7. Clasificación de documentos basada en la Factorización No Negativa de Matrices

7.1 Introducción

En este capítulo, desarrollaremos una metodología que permite identificar y clasificar automáticamente por temática o categoría un conjunto dado de textos o documentos. En el caso de que dicho conjunto sea relativamente pequeño, la clasificación se podría realizar manualmente, aunque si se dispone de una cantidad de documentos medianamente elevada, esta tarea podría resultar bastante tediosa. Esta será la motivación principal de la aplicación, es decir, la *automatización.*

Cuando las categorías de clasificación están predefinidas, el proceso se denomina *supervisado* y está constituido por diferentes métodos que permiten automatizar la tarea de clasificar los documentos. Por el contrario, si la clasificación de los textos es del tipo *no supervisado*, la única suposición válida radica en que la colección de documentos está completamente desestructurada. La tarea de clasificación se convierte ahora en la de organizar los documentos de forma estructurada basándonos en los patrones *adquiridos* del propio conjunto de documentos. Esta estructuración puede ser *jerárquica* o *no jerárquica*. En el primero de los tipos, la organización de los documentos se realiza según una estructura en forma de árbol, de modo que colocando el conjunto completo de documentos en la raíz, irán apareciendo ramas representativas de las diferentes categorías existentes. Al final, cada documento formará parte de uno de esos grupos.

Cuando la estructura es *no jerárquica*, los documentos se colocan en grupos no solapados que no guardan ninguna relación de dependencia entre sí. En esta aplicación, realizaremos la clasificación siguiendo este modelo y haciendo uso de la *Factorización No Negativa de Matrices (NMF)*, de forma que se irán clasificando los documentos en función de una serie de característica semánticas que irán definiendo cada uno de los grupos temáticos o *clusters*.

Dado que la clasificación de documentos es una técnica bastante extendida, son numerosas las referencias bibliográficas relacionadas con la materia, si bien cabe destacar que gran parte de la información consultada se ha extraído de [Lee], [Berry] y [Berry2]. Por su enorme utilidad en el tratamiento de los datos, hay que mencionar la aplicación *TMG (Text to Matriz Generator)*.

7.2 Descripción general del problema de la clasificación de documentos en categorías

7.2.1 Representación de datos

Sea una colección de documentos, representada por una matriz que llamaremos **V** (matriz de datos), de dimensiones '*n x m*', donde *m* representa el número total de

documentos existentes en la colección y *m* el número de palabras contenidas en el diccionario de palabras que se tendrán en cuenta en la aplicación de cara a la clasificación. Por tanto, en cada columna de **V**, $\mathbf{V_j}$ se codificará el número de ocurrencias de cada una de las palabras del diccionario en el j-ésimo documento. De forma análoga resulta sencillo deducir, que en cada fila de la matriz **V** se contabilizará el número de veces que una cierta palabra del diccionario aparece en cada uno de los documentos de la colección.

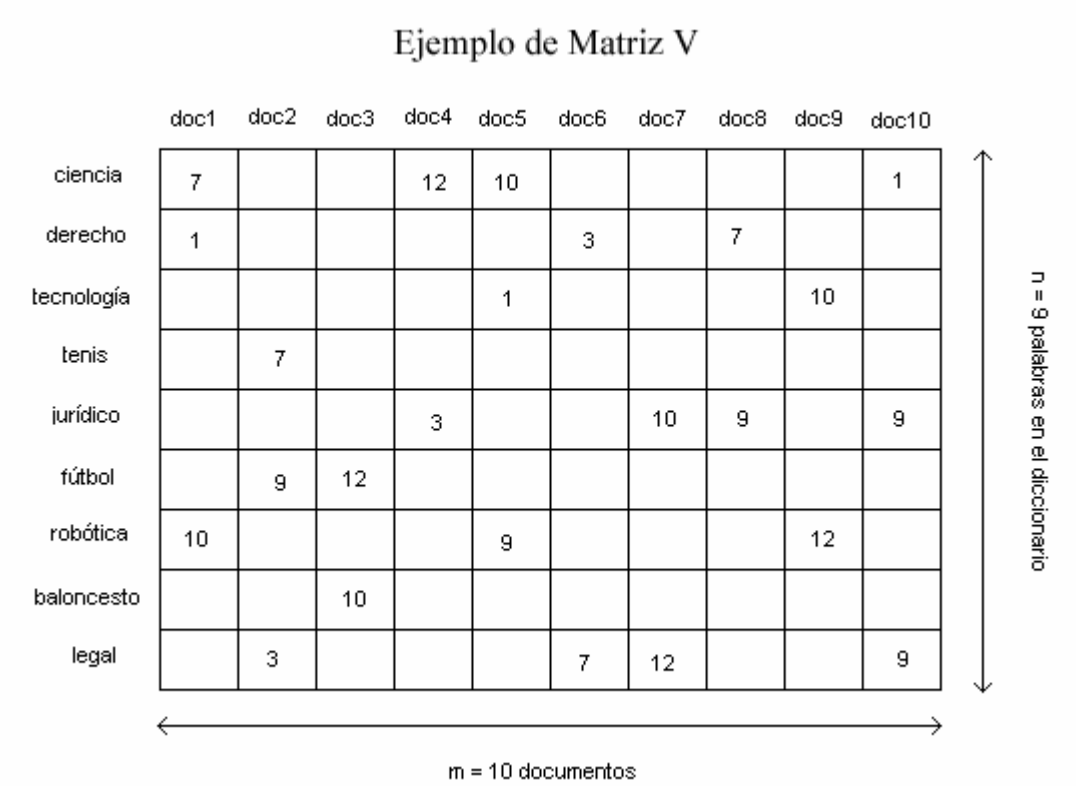

Figura 7.1 *Matriz de datos. En la figura se muestra una hipotética matriz **V** que contiene diez documentos y nueve palabras del diccionario representativas de tres temáticas claramente diferenciadas (temas deportivo, científico y jurídico). En cada elemento de la matriz se refleja el número de ocurrencias de cada una de las palabras en los documentos.*

El problema de clasificación basado en NMF pretende encontrar una aproximación de la matriz **V** en términos de dos matrices de menores dimensiones que llamaremos **W** y **H,** de forma que se pueda escribir como $\mathbf{V} \approx \mathbf{WH}$. Dimensionalmente, dichas matrices obtenidas al aplicar el algoritmo NMF tendrán dimensiones '*n x r*' para **W** y '*r x m*' en el caso de **H**.

La matriz **W** se conoce como *matriz de la* base. Cada columna de la matriz **W** es un vector de la base y representa un cierto concepto semántico extraído de los documentos contenidos en **V**. El algoritmo se ha diseñado de forma que en cada uno de los elementos de la base aparecerá reflejada la probabilidad de ocurrencia de cada palabra del diccionario relacionada con la temática que representa dicho vector.

