UNIVERSIDAD DE GRANADA

Máster Universitario en Estadística Aplicada

TRABAJO FIN DE MÁSTER

Análisis de datos de proximidad para exploración y clasificación de textos

J. David Fernández Romero

Granada, Septiembre de 2021

Índice general

Resumen

El objetivo de este trabajo es mostrar la aplicación del análisis de datos de proximidad al análisis estadístico de textos.

El trabajo se encuentra dividido en 4 capítulos. En el primer capítulo se definen los conceptos mas importantes en el campo de la estadística textual, se realiza una breve revisión de su desarrollo histórico, se muestran sus principales aplicaciones y se realiza una pequeña introducción sobre las técnicas estadísticas de análisis multivariante.

El segundo capítulo se dedica a una etapa propia de la estadística de variables textuales: el necesario tratamiento previo de estos datos para obtener un conjunto estructurado sobre el que poder aplicar las correspondientes técnicas estadísticas, describiendo los diversos tratamientos aplicables para codificar el corpus lingüístico en tablas numéricas.

El tercer capítulo trata acerca de las medidas de proximidad para datos multivariantes. Se definen los conceptos de distancia y similaridad, se introducen algunas de las métricas mas importantes y se describe la medida de similaridad mas utilizada para comparar textos: el coseno del ángulo entre dos vectores.

En el capítulo 4 se describen algunas técnicas estadísticas exploratorias para datos multivariantes:

- El análisis de componentes principales, técnica de aplicación muy generalizada por su potencia para abordar la reducción de la dimensionalidad y permitir la detección de variables latentes.
- El escalado multidimensional, técnica de aplicación específica sobre datos de proximidad que permite abordar el análisis desde el punto de vista de la similaridad.
- El análisis de correspondencias, técnica históricamente ligada al análisis estadístico de textos.

El capítulo 5 se dedica al problema de la clasificación mediante metodología del análisis discriminante con una breve descripción inicial del problema de clasificación o discriminación y una parte de desarrollo teórico centrada en el análisis discriminante clásico desarrollado por Fisher.

Por último, en el capítulo 6 se muestra la aplicación de los distintos conceptos y técnicas estadísticas expuestos sobre un corpus lingüístico real realizando un análisis exploratorio y resolviendo un problema de clasificación.

Abstract

The aim of this paper is to show the application of proximity data analysis to the statistical analysis of texts.

The paper is divided into 4 chapters. In the first chapter, the most important concepts in the field of textual statistics are defined, a brief review of its historical development is made, its main applications are shown and a brief introduction is made on the statistical techniques of multivariate analysis.

The second chapter is devoted to a stage typical of the statistics of textual variables: the necessary preliminary treatment of these data to obtain a structured set on which to apply the corresponding statistical techniques, describing the various treatments applicable to codify the linguistic corpus in numerical tables.

The third chapter deals with proximity measures for multivariate data. The concepts of distance and similarity are defined, some of the most important metrics are introduced and the most commonly used similarity measure for comparing texts is described: the cosine of the angle between two vectors.

Chapter 4 describes some exploratory statistical techniques for multivariate data:

- Principal component analysis, a technique of widespread application because of its power to address dimensionality reduction and to allow the detection of latent variables.
- Multidimensional scaling, a technique of specific application on proximity data that allows to approach the analysis from the point of view of similarity.
- Correspondence analysis, a technique historically linked to the statistical analysis of texts.

Chapter 5 is devoted to the problem of classification using the methodology of discriminant analysis with a brief initial description of the classification or discrimination problem and a theoretical development focused on the classical discriminant analysis developed by Fisher.

Finally, chapter 6 shows the application of the different concepts and statistical techniques presented on a real linguistic corpus by performing an exploratory analysis and solving a classification problem.

Capítulo 1

El nálisis estadístico de textos. Origen y aplicaciones.

