Análisis de datos textuales con DtmVic

Campo Elías Pardo, Jorge Eduardo Ortiz, Daniel Leonardo Cruz
Universidad Nacional de Colombia Bogotá. Universidad Santo Tomás Bogotá. [1]

XXII Simposio Internacional de Estadística
Bucaramanga, julio 17 al 21 de 2012

[1]E-mail:cepadot@unal.edu.co; jorgeortiz@usantotomas.edu.co; dlcruzc@unal.edu.co

Índice general

1. Introducción **1**

1.1. Métodos de análisis 2

1.2. Un ejemplo de un corpus de datos textuales 2

2. Pretratamiento del texto y construcción de tablas **5**

2.1. Las unidades estadísticas textuales 5

2.1.1. Alfabeto: conjunto de caracteres 5

2.1.2. Palabra 5

2.1.3. Lema 6

2.1.4. Segmentos repetidos 6

2.1.5. Textos 7

2.2. Pretratamiento del corpus textual 7

2.2.1. Concordancias 7

2.2.2. Reducción del vocabulario 8

2.2.3. Lematización 8

2.3. Construcción de tablas 9

2.3.1. Tabla léxica 10

2.3.2. Tabla léxica agregada 10

3. Análisis de tablas léxicas **13**

3.1. Palabras características 13

3.2. Respuestas características 13

3.2.1. Criterio del ji-cuadrado 13

3.2.2. Criterio del promedio de los valores test 14

3.3. Análisis de correspondencias 15

3.3.1. AC de una tabla léxica 16

3.3.2. AC de una tabla léxica agregada 16
3.4. Clasificación automática 19
3.4.1. El método de Ward 19
3.4.2. El método $K - means$ 20
3.4.3. Combinación de análisis de correspondencias y clasificación 20

4. DtmVic **25**
4.1. Instalación 25
4.1.1. Windows 25
4.1.2. Linux 25
4.1.3. Posibles problemas 25
4.2. Entorno visual y herramientas 26
4.3. Archivos Principales 27
4.3.1. Archivos de entrada 27
4.3.2. Archivos de Salida 30
4.4. Datos textuales 30
4.4.1. Pre procesamiento del texto 31
4.4.2. Herramientas lexicométricas 31
4.4.3. Análisis tablas léxicas 33
4.5. Importación desde Excel® 34
4.6. Dimensión de textos y datos 35
4.7. Lematizadores. TreeTagger 35
4.7.1. Acerca de TreeTagger 35
4.7.2. Instalación 35
4.7.3. Creación de un archivo lematizado 36

Capítulo 1

Introducción

La mayoría de estudios de tipo social se encuentran en la necesidad de analizar datos textuales provenientes de documentos, entrevistas, o encuestas con preguntas abiertas. En muchos casos, los investigadores necesitan sintetizar, clasificar y relacionar esta información con características específicas de los autores, de los entrevistados o de los encuestados, o incluso con condiciones de diversa índole (social, económica, ambiental) que contextualizan su producción.

El texto completo que se somete a un análisis se denomina *corpus*. Algunos ejemplos de tipos de corpus son: (1) el conjunto de respuestas a una o varias preguntas abiertas en una encuesta, (2) el conjunto de palabras claves de una serie de documentos científicos de interés en un estudio, (3) los discursos de un presidente durante su período de gobierno, (4) las editoriales de uno o varios periódicos o revistas, (5) la obra literaria de un autor o de una época, etc. Haremos énfasis en el análisis de datos de encuestas con respuestas a preguntas abiertas.

Existen por lo menos tres razones para utilizar preguntas abiertas: disminuir el tiempo de entrevista, recolectar información que debe ser espontánea y explicar y comprender la respuesta a una pregunta cerrada. Lo tradicional es *poscodificar* las respuestas con el riesgo frecuente de analizar las interpretaciones de quien aplica el procedimiento y no el mensaje de las personas encuestadas. Además, las respuestas raras se eliminan a priori. Sin embargo, con el debido cuidado, puede ser un procedimiento útil. Otra opción consiste en grabar estas respuestas en su forma original sobre un soporte informático y hacer su lectura en asociación con características específicas de los respondientes, por ejemplo, reagrupando las respuestas por categorías socioprofesionales y, luego, leer las respuestas de los agricultores, de los obreros, de los ejecutivos, etc.

En este documento se presenta el análisis de datos textuales como una aplicación específica de algunos métodos de la estadística descriptiva multivariada, en particular, del análisis de correspondencias y de la clasificación automática (Lebart, Piron & Morineau 2006). Lebart & Salem (1994) proponen esta metodología, complementada con el pre-tratamiento de los textos y algunas herramientas clásicas de los análisis lexicométricos. Se encuentran textos similares en inglés (Lebart, Salem & Berry 1997) y en español (Lebart, Salem & Bécue 2000). Haremos uso del software DtmVic (Lebart 2012) que incluye algunos de los aportes desarrollados en la tesis doctoral de Bécue (1991). Se encuentra disponible, en forma gratuita para fines académicos.

1.1. Métodos de análisis

Para el tratamiento, el corpus se segmenta en palabras o expresiones que se consideran indivisibles. Los segmentos se conocen como *formas gráficas* (palabra escrita). Por lo general, se eliminan las formas gráficas irrelevantes o que aparezcan muy pocas veces. Se suele utilizar la lematización, como un proceso de homogeneización, que consiste en llevar las formas verbales al infinitivo, los sustantivos al singular y los adjetivos a singular masculino. También se sustituyen algunas palabras o expresiones por otras equivalentes.

Se define una *variable léxica* cuyas modalidades son las formas gráficas del corpus tratado. Con esta variable se construyen tablas de contingencia particulares:

1. La *tabla léxica* que contiene la frecuencia relativa con la que cada forma gráfica ha sido empleada por cada individuo; la tabla léxica es una tabla de contingencia que contiene los perfiles léxicos de los individuos.
2. Cuando existen particiones del corpus, se calcula, para cada una de ellas, la frecuencia de cada forma gráfica. Estas tablas se llaman *tablas léxicas agregadas*.
3. Se pueden obtener tablas similares sustituyendo las palabras por segmentos de frase repetidos.

En el análisis textual, los individuos se representan en el espacio referenciados por las formas léxicas. Los métodos de análisis de datos aplicados a las tablas léxicas permiten una aproximación diferenciadora de las respuestas individuales o de las partes del corpus. Se procede por comparación de perfiles léxicos. El análisis de correspondencias da una visualización de las proximidades entre individuos y entre formas y permite observar que formas y/o expresiones diferencian a los individuos. Alternativamente, si se utiliza conjuntamente información textual y no textual se puede observar las características objetivas de los individuos asociadas a un tipo de vocabulario. Por ejemplo se podría ver si un mismo contenido semántico se expresa con formas distintas, según el grupo socioeconómico, el sexo, la edad, etc.

La clasificación automática de los individuos en función de su vocabulario completa y enriquece los resultados anteriores. Se puede caracterizar cada clase en función de la información objetiva que se tiene sobre los individuos que la componen.

1.2. Un ejemplo de un corpus de datos textuales

El corpus que utilizaremos como ejemplo es el conjunto de respuestas a la pregunta abierta: "En su opinión, ¿por qué le ha ido bien con el café?", realizada en una encuesta a fincas cafeteras colombianas. El corpus corresponde a los 93 encuestados que respondieron a la pregunta abierta. En la tabla 1.1 se presentan las 10 primeras respuestas, separadas por "—-" y un identificador.

En el análisis de este tipo de corpus se busca dar respuesta principalmente a dos preguntas: (1) 'Qué dicen los encuestados? y (2) ¿Quién dice qué?. La segunda pregunta hace referencia a la comparación de respuestas según algunas características de los que responden, que se obtienen de las preguntas cerradas de la encuesta.

Tabla 1.1: Primeras 10 respuestas del corpus café

```
 ---- N1
  por llevar una excelente administración de los cultivos
---- N2
  porque es agrónomo y realiza una administración directa de la finca
---- N3
  porque lleva una administración directa y realiza las labores oportunamente
---- N4
  por vivir en la finca y llevar una administración directa.
---- N5
  porque se ha dedicado siempre al cultivo del café y esto le ha dado para vivir
---- N6
  por realizar administración directa de los cultivos.
---- N7
  porque tiene buena capacidad de endeudamiento.
---- N8
  porque vive del cultivo de café y siempre se ha dedicado a esta actividad
---- N9
  es una actividad que le gusta mucho y lleva una administración directa
---- N10
  porque ha vivido de este cultivo toda la vida.
```

Para el ejemplo se utilizan, como preguntas cerradas:

- ¿Tiene cultivos de diversificación?: no/si.
- ¿Vende su mano de obra?: no/si.
- ¿Posee créditos?: no/si.
- Tipo de caficultor:
  1. Empresario tecnificado moderno.
  2. Tecnificado moderno.
  3. Campesino tecnificado moderno.
  4. Campesino tradicional.

Capítulo 2

Pretratamiento del texto y construcción de tablas

El tratamiento estadístico del texto requiere de una codificación que facilite los conteos y la construcción de tablas para futuros análisis. La opción tomada en los programas SPAD-T y DtmVic es la de codificar el texto mediante números asignados a cada palabra según su orden alfabético. La tabla 2.1 a la izquierda muestra las primeras palabras del corpus café en orden alfabético, cuyo número de orden es el código que utilizan estos programas para los procesamientos de análisis.

2.1. Las unidades estadísticas textuales

2.1.1. Alfabeto: conjunto de caracteres

Por defecto, el alfabeto del lenguaje en el cual está escrito el corpus se define como el conjunto de caracteres del teclado del computador. Esta definición se hace por motivos prácticos y no teóricos, debido a que el corpus debe ser grabado en un medio magnético para su procesamiento en el computador. Los delimitadores se definen explícitamente: espacio, punto, coma, dos puntos, punto y coma, etc.

2.1.2. Palabra

La forma gráfica se define como una sucesión de caracteres definidos entre dos delimitadores. La forma gráfica es la representación escrita de una palabra. En estas notas utilizaremos palabra para referirnos a forma gráfica. La palabra es la unidad estadística básica que se utiliza en la Estadística Textual propuesta por Lebart & Salem (1994). Cada presencia de una palabra en un corpus se denomina *ocurrencia*. El número de ocurrencias en un corpus se denomina *tamaño del corpus* y el número de palabras distintas es el *vocabulario*.