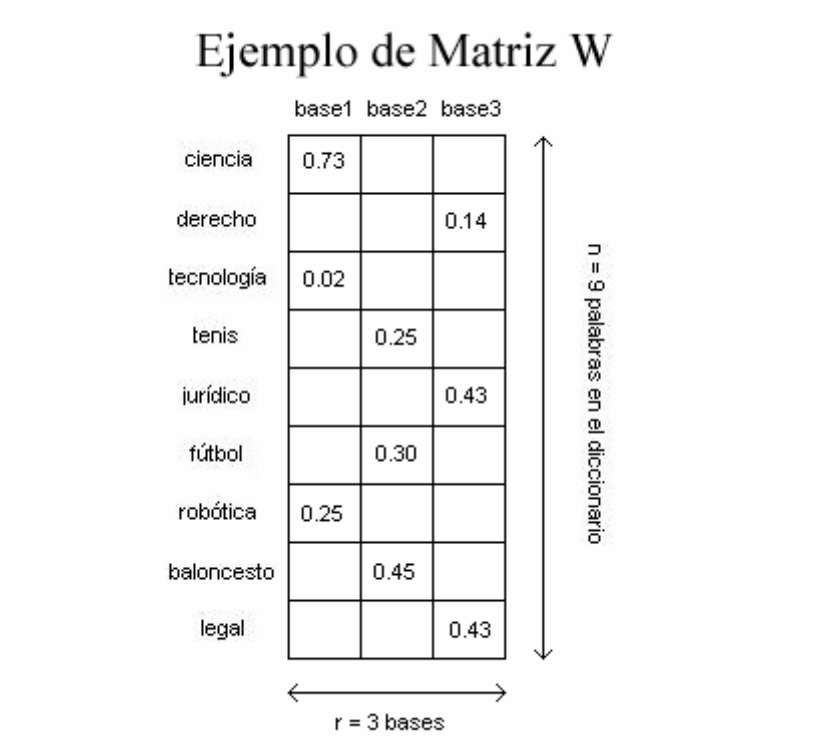

Figura 7.2 *Matriz de las bases. En la figura se muestra el resultado obtenido tras aplicar el algoritmo NMF a la matriz* ***V*** *dada en la figura 7.1. Se puede comprobar a partir de la distribución de probabilidades de ocurrencia en cada una de las tres bases que se ha realizado una correcta separación en función de las temáticas de los documentos.*

Por otro lado, i-ésima columna de la matriz **H**, conocida como *matriz de codificación*, contiene los coeficientes correspondientes a la combinación lineal de los vectores de la base que dan lugar a la i-ésima columna de la matriz **V**.

Una vez realizada la descomposición de la matriz de datos original hay que comprobar que es correcta. Para ello tendremos que ver si la clasificación original de los documentos coincide con la que se obtiene al aproximar $\mathbf{V} \approx \mathbf{WH}$.

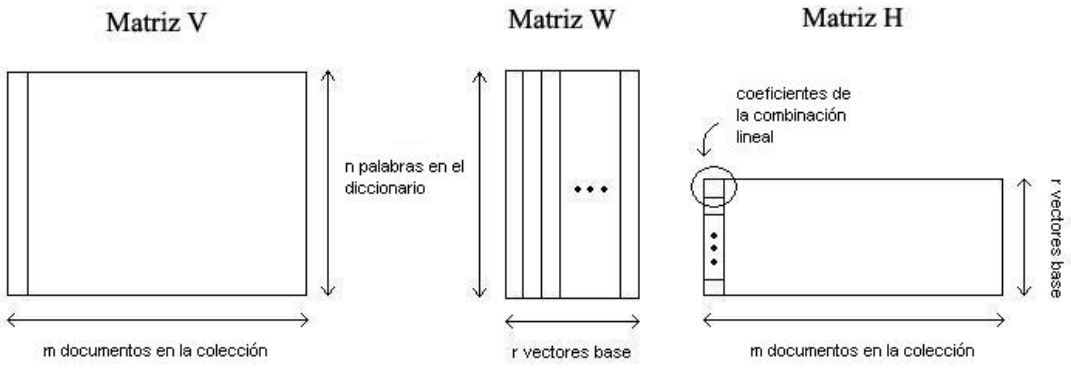

Figura 7.3 *Esquema general del problema de la clasificación de documentos.*

7.2.2 Valoración de los resultados obtenidos

¿Hasta qué punto los resultados son satisfactorios? Si queremos valorar de forma objetiva la bondad de los resultados, es necesario establecer una medida que permita indicar si la clasificación resultante se corresponde con la realidad.

En primer lugar tendremos que ver qué categoría le ha sido asignada a cada documento tras aplicar el algoritmo NMF. Definiremos dos formas de clasificación según empleemos la matriz de codificación o usemos el método basado en las proyecciones sobre la matriz de las bases.

7.2.2.1 Clasificación basada en la matriz de codificación H

Para ello, recordemos que la matriz de codificación **H** tiene por columnas a los documentos y sus filas representan cada una de las bases obtenidas, por lo que basta ver para cada columna (es decir, para cada documento) qué coeficiente toma un valor mayor. Dicho coeficiente representará para cada documento la base, y por tanto la categoría que se le asignará en la clasificación definitiva.

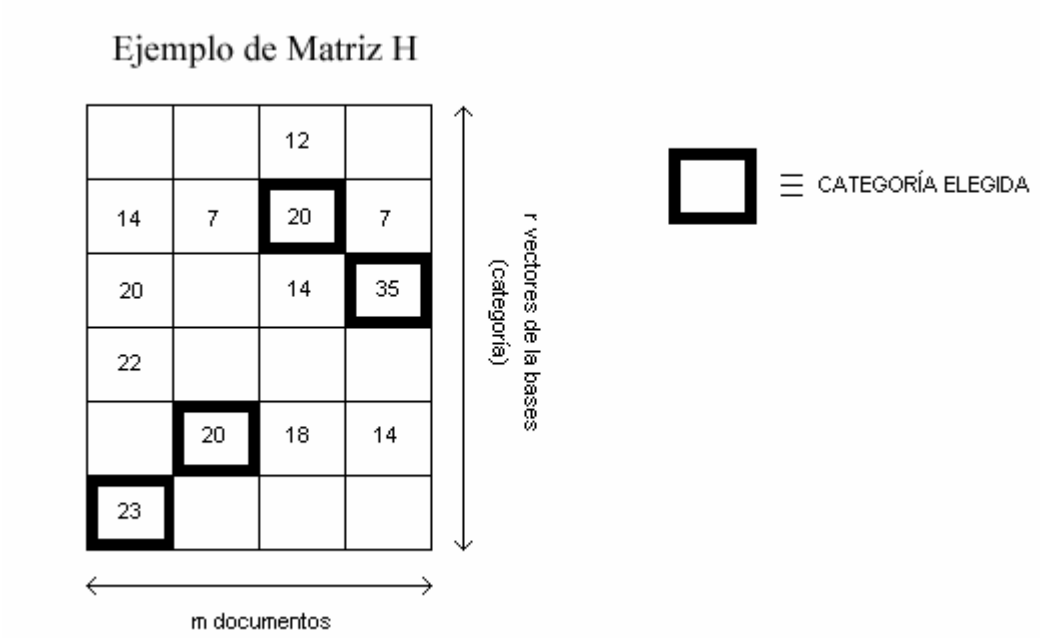

Figura 7.4 *Para decidir a qué categoría corresponde cada documento se comprueba en cada una de las columnas de la matriz de codificación **H** el coeficiente de mayor valor. Dicho coeficiente se asocia a una base representativa de la categoría asignada a dicho documento.*

Esta idea se puede expresar de forma matemática con el objeto de dotarla de una mayor generalidad. Para ello consideramos que la matriz de codificación **H** tiene unas dimensiones correspondientes a r filas y m columnas. De esta forma resulta sencillo verificar que existen $i = 1,...,m$ documentos a clasificar y $j = 1,...,r$ categorías disponibles, por lo que los coeficientes de **H** se pueden nombrar como h_{ji}. Fijando nuestra atención en un cierto documento i, si el máximo valor del i-ésimo vector

columna de **H** se localiza en la j-ésima entrada, entonces la categoría asignada a dicho documento es *j*.

7.2.2.2 Clasificación basada en las proyecciones sobre la matriz W

Una alternativa al método anterior de clasificación consiste en proyectar cada uno de los *m* vectores que representan a los documentos sobre la matriz de las bases **W**, y a partir de los coeficientes de las proyecciones reconstruir los vectores. La categoría elegida se determinará en función de la palabra de mayor peso dentro de cada documento.