Se suele definir la minería de datos como el proceso que, utilizando técnicas estadísticas y de las ciencias de la computación, pretende descubrir patrones no triviales de información desconocida en conjuntos de datos estructurados. En la actualidad este concepto se encuentra íntimamente ligado al de Big Data debido al desarrollo experimentado en las últimas décadas en la capacidad de almacenamiento y procesamiento de las computadoras que ha ha permitido la aplicación de determinadas técnicas estadísticas a volúmenes de datos que hasta hace poco resultaban imposibles de abordar.

Por su parte, se conoce como minería de textos al proceso de extracción, a partir de textos no estructurados, de patrones no triviales que proporcionan nueva información y conocimientos.

Podríamos entonces decir que la principal diferencia entre la minería de textos y la minería de datos es que la primera se aplica sobre información no estructurada mientras que la segunda lo hace sobre información estructurada. En realidad la minería de textos necesita una fase previa de procesamiento de los datos textuales (no estructurados) que los transforma en un conjunto de datos estructurados sobre los que se aplican las técnicas estadísticas que dan sustento a la minería de datos.

Los importantes avances en tecnología, tanto a nivel de hardware como de software, han propiciado un fuerte desarrollo de las técnicas de minería de datos, incidiendo especialmente en el caso de datos de origen textual.

Paralelamente al desarrollo tecnológico, la implantación de Internet ha resultado en la disponibilidad de una gran cantidad de contenido textual, de carácter muy diverso y fácilmente almacenable y procesable, lo que ha contribuido también al desarrollo de métodos y algoritmos para la detección de patrones no triviales de interés en los datos textuales.

Además, el hecho de que la forma mas común de almacenamiento de información sea en forma de texto ha provocado que la minería de textos se haya convertido en una de las áreas con mayor potencial de la minería de datos.

Sin embargo, la minería de textos es mucho más compleja que la minería de datos, fundamentalmente por el hecho de que esta última trata conjuntos de datos estructurados mientras que la primera lo hace con datos inicialmente desestructurados y en consecuencia mas confusos.

Una de las características más destacada de los datos textuales es su dispersión y alta dimensionalidad. Un corpus textual se puede representar como una matriz $n\ x\ d$, donde n es el número de documentos que componen el corpus y d es el número de términos distintos que aparecen en el conjunto de documentos. Así, el elemento (i, j) de la matriz representará la frecuencia normalizada del j-ésimo término en el i-ésimo documento. Esta representación numérica de los corpus textuales da lugar a matrices huecas y de muy elevada dimensionalidad, lo que condiciona en gran medida la aplicación de determinadas técnicas estadísticas.

1.1. Origen y desarrollo de la estadística textual.

El origen de la analítica de textos se remonta a la antigua Alejandría, cuando los gramáticos elaboraron el inventario de palabras de la Biblia, así como listados de los hapax de Homero. Este primitivo análisis textual se centraba fundamentalmente en el tipo y volumen del vocabulario empleado, sin acometer el análisis del contenido de los documentos.

Durante las primeras décadas del siglo XX, los lingüistas anglosajones abordaron el estudio de las concordancias de determinados vocablos en grandes autores literarios. Se trataba de inferir el contexto común alrededor de cada palabra.

A partir de la década de 1930, Zipf, Yule y Guiraud entre otros, realizaron importantes aportes teóricos al análisis estadístico de datos textuales, introduciendo leyes empíricas sobre la distribución de las palabras y resolviendo así algunos problemas planteados por los estilistas franceses.

Jean-Paul Benzécri, gran estadístico francés considerado el padre de la escuela francesa de análisis de datos, publicó en 1964 un curso de lingüística matemática que venía impartiendo en la Facultad de Ciencias de Rennes desde 1960. En él, Benzécri plantea un nuevo método de análisis descriptivo multivariante: la técnica del Análisis Factorial de Correspondencias. Al año siguiente y en esa misma facultad, Escofier defendía su tesis doctoral resaltando las principales propiedades de este método.

Con el Análisis de Correspondencias, Benzécri aporta un método estadístico para solucionar los problemas fundamentales de los lingüistas. El Análisis de Correspondencias puede ser aplicado en muchos campos, pero debemos tener en cuenta que, en comparación, el tratamiento de datos lingüísticos presenta particularidades propias debidas a la multidimensionalidad intrínseca de esta materia.