En el corpus café aparece, por ejemplo, la palabra *administración* con 26 ocurrencias, es decir que se utilizó 26 veces en todo el corpus. El tamaño del corpus es de 1017 ocurrencias

y su vocabulario tiene 296 palabras distintas. Entonces la riqueza del vocabulario es de $296/1017 = 29.1\,\%$.

Tabla 2.1: Primera parte del vocabulario del corpus en orden alfabético y de frecuencias

words (alphabetical order)			words (frequency order)		
num.	used words	freq.	num.	used words	freq.
1	a	11	20	por	72
2	administración	26	8	el	51
3	bien	12	21	que	45
4	buena	21	13	la	44
5	café	15	24	y	44
6	cultivo	19	7	de	34
7	de	34	2	administración	26
8	el	51	12	ha	24
9	en	14	4	buena	21
10	es	12	19	para	20
11	finca	15	6	cultivo	19
12	ha	24	14	le	17
13	la	44	16	manejo	16
14	le	17	18	no	15
15	los	14	11	finca	15
16	manejo	16	5	café	15
17	me	11	15	los	14
18	no	15	9	en	14
19	para	20	22	se	13
20	por	72	23	una	13
21	que	45	10	es	12
22	se	13	3	bien	12
23	una	13	17	me	11
24	y	44	1	a	11

2.1.3. Lema

El lema es otra de las unidades estadísticas utilizadas, ya que permite reducir el vocabulario y seleccionar palabras por su tipo gramatical. El lema es la entrada al diccionario, es decir que es la palabra considerada como raíz, que por convención es:

- Singular para sustantivos.
- Singular masculino para adjetivos.
- Infinitivo para verbos.

2.1.4. Segmentos repetidos

Los separadores de la palabras se suelen dividir en fuertes y débiles. Los delimitadores fuertes separan frases. Se pueden construir todos los segmentos para las frases del corpus y hacer un conteo de ellos. El número de palabras del segmento es su tamaño. Los segmentos

se pueden tomar como unidades léxicas y realizar sobre ellos tratamiento similares a las de las palabras.

2.1.5. Textos

En el análisis de preguntas abiertas en encuestas se suelen agrupar las respuestas individuales en textos utilizando las preguntas cerradas en las encuestas. Se puede tener, por ejemplo, el corpus dividido en los textos de los hombres y de las mujeres o en cinco textos asociados a los niveles educativos: primaria, básica, secundaria, tecnológica y universitaria.

En el análisis de una novela se puede dividir en capítulos, en la producción literaria de un autor las obras pueden ser los textos, etc.

2.2. Pretratamiento del corpus textual

Desde el punto de vista estadístico, un corpus textual se constituye en una información dispersa y requiere un procesamiento para llevarlo a tablas de datos que se puedan analizar mediante métodos estadísticos apropiados. En las frases del corpus las palabras no tienen un orden aleatorio sino que obedecen a normas gramaticales y sintácticas propias del lenguaje. El procesamiento automático del lenguaje natural es materia de constante investigación y es el objetivo de la *lingüística computacional* (Gelbukh & Sidorov 2006).

En el análisis de datos textuales de este cursillo se pretende responder a objetivos de análisis mediante la estadística descriptiva mono y multidimensional. Se requiere, entonces reducir el número de palabras buscando perder poca información. Algunas herramientas de la lingüística computacional son útiles para este fin. Un analizador morfosintáctico se puede utilizar para lematizar el corpus y realizar una análisis de lemas en lugar de palabras.

Las palabras tienen problemas difíciles de solucionar con procedimientos automáticos, lo que hace inevitable la intervención del cerebro humano. La presencia de palabras homógrafas (la misma palabra con significados distintos), hace necesario recurrir al contexto para diferenciarlas. Esta tarea recibe el nombre de*desambiagüación*. Por ejemplo la palabra banco puedes ser un banco para sentarse o una entidad bancaria; la palabra estado puede referirse a un País o a un estado de la materia. La búsqueda de concordancias son una herramienta de la lingüística que sirve para observar el significado de las palabras en el contexto.

2.2.1. Concordancias

Usualmente es interesante listar todos los contextos de una misma palabra, limitándolos a una cierta dimensión en función de las necesidades particulares. El conjunto de los contextos de una cierta palabra, llamada palabra-polo se denomina concordancia de la palabra. En la tabla 2.2 se muestran las cuatro concordancias de la palabra-polo producción: las dos primeras se refieren a la cantidad y calidad de la producción y las dos últimas a los costos de producción.

Tabla 2.2: Concordancias de la palabra producción

```
Concordance of words equivalent with:      producción
 ---------------------------------------
           frequency of repetition    4

  finca y los buenos métodos que utiliza para una buena producción
                              por la buena calidad de la producción
                                        por los costos de producción
    por que los ingresos no compensan con los costos de producción
```

2.2.2. Reducción del vocabulario

Para que el análisis estadístico tenga sentido, será necesario que las palabras aparezcan con una frecuencia mínima, por ello normalmente se eliminan las palabras poco frecuentes del corpus, escogiendo un umbral de frecuencias por encima del cual conservamos las palabras. Sin embargo, se debe buscar aumentar la frecuencia de las palabras con herramientas propias del lenguaje, algunas de las estrategias son:

- Dejar en el corpus una sola palabra para todos sus sinónimos, cuando los hay.
- Escribir las palabras compuestas (varias palabras asociadas a un significado) como una sola palabra.

Por otras razones también se suelen eliminar palabras por su función gramatical, por ejemplo las denominadas palabras herramientas que son generalmente las de mayor frecuencia.

La eliminación de palabras se hace únicamente para los análisis estadísticos, es decir que se conserva el corpus original, el cual se puede combinar con algunos resultados de los análisis estadísticos.

En la tabla 2.3 se muestra el vocabulario retenido para el análisis con un umbral de frecuencia de 3, una vez eliminadas las *palabras herramientas*. Es decir que se elimina las palabras con frecuencia 3 o inferior. Quedan 218 ocurrencias de 21 palabras distintas.

2.2.3. Lematización

En un corpus lematizado las palabras del corpus en estudio se cambian por sus lemas. Esta tarea se realiza automáticamente mediante programas de análisis morfológico. Para cada palabra se presenta su categoría gramatical y su lema. En la tabla 2.4 se presentan los resultados del proceso para la respuesta 3 del corpus café. La primera linea es la respuesta, luego aparece el etiquetado gramatical de las palabras y al final la respuesta lematizada.

Un analizador morfosintáctico no puede realizar su tarea al 100 % y puede ser necesario realizar un afinamiento manual. Sobre el texto lematizado se pueden hacer análisis parciales del corpus, por ejemplo: de los sustantivos, adjetivos, verbos, etc.

Tabla 2.3: Palabras retenidas para el análisis del ejemplo café

```
selection of words
 ------------------
                frequency threshold  =       3
                         kept words  =     218
                  distinct kept word =      21

 !--------------------------------------!
 !      words (alphabetical order)      !
 !-------!--------------------!--------!
 !  num. !    used words       ! freq.  !
 !-------!--------------------!--------!
 !     1 ! actividad           !     4  !
 !     2 ! administración      !    26  !
 !     3 ! apta                !     8  !
 !     4 ! año                 !     4  !
 !     5 ! bien                !    12  !
 !     6 ! buen                !     8  !
 !     7 ! buena               !    21  !
 !     8 ! café                !    15  !
 !     9 ! cultivo             !    19  !
 !    10 ! dado                !     7  !
 !    11 ! directa             !     5  !
 !    12 ! finca               !    15  !
 !    13 ! ido                 !     9  !
 !    14 ! manejo              !    16  !
 !    15 ! no                  !    15  !
 !    16 ! producción          !     4  !
 !    17 ! rentable            !     4  !
 !    18 ! ser                 !     8  !
 !    19 ! siempre             !     5  !
 !    20 ! tecnificación       !     4  !
 !    21 ! zona                !     9  !
 !-------!--------------------!--------!
```

Tabla 2.4: Marcaje morfosintáctico de la respuesta 3

```
por que lleva una administración directa y realiza las labores oportunamente

por PREP por
que CQUE que
lleva VLfin llevar
una ART un
administración VLfin administración
directa ADJ directo
y CC y
realiza VLfin realizar
las ART el
labores NC labor
oportunamente ADV oportuno

por que llevar un administración directo y realizar el labor oportuno
```

2.3. Construcción de tablas

El corpus es una sucesión de ocurrencias de palabras y de delimitadores. Esta sucesión puede ser particionada de diferentes maneras . Básicamente se habla de dos particiones

jerarquizadas a saber: el corpus está compuesto de "respuestas individuales" que se pueden agrupar en "textos".

La partición del corpus en "respuestas individuales" se define en la entrada de los datos. Esta partición puede corresponder a una realidad "a priori" , como es el caso de las preguntas abiertas de encuesta, o ser decidida en forma arbitraria, como por ejemplo frases o párrafos de un texto literario.

2.3.1. Tabla léxica

Después de que el corpus ha sido codificado, es posible construir una tabla léxica **Z** en donde cada fila corresponde a una respuesta y cada columna a una palabra. La celda (i, j) de esta tabla, contiene la frecuencia con la cual la palabra j ha sido utilizada en la respuesta i. **Z** es la tabla de contingencia $Respuestas \times Palabras$. Si las respuestas son cortas y numerosas, esta tabla es dispersa. El objetivo al construir esta tabla es comparar los perfiles léxicos de cada una de las respuestas.

En el ejemplo café con el vocabulario retenido para el análisis la tabla léxica tiene 93 respuestas (filas) y 21 palabras (columnas), lo que corresponde a $93 \times 21 = 1953$ celdas para llenar con 218 ocurrencias, es decir que hay por lo menos $1953 - 218 = 1735$ celdas con frecuencia 0. Las tablas con estas características se denominan dispersas y requieren procedimientos específicos para el análisis de correspondencias, los cuales están programados en Spad-T y DtmVic. En esta tabla cada fila tiene la frecuencia de las palabras retenidas que están en la respuesta respectiva. La columna tiene la frecuencia con que cada palabra se utiliza en cada una de las respuestas.

