	Utiliza.Centro.Documentación.Judicial.CGPJ	Mod1	174
-3.08	0.001	18.96	87.73	92.97	Consulta bases SI	Doc.juris..base.datos	Mod2	1018
-3.20	0.001	12.26	11.82	19.36	Juez Mayor Mujer	JMSEX	MayM	212
-3.23	0.001	5.80	1.82	6.30	VAL. VIDA BAJA	Valoración.de.calidad.de.vida	BAJ	69
-3.28	0.001	12.50	13.18	21.19	Juez Joven Hombre	JMSEX	JovH	232
-3.58	0.000	12.97	17.27	26.76	Juez Joven Mujer	JMSEX	JovM	293
-4.05	0.000	0.00	0.00	4.20	EXP. DEFRAUDADO	La.carrera.judicial.responde.a.sus.expectativas	DEF	46
-4.19	0.000	11.32	13.64	24.20	Mujer menor de 35	EDSEX	Ed1M	265
-4.59	0.000	3.37	1.36	8.13	VAL. OPOS. NC	Valoración.oposición	NC	89
-5.07	0.000	12.67	25.45	40.37	menor de 35	EDACLA	Eda1	442
-5.47	0.000	4.79	3.18	13.33	Ayudaria red tel. NC	Ayudaría.red.telemática.a.tomar.decisiones	C103	146
-6.14	0.000	9.54	15.00	31.60	NO MAGISTRADO	Magistrado	NO	346
-6.60	0.000	11.94	29.09	48.95	SI FORMACION EJB	Formación.EJB.del.CGPJ	SI	536
-8.45	0.000	12.90	44.09	68.68	Frec.uso portalNUNCA	Frecuencia uso portal web CGPJ 2	AI01	752
-10.19	0.000	11.49	39.55	69.13	Consulta doctrina NO	Consulta.documentación.doctrina	Mod1	757
-14.70	0.000	6.16	19.09	62.28	Frec.uso bases FREC	Frecuencia uso bases CGPJ 2	AF03	682

ANEXO F: ABREVIACIONES

AC: Análisis de Correspondencias Simple (**CA** en inglés)

ACC: Análisis Canónico de Correspondencias (**CCA** en inglés)

ADE4: Paquete que se usa para analizar datos ecológicos que incluye la función CCA

PAM: Método de clasificación (Partition Around Medoids)

CLARA: Método de clasificación (Clustering Large Application)

R: Software estadístico libre

tm: Text Mining

Vegan: Paquete para analizar datos de comunidad ecológica que incluye la función CCA

ANEXO G: EL CÓDIGO DE LA FUNCIÓN "SORTTERMDOCMATRIX"

```
SortTermDocMatrix<-function(tdm,minfreq="d")
{
	nr<-nrow(tdm)
	nc<-ncol(tdm)
	dic<-createDictionary(tdm)
	dico<-sort(dic)
	m<-matrix(nr=nr,ncol=nc)
	w<-matrix(nr=nr,ncol=nc)
	guia<-matrix(nr=nc,ncol=3)
	vf<-numeric()
	fd<-matrix(nrow=nr+1,ncol=2)
	guia[1:nc,1]=c(1:nc)
	for (i in 1:nc) guia[i,2] <- which(dico[i]==dic)
	if(minfreq=="d") minfreq<-nr*0.02
	for(i in 1:nr){
		m[i,]<-tdm[i,]
		fd[i,1]<-sum(m[i,])
	}
	for(j in 1:nc) {
		k<-guia[j,2]
		w[1:nr,j]<-m[1:nr,k]
		vf[j]<-sum(w[,j])
	}
	k<-1
	l=length(vf[vf>=minfreq])
	df=matrix(nr=nr,ncol=l)
	vf2<-matrix(nr=1,ncol=l)

	for(j in 1:nc){
		if (vf[j]>=minfreq){
		df[1:nr,k]<-w[1:nr,j]
			vf2[1,k]<-vf[j]
			guia[j,3]<-1
			k<-k+1
```

```
            }
            else guia[j,3]<-0
        }
        for(i in 1:nr) fd[i,2]<-sum(df[i,])
        fd[nr+1,1]<-sum(fd[1:nr,1])
        fd[nr+1,2]<-sum(fd[1:nr,2])
        fd<-as.data.frame(fd)
        dimnames(fd)[[2]][1]<- "Frequency of used words"
        dimnames(fd)[[2]][2]<- "Frequency of conserved words"
        dimnames(fd)[[1]][nr+1]<- "Total"
        df<-as.data.frame(df)
        vf2<-as.data.frame(vf2)
        k<-1
        for(i in 1:nc){
            if (guia[i,3]==1){
                dimnames(df)[[2]][k]<-dico[guia[i,1]]
                dimnames(vf2)[[2]][k]<-dico[guia[i,1]]
                k<-k+1
            }
        }
        res<-list(dataframe=df,tfrec=vf2,dfrec=fd)
        return(res)
}
```