Las primeras aproximaciones a la minería de textos en el sentido ya de analizar la información contenida en los documentos tiene su origen en las tareas de catalogación de documentos que se extendieron rápidamente a la extracción de información gracias al desarrollo de técnicas de procesamiento del lenguaje natural, en el que tuvo y sigue teniendo gran importancia la necesidad de extraer información de los documentos de texto de manera automatizada. De esta forma, el análisis estadístico de textos y el procesamiento del lenguaje natural se han convertido en áreas complementarias y muy interrelacionadas, cuyas investigaciones se retroalimentan permanentemente.

En la segunda mitad del siglo XX el desarrollo de programas informáticos corre paralelo al de los métodos y aplicaciones a grandes conjuntos de datos. L. Lebart y A. Morineau desarrollan, en 1984, un módulo de tratamiento de textos en el sistema SPAD. Posteriormente, en 1988, M. Bécue Bertaut presenta su tesis doctoral titulada "Un sistema informático para el Análisis de Datos Textuales", en la que desarrolla el programa SPAD.T. A partir de entonces se realizan grandes progresos en el análisis de respuestas libres a cuestiones abiertas y su relación con el resto de variables informadas en las encuestas.

En estas últimas décadas del siglo XX la minería de textos buscaba automatizar, en cierta medida, el acceso a información concreta entre una gran cantidad de documentos de texto, por lo que las técnicas desarrolladas buscaban optimizar métodos para resumir los documentos manteniendo la información mas relevante de manera que se consiguiera disminuir el tamaño de los textos a tratar facilitando con ello las tareas de búsqueda, catalogación y clasificación.

En la actualidad, debido al desarrollo de la informática y las telecomunicaciones (especialmente Internet, la inteligencia artificial y el big data), son muchas las posibilidades de profundizar en el análisis de textos desde el punto de vista de su contenido y no únicamente del vocabulario. En este punto, es importante resaltar la diferencia entre el análisis textual como concepto general y el análisis estadístico de textos o simplemente análisis de datos textuales que consiste en la aplicación a la lingüística de técnicas propias de la estadística.

1.2. Aplicaciones mas importantes.

En la actualidad son muchas las aplicaciones de la estadística textual siendo uno de sus objetivos principales el descubrimiento y análisis de patrones de interés.

Entre las aplicaciones mas habituales podemos citar el análisis de sentimientos, el análisis de las respuestas a preguntas abiertas en encuestas o el filtrado de correos electrónicos.

El objetivo del análisis de sentimientos es, a partir de una opinión expresada en forma textual, deducir la actitud del individuo hacia el objeto de dicha opinión. Es una técnica de gran utilidad para personas o empresas cuyos productos o actividades tienen una dependencia importante de su proyección social por lo que ha experimentado un fuerte desarrollo e incrementado en gran medida su presencia en paralelo al auge de las redes sociales siendo uno de los campos de la minería de textos mas relacionados y que mas se beneficia de las técnicas de procesamiento del lenguaje natural.

La utilización de preguntas de respuesta libre es muy frecuente destacando su empleo habitual en encuestas de satisfacción de clientes, encuestas de opinión y estudios de preferencia en artículos de consumo. Estas preguntas permiten captar información que no es posible obtener mediante preguntas cerradas y cuentan con algunas ventajas sobre estas últimas como el hecho de un menor condicionamiento del encuestado y mayor fiabilidad acerca de sus opiniones reales, así como la eliminación de las limitaciones que conlleva el diseño de preguntas cerradas en las que demasiadas veces se incluye una categoría "otros" que no aporta información válida sobre la verdadera respuesta del encuestado mas allá de no coincidir con ninguna de las opciones tenidas en cuenta en el diseño de las preguntas. Las técnicas de minería de textos permiten extraer la información contenida en las respuestas abiertas operando sobre las respuestas originales en los términos en que han sido expresadas por el encuestado al tiempo que permiten aumentar la calidad de la interpretación de las respuestas a preguntas cerradas en la misma encuesta.