2.3.2. Tabla léxica agregada

La tabla léxica agregada se construye cuando el corpus es particionado en textos que se desean comparar, de acuerdo a lo expresado. El propósito al construir la tabla es comparar los perfiles léxicos de los textos en los cuales se particiona el corpus. En el caso de las respuestas a preguntas abiertas en encuestas, se compararan los perfiles léxicos de cada grupo, según las categorías, de la variable categórica utilizada para particionar el corpus. La tabla léxica agregada es una tabla de contingencia que contiene las frecuencias de las palabras en cada uno de los textos; es la tabla de contingencia $Palabras \times Textos$ **T** La celda (i, j) de **T** es la frecuencia con la que la palabra i se encuentra el texto j.

En la tabla 2.5 se muestra la tabla léxica agregada de las 21 palabras retenidas por 5 textos (columnas) derivados de la pregunta cerrada tipo de productor. Arriba se muestra la información asociada: la distribución de las respuestas entre las 5 categorías, donde la primera corresponde a 19 caficultores que no respondieron a esa pregunta; y la repartición de las ocurrencias y palabras entre los 5 textos. Los empresarios tecnificados modernos (cat. 2) respondieron con 167 palabras (31.4 % de las 532 palabras), con 6.4 palabras por respuesta, en promedio; de las palabras retenidas usaron 19 palabras distintas (11.4 % de las 167 palabras) con 66 ocurrencias. Obsérvese, por ejemplo, que la palabra tecnificación fue usada 2 veces por los empresarios tecnificados modernos y 2 veces por los caficultores tecnificados modernos. Solo hay dos caficultores categorizados como campesinos tradicionales, con 9 ocurrencias de las palabras retenidas.

Tabla 2.5: Distribución de palabras y tabla léxica agregada según los tipos de caficultor

```
-------------------------------------------------------------------------------
 grouping  responses into   5 texts
 using categorical variable      4   =   tipo
 ------------------------------------------------------------------------------

           ------------------------------------------------------------
             number       identifier            number         number of
            of text                          of individ.       responses
           ------------------------------------------------------------
                1    cat0b_Tipo                    19              19
                2    cat1EmpTecModer               26              26
                3    cat2TecModerno                29              29
                4    cat3CampTecModer              17              17
                5    cat4CampTradicional            2               2
           ------------------------------------------------------------
           t o t a l                                93              93


 repartition  of  terms  in  texts/     -------------------------------------

 ---------------------------------------------------------------------------------------------------
  number                          *  number  /1000     mean    * number of    /1000   * number  *
    of        identifier          *    of      of       per    *   words    words of  * of words*
  text                            *   words  total    response * (distinct)  text     *  kept   *
 ---------------------------------------------------------------------------------------------------
     1 = cat0b_Tipo               *    115   216.2       6.1   *     19      165.2    *      51 *
     2 = cat1EmpTecModer          *    167   313.9       6.4   *     19      113.8    *      66 *
     3 = cat2TecModerno           *    155   291.4       5.3   *     16      103.2    *      58 *
     4 = cat3CampTecModer         *     76   142.9       4.5   *     15      197.4    *      34 *
     5 = cat4CampTradicional      *     19    35.7       9.5   *      8      421.1    *       9 *
 ---------------------------------------------------------------------------------------------------
                 g l o b a l      *    532 1000.0       5.7   *                       *     218 *
 ---------------------------------------------------------------------------------------------------

table  Words - Texts
 ------------------------------------------------------
                            cat0 cat1 cat2 cat3 cat4
 ------------------------------------------------------
  actividad                i   1.   1.   0.   2.   0.
  administración           i   6.   7.  12.   1.   0.
  apta                     i   4.   1.   2.   1.   0.
  año                      i   1.   2.   0.   1.   0.
  bien                     i   2.   5.   3.   0.   2.
  buen                     i   2.   4.   0.   1.   1.
  buena                    i   2.   6.  13.   0.   0.
  café                     i   2.   3.   5.   5.   0.
  cultivo                  i   5.   2.   3.   9.   0.
  dado                     i   1.   2.   1.   2.   1.
  directa                  i   1.   3.   0.   1.   0.
  finca                    i   3.   6.   4.   1.   1.
  ido                      i   2.   4.   2.   0.   1.
  manejo                   i   5.   6.   2.   2.   1.
  no                       i   2.   9.   3.   0.   1.
  producción               i   3.   1.   0.   0.   0.
  rentable                 i   1.   0.   2.   1.   0.
  ser                      i   4.   1.   1.   2.   0.
  siempre                  i   0.   1.   1.   2.   1.
  tecnificación            i   0.   2.   2.   0.   0.
  zona                     i   4.   0.   2.   3.   0.
 ------------------------------------------------------
                            cat0 cat1 cat2 cat3 cat4
```

Capítulo 3

Análisis de tablas léxicas

3.1. Palabras características

La detección de las palabras con frecuencias particularmente altas o particularmente bajas dentro de los textos, en los que se ha dividido un corpus, son usualmente de importancia para el investigador, pues representan características distintivas de los textos entre sí. Esta información completada con cálculos probabilísticos permite tener una idea sobre las diferentes frecuencias de una misma forma en los distintos textos.

El modelo estadístico utilizado usualmente para detectar las palabras características en los textos es el siguiente: se considera cada texto como una muestra del corpus y se sitúa en el conjunto de todas la muestras posibles de la misma longitud del texto que pueden ser obtenidas.

El valor test es un índice que sirve para ordenar las palabras características, se interpreta como un cuantil de la distribución normal estándar. Valores superiores a 2 indican que la frecuencia relativa de la palabra en el texto es superior a la frecuencia relativa en todo el corpus. Valores test inferiores a -2 son indicadores de frecuencias relativas inferiores dentro del texto comparadas con las del corpus. Por ejemplo los caficultores tecnificados modernos usan con más frecuencia las palabras buen (18.8 % vs 7.9 % y administración (21.9 % vs 10.0 %).

3.2. Respuestas características

Las respuestas características no son respuestas artificiales construidas a partir de las palabras características, sino respuestas reales, escogidas según un criterio como representantes del texto.

3.2.1. Criterio del ji-cuadrado

Cada respuesta puede considerarse como un vector fila cuyas componentes son las frecuencias de cada una de las palabras en esta respuesta. Un texto es un conjunto de vectores

Tabla 3.1: Palabras características para los tipos de caficultores

```
Selection of characteristic words
 ------------------------------------------------
 ------------------------------------------------------------------------------------------------
    spelling of                    --- percentage---        frequency        test.v   proba
        word                       within      global   within    global
 ------------------------------------------------------------------------------------------------

 text number    1   cat0b_Tipo
 ----------------
  1 producción                    8.57        2.86         3.        4.       1.662    .048
 ------------------------------
 ------------------------------------------------------------------------------------------------
 text number    2   cat1EmpTecModer
 ----------------
 ------------------------------
  1 zona                           .00        5.71         0.        8.      -1.864    .031
 ------------------------------------------------------------------------------------------------
 text number    3   cat2TecModerno
 ----------------
  1 buena                        18.75        7.86         6.       11.       2.099    .018
  2 administración               21.88       10.00         7.       14.       2.096    .018
 ------------------------------
 ------------------------------------------------------------------------------------------------
 text number    4   cat3CampTecModer
 ----------------
  1 cultivo                      29.41        8.57         5.       12.       2.467    .007
  2 siempre                      11.76        2.14         2.        3.       1.764    .039
 ------------------------------
 ------------------------------------------------------------------------------------------------
 text number    5   cat4CampTradicional
 ----------------
 ------------------------------
 ------------------------------------------------------------------------------------------------
```

fila. El perfil léxico promedio del texto es la media de los perfiles de las respuestas del texto. Es legítimo calcular distancias entre respuestas y textos. La distancia seleccionada entre textos y respuestas es precisamente la utilizada en los cálculos del análisis de correspondencias, es decir la distancia ji-cuadrado. La respuesta más característica será aquella más cercana la perfil medio del texto. Lo que se hace es ordenar las respuestas en orden decreciente de distancia al perfil medio. Este criterio tiende a favorecer a las respuestas largas.

3.2.2. Criterio del promedio de los valores test

Recuérdese que al calcular las palabras características se ha asociado a cada par "palabra, texto" un valor "test", que puede ser positivo o negativo. Según la pertenencia de una palabra a un texto, se le puede atribuir la media de los valores "test" correspondientes a las palabras que componen la respuesta. La respuesta más característica será aquella cuya media sea más alta. Este criterio tiende a favorecer a las respuestas cortas.

Las respuestas características son respuestas originales pronunciadas por los individuos entrevistados. En general se extraen varias respuestas características para cada texto (10

a 20, según el caso). Una sola respuesta en general no resume en general todo el texto. Tampoco un único individuo es un buen representante de todo un grupo de individuos.

La tabla 3.2 muestra las respuestas más características para las cinco categorías. Según estas respuestas, a los empresarios tecnificados modernos no les ha ido bien; a los caficultores tecnificados modernos les ha ido bien por la buena administración, manejo y por estar en una zona apta para el cultivo. Con respecto a los dos campesinos tradicionales, a uno le ha ido bien y al otro no.

Tabla 3.2: Respuestas características según los tipos de caficultores

```
Selection of characteristic individuals or responses (criterion: frequency of words)
 ---------------------------------------------------------------------------------------------
   criterion             characteristic response/individual
    of selection
 ---------------------------------------------------------------------------------------------
 text number    1   cat0b_Tipo
 ---------------
    .35 -  1   por ser zona apta para el cultivo
    .31 -  2   por ser una zona apta para el cultivo.
    .28 -  3   por manejo, administración y zona apta para el cultivo.
    .21 -  4   por manejo, variedad, y por ser una zona apta para el cultivo.
 ---------------------------------------------------------------------------------------------
 text number    2   cat1EmpTecModer
 ---------------
    .28 -  1   no le ha ido bien.
    .14 -  2   no cree que le ha ido bien con el café.
    .14 -  3   no le ha ido bien, por eso tienda a diversificar.
    .12 -  4   no estoy bien, me he sostenido por ser mesurado con los créditos
 ---------------------------------------------------------------------------------------------
 text number    3   cat2TecModerno
 ---------------
   1.05 -  1   buena administración y manejo.
    .60 -  2   buena administración, zona apta para el cultivo.
    .52 -  3   por la buena administración la ha ido bien .
    .47 -  4   por la buena tecnificación y empeño en la administración.
 ---------------------------------------------------------------------------------------------
 text number    4   cat3CampTecModer
 ---------------
    .58 -  1   el cultivo del café le ha dado para sobrevivir.
    .49 -  2   zona apta para el cultivo.
    .41 -  3   por que se ha dedicado siempre al cultivo del café y esto le ha dado para vivir
    .39 -  4   por que vive del cultivo de café y siempre se ha dedicado ha esta actividad
 ---------------------------------------------------------------------------------------------
 text number    5   cat4CampTradicional
 ---------------
    .10 -  1   si me ha ido bien por el buen manejo administrativo que le he dado a la finca y
               el empeño en hacer siempre las cosas bien

    .00 -  2   no ya que estamos muy endeudados con los bancos.
 =============================================================================================
```

3.3. Análisis de correspondencias

En el análisis de correspondencias simples (ACS) se busca una representación simple pero optimizada para analizar simultáneamente los perfiles fila y columna obtenidos a partir de una tabla de contingencia. Los perfiles fila, definidos como las distribuciones condicionales

de las filas de la tabla, se consideran inmersos en un espacio multidimensional donde los ejes corresponden a las columnas y pueden verse como una nube de puntos en este espacio. Simétricamente, los perfiles columna, definidos como las distribuciones condicionales de las columnas de la tabla, forman la nube de perfiles columna en un espacio multidimensional diferente donde las filas son los ejes.