El procedimiento de obtención de las proyecciones es análogo al estudiado en el capítulo anterior (referente a la reconstrucción de imágenes) y se basa en multiplicar cada uno de los vectores de la matriz de datos por los vectores de las bases (lo cual equivale a proyectar sobre el subespacio generado por las r bases). Una vez determinada la matriz de proyecciones es preciso normalizarla.

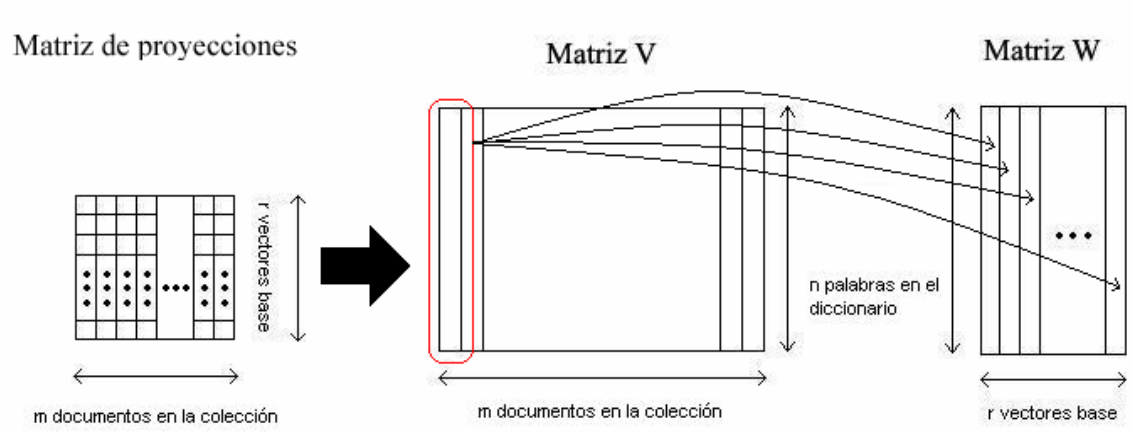

Figura 7.5 *Esquema general de la obtención de proyecciones a partir de la matriz de datos y de la matriz de las bases. La matriz de proyecciones se obtiene multiplicando cada vector (documento) de la matriz **V** por cada uno de los vectores de la base.*

A partir de las proyecciones, la reconstrucción del i-ésimo vector de datos consistirá en sumar cada uno de los *r* vectores de la base, que serán ponderados por su correspondiente coeficiente de proyección asociado a dicha base y al documento *i*. La asignación de categorías se realiza tomando la entrada (palabra) con mayor peso y se busca la temática en la que dicho término aparece con una frecuencia mayor.

Este método de reconstrucción es mucho más general que el anterior ya que permitiría clasificar incluso textos que no formaran parte de la base de datos original. Este hecho permitiría disponer de un sistema de clasificación en el que cualquier texto podría ser incluido en la categoría que guardara una mayor relación semántica con alguna de las temáticas de la base.

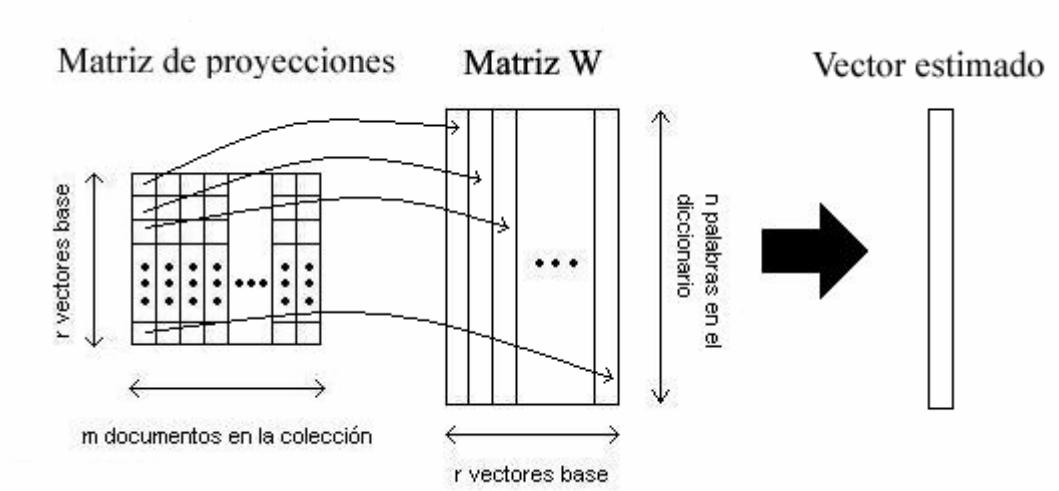

Figura 7.6 *Reconstrucción del primer vector de datos de la matriz V a partir de la matriz de bases W y de la matriz de proyecciones calculada previamente. En este caso, la reconstrucción se realizará mediante la suma de los vectores de la base ponderados por su correspondiente coeficiente de la proyección asociado al primer documento.*

7.2.2.3 Tasa de aciertos

Una vez los *m* documentos han sido clasificados según las *r* categorías, se establece una comparación con la correspondiente clasificación original previa a la aplicación del algoritmo NMF. De este modo, podremos comprobar como la *tasa de aciertos (TA)* en la nueva clasificación se definirá como:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m \tag{7.1}$$

donde δ_i con $i = 1,...,m$ es una variable binaria que toma valor '1' si la nueva clasificación del documento *i* coincide con la original, y '0' en caso contrario. Esta expresión proveerá valores entre cero y uno y servirá como medida de la bondad del proceso de clasificación realizado.

Como ejemplo de aplicación, consideremos la matriz **H** obtenida en la *figura 7.4*. Atendiendo a la forma de la matriz, existen $m = 4$ documentos a clasificar y $r = 6$ categorías disponibles. Sean d_i con $i = 1,...,m$ los documentos y C_j con $j = 1,...,r$ las categorías, de modo que la clasificación obtenida se representa en la siguiente tabla:

Documento	Categoría asignada
d_1	C_6
d_2	C_5
d_3	C_2
d_4	C_3

Tabla 7.1 *Asignación entre documentos y categorías para el ejemplo de la figura 7.4.*

Supongamos ahora que la clasificación original coincide con la asignada salvo en el valor estimado para d_3, por lo que ya será posible obtener el valor que tienen las variables binarias δ_i para cada documento. Seguidamente se muestra una tabla comparativa que resume el ejemplo:

Documento	Categoría asignada	Categoría original	δ_i
d_1	C_6	C_6	1
d_2	C_5	C_5	1
d_3	C_2	C_6	0
d_4	C_3	C_3	1

Tabla 7.2 *Tabla resumen con los resultados obtenidos en el ejemplo de la figura 7.4.*

Una vez tenemos los valores de δ_i podemos calcular la tasa de aciertos para este ejemplo según la expresión (7.1):

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{1+1+0+1}{4} = 0.75 \qquad (7.2)$$

por lo que podemos ver como el algoritmo proporciona una clasificación correcta en el 75% de los casos.

7.3 Experimentos

En este apartado se presentarán los resultados obtenidos al simular el problema de la clasificación de documentos mediante técnicas basadas en la Factorización No Negativa de Matrices. En primer lugar realizaremos una descripción de la base de datos que se va a emplear en los experimentos, luego se presentará la herramienta *TMG* que permite editar los datos para su tratamiento en Matlab y por último se realizará un análisis de los resultados que arroja el algoritmo y comprobaremos hasta qué punto han sido satisfactorios.