La potencia de las técnicas estadísticas aplicadas al análisis de textos ha posibilitado, por ejemplo, el desarrollo de herramientas informáticas que permiten automatizar la organización del correo electrónico en carpetas y específicamente la clasificación o no como spam de los correos recibidos por los usuarios. Para ello, se han construido clasificadores que categorizan en spam o no spam el correo entrante mediante la extracción de cada correo entrante de los términos más relevantes (en el sentido de contar con una menor probabilidad de constituir spam).

1.3. Técnicas estadísticas multivariantes.

Las técnicas de análisis estadístico multivariante son herramientas de uso común en muchas disciplinas: desde la psicología hasta la inteligencia artificial pasando por la sociología, la economía, la ingeniería, la medicina o las ciencias ambientales. En la actualidad, muchas de estas técnicas soportan los procesos de extracción de conocimiento que se encuentran detrás de los métodos conocidos como minería de datos y constituyen la base de las técnicas de inteligencia artificial.

El análisis de datos multivariantes comprende el estudio estadístico de varias variables medidas sobre elementos de una población y puede plantearse a dos niveles:

- Extraer la información contenida en los datos, para lo que se utilizan métodos exploratorios que extienden al caso multivariante las técnicas estadísticas descriptivas habituales para resumir los valores de las variables y describir su estructura de dependencia, así como realizar representaciones gráficas y elegir transformaciones de las variables que simplifiquen su descripción.
- Obtener conclusiones sobre la población que ha generado los datos, para lo que es preciso construir un modelo que explique dicha generación y permita realizar predicciones sobre datos futuros. Se utilizan para ello métodos inferenciales con los que se pretende generar conocimiento sobre el problema que subyace en los datos disponibles.

El objetivo último de las técnicas multivariantes exploratorias puede consistir en uno o varios de los siguientes:

- Resumir los datos en un conjunto de nuevas variables que resulten de aplicar determinadas transformaciones a las variables originales
- Identificar, si existen, grupos homogéneos de individuos o variables
- Clasificar nuevas observaciones en grupos predefinidos
- Relacionar conjuntos de variables entre sí

Transformar las variables originales posibilita simplificar la descripción de los datos reduciendo su dimensionalidad. En el caso mas extremo, reducir la dimensionalidad de un conjunto de datos multivariante a dos únicas variables indicadoras permite la representación bidimensional de los individuos y con ello la visualización e interpretación de las relaciones entre ellos.

Entre los métodos exploratorios multivariantes orientados a la obtención de nuevas variables indicadoras que sinteticen la información de las variables originales destaca, cuando los datos son continuos, el análisis de componentes principales. Esta técnica permite determinar las dimensiones necesarias para representar adecuadamente los datos. Cuando se cuenta con información sobre similitudes o semejanzas entre los individuos e interesa encontrar las dimensiones de dichas similitudes el concepto de componentes principales se generaliza en las técnicas de escalado multidimensional, mientras que en el caso de los datos textuales se ha utilizado tradicionalmente otra generalización de dicho concepto, en este caso dirigida al análisis de datos cualitativos, denominada análisis de correspondencias.

Por otra parte, el análisis de las similitudes entre los individuos permite encontrar grupos homogéneos cuando el sentido o significado de la similaridad es, a priori, desconocido, aunque en muchos casos las motivaciones o causas de esta similitud puedan permanecer ocultas a la interpretación del analista. Para estudiar si los datos forman o no un grupo homogéneo y, si existen varios grupos, identificar los elementos que pertenecen a cada uno de ellos, se utilizan los métodos del análisis de conglomerados o análisis cluster, conocidos también como métodos de clasificación automática o no supervisada que permiten, a través del agrupamiento de los datos, abordar distintos objetivos:

- Cuando se conoce o sospecha que los datos son heterogéneos así como el número de grupos en que se podrían dividir: particionar los datos en un número de grupos preestablecido de forma que cada elemento pertenezca a uno y sólo uno de los grupos, todos los elementos pertenezcan a algún grupo y cada grupo sea internamente homogéneo.
- Cuando no existe información alguna acerca del grado de heterogeneidad de los datos: obtener una estructura de los elementos ordenada en niveles de forma jerárquica en función de su grado de similitud, de manera que los niveles superiores contienen a los inferiores, lo que permite al analista particionar los datos a posteriori en el número de grupos que considere mas apropiado.
- Cuando se abordan problemas con un elevado número de variables y resulta de interés realizar un estudio exploratorio previo para dividir las variables en grupos: clasificar las variables, lo que a su vez puede facilitar el planteamiento posterior de modelos de reducción de dimensionalidad.