En cada espacio se hace uso de la distancia ji-cuadrado entre distribuciones. Sin embargo, las representaciones geométricas de las nubes son imposibles si se tienen más de dos dimensiones. Entonces, para cada espacio, es necesario hacer proyecciones sobre planos, buscando que se conserven al máximo los conjuntos de distancias originales. Este es el mismo problema que buscan resolver todos los métodos factoriales. La lectura en los subespacios proyectados es aproximada pero se tiene lo más relevante de la información de la tabla de contingencia. Además, las fórmulas de transición hacen posible una proyección simultánea de las nubes de perfiles fila y columna en un mismo plano, permitiendo interpretar la posición de un punto de un espacio utilizando como referencia toda la nube del otro espacio. Se pueden también construir algunos indicadores que complementan los gráficos y evitan lecturas erróneas.

3.3.1. AC de una tabla léxica

Una tabla léxica es una tabla de contingencias que cruza los textos con las palabras retenidas para el análisis. Cada fila es una respuesta individual y la frecuencia de uso de las palabras se constituye en el perfil léxico. El ACS permite comparar simultáneamente todos estos perfiles léxicos. Cada columna es el perfil de uso de una palabra en las respuestas. El conjunto de estos perfiles también se describe con el análisis, además de las posiciones de cada uno de los perfiles de una nube con relación al conjunto de perfiles de la otra, generándose la posibilidad de realizar análisis en profundidad de las relaciones entre las dos nubes de puntos.

El la tabla 3.3 se muestra el primer plano factorial del ACS de la tabla léxica del ejemplo café. Las palabras se sitúan cerca cuando se usan más o menos en las mismas respuestas. Por ejemplo, en la parte derecha inferior del plano se muestran las palabras manejo, administración, buena, directa, finca, lo que indica que hay un grupo importante de caficultores que utilizan estas palabras más o menos con la misma frecuencia.

3.3.2. AC de una tabla léxica agregada

En el caso de encuestas, la elaboración de tablas léxicas agregadas provee el insumo apropiado para analizar las preguntas abiertas en relación con las cerradas. Para este fin, se suelen combinar varias preguntas cerradas en una, por ejemplo: sexo por grupo de edad.

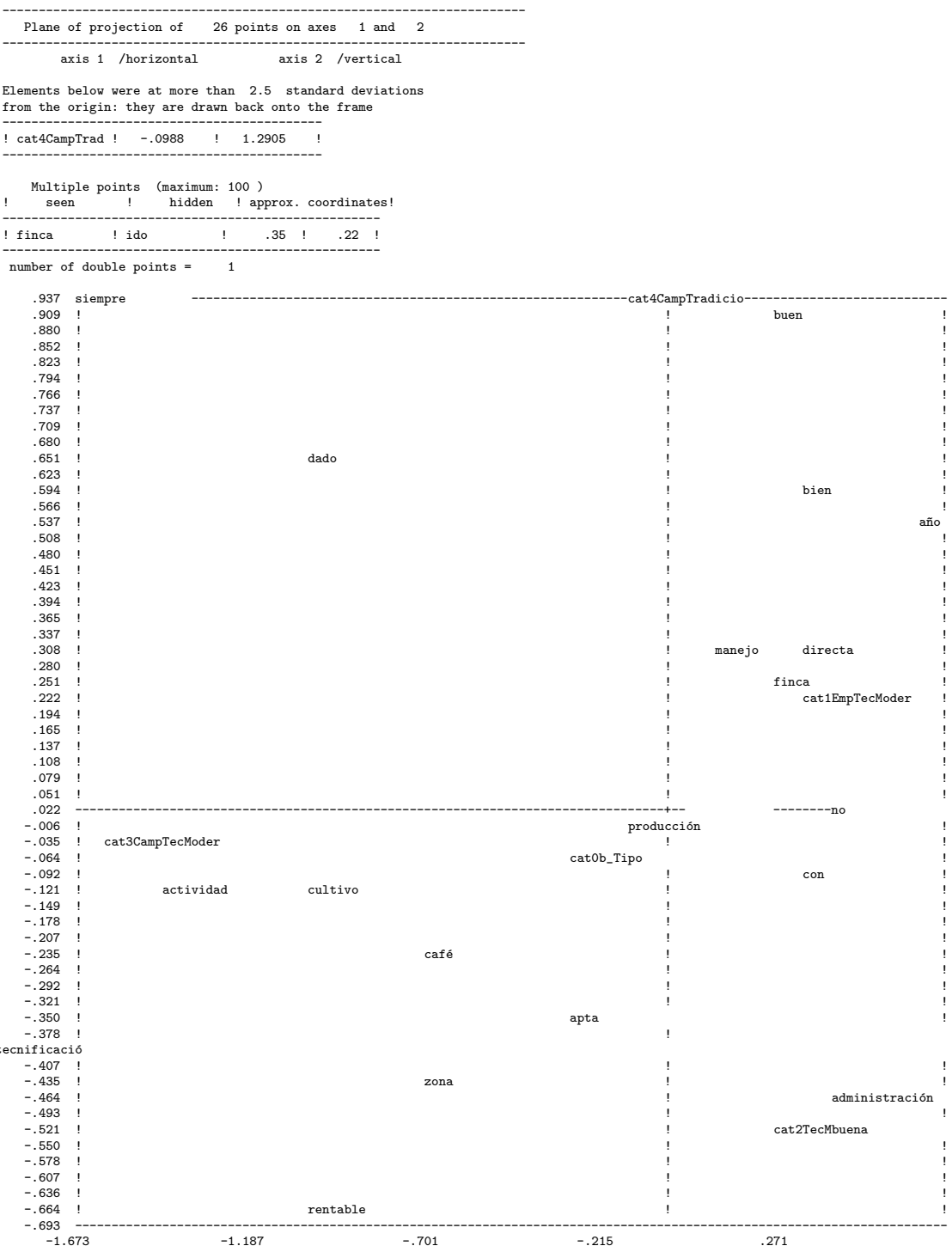
El la tabla 3.4 se muestra el primer plano factorial del ACS de la tabla léxica agregada 2.5. La palabras próximas tiene perfiles de uso similares (frecuencias relativas de uso según tipos de caficultores. La palabras cercanas a un tipo de caficultor pero más alejadas del centro tienen más frecuencia de uso dentro de esa categoría con respecto a la frecuencia global. Por ejemplo las palabras administración y buena están asociadas a los caficultores tecnificados modernos.

Tabla 3.3: Primer plano factorial del ACS de la tabla léxica del corpus café

```
-----------------------------------------------------------------------
   Plane of projection of    21 points on axes   1 and   2
 ----------------------------------------------------------------------
        axis 1  /horizontal          axis 2  /vertical

 Elements below were at more than  2.5  standard deviations
 from the origin: they are drawn back onto the frame
 -------------------------------------------
 ! año          !  -3.3981    !   -.3414    !
 -------------------------------------------
   1.449  ------------------------------------------------------zona           ------------------------------------------------
   1.412  !                                                     apta                                                          !
   1.375  !                                                      !                                                            !
   1.339  !                                                      !  rentable                                                  !
   1.302  !                                                      !                                                            !
   1.265  !                                        cultivo                                                                     !
   1.228  !                                                      !                                                            !
   1.192  !                                                      !                                                            !
   1.155  !                                            ser                                                                     !
   1.118  !                                                      !                                                            !
   1.081  !                                            café                                                                    !
   1.045  !                                                      !                                                            !
   1.008  !                                                      !             actividad                                      !
    .971  !                                                      !                                                            !
    .934  !                                                      !                                                            !
    .897  !                                                      !                                                            !
    .861  !                                        siempre                                                                     !
    .824  !                                                      !                                                            !
    .787  !                                                      !                                                            !
    .750  !                                                      !                                                            !
    .714  !                                                      !                                                            !
    .677  !                                                      !                                                            !
    .640  !                                                      !                                                            !
    .603  !                                                      !                                                            !
    .567  !                                                      !                                                            !
    .530  !                                                      !                                                            !
    .493  !                                                      !                                                            !
    .456  !                                                      !                                                            !
    .420  !                                                      !                                                            !
    .383  !                                                      !                                                            !
    .346  !                                                      !                                                            !
    .309  !                                                      !                                                            !
    .272  !                                                      !                                                            !
    .236  !                                                      !                                                            !
    .199  !                                                      !                                                            !
    .162  !                                                      !                                                            !
    .125  !                                                      !                                                            !
    .089  !                                            dado                                                                    !
    .052  !                                                      !                                                            !
    .015  ------------------------------------------------------+--             ------------------------------------------------
   -.022  !                                                      !                                                            !
   -.058  !                                                      !                                                            !
   -.095  !                                                      !                                                            !
   -.132  año                                                    !                                                            !
   -.169  !                                                      !                                                            !
   -.206  !                                                      !                                                            !
   -.242  !                                                      !                        manejo                              !
   -.279  !                                                      !                                                            !
   -.316  !                                                      !                                                            !
   -.353  !                                                      !                                                            !
   -.389  !                                                      !                                  administración            !
   -.426  !                                                      !                                                            !
   -.463  !                                                      !                        buenbuena               directa
   -.500  !                                                      !                                                            !
   -.536  !                                                      !                                                            !
   -.573  !                                                      !                                                            !
   -.610  !                                                      !                                                            !
   -.647  !                                                      !                                                            !
   -.683  !                                                      !                           finca                            !
   -.720  !                                                      !                                                            !
   -.757  !                                                      !                                                            !
   -.794  !                                                      !                                                            !
   -.831  !          bien                                        !                                                            !
   -.867  !                                                      !                                                            !
   -.904  !                                                      !                                                            !
   -.941  !                                                      !
tecnific
   -.978  !                                     producción       !                                                            !
  -1.014  !                                                      !                                                            !
  -1.051  no      ido                                           !                                                            !
  -1.088  ------------------------------------------------------------------------------------------------------------------------
      -1.448                  -.885                 -.322                  .242                  .805
```

Tabla 3.4: Primer plano factorial del ACS de la tabla léxica agregada según tipos de caficultores



3.4. Clasificación automática

Una forma de sintetizar la información contenida en una tabla multidimensional (por ejemplo una tabla léxica agregada), es mediante la conformación y caracterización de grupos. Los grupos o clases se conforman de manera que los elementos dentro de cada grupo sean lo más homogéneos posibles y que, en cambio, los elementos de diferentes grupos sean lo más diferentes posibles.