7.3.1 Descripción de la base de datos

Una cuestión esencial que cabe plantearse al iniciar el estudio de este problema consiste en decidir el conjunto de datos con el que trabajar. En principio, se consideró la posibilidad de trabajar con una base de datos extraída de alguno de los múltiples sitios web que hablan sobre esta técnica, si bien esta opción presentaba varios problemas que no la hacían del todo recomendable. Por un lado, resulta bastante difícil encontrar una base de datos que esté ya en formato Matlab (generalmente en .mat) y que permita su inserción directa en el algoritmo ya que la mayoría, como la base de datos *Reuters21578*, se encuentran codificadas en lenguajes cuya adaptación a formato Matlab se saldría de los objetivos de este Proyecto. Por otro lado, el tamaño de la mayoría de bases de datos, como la antes mencionada, es excesivamente grande, lo que hace que la carga computacional se elevara de forma que resultara inviable su ejecución

en un PC de sobremesa. Se podría pensar en intentar recortar la base de datos y tomar tan sólo un número de documentos que hiciera viable su procesado, sin embargo esta situación podría provocar que la información que aportara cada una de las categorías estuviera descompensada y por tanto las bases generadas no fueran del todo óptimas para una correcta reconstrucción. Además incluso en un caso extremo se podría dar la situación de que al reducir la base no se tomara ningún texto de alguna de las categorías, por lo que la base asignada a esa categoría contendría información de otras y podría inducir a errores en la clasificación final.

Por este motivo, se pensó en crear una *base de datos propia*, hecha a medida para el problema y cuyo tamaño no supusiera un obstáculo para la obtención de resultados. Si bien el hecho de que todas las bases encontradas en Internet estuvieran en inglés no fue un motivo crítico de cara a optar por la opción de crear una propia, el hecho de contar con una base de datos en español ayudaría en cierta medida a comprender los resultados de forma más clara. El proceso de documentación resultó arduo en cuanto a que fue necesario encontrar una gran cantidad de textos relativos a diferentes temáticas y editarlos de forma adecuada para que pudieran ser procesados por la herramienta *TMG* que presentaremos en un apartado posterior.

La base de datos consta de 120 documentos agrupados en 20 temáticas, de forma que se cuenta con 6 textos representativos de cada tema o categoría. Seguidamente se muestra la relación de temas que conforman la base, junto con el identificador que se le asignará de aquí en adelante para referenciarlas.

Tabla descriptiva de las temáticas de la base de datos

Identificador	Descripción	Nº palabras
aceite	Textos relativos al aceite de oliva: fabricación, propiedades, régimen económico, etc	5419
américa	Documentos extraídos de condenas inquisitoriales en América en el siglo XVII.	10047
aviones	Textos referentes a accidentes de aviones: rescate y atención de pasajeros, etc.	7469
coches	Descripción funcional de los diferentes sistemas que componen un coche.	6835
comercial	Nociones básicas de marketing y empresas.	5275
constitución	Varios artículos extraídos de la Constitución española.	6019
derecho	Textos jurídicos que reglamentan aspectos de la propiedad privada en Perú.	6678
dinosaurios	Documentos que describen cómo eran los dinosaurios: hábitos de vida, tipos, etc.	4466
judíos	Historia del pueblo judío a lo largo de diferentes etapas de su existencia.	6213
marx	Selección de varios textos de Marx.	8548
matemáticas	Documentos relativos a la enseñanza de las matemáticas.	4193
mitología	Relatos que describen algunos rasgos característicos de la mitología griega.	6595

Identificador	Descripción	Nº palabras
plantas	Descripción de los efectos beneficiosos sobre la salud de algunos tipos de plantas.	8154
pregón	Textos extraídos del Pregón de la Semana Santa de Sevilla del año 2000.	11475
psico	Textos introductorios a la psicología.	6106
relax	Técnicas y terapias de relajación.	3453
sacramento	Descripción de algunos Sacramentos.	13423
turismo	Conceptos básicos y fundamentos relativos al turismo en España.	4911
voz	Documentos relativos a la síntesis y reconocimiento de la voz humana.	5680
windows	Textos extraídos de un manual sobre el sistema operativo Windows.	8033

Tabla 7.3 *Categorías de la base de datos creada. En la tabla figura el identificador con el que se denota cada categoría, una breve descripción de su contenido y el número de palabras que contienen los documentos asociados.*

Como podemos constatar en la tabla anterior, los temas sobre los que tratan los textos que conforman la base de datos son variados, si bien se puede observar cierta similitud o cercanía entre algunos de ellos, lo cual provocará errores en la clasificación como comprobaremos posteriormente.

Otro aspecto importante a destacar es la *extensión de los textos*. Como parece lógico pensar, mientras mayor sea el tamaño global de los documentos correspondientes a una cierta temática, mayor será el número de ocurrencias de las palabras significativas relativas a dicho tema. Esto provocará que la correspondiente entrada de la matriz de datos referente a ese documento y a una cierta palabra que aparezca con mucha frecuencia tenga un gran peso de cara a la posterior descomposición en matrices **W** y **H**. Además mientras mayor sea el tamaño de los textos, mayor será la probabilidad de que aparezcan palabras representativas de otras categorías, lo cual conllevará errores en la obtención de las bases y por tanto provocará fallos en la clasificación definitiva.

7.3.2 La herramienta Text to Matriz Generator (TMG)

Una vez definido el conjunto de textos que conforman la base de datos nos planteamos un problema cuya solución no parece trivial a priori: ¿cómo transformamos los datos al formato de la matriz **V**?, ¿cómo se generará el diccionario?

Afortunadamente contamos con la ayuda de una aplicación de Matlab conocida como *Text to Matriz Generator (TMG)* que proporcionará los datos de forma adecuada para su procesado. Esta herramienta, desarrollada por los profesores *Dimitrios Zeimpekis* y *Efstratios Gallopoulos* de la Universidad de Patras se muestra en forma de GUI de Matlab de forma que su funcionamiento resulta bastante sencillo.