La obtención de modelos para generar conocimiento sobre la población de la que provienen los datos se puede abordar también mediante la reducción del número de variables con la metodología del análisis factorial que, como en el caso del análisis de correspondencias y el escalado multidimensional, se puede interpretar como una generalización de las componentes principales. Se trata de reemplazar un determinado conjunto de variables por un número menor de factores o variables latentes, no observables, que permitan predecir los valores de las variables originales.

El análisis de la heterogeneidad de la población desde el punto de vista inferencial se realiza mediante lo que se conoce como técnicas de clasificación supervisada. El calificativo de supervisada hace referencia a que se parte de una muestra de elementos cuya clasificación es conocida a partir de los que se obtendrá el modelo para la clasificación de futuras observaciones. Para ello se pueden utilizar distintos métodos estadísticos multivariantes cuya elección dependerá de las características concretas del problema a resolver. Por ejemplo, si todas las variables son continuas, es frecuente aplicar el análisis discriminante clásico de Fisher, puesto que resulta óptimo bajo el supuesto de normalidad multivariante y, aunque los datos originales no sean normales es posible aplicar una transformación que para obtener normalidad. Si no todas las variables son continuas el problema de clasificación se abordará con otros métodos de clasificación, como las basadas en modelos de respuesta cualitativa, árboles de clasificación, máquinas de los vectores soporte, redes neuronales, etc.

Capítulo 2

Tratamiento previo de los datos textuales.

La variable textual, expresada en tablas de recuentos, implica una mayor complejidad en su forma que las variables cuantitativas o cualitativas "puras". Sin embargo, esta complejidad junto a la dificultad de tratamiento de este tipo de variables aporta un mayor apego a la realidad de los resultados obtenidos.

2.1. Preprocesamiento del corpus y segmentación en unidades léxicas.

Llamamos *corpus* al conjunto de textos a analizar. Puede tratarse de artículos de opinión, relatos de diversos autores o épocas, comentarios en una red social, respuestas libres a una pregunta abierta en una encuesta, etc.

Es fundamental realizar un cuidadoso procesamiento previo del corpus inicial para una correcta identificación de las unidades léxicas cuyas ocurrencias se van a contar. Este preprocesamiento deberá estar formado por una serie de reglas bien definidas que aporten estabilidad, facilidad de comprensión y reproducibilidad. Es habitual seguir la norma lexicométrica desarrollada por Muller en 1977 y posteriormente completada por Labbé en 1990, que contempla:

- Utilizar un corrector ortográfico automático potente que tenga en cuenta las reglas gramaticales.

- Realizar una limpieza de notaciones normativas (eliminando por ejemplo, las mayúsculas al inicio de frases o las abreviaciones ambiguas). Se trata, en definitiva, de dotar de un estatus único a cada carácter del texto (por ejemplo, el punto que indica el fin de una frase es diferente del punto que separa ciertas abreviaciones como D.N.I. o N.I.E.).

- Definir los signos considerados de puntuación, de manera que todos los demás signos son tratados como parte del conjunto de las letras.

- Si se considera necesario, *lematizar* el corpus, es decir, transformar cada palabra en la *entrada del diccionario* a la que se asocia.

- Definir lo que se conoce como *stoplist* (o conjunto de *stopwords*), que no es mas que la lista de palabras que se eliminan del estudio por considerar que no aportan información (resulta habitual eliminar las preposiciones, artículos y conjunciones).