En el análisis de datos textuales se puede hacer clasificación de las filas de una tabla léxica, en cuyo caso se obtienen grupos de respuestas (individuos), que se parecen en el vocabulario que utilizan. "Ya que los individuos no se expresan de la misma forma según su pertenencia a un grupo socioeconómico, su edad, su nivel de educación, sus opiniones,..., parece tener sentido agrupar los individuos según su vocabulario para, después, caracterizar las clases así obtenidas por la información conocida sobre los individuos." (Bécue 1991, p.61)

También se pueden clasificar las columnas, con lo que se obtienen grupos de palabras, que son utilizadas más o menos por los mismos individuos. "Unas palabras tenderán a pertenecer a la misma clase, si son pronunciadas con frecuencia por los mismos individuos. La clasificación automática de las palabras describe sistemáticamente las asociaciones que existen entre ellas. Esas asociaciones dejan intuir cadenas, es decir, sucesiones de palabras no forzosamente consecutivas, empleadas en las mismas respuestas. En cierta manera sugiere la repetición de ciertas respuestas" (Bécue 1991, p.61).

La aplicación de la clasificación a la tabla léxica agregada conlleva a la clasificación de las palabras (filas) según sus perfiles de utilización en los textos y de los textos (columnas), según los perfiles de las palabras que aparecen.

Los métodos de clasificación se pueden dividir en jerárquicos y no jerárquicos. En los no jerárquicos el número de clases se establece previamente y el algoritmo de clasificación asigna los individuos a las clases, partiendo de algunos valores iniciales y buscando optimizar algún criterio establecido de antemano.

En la clasificación jerárquica se construye un "árbol" o "dendrograma", (del griego dendron = árbol), cuyas ramas terminales representan a cada uno de los individuos y el tronco es la clase conformada por todos los individuos. Un dendrograma representa una serie de particiones embebidas, en donde el número de clases decrece a medida que se aumenta la altura del árbol. Para obtener alguna clasificación particular se hace "un corte" en el árbol.

Un árbol se puede construir partiendo de las ramas terminales (cada uno de los individuos) y haciendo uniones sucesivas hasta llegar a un grupo con todos los individuos. Este método se denomina "clasificación jerárquica aglomerativa".

Los métodos de clasificación requieren de una definición de la distancia o un índice de disimilitud entre los elementos que se van a clasificar. Si las variables son de tipo continuo la distancia más utilizada es la euclidiana canónica

3.4.1. El método de Ward

Los métodos de clasificación jerárquica requieren, además de la distancia entre individuos, una distancia entre grupos de individuos, que se denomina también criterio de agregación

y es la que da el nombre al método de clasificación jerárquica. El método de Ward es el que más sentido estadístico tiene, en el caso de variables continuas, pues en cada paso del algoritmo se obtienen grupos de la manera que la inercia dentro de los grupos es mínima y por ende la inercia entre los grupos es máxima.

La distancia de Ward entre dos grupos A y B se define como el aumento de la inercia intra grupos al unir A y B en un solo grupo. El algoritmo para construir un árbol de clasificación utilizando el método de Ward se puede consultar en Montenegro & Pardo (1996) y de manera más detallada en Pardo (1992).

3.4.2. El método $K - means$

El método $K - means$ permite construir una partición directa de los elementos a clasificar, pero requiere como información de partida los puntos iniciales de las clases. A partir de los K puntos iniciales se construye una partición en K clases, se calculan los centros de gravedad de las clases, los que se convierten en los nuevos puntos para construir una nueva partición. El algoritmo termina cuando no hay cambios en la partición o la disminución de la inercia intra clases entre dos etapas sucesivas del algoritmo está por debajo de un umbral. Ver por ejemplo Cabarcas & Pardo (2001) o Lebart et al. (2006).

3.4.3. Combinación de análisis de correspondencias y clasificación

La estrategia implementada en SPAD y DtmVic es la de realizar la clasificación sobre las coordenadas factoriales de análisis correspondencias simples. En estos programas se utiliza una combinación del método de clasificación jerárquica utilizando el criterio de Ward y del método de nubes dinámicas (método no jerárquico). La estrategia contempla los siguientes pasos:

1. Seleccionar el número de ejes factoriales a usar en la clasificación. Puesto que la clasificación se realiza sobre las coordenadas factoriales de un análisis factorial previo, es posible seleccionar todos los ejes, lo cual equivale a hacer una clasificación directa, o seleccionar un menor número de ejes. Al seleccionar un menor número de ejes se esta filtrando posiblemente ruido, es decir inercia que puede deberse al azar y que no contiene información. Con esta opción se obtienen, a menudo, clasificaciones que son más claras que las obtenidas con toda la información. El número de ejes es una opción del usuario, el valor por defecto es 10.

2. Realizar una clasificación jerárquica partiendo de las coordenadas factoriales sobre los ejes retenidos para la clasificación..

3. Obtener una partición de del árbol obtenido en el paso 2. El número de clases es una decisión del usuario, para la cual es muy útil es histograma de índices de nivel.

4. Hacer una optimización de la partición obtenida en el paso 2, haciendo uso del procedimiento de nubes dinámicas, en este caso los centros de gravedad de la participación obtenida con la clasificación jerárquica son los núcleos iniciales del procedimiento.

5. Descripción de las clases obtenidas. Produce las salidas más útiles para caracterizar las clases obtenidas. En el caso de clasificar respuestas las clases se caracterizan con las elementos característicos, es decir, palabras características, segmentos característicos o respuestas características.

El procedimiento de clasificación automática genera una variable categórica a partir de la tabla léxica analizada. Para describir las clases se construye una tabla léxica agregada y se buscan sus elementos característicos. En el ejemplo café se realiza una partición de 5 clases. En la tabla 3.5 se muestra la distribución de palabras entre los cinco textos y la tabla léxica agregada que tiene 21 palabras (filas) y los 5 textos (columnas), correspondientes a las 5 clases obtenidas. La clase 1 tiene 217 ocurrencias de palabras del corpus original sin las palabras herramientas, que son el 40.8 % de las ocurrencias con un promedio de 5 palabras por respuesta; de las palabras retenidas tiene 16 palabras distintas (7.4 % del corpus) y 91 ocurrencias retenidas en la tabla léxica.

En la tabla 3.6 se muestran las palabras características de las cinco clases. La clase 1 explica su buen resultado por la buena y directa administración y el buen manejo; a la clase 2 no le ha ido bien; la clase 3 agrupa a los 4 caficultores que usan la palabra tecnificación; con las primeras palabras características de la clase 4 son: *zona apta cultivo café*; finalmente, la clase 5 esta formada por los 3 caficultores que utilizan la palabra año (hay uno que lo utiliza dos veces).

La descripción de las clases se complementa con las palabras características que se muestran en la tabla 3.7 que son las respuestas tal como están en el corpus más representativas de cada clase, utilizando el promedio de los valores test y el corpus sin las *palabras herramienta*. En el caso de las clases 3 y 5 aparecen las 4 y 3 respuestas, respectivamente.

Tabla 3.5: Distribución de palabras y tabla léxica agregada según la partición en 5 clases

```
grouping  responses into   5 texts
 using classification      1   =    cut a of the tree into  5 classes
 ------------------------------------------------------------------------------
           ---------------------------------------------------------
            number       identifier            number         number of
           of text                           of individ.       responses
           ---------------------------------------------------------
               1    aa1aclass   1 /  5            43              43
               2    aa2aclass   2 /  5            16              16
               3    aa3aclass   3 /  5             4               4
               4    aa4aclass   4 /  5            27              27
               5    aa5aclass   5 /  5             3               3
           ---------------------------------------------------------
           t o t a l                              93              93

 repartition  of  terms  in  texts/     ------------------------------------
 ------------------------------------------------------------------------------------------------
  number                          *  number  /1000     mean    * number of    /1000   * number  *
    of        identifier          *    of      of       per    *   words    words of  * of words*
  text                            *   words  total    response * (distinct)  text     *  kept   *
 ------------------------------------------------------------------------------------------------
     1 = aa1aclass   1 /  5       *    217   407.9       5.0   *     16       73.7    *      91 *
     2 = aa2aclass   2 /  5       *    106   199.2       6.6   *     10       94.3    *      40 *
     3 = aa3aclass   3 /  5       *     32    60.2       8.0   *      5      156.3    *       9 *
     4 = aa4aclass   4 /  5       *    144   270.7       5.3   *     12       83.3    *      72 *
     5 = aa5aclass   5 /  5       *     33    62.0      11.0   *      3       90.9    *       6 *
 ------------------------------------------------------------------------------------------------
                 g l o b a l     *     532 1000.0       5.7   *                      *     218 *
 ------------------------------------------------------------------------------------------------

table  Words - Texts
 ------------------------------------------------------
                            aa1a aa2a aa3a aa4a aa5a
 ------------------------------------------------------
  actividad                 i   1.   0.   0.   3.   0.
  administración            i  21.   1.   1.   3.   0.
  apta                      i   0.   0.   0.   8.   0.
  año                       i   0.   0.   0.   0.   4.
  bien                      i   2.  10.   0.   0.   0.
  buen                      i   7.   0.   1.   0.   0.
  buena                     i  15.   2.   2.   2.   0.
  café                      i   1.   1.   0.  13.   0.
  cultivo                   i   1.   1.   0.  16.   1.
  dado                      i   4.   1.   0.   2.   0.
  directa                   i   5.   0.   0.   0.   0.
  finca                     i  14.   0.   1.   0.   0.
  ido                       i   1.   8.   0.   0.   0.
  manejo                    i  14.   0.   0.   2.   0.
  no                        i   1.  13.   0.   0.   1.
  producción                i   2.   2.   0.   0.   0.
  rentable                  i   0.   0.   0.   4.   0.
  ser                       i   1.   1.   0.   6.   0.
  siempre                   i   1.   0.   0.   4.   0.
  tecnificación             i   0.   0.   4.   0.   0.
  zona                      i   0.   0.   0.   9.   0.
 ------------------------------------------------------
                            aa1a aa2a aa3a aa4a aa5a
```