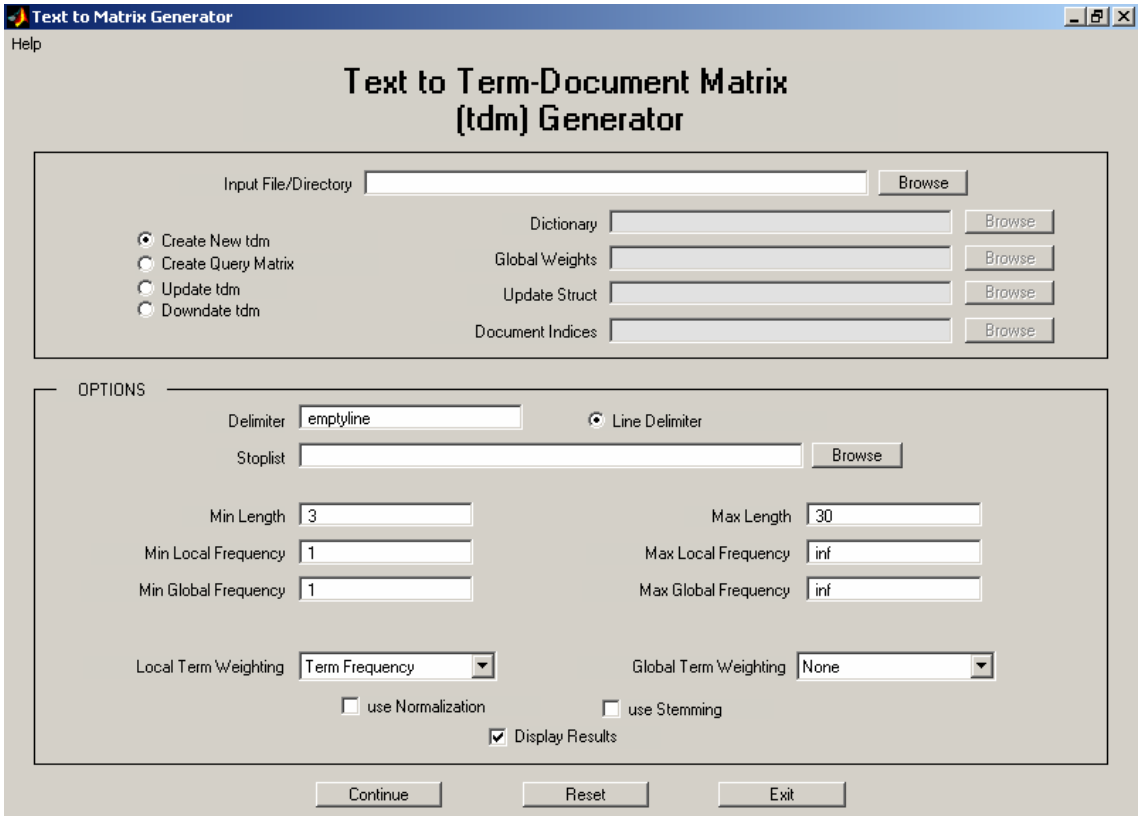

Figura 7.7 *Interfaz gráfica de la aplicación Text to Matriz Generator (TMG).*

Aunque permite multitud de opciones, para la aplicación de clasificación que se está realizando tan sólo emplearemos la obtención de la matriz **V** (*tdm matrix* si seguimos la nomenclatura empleada en el TMG*)* y del diccionario.

Text to Matrix Generator proporciona algunas características que facilitan la tarea de obtención de un diccionario de palabras óptimo en el sentido de que no contenga palabras comunes que podrían distorsionar en exceso la clasificación (véase preposiciones, algunos adverbios, artículos, etc). Para ello se permite incluir en el proceso de generación de los datos un *fichero de 'palabras prohibidas'* que en ningún caso podrán ser incluidas en el diccionario. La primera vez que se ejecutó el algoritmo, este fichero contenía preposiciones, adverbios, artículos, determinantes y una recopilación de verbos conjugados ya que la clasificación se hará tan sólo a partir de sustantivos y adjetivos en su mayoría.

Como opciones adicionales se permite seleccionar un tamaño mínimo y máximo de palabras a tener en cuenta, que en nuestro caso estará fijado en 3 y 18 caracteres respectivamente, y además cribar aquellas palabras cuya frecuencia de aparición local (en cada documento) y/o global (en el conjunto de los documentos) no se encuentre comprendida en un cierto rango prefijado. Además la selección de los textos a transformar a formato Matlab (.mat) se hace simplemente colocándolos en una misma carpeta cuya selección se hace en la misma interfaz gráfica de TMG.

Una vez ejecutado este programa con los textos de nuestra base de datos se comprobó si realmente los resultados eran satisfactorios. Dado que la versión disponible en la actualidad es una versión de las conocidas como *beta* (o en fase de pruebas),

salieron a la luz algunos fallos a los que se les ha intentado dar solución de la mejor forma posible. Entre todos el más grave consiste en que TMG trunca algunas palabras a partir de la vocal que contiene la tilde, para la cual se optó por sustituir en todos los textos las vocales con tilde por la correspondiente sin acentuar. Además algunos caracteres ASCII que aparecen junto a algunas palabras no eran reconocidos por el programa, para lo cual se tuvo que realizar una corrección manual.

El programa ofrece la posibilidad de mostrar una serie de datos estadísticos por pantalla que permitirán ver cómo ha ido transcurriendo el proceso de generación de las matrices de datos y del diccionario.

```
================================================================================
Results:
================================================================================
Number of documents = 120
Number of terms = 11516
Average number of terms per document (before the normalization) = 1158.27
Average number of indexing terms per document = 393.433
Sparsity = 2.12899%

Removed 5711 stopwords...
Removed 150 terms using the term-length thresholds...
Removed 0 terms using the global thresholds...
Removed 0 elements using the local thresholds...
Removed 0 empty terms...
Removed 0 empty documents...
================================================================================
```

Figura 7.8 *Resultados mostrados por pantalla tras la ejecución de TMG aplicada al conjunto de 120 textos.*

La matriz de datos **V** generada en un archivo con formato .mat consta de 120 columnas (una por documento) y 11516 filas correspondientes a cada una de las palabras. Por su parte, la matriz representativa del diccionario constará de 11516 filas (una por palabra) y 18 columnas (una por cada carácter de la palabra).

7.3.3 Aplicación del algoritmo NMF

Una vez se dispone del conjunto de datos, procedemos a la aplicación del algoritmo de Factorización No Negativa de Matrices con el objeto de obtener las matrices de las base (**W**) y de codificación (**H**). La ejecución del algoritmo se realizó empleando como máximo 2000 iteraciones o bien tomando como condición de finalización que la diferencia entre la función objetivo entre dos pasos consecutivos fuese menor que 10^{-9}. Con estos valores podemos garantizar que hemos llegado a una situación de convergencia en torno a un máximo de la función objetivo, por lo que la aproximación ha de ser en teoría aceptable. Sin embargo, la imposición de unas condiciones tan restrictivas provocará que la ejecución se alargue en el tiempo, si bien se encuentra aún en unos límites asumibles.

7.3.4 Resultados obtenidos

Con las matrices **W** y **H** calculadas, podremos comenzar ya a clasificar los diferentes documentos en categorías y comprobar qué medida la descomposición realizada arroja resultados satisfactorios. Para ello se realizarán diferentes experimentos en los que se pondrán de manifiesto las posibles causas que motivarán los errores en la clasificación.

Experimento 1: Obtención de las bases

En primer lugar y como experimento más básico, obtendremos $r = 20$ bases (una por cada uno de los temas que conforman la base de datos) y mostraremos las 5 palabras más representativas de cada una de ellas, para lo cual se tomarán en orden creciente las 5 entradas del diccionario que tengan asociado un mayor peso en cada una de las columnas de **W**. A partir de estas palabras, se intentará realizar una clasificación por temáticas tomando la palabra de la base con un mayor peso e identificándola con el texto en el que aparece más veces, asignándole ese tema. Esta clasificación se podría haber realizado también de forma subjetiva viendo las palabras que conforman las bases y asignándoles la temática más apropiada, si bien se ha considerado que lo ideal sería que todo el proceso fuera automático evitando así posibles confusiones en el caso de que la base de datos fuese mayor.

En la siguiente tabla se muestran tanto las bases obtenidas como la clasificación realizada siguiendo el criterio anteriormente expuesto, representando las palabras dominantes para cada una de las bases, así como la categoría que le ha sido asignada y la que realmente le correspondería. La decisión acerca de la categoría real de cada base se realizará por mayoría entre las cinco palabras de la base (en caso de empate se decide en función de la palabra de mayor peso). Para ver hasta qué punto se ha realizado de forma correcta la clasificación, incluiremos el parámetro δ_i que permitirá definir la tasa de aciertos. Como se verá a continuación, existen algunas categorías que aparecen repetidas y por tanto otras que no aparecen en la clasificación. En estos casos, la variable δ_i se colocará a cero para aquellas categorías en las que se produzca repetición de una de las anteriores. De esta forma se pretende modelar el hecho de que algunas temáticas aparezcan representadas en los términos dominantes de las bases y otras sin embargo no.