- En estudios *comparativos*, tales como en el análisis de emociones o el análisis de encuestas con preguntas abiertas, es común establecer un *umbral de frecuencia*, ya que se considera que la comparación sólo tendrá sentido entre palabras que se utilicen con al menos una determinada frecuencia.

2.1.1. Lematización.

La principal dificultad al acometer el preproceso de todo corpus es la definición de la unidad léxica, o *unidad de segmentación* del corpus, puesto que será la base del análisis estadístico que vamos a realizar.

Podríamos adoptar la *palabra*, en su concepción lexicográfica como secuencia de letras delimitada a izquierda y derecha por espacio en blanco o signo de puntuación, (también denominada forma gráfica), pero así no obtendríamos una unidad léxica claramente determinada, por lo que se suele optar por el *lema*[1].

En el caso del idioma *castellano* la lematización de un texto requiere convertir:

- las flexiones verbales al infinitivo

- los sustantivos a singular

- los adjetivos al masculino singular

Aun cuando se utilice un analizador morfo-sintáctico de alta calidad para automatizar el proceso de lematización, pueden subsistir ambigüedades únicamente solventables mediante una operación manual.

La lematización del corpus previa a su análisis estadístico proporciona las siguientes ventajas:

- Se reduce la variabilidad entre respuestas: las frases "*me gusta ver series*" y "*por la noche veo una serie*" tienen en común los lemas "*ver*" y "*serie*", pero en su forma original no comparten ninguna forma gráfica

- Se limita la pérdida de unidades textuales, puesto que la frecuencia de cada lema considerado será la suma de las frecuencias de las formas gráficas reagrupadas en el lema: en un texto en el que aparezca 18 veces la forma gráfica *permanencia*, 1 vez *permanecen*, 3 veces *permaneciendo* y 5 veces *permaneceremos*, el lema *permanecer* tendrá una frecuencia de 27 ocurrencias, cuando ninguna de las formas gráficas alcanza este umbral

[1] *lema*: entrada del diccionario asociada a una palabra.

- Se asocia a cada lema (o forma gráfica) su categoría sintáctica, aspecto fundamental en diversos momentos del tratamiento, puesto que hace posible seleccionar palabras por su categoría o utilizar la categoría para *apoyar* la interpretación de los resultados

Es aconsejable tomar en consideración ambos tipos de unidad léxica (palabra o forma gráfica, y lema) repitiendo incluso las distintas fases del tratamiento en cada caso puesto que los resultados obtenidos se enriquecerán mutuamente.

2.1.2. Stoplist.

En determinados casos se separan las unidades léxicas consideradas en dos grupos: las consideradas *llenas* (que aportan significado por sí solas, como los sustantivos y los verbos, y a veces pero no siempre, los adjetivos y los adverbios) y las consideradas *herramientas* (o gramaticales, no consideradas informativas, tales como los artículos, las preposiciones y las conjunciones). En algunas aplicaciones las unidades léxicas gramaticales se consideran no útiles.

El conjunto de palabras consideradas no útiles se denomina *stoplist*, pero debe tenerse en cuenta que en determinados casos, como el caso de las respuestas libres de encuestas, los adverbios, las negaciones o los adjetivos, no deben eliminarse del tratamiento puesto que pueden ser extremadamente importantes.

2.1.3. *Stematización* y/o reagrupación de sinónimos.

Puede resultar de utilidad realizar una *stematización*[2], consistente en el reagrupamiento de varios lemas provenientes de una misma raíz.

También, en determinados estudios, puede ser apropiada la unificación de sinónimos.

Debemos tener en cuenta que estos reagrupamientos deben realizarse con posterioridad a un primer tratamiento para garantizar que no se distorsionen los resultados.

2.1.4. Umbral de frecuencia.

El último concepto importante relacionado con el preproceso del corpus es el que hace referencia a la determinación de una frecuencia mínima de presencia en el corpus para que la unidad léxica sea considerada en el estudio. A veces se sustituye por el establecimiento *a priori* del número de palabras a incluir.

2.2. Codificación del texto y tablas léxicas.

Una vez realizada la fase de preproceso mediante la segmentación del corpus en unidades léxicas y, en su caso, operadas la