Tabla 3.6: Palabras características de las cinco clases

Selection of characteristic words

spelling of word	--- percentage within	--- global	frequency within	frequency global	test.v	proba
text number 1 aa1aclass 1 / 5						
1 administración	23.08	11.93	21.	26.	4.115	.000
2 finca	15.38	6.88	14.	15.	4.053	.000
3 manejo	15.38	7.34	14.	16.	3.650	.000
4 buena	16.48	9.63	15.	21.	2.659	.004
5 buen	7.69	3.67	7.	8.	2.326	.010
6 directa	5.49	2.29	5.	5.	2.261	.012
5 apta	.00	3.67	0.	8.	-2.255	.012
4 zona	.00	4.13	0.	9.	-2.466	.007
3 café	1.10	6.88	1.	15.	-2.778	.003
2 no	1.10	6.88	1.	15.	-2.778	.003
1 cultivo	1.10	8.72	1.	19.	-3.415	.000
text number 2 aa2aclass 2 / 5						
1 no	32.50	6.88	13.	15.	5.792	.000
2 bien	25.00	5.50	10.	12.	4.804	.000
3 ido	20.00	4.13	8.	9.	4.401	.000
3 finca	.00	6.88	0.	15.	-1.721	.043
2 manejo	.00	7.34	0.	16.	-1.822	.034
1 administración	2.50	11.93	1.	26.	-1.915	.028
text number 3 aa3aclass 3 / 5						
1 tecnificación	44.44	1.83	4.	4.	4.688	.000
text number 4 aa4aclass 4 / 5						
1 cultivo	22.22	8.72	16.	19.	4.585	.000
2 café	18.06	6.88	13.	15.	4.185	.000
3 zona	12.50	4.13	9.	9.	3.991	.000
4 apta	11.11	3.67	8.	8.	3.700	.000
5 rentable	5.56	1.83	4.	4.	2.282	.011
6 ser	8.33	3.67	6.	8.	2.122	.017
7 siempre	5.56	2.29	4.	5.	1.729	.042
7 buen	.00	3.67	0.	8.	-1.775	.038
6 ido	.00	4.13	0.	9.	-1.961	.025
5 buena	2.78	9.63	2.	21.	-2.306	.011
4 administración	4.17	11.93	3.	26.	-2.389	.008
3 bien	.00	5.50	0.	12.	-2.459	.007
2 finca	.00	6.88	0.	15.	-2.894	.002
1 no	.00	6.88	0.	15.	-2.894	.002
text number 5 aa5aclass 5 / 5						
1 año	66.67	1.83	4.	4.	5.107	.000

Tabla 3.7: Respuestas características de las cinco clases, sin las palabras herramienta

```
Selection of characteristic individuals or responses (criterion: frequency of words)
 ------------------------------------------------------------------------------------
 ------------------------------------------------------------------------------------------------
   criterion             characteristic response/individual
    of selection
 ------------------------------------------------------------------------------------------------
 text number    1   aa1aclass   1 /  5
 ---------------
   3.88 -  1  manejo administración.
   3.88 -  2  administración manejo.
   3.65 -  3  manejo.
   3.65 -  4  manejo.
   3.47 -  5  buena administración manejo.

 ------------------------------------------------------------------------------------------------
 text number    2   aa2aclass   2 /  5
 ---------------
   5.00 -  1  no ido bien.
   5.00 -  2  no ido bien.
   3.00 -  3  no cree ido bien  café.
   2.50 -  4  no ido bien, eso tienda diversificar.
   2.14 -  5  no ido bien solo producido necesario subsistir.
 ------------------------------------------------------------------------------------------------
 text number    3   aa3aclass   3 /  5
 ---------------
   1.17 -  1  buena tecnificación empeño administración.
    .78 -  2  buena tecnificación maravillosas condiciones suelo clima
    .43 -  3  gran esfuerzo personal tocado desarrollar obtener logro mis objetivos
              tecnificación implantada
    .43 -  4  fácil llevada finca, dar buenos manejos esta tener buen grado tecnificación
 ------------------------------------------------------------------------------------------------
 text number    4   aa4aclass   4 /  5
 ---------------
   4.09 -  1  zona apta cultivo.
   3.96 -  2  zona apta café.
   3.60 -  3  ser zona apta cultivo
   3.60 -  4  ser zona apta cultivo.
   2.86 -  5  zona optima cultivo.
 ------------------------------------------------------------------------------------------------
 text number    5   aa5aclass   5 /  5
 ---------------
    .85 -  1  único cultivo da rentabilidad todo año.
    .85 -  2  hasta año pasado mal malos precios.
    .49 -  3  sostenido ya tarjetas usadas 94 tienen fecha vencimiento año 95,
              no fueron incluidas refinanciación debido uso hizo primer semestre año
              94
 ================================================================================================
```

Capítulo 4

DtmVic

Este programa nace del esfuerzo del doctor Ludovic Lebart, como un servicio académico para la aplicación de métodos estadísticos multidimensionales al análisis de archivos numéricos o textuales, en la investigación de los estudiantes de doctorado. El DtmVic (Lebart 2012) al igual que SPAD (Cisia-Ceresta 2000) se basa en lenguaje Fortran 77 y fue creado para sistemas operativos Windows, aunque se puede usar en *Linux* mediante *Wine*.

4.1. Instalación

4.1.1. Windows

En la pagina oficial de DtmVic (www.dtmvic.com) se presentan dos formas de instalación del software. En este curso usaremos la version portable, la cual se encuentra dentro del CD de memorias de este evento. Al pulsar sobre *DtmVic.exe* se abrirá el programa en la pantalla principal.

4.1.2. Linux

Se debe instalar Wine desde la terminal mediante la sentencia *apt-get install wine*, luego hacer click derecho sobre *DtmVic.exe* y *ejecutar con Wine*.

4.1.3. Posibles problemas

1. Aveces es necesario ejecutar la aplicación como administrador para que pueda generar los archivos de salida.

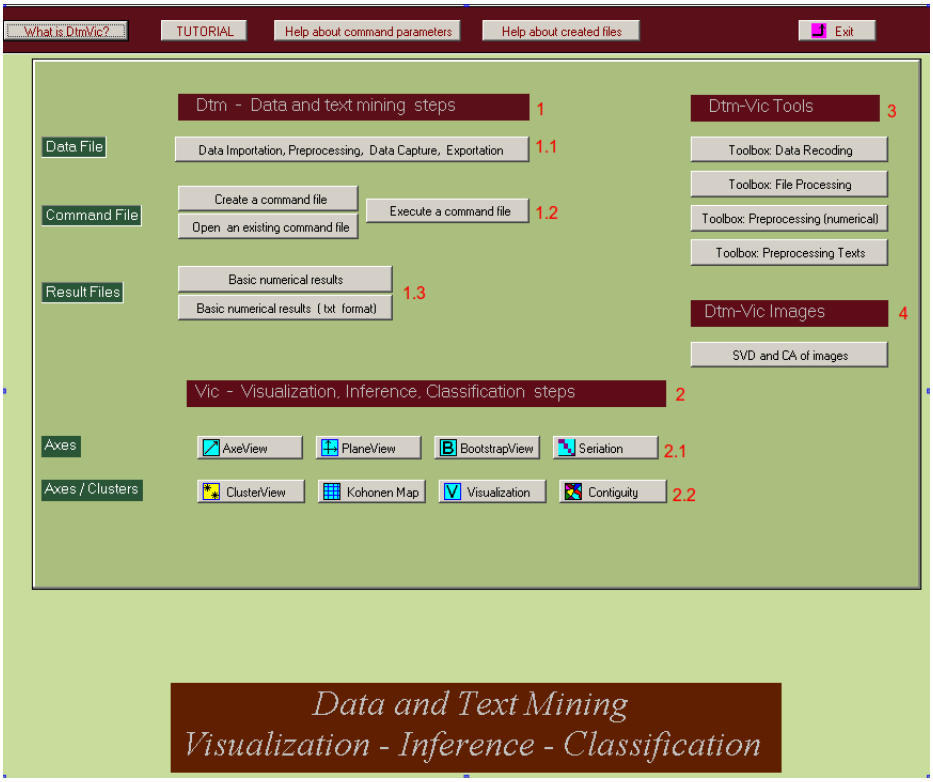
2. En XP se debe copiar el ejecutable en una carpeta del sistema con permisos de usuario.

3. Cuando se usa Wine el programa puede ser inestable.

4. Es recomendable crear una carpeta para cada proyecto, puesto que DtmVic trabajara sobre esta.

4.2. Entorno visual y herramientas

En esta sección hacemos un recorrido por las herramientas de DtmVic desde la primera pantalla del software.



1. Procedimientos de minería de datos y textos

   1.1 *Importación de datos, preprocesamiento, captura de datos y exportación:* son herramientas que nos permiten traer datos desde otros formatos, crear los archivos de diccionarios y datos manualmente, exportar las salidas a R (R Core Team 2012) o Excel®.

   1.2 *Archivos de comando:* son herramientas de estadística básica, como creación de tablas, medidas de dispersión, variabilidad (media, desviación estándar,máximo, mínimo,...) y descripción de variables categóricas (frecuencias). También están los procedimientos descriptivos multivariados: análisis en componentes principales, análisis de correspondencias simples y múltiples. Por último los procedimientos para datos textuales que serán abordados detalladamente en sección 4.4

   1.3 *Archivos de resultados:* se presentan los dos formatos del archivo *imp* que contiene los resultados básicos y la lista de parámetros.