Base	Palabras dominantes	Tema asignado	Tema correcto	δ_i
1	clase derecho windows dominante carpeta	*marx*	*marx*	1
2	articulo camaras rey generales matematicas	*constitución*	*constitución*	1

Base	Palabras dominantes	Tema asignado	Tema correcto	δ_i
3	windows sistema carpeta juegos velocidad	*windows*	*windows*	1
4	semillas planta luz gloria semilla	*plantas*	*plantas*	1
5	bautismo cristo agua dios mision	*sacramento*	*sacramento*	1
6	cannabis naturaleza hombre feuerbach planta	*plantas*	*plantas*	0
7	dios aprendizaje sevilla psicologia enseñanza	*pregón*	*pregón*	1
8	enfermos sacramento dios cristo eucaristia	*sacramento*	*sacramento*	0
9	plan emergencia heridos victima empresa	*aviones*	*aviones*	1
10	imagenes leccion situaciones vida tension	*relax*	*relax*	1
11	aeronave dinosaurios evacuacion equipos incendio	*aviones*	*aviones*	0

Base	Palabras dominantes	Tema asignado	Tema correcto	δ_i
12	servicio catering empresa oficio servicios	*comercial*	*comercial*	1
13	dios matrimonio hombre mujer inquisicion	*sacramento*	*sacramento*	0
14	reconocimiento voz palabras sistema sistemas	*voz*	*voz*	1
15	aceite oliva aceites campaña flavor	*aceite*	*aceite*	1
16	turismo servicios empresas agencias publicos	*turismo*	*turismo*	1
17	dinosaurios helios grandes animales tierra	*dinosaurios*	*dinosaurios*	1
18	espiritu santo confirmacion uncion bautismo	*sacramento*	*sacramento*	0
19	propiedad sistema transferencia contrato articulo	*derecho*	*derecho*	1
20	ley israel derecho kneset justicia	*constitución*	*constitución*	0

Tabla 7.4 *Palabras dominantes y categorías para cada una de las 20 bases obtenidas.*

A raíz de la asignación realizada de las variables δ_i, podemos concluir que la tasa de aciertos ha resultado:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{14}{20} = 0.70$$

A pesar de obtener una tasa de aciertos relativamente elevada en relación al número de bases calculadas, muchas temáticas no cuentan con una base en la que sus principales términos representativos tengan un peso elevado. Concretamente, en el problema que estamos tratando, dichas categorías son *américa*, *coches*, *judíos*, *matemáticas*, *mitología* y *psico*. Por el contrario, otras temáticas aparecen representadas en varias ocasiones, como es el caso de *aviones* (en dos ocasiones), *constitución* (en dos ocasiones), *plantas* (en dos ocasiones) y *sacramento* (en cuatro ocasiones).

¿A qué puede ser debido este efecto? En primera aproximación, se puede deber a tres motivos fundamentales: la extensión de los textos, la frecuencia de ocurrencia de las palabras dominantes en los documentos y la existencia de palabras comunes a varias categorías.

1. *Extensión de los textos*

   Como se comentó anteriormente, mientras mayor sea la extensión de cada documento mayor será el número de ocurrencias de las palabras dominantes y por tanto, más importancia tendrán de cara a la descomposición NMF, frente a otras. En la siguiente gráfica se intentará comprobar si realmente este efecto es realmente en parte causante de los errores en la clasificación, para lo cual en el eje de abcisas se colocará el número de palabras totales que conforman cada una de las veinte temáticas y en el de ordenadas se representará un ‘1’ si dicha categoría presenta base propia y un ‘0’ en caso contrario.

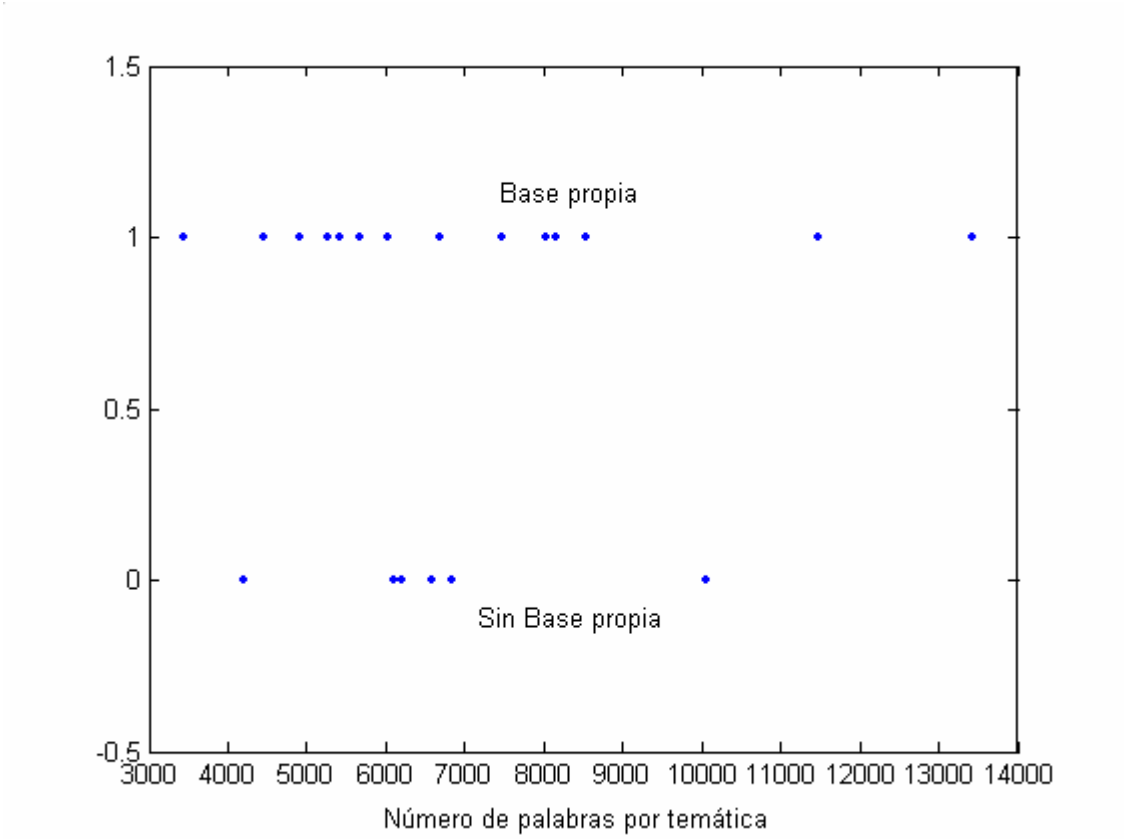

Figura 7.9 *En esta figura se representa la existencia de base asociada a la temática, frente al número de palabras que conforman los documentos de dicha categoría.*

A tenor de los resultados se podría considerar que la extensión de los textos no tiene influencia en la existencia de una base propia relativa a la temática, ya que las categorías que carecen de ella no se localizan en la zona de menor número de palabras. Sin embargo si hay que destacar el hecho de que las temáticas que presentan más de una base asociada suelen ser las que tienen un mayor número de términos.

2. Frecuencia de ocurrencia de las palabras dominantes

Por otro lado, existe la posibilidad de que el mayor peso de unas entradas del diccionario frente a otras no venga provocada por el tamaño de los documentos, sino porque el número de ocurrencias de una misma palabra en un documento es excesivamente elevada. En la siguiente figura se pondrá este hecho de manifiesto al representar por '1' o '0' la existencia o no de base propia, frente a la frecuencia de la palabra dominante de cada clase.