ANEXO H: EL CÓDIGO DE LA FUNCIÓN "FILTER"

```
filter<-function(df,sw)
{
	nr<-nrow(df)
	nc<-ncol(df)
	df2<-data.frame()
	sw<-sort(sw)
	k<-1
	i<-1
	for (j in 1:nc){
		while (dimnames(df)[[2]][j]>sw[k]) k<-k+1
		if (dimnames(df)[[2]][j]<sw[k]){
			df2[1:nr,i]<-df[1:nr,j]
			dimnames(df2)[[2]][i]<-dimnames(df)[[2]][j]
			i<-i+1
		}
		else k<-k+1
	}
	df2
}
```

ANEXO I: EL CÓDIGO DE LA FUNCIÓN "CALCULARCONTRIBUCIONES"

```
Calcularcontribuciones<-function(coordenadas,frecuencias)
{
	frectotal<-sum(frecuencias)
	contribuciones<-data.frame()
	for (i in 1:nrow(coordenadas))
	{
		contribuciones[i,1]<-coordenadas[i,1]^2*(sum(frecuencias[,i])/frectotal)
		contribuciones[i,2]<-coordenadas[i,2]^2*(sum(frecuencias[,i])/frectotal)
		dimnames(contribuciones)[[1]][i]<-dimnames(coordenadas)[[1]][i]
	}
	media1<-mean(contribuciones[,1])
	media2<-mean(contribuciones[,2])
	palabras1<-numeric()
	palabras2<-numeric()
	k1<-1
	k2<-1
	for (i in 1:nrow(contribuciones))
	{
		if (contribuciones[i,1]>=3*media1)
		{
			palabras1[k1]<-dimnames(contribuciones)[[1]][i]
			k1<-k1+1
		}
		if (contribuciones[i,2]>=2*media2)
		{
			palabras2[k2]<-dimnames(contribuciones)[[1]][i]
			k2<-k2+1
		}
	}
	palabras1
	palabras2
	contribuciones
}
```

ANEXO J: EL CÓDIGO DE LA FUNCIÓN "FEATUREWORDS" – PALABRAS CARACTERÍSTICAS

```
FeatureWords<-function(dframe,vdf,cv,tc,prob=TRUE,value=0.05)
{
	cat<-levels(factor(dframe[,cv]))
	ncat<-length(cat)
	tci<-tc[1]
	tcf<-tc[2]
	aux<-matrix(nrow=2,ncol=ncat)
	aux[]<-0
	nr<-nrow(dframe)
	for (i in 1:nr) aux[2,which(cat[]==dframe[i,cv])]<-vdf[i]+aux[2,which(cat[]==dframe[i,cv])]
	k<-sum(vdf)
	fw<-data.frame()
	if (prob==FALSE)
	if (value>=0) pr=pnorm(value, mean=0, sd=1, lower.tail = FALSE, log.p = FALSE)
	else stop("Valor test must be positive",call.=FALSE)
	else pr=value
	pc<-1
	for (z in tci:tcf){
		for(l in 1:nr) aux[1,which(cat[]==dframe[l,cv])]<-
dframe[l,z]+aux[1,which(cat[]==dframe[l,cv])]
		for(i in 1:ncat){
			ki<-sum(dframe[,z])
			kij<-aux[1,i]
			kj<-aux[2,i]
			if (kj!=0){
				if((kij/kj)>(ki/k)){
					p<-phyper(kij,ki,k-ki,kj,lower.tail=FALSE)+dhyper(kij,ki,k-ki,kj)
					if(p<=pr){
						fw[pc,1]<-dimnames(dframe)[[2]][z]
						fw[pc,2]<-cat[i]
						fw[pc,3]<-kij
						fw[pc,4]<-ki
						fw[pc,5]<-(kij/kj)*100
						fw[pc,6]<-(ki/k)*100
						fw[pc,7]<-qnorm(p,0,1,lower.tail=FALSE)
						fw[pc,8]<-p
						pc<-pc+1
					}
```

```
                        }
                        else{
                                p<-phyper(kij,ki,k-ki,kj,lower.tail=TRUE)
                                if(p<=pr){
                                        fw[pc,1]<-dimnames(dframe)[[2]][z]
                                        fw[pc,2]<-cat[i]
                                        fw[pc,3]<-kij
                                        fw[pc,4]<-ki
                                        fw[pc,5]<-(kij/kj)*100
                                        fw[pc,6]<-(ki/k)*100
                                        fw[pc,7]<-qnorm(p,0,1)
                                        fw[pc,8]<-p
                                        pc<-pc+1
                                }
                        }
                    }
                }
                aux[1,]<-0
        }
        if (nrow(fw)==0) stop(" There isn't any feature word ",call.=FALSE)
        else
        dimnames(fw)[[2]]<-c("PALABRA","CATEGORIA","FRECUENCIA INTERNA", "FRECUENCIA
GLOBAL", "PORCENTAJE INTERNA", "PORCENTAJE GLOBAL", "VALOR TEST",
"PROBABILIDAD")
        fw<-fw[ order(fw[,2]),]
        aux<-as.matrix(table(fw[,2]))
        f<-0
        for (j in 1:nrow(aux)){
                i<-f+1
                f<-i+aux[j,1]-1
                aux2<-data.frame()
                aux2<-fw[i:f,]
                fw[i:f,]<-aux2[order(-aux2[,7]),]
                cat("\n")
                print(fw[i:f,])
        }
        res<-list(fw=fw)
        return(res)
}
```