2. Vic, procedimientos de clasificación, visualización e inferencia.

   2.1 *Axes:* en estas opciones se pueden visualizar los planos factoriales, las coordenadas en los ejes, las diferentes formas de bootstrap (Lebart 2004) y por último la seriación que es una forma muy antigua de presentar tablas, donde se puede ver una forma gráfica de los ejes principales.

   2.2 *Axes/clusters:* son las opciones de clasificación. ClusterView, es la visualización de los conglomerados usando la metodología de Lebart, Morineau & Piron (1995). *Kohonen Map* utiliza redes neuronales para hacer clasificaciones mediante cuadriculas. *Visualization* permite ver las diferentes clasificaciones en los planos factoriales, además se puede hacer agrupaciones mediante *k-means* instantáneamente.

3. *DtmVic Tools:* son herramienta que ayudan al procesamiento de archivos numéricos y textuales. Encontramos opciones de guardar ejes factoriales, crear tablas de contingencia, tomar subgrupos de variables e individuos y otras herramientas de procesamiento de textos.

4. *DtmVic Images:* es una herramienta académica en la que se puede ver el método de componentes principales aplicado en imágenes a color o en escala de grises. Es útil en estudios geográficos, en los cuales se tiene interés por resumir información de mapas. Las imágenes deben estar en formato pgm o ppm. La conversión de imágenes jpg o png a ppm se puede hacer mediante *Image Converter Plus*, programa que se puede descargar de http://www.imageconverterplus.com

4.3. Archivos Principales

Para hacer cualquier análisis en este software es necesario proporcionar tres tipos de *archivos de entrada*, y en cada una de sus tareas el genera varios *archivos de salida*, los cuales sirven de comunicación entre etapas y pueden ser modificados por el usuario. Por ejemplo, si usamos la herramienta Visutex (1.2 de la anterior sección) el software solicitara un archivo de texto con ciertas condiciones.

A continuación presentaremos detalladamente los archivos de entrada y de salida, haciendo especial énfasis en su construcción.

4.3.1. Archivos de entrada

Son tres archivos en formato .txt; Diccionario, datos y textos. Estos deben ser proporcionados por el usuario o importados directamente desde un archivo Excel. A continuación se presenta la forma manual como se construyen:

Diccionario

Tiene como función identificar con tipo y nombre las variables asociadas al texto. Los tipos de variables se deben diferenciar por categóricas o numéricas, las cuales deben ir en el mismo orden que en el archivo de datos.

Variables Categóricas: Se escribe el número de categorías, seguido por el nombre. Debajo un identificador de cuatro dígitos y luego de un espacio el nombre real de la variable. En caso de no respuesta, por default el programa recibe catb como identificador de clase.

```
   3 diversifica
cat1 diver_NO
cat2 diver_SI
catb diver_catb
   3 jornalea
cat1 jornalea_NO
cat2 jornalea_SI
catb jornalea_catb
   3 crédito
cat1 crédito_NO
cat2 crédito_SI
catb crédito_catb
   5 tipo
cat1 EmpTecModer
cat2 TecModerno
cat3 CamTecModer
cat4 CamTradic
catb tipo_catb
```

Variables Numéricas: Se identifican con un cero, seguido por el nombre de la variable.

```
   0 Edad
```

Datos

Contiene los datos numéricos ordenados por variables, cuya primera columna identifica a cada individuo y debe estar entre comillas simples. No debe contener los nombres de las variables, puesto que ya están guardadas en el diccionario.

```
'N1'  1 3 2 1
'N2'  1 3 2 2
'N3'  3 1 2 4
'N4'  2 1 1 2
```

En este caso el primer productor de café respondió que no tenia cultivos de diversificación, no respondió si jornaleaba, posee crédito y es empresario tecnificado moderno.

Texto

El archivo de texto debe crearse de acuerdo al tipo de análisis que se va a efectuar. A continuación se presentan estas posibilidades.

Sin archivos numéricos asociados: análisis de textos líricos divididos por individuos o capítulos. Esta división se hace precediendo 4 asteriscos a cada respuesta o sección y ==== para el final del texto. El siguiente ejemplo son dos poemas de Carmen feito Maeso y Nicomedes Santa Cruz sobre el café.

```
**** El olor del café, CARMEN FEITO MAESO
Octubre. Otoño las hojas se vuelven rojas,
el color del otoño se acentúa en ellas.
En la mesa del balcón viendo el nuevo octubre y
saboreando el humeante café , oloroso café, el café.
La conversación alrededor del café fluye intima.
El amor de Octubre huele y sabe a café, dulce y tranquilo.
```

```
La luz roja de las hojas, se refleja en la taza de café.
¡Ah! El café.

¿Tú también tomas café? Se refleja también en tu otoño,
el olor de mi café.
Aire, agua y sol y café.
El café que da vida al espíritu.
El olor del café hace recordar el pasado.
El amor perdido, el dolor que se siente al perderlo.
Olor a café, olor a calor a ternura a vida.
En la mesa del café renace la inquietud.
Emerge del alma el deseo de vivir.
El café calienta el corazón.

**** EL CAFÉ, Nicomedes Santa Cruz

Tengo tu mismo color
Y tu misma procedencia.
Somos aroma y esencia
Y amargo es nuestro sabor.
Tú viajaste a Nueva York
Con visa en Bab-el-Mandeb,
Yo mi Trópico crucé
De Abisinia a las Antillas.
Soy como ustedes semillas.
Son un grano de café.

En los tiempos coloniales
Tú me viste en la espesura
Con mi liana a la cintura
Y mis abóreos timbales.
Compañero de mis males,
Yo mismo te trasplanté.
Surgiste y yo progresé:
En los mejores hoteles
Te dijeron ¡qué bien hueles!
Y yo asentí ?¡uí, mesié!?.
====
```

Con archivos numéricos asociados: hacer análisis de preguntas abiertas o literatura, en la cual se tienen información adicional acerca del escritor, año, editorial, entre otros. Muy útil en encuestas y análisis de textos de diferentes autores. Para separar las respuestas se usan cuatro guiones y para el final del texto ====. Para el ejemplo de la sección 1.1 seria.

```
---- N1
  por llevar una excelente administración de los cultivos
---- N2
  porque es agrónomo y realiza una administración directa de la finca
---- N3
  por que lleva una administración directa y realiza las labores oportunamente
---- N4
  por vivir en la finca y llevar una administración directa.
---- N5
  por que se ha dedicado siempre al cultivo del café y esto le ha dado para vivir
---- N6
  por realizar administración directa de los cultivos.
---- N7
  por que tiene buena capacidad de endeudamiento.
---- N8
  por que vive del cultivo de café y siempre se ha dedicado ha esta actividad
---- N9
  es una actividad que le gusta mucho y lleva una administración directa
---- N10
  por que ha vivido de este cultivo toda la vida.
====
```

4.3.2. Archivos de Salida

Cada uno de los procedimientos genera archivos que se guardan en la carpeta de trabajo. En esta sección solo vamos a detallar los archivos imp (1.3 Resultados básicos) y las listas de parámetros.

Imp (resultados básicos)

En las ultimas versiones de DtmVic el archivo Imp se encuentra en formatos txt y html, para facilitar la búsqueda de los resultados. Dentro de estos se ven cada uno de los procedimientos utilizados precedidos de la palabra Step. En el caso de datos textuales encontraremos frecuentemente los procedimientos Ardat (lectura de diccionario y datos), Artex (construcción del archivo de texto), Selox (selección de las preguntas abiertas) y Numer (numeración del texto), entre otros. Cada uno de estos procedimientos generan archivos que son conectados, por ejemplo ndicz, ndonz y ntexz son recibidos por Ardat y Artex, y generan los archivos ndica y ndona que son usados en otros tratamientos.

En este archivo también encontraremos los valores y vectores propios, las coordenadas de los individuos, el dendograma, las clases, las palabras asociadas, y en general los resultados de cada metodología que usemos.

Observación: si hubiese un error en algún procedimiento, seria indicado en este archivo.

Listas de parámetros

Cada vez que usamos un procedimiento, el software genera un archivo con nombre *ParamX* con X el nombre del método (Visutex, Visuresp, Analex,...). Este archivo contiene los parámetros requeridos por el procedimiento para estos datos específicos y algunas opciones escogidas por el usuario.

Estos archivos son muy importantes, ya que pueden ser modificados directamente y retomados por el software con nuevas propiedades. Para modificarlos y retomarlos en DtmVic se deben seguir los siguientes pasos:

1. Open a existing command file
2. Abrir el archivo de parámetros
3. Modificar los parámetros y pulsar en *Return to execute*
4. Execute a command file

4.4. Datos textuales

En esta sección se muestran los procedimientos que permite DtmVic para el análisis de textos. Estos pueden ser divididos en tres, pre procesamiento del texto, herramientas lexicométricas y análisis de tablas léxicas, el cual a su vez se divide en dos, con datos asociados y sin ellos.

4.4.1. Pre procesamiento del texto

Procedimiento Cortex: Corte y fusión de palabras

A partir del glosario del texto, es decir una tabla de frecuencias léxicas, el procedimiento Cortex permite poner un umbral mínimo de frecuencia desde el cual se tendrán en cuenta las palabras en los métodos. También permite eliminar palabras que no sean relevantes para el análisis, como en algunos casos las preposiciones.

Este procedimiento recibe el archivo numerado y genera otro archivo de texto con las palabras que quedaron encima del umbral y no fueron borradas.

Para hacer uso de esta herramienta los pasos son:

1. Create a command file
2. Cortex
3. Open the text file to be preprocessed
4. Seleccionar el texto a recortar
5. Characters separating the words, OK
6. Basic Vocabulary
7. Seleccionar las palabras a eliminar o unir
8. Crear el nuevo archivo de texto

Finalmente obtendremos un texto recortado con el cual usaremos los demás métodos.

Uso de Lematizadores

Bolasco (1992, p.70) propone como unidad textual la raíz léxica, permitiendo identificar ciertas equivalencias entre formas. Ejemplos de raíces léxicas son, los verbos en infinitivo, los sustantivos en singular, entre otros.

Una herramienta de lematización, sera presentada dos capítulos mas adelante, con lo que sera mas claro.

4.4.2. Herramientas lexicométricas

Los resultados de cada uno de los procedimientos que presentaremos se pueden ver en el archivo de resumen(Imp).

Procedimiento Corda, concordancias

Las concordancias muestran el contexto de una forma (palabra), en el orden de aparición del texto. Es decir, muestra las lineas en las cuales una palabra es encontrada, lo cual permite identificar la idea que representa. Es muy importante para el análisis de los resultados, ya que las palabras pueden cambiar de significado según el contexto o el autor, ademas se pueden separar aquellas homografas (que se escriben igual pero tienen significados distintos).