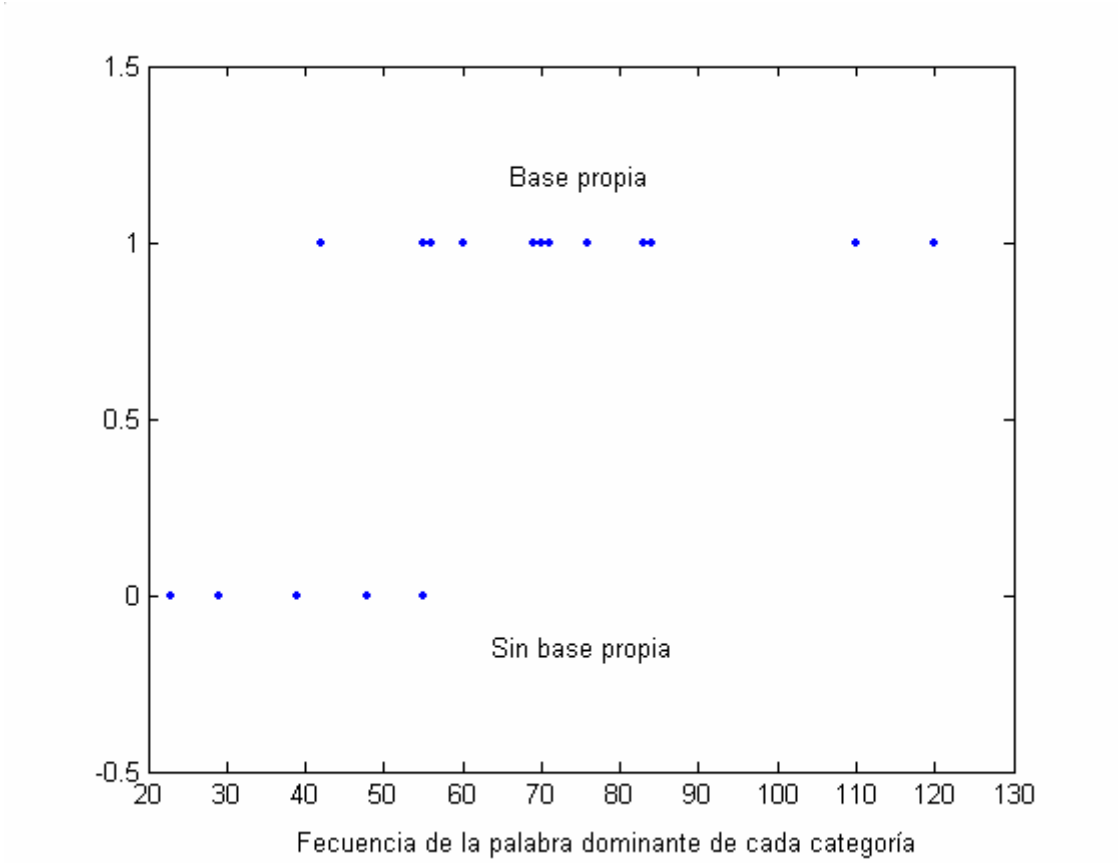

Figura 7.10 *En esta figura se representa la existencia de base asociada a la temática, frente a la frecuencia de la palabra dominante de dicha categoría. Se puede comprobar como las temáticas que carecen de base propia se localizan en la zona de frecuencias bajas.*

Como se observa en la figura anterior, aquellas categorías que *no* han logrado obtener una base propia en el proceso de descomposición NMF se encuentran en la zona de baja frecuencia de repetición de la palabra dominante.

3. *Existencia de palabras comunes a varias categorías*

Se puede dar el caso de que existan palabras que presenten diferentes significados según estemos en una aplicación u otra. Como no es posible establecer una clasificación atendiendo al contexto en el que se encuentre la palabra, este hecho será una nueva fuente de posibles errores en la descomposición y obtención de las bases.

Al realizar un análisis de las diferentes palabras propias de una categoría que se repiten en otras comprobamos como al menos 183 palabras eran comunes a diez o más categorías y que un total de 3315 palabras (de las 11516 que consta el diccionario) se repiten al menos una vez. Como ejemplo sirvan las palabras más repetidas entre todas las categorías:

Palabra	Nº repeticiones (categ)	Palabra	Nº repeticiones (categ)
tiempo	192 (20)	siempre	128 (19)
gran	160 (20)	posible	107 (19)
asi	151 (20)	manera	106 (19)
vez	132 (19)	donde	106 (19)

Tabla 7.5 *Palabras más repetidas entre todas las categorías. En la tabla se representa el número de veces que se repite cada una de las palabras y entre paréntesis, el número de categorías distintas en las que aparece.*

Como observamos, este hecho es extremadamente crítico ya que va a suponer la existencia de cierta correlación entre las categorías, por lo que cabría la posibilidad de que la separación por clases no se hiciera de manera correcta. Una posible forma de corregir este problema podría consistir en incluir todas aquellas palabras que formaran parte del grupo de términos con exceso de repetición por temáticas, en la lista de palabras prohibidas y de esta forma volver a determinar las bases. Sin embargo, es posible que muchas de estas palabras eliminadas fuesen las palabras más representativas de algunas categorías, por lo que de nuevo estaríamos induciendo una desviación en la obtención de las bases.

Experimento 2: Clasificación de textos

Una vez calculadas las bases procederemos a clasificar cada uno de los 120 documentos que conforman la base de datos en sus diferentes categorías y así verificar hasta qué punto la factorización de la matriz **V** es correcta. Como se estudió en el *apartado 7.2.2*, la clasificación se puede realizar atendiendo a dos criterios según se emplee la matriz de codificación **H** o las proyecciones de los vectores de datos sobre la matriz de las bases **W**.

1. *Clasificación a partir de la matriz de codificación **H***

El primer problema que se plantea a la hora de clasificar los documentos mediante este procedimiento, radica en el hecho de que la matriz **H**, al igual que ocurre en el caso de la matriz de las bases, no presenta una ordenación por filas adecuada a la de las temáticas originales. A ello hay que añadir que como vimos en el experimento 1, existen temáticas que constan de más de una base asignada, mientras que otras categorías no tienen ninguna, lo cual podría provocar errores en la clasificación.

Tras implementar el algoritmo de clasificación de categorías basado en la matriz de codificación y haciendo uso de la *tabla 7.4* para la identificación de las bases, se extrajeron los siguientes resultados que se presentan en la siguiente tabla:

Tabla de resultados obtenidos

[1] Número de orden del documento dentro de su categoría.
[2] Categoría asignada tras la clasificación.

Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i	Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i
aceite	1	aceite	1	dinosaurios	1	aviones	0
	2	aceite	1		2	comercial	0
	3	aceite	1		3	aviones	0
	4	plantas	0		4	dinosaurios	1
	5	turismo	0		5	dinosaurios	1
	6	aceite	1		6	dinosaurios	1
america	1	sacramento	0	judios	1	constitución	0
	2	sacramento	0		2	sacramento	0
	3	sacramento	0		3	voz	0
	4	sacramento	0		4	constitucion	0
	5	comercial	0		5	constitucion	0
	6	sacramento	0		6	sacramento	0
aviones	1	aviones	1	marx	1	derecho	0
	2	aviones	1		2	derecho	0
	3	aviones	1		3	relax	0
	4	aviones	1		4	sacramento	0
	5	aviones	1		5	relax	0
	6	aviones	1		6	marx	1
coches	1	constitución	0	matematicas	1	pregon	0
	2	pregón	0		2	pregon	0
	3	sacramento	0		3	pregon	0
	4	windows	0		4	constitucion	0
	5	windows	0		5	constitucion	0
	6	constitución	0		6	plantas	0
comercial	1	aviones	0	mitologia	1	sacramento	0
	2	comercial	1		2	turismo	0
	3	comercial	1		3	dinosaurios	0
	4	plantas	0		4	dinosaurios	0
	5	aceite	0		5	windows	0
	6	marx	0		6	voz	0
constitucion	1	marx	0	plantas	1	plantas	1
	2	constitucion	1		2	plantas	1
	3	constitucion	1		3	plantas	1
	4	turismo	0		4	plantas	1
	5	constitucion	1		5	plantas	1
	6	constitucion	1		6	relax	0
derecho	1	derecho	1	pregon	1	plantas	0
	2	derecho	1		2	plantas	0
	3	derecho	1		3	pregon	1
	4	derecho	1		4	turismo	0
	5	derecho	1		5	constitucion	0
	6	derecho	1		6	relax	0

Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i	Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i
psico	1	pregon	0	turismo	1	turismo	1
	2	aviones	0		2	turismo	1
	3	aviones	0		3	turismo	1
	4	plantas	0		4	turismo	1
	5	sacramento	0		5	turismo	1
	6	plantas	0		6	turismo	1
relax	1	relax	1	voz	1	voz	1
	2	relax	1		2	voz	1
	3	relax	1		3	voz	1
	4	relax	1		4	voz	1
	5	relax	1		5	voz	1
	6	relax	1		6	voz	1
sacramento	1	sacramento	1	windows	1	windows	1
	2	sacramento	1		2	windows	1
	3	sacramento	1		3	marx	0
	4	sacramento	1		4	windows	1
	5	sacramento	1		5	windows	1
	6	sacramento	1		6	marx	0

Tabla 7.6 *Resultados obtenidos tras realizar la clasificación basada en la matriz de codificación* ***H****.*

Una vez realizada la clasificación, el siguiente paso consiste en verificar la tasa de aciertos sobre el conjunto de los documentos clasificados:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{60}{120} = 0.50 \tag{7.3}$$

A priori esta cifra parece muy baja pero hay que tener en cuenta que en ella van incluidos documentos que no tienen bases propia, lo cual provoca que se introduzca un error considerable de entrada. Por tanto, si excluimos dichas categorías y nos centramos en las que tienen base propia, la tasa de aciertos pasa a tomar un valor:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{60}{84} = 0.7143 \tag{7.4}$$

Con todo, los porcentajes de acierto están en la línea de los obtenidos por otros autores en sus experimentos de clasificación de documentos mediante técnicas NMF, como por ejemplo en [Berry], donde la tasa de aciertos para este mismo número de bases ronda el 55%. Además, en todos ellos el tamaño del 'cluster' es muy inferior (en torno a 2000 palabras) al empleado en esta aplicación.

2. *Clasificación a partir de las proyecciones de los vectores de datos sobre la matriz de las bases **W***

Siguiendo el procedimiento estudiado en el *apartado 7.2.2.2*, se procedió a la clasificación de los 120 documentos en función de las categorías existentes. Al igual que ocurre en el caso de la clasificación basada en la matriz de codificación, existen categorías que por diferentes motivos no tienen la suficiente relevancia con respecto a otras que permitan su correcta clasificación. A la luz de los resultados arrojados en la siguiente tabla, podremos comprobar como dichas temáticas coinciden en su mayoría con las obtenidas en el experimento anterior.

Tabla de resultados obtenidos

[1] Número de orden del documento dentro de su categoría.
[2] Categoría asignada tras la clasificación.

Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i	Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i
aceite	1	aceite	1	dinosaurios	1	dinosaurios	1
	2	aceite	1		2	dinosaurios	1
	3	aceite	1		3	dinosaurios	1
	4	aceite	1		4	dinosaurios	1
	5	aceite	1		5	dinosaurios	1
	6	aceite	1		6	dinosaurios	1
america	1	pregón	0	judios	1	pregon	0
	2	pregón	0		2	sacramento	0
	3	sacramento	0		3	voz	0
	4	pregón	0		4	constitucion	0
	5	pregón	0		5	constitucion	0
	6	sacramento	0		6	sacramento	0
aviones	1	dinosaurios	0	marx	1	derecho	0
	2	aviones	1		2	derecho	0
	3	aviones	1		3	sacramento	0
	4	aviones	1		4	pregon	0
	5	dinosaurios	0		5	pregon	0
	6	aviones	1		6	constitucion	1
coches	1	comercial	0	matematicas	1	pregon	0
	2	pregón	0		2	pregon	0
	3	sacramento	0		3	pregon	0
	4	derecho	0		4	pregon	0
	5	derecho	0		5	pregon	0
	6	constitución	0		6	aceite	0
comercial	1	comercial	1	mitologia	1	pregon	0
	2	comercial	1		2	pregon	0
	3	comercial	1		3	dinosaurios	0
	4	comercial	1		4	dinosaurios	0
	5	aceite	0		5	derecho	0
	6	aceite	0		6	voz	0

Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i	Categoría	Nº doc.[1]	Cat. asign.[2]	δ_i
constitucion	1	constitucion	1	plantas	1	plantas	1
	2	constitucion	1		2	derecho	0
	3	constitucion	1		3	plantas	1
	4	constitucion	1		4	plantas	1
	5	constitucion	1		5	sacramentos	0
	6	constitucion	1		6	aceite	0
derecho	1	derecho	1	pregon	1	pregon	1
	2	derecho	1		2	pregon	1
	3	derecho	1		3	pregon	1
	4	derecho	1		4	pregon	1
	5	derecho	1		5	pregon	1
	6	derecho	1		6	pregon	1
psico	1	pregon	0	turismo	1	turismo	1
	2	pregon	0		2	turismo	1
	3	pregon	0		3	turismo	1
	4	pregon	0		4	turismo	1
	5	pregon	0		5	turismo	1
	6	sacramento	0		6	turismo	1
relax	1	relax	1	voz	1	voz	1
	2	relax	1		2	voz	1
	3	relax	1		3	voz	1
	4	relax	1		4	voz	1
	5	relax	1		5	voz	1
	6	relax	1		6	voz	1
sacramento	1	sacramento	1	windows	1	windows	1
	2	sacramento	1		2	windows	1
	3	sacramento	1		3	windows	1
	4	sacramento	1		4	windows	1
	5	sacramento	1		5	derecho	0
	6	sacramento	1		6	derecho	0

Tabla 7.7 *Resultados obtenidos tras realizar la clasificación basada en las proyecciones de los vectores de datos sobre la matriz de las bases* ***W****.*

Como se puede comprobar al calcular la tasa de aciertos que el porcentaje de éxitos en este caso es mayor que en el caso de la clasificación basada en la matriz **H**:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{68}{120} = 0.567 \tag{7.5}$$

Al igual que antes, se podría realizar una medida del éxito de la clasificación sin tener en cuenta aquellas categorías en las que no se ha podido clasificar ningún documento correctamente. Esta medida, aunque pueda resultar un tanto engañosa, permite verificar como para la mayoría de las categorías (sin tener en cuenta aquellas cuya palabra dominante tenga una frecuencia de

ocurrencia relativamente baja), se consigue clasificar los documentos de forma correcta en un porcentaje bastante elevado:

$$TA = \sum_{i=1}^{m} \delta_i \Big/ m = \frac{68}{78} = 0.872 \tag{7.6}$$

7.4 Conclusiones

En esta sección se ha mostrado la aplicación de la Factorización No Negativa de Matrices de cara a la clasificación de documentos en función de su temática. En definitiva, la Factorización No Negativa de Matrices permite obtener una representación de datos basada en partes de gran utilidad en procesos de clasificación como el estudiado en esta aplicación. La percepción de un conjunto, en este caso la base de datos, se puede descomponer como una combinación de sus partes, que serán las bases representativas de cada una de las temáticas o categorías.

El problema se fundamenta a partir de la representación de la matriz de datos (que contiene el número de ocurrencias de cada una de las palabras del diccionario) como el producto de una matriz de bases y otra de codificación. Ello arroja como ventaja fundamental un ahorro de información a almacenar si el número de bases elegido es relativamente pequeño.

Dado que en principio se desconoce la naturaleza de los textos a clasificar, se diseñó una base de datos propia a partir de documentos relativos a distintas materias e intentado en la medida de lo posible que tuvieran diferentes extensiones con el objeto de asemejarla en la medida de lo posible a un ejemplo real. Este hecho provocó que en el conjunto de temáticas obtenidas en la factorización existieran categorías *dominantes* (cuentan con más de una base propia) y *excluidas* (carecen de base propia). Aún así, los porcentajes de acierto en la clasificación de la aplicación se mantienen en la línea de los obtenidos por diferentes autores en sus experimentos.