Los pasos para usar esta herramienta son:

1. Create a command file
2. Other analyses
3. Corda
4. Open a text file
5. Select open question and separators
6. Seleccionar la primera pregunta abierta
7. Seleccionar las palabras a eliminar o unir
8. Vocabulary and counts
9. Seleccionar las palabras para las concordancias
10. Create the command file
11. Execute

Procedimiento Segme: segmentos repetidos

Son secuencias de dos o mas palabras ,no separadas por un delimitador, que aparecen mas de una vez en un corpus de datos textuales (Etxeberria, García, Gil & Rodríguez 1995). A partir de estos segmentos se puede crear una tabla de frecuencias y hacer el análisis sobre esta.

Los pasos para usar esta herramienta son:

1. Create a command file
2. Other analyses
3. Segme
4. Open a text file
5. Select open question and separators
6. Vocabulary and counts
7. Continue: Create a parameter file
8. Options: Minimo y maximo de frecuencia y tamaño
9. Confirm (Se debe hacer para cada opción)
10. Continue
11. Create a command file
12. Execute

4.4.3. Análisis tablas léxicas

En general los procedimientos requieren de una tabla de frecuencias sobre la cual se usará análisis de correspondencias simples o múltiples. Cada uno de estos procedimientos puede ir acompañado de pre procesamientos como Cortex, los cuales ya están unidos en el software mediante la pestaña *Other Analyses* de *Create a command file* en la pantalla inicial.

Los pasos para usar estas herramientas son muy parecidos:

1. Create a command file
2. Visutex,Visuresp, Analex, Visureca, MCA text
3. Open a text file (Seleccionar el archivo de texto 4.3.1.3)
4. Select open question and separators
5. Vocabulary and counts
6. Seleccionar un umbral: Confirm
7. Continue: create a command files
8. Open a dictionary (4.3.1.1), open a data file (4.3.1.2)
9. Continue: Select active and supplementary variables
10. Seleccionar los individuos suplementarios
11. Seleccionar algunas opciones: Bootstrap y cluster
12. Create a first parameters file
13. Execute

Sin datos numéricos asociados

i *Visutex:* Visualización del texto. Crea un glosario y hace ACS sobre este.

ii *Visuresp:* Visualización de respuestas. Crea una tabla de *Palabras de la respuesta X* y hace ACS sobre esta. Ademas hace clasificación de respuestas.

Con datos numéricos asociados

i *Analex:* Análisis de correspondencias simples, a través de una tabla léxica construida de una variable categórica especifica, caracterizando las respuestas.

ii *Visureca:* Visualización y clasificación de respuestas con datos categóricos y elementos suplementarios.

iii *MCA Text:* Análisis de correspondencias múltiples, clasificación y descripción de clases desde variables numéricas, categóricas y textuales.

4.5. Importación desde Excel®

La forma mas simple de introducir los datos a DtmVic es mediante un archivo de Excel guardado en formato .csv delimitado por punto y coma. Para tal fin debemos construir una base de datos, en la cual las columnas son variables categóricas o textuales. La primera columna debe ser el nombre de los individuos o capítulos. De haber solo dos columnas, nombre de los individuos y texto, la herramienta de importación dará formato para visualización (Visutex).

Si se tiene el sistema operativo en español se debe guardar la base de datos con la opción *Guardar como* y *CSV delimitado por comas*, el cual separara cada columna por punto y coma. Se debe prestar gran atención al archivo resultante, ya que puede tener algunas filas de más.

Luego de guardar la base de datos en CSV se deben seguir los siguientes pasos:

1. Data Importation, Preprocesing, Data Capture, Exportation
2. Importing, Dictionary, Data, and Text
3. Excel® type file
4. Star the importation process
5. Select imput data file: Abrir el archivo cvs
6. Identificar el tipo de variables: Categórica, Numérica, Textual y descartar
7. Update and continue
8. Values and counts
9. Create a dictionary and data
10. Nombrar los archivos con la extension .txt
11. Create a new dictionary
12. Create data and text file
13. Create a DtmVic parameter file
14. Create a first parameter file
15. Execute

Los archivos de entrada creados se encuentran en la carpeta de trabajo. El diccionario generalmente no tiene nombres para las variables que complazcan a los usuarios, pero estos pueden ser cambiados directamente en el txt resultante.

Observación: el texto a importar no debe tener punto y coma ni tabulaciones, ya que estos son cambios de columna para el programa.

4.6. Dimensión de textos y datos

1. Existen algunas restricciones en cuanto a las dimensiones del texto. La cantidad máxima de lineas es de 1 millón y el tamaño máximo de linea es 200. La primera restricción generalmente no se incumple, pero la segunda se puede volver un problema. Afortunadamente el programa cuenta con una herramienta que hace estos recortes de forma automática, los pasos para usarla son:

   i Toolbox: Preprocesing Text

   ii Changing the sizes of the lines in a DtmVic text file

   iii Seleccionar el archivo a recortar y escoger el tamaño de las lineas

2. En algunos casos el archivo de datos viene de R, luego sus separadores son tabulaciones o comas. Para poder usarlo en Dtm es necesario que estos sean cambiados por punto y coma, lo cual se puede hacer como sigue:

   i Toolbox: Preprocesing Numerical

   ii Buscar la opción necesaria

   iii Seleccionar el archivo

3. Si el texto llegase a ser muy largo, existe la opción de fragmentarlo en varios textos. Para usar esta herramienta se sigue:

   i Toolbox: Preprocesing Text

   ii Fragmentation of a Dtm text

   iii Seleccionar el archivo a fragmentar

4.7. Lematizadores. TreeTagger

4.7.1. Acerca de TreeTagger

TreeTagger es una herramienta para anotar texto con parte de su discurso y la información lematizada. Fue desarrollado por Helmut Schmid en el proyecto de cooperación técnica en el Instituto de Lingüística Computacional de la Universidad de Stuttgart. Tiene licencia libre para objetivos académicos. Ha sido utilizado con éxito para etiquetar Alemán, Inglés, francés, italiano, holandés, español, búlgaro, ruso, griego, portugués, chino, swahili, el latín, estonio y antiguos textos en francés y es adaptable a otros idiomas si un léxico y una corpus etiquetado manualmente de formación están disponibles (Schmid 1994).

4.7.2. Instalación

En Windows Para su instalación solo es necesario copiar la carpeta *TreeTagger* del CD en *c:Archivos del Programa* y enviar al escritorio un acceso directo del ejecutable *WinTreeTagger* que esta dentro de la carpeta *BIN*.

4.7.3. Creación de un archivo lematizado

El procedimiento es el siguiente:

1. Importar el texto de Excel a DtmVic como en 4.5
2. Abrir TreeTagger
3. Elegir el Lenguaje español
4. Pulsar en *The token in place of unknown lemma*
5. Input: cargar el archivo resultante de la importación en 1.
6. Pulsar *Verify* y *Run*
7. Ir a DtmVic, Toolbox: Preprocessing text
8. Re importing a Dtm file text after WinTreeTagger
9. Cargar el archivo generado en 6.
10. Eliminar las palabras que no son útiles en el análisis
11. Crear el nuevo archivo lematizado

Finalmente obtenemos un archivo lematizado el cual se puede usar en los métodos comentados en 4.4.

Referencias

Bécue, M. (1991), *Análisis de datos textuales. Métodos estadísticos y algoritmos*, CISIA, Paris.

Bolasco, S. (1992), Sur différentes stratégies dans une analyse des formes textuelles: une experimentation à partir de données d'enquête, *in* M. Bécue, L. lebart & N. Rajadell, eds, 'Jornades Internacionals d'Anàlisi de Dades Textuals', Servicio de Publicaciones de la UPC, Barcelona, pp. 69–88.

Cabarcas, G. & Pardo, C.-E. (2001), 'Métodos estadísticos multivariados en investigación social', Cursillo del Simposio de Estadística. Santa Marta, Bogotá.
URL: *http://www.docentes.unal.edu.co/cepardot/docs/SimposiosEstadistica/*

Cisia-Ceresta (2000), *SPAD. Versión 4.5. Manuel de prise en main*, Montreuil.

Etxeberria, J., García, E., Gil, J. & Rodríguez, G. (1995), *Análisis de datos y textos*, RA-MA Editorial, Madrid, España.

Gelbukh, A. & Sidorov, G. (2006), *Procesamiento automático del español con enfoque en recursos léxicos grandes*, Instituto Politécnico Nacional. Dirección de Publicaciones, Centro de Investigación en Computación. México.
URL: *www.gelbukh.com/libro-procesamiento/LibroPLN.pdf*

Lebart, L. (2004), Validité des visualisations de données textuelles, *in* '6eme International Conference on the Statistical Analysis of Textual Data', pp. 708–715.

Lebart, L. (2012), 'DtmVic: Data and Text Mining - Visualization, Inference, Classification. Exploratory statistical processing of complex data sets comprising both numerical and textual data.', Web.
URL: *http://www.dtmvic.com/*

Lebart, L., Morineau, A. & Piron, M. (1995), *Statisitique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart, L., Piron, M. & Morineau, A. (2006), *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données*, 4 edn, Dunod, Paris.

Lebart, L. & Salem, A. (1994), *Statistique textuelle*, Dunod, Paris.

Lebart, L., Salem, A. & Bécue, M. (2000), *Análisis estadístico de textos*, Milenio, Lleida (Espaïa).

Lebart, L., Salem, A. & Berry, L. (1997), *Exploring Textual Data*, Kluwer Academic Publishers.

Montenegro, A. & Pardo, C. (1996), Introducción al análisis de datos textuales, Folleto, Universidad Nacional de Colombia. Departamento de Matemáticas y Estadística, Bogotá.
URL: *http://www.docentes.unal.edu.co/cepardot/docs/Notas/CursoTex.zip*

Pardo, C. E. (1992), Análisis de la aplicación del método de Ward de clasificación jerarquica al caso de variables cualitativas, Tesis Magister Scientiae en Estadística, Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Matemáticas y Estadística, Bogotá. Clas. Local 1.96 P226a 1992.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Schmid, H. (1994), Probabilistic part-of-speech tagging using decision trees, *in* 'Proceedings of international conference on new methods in language processing', Vol. 12, Manchester, UK, pp. 44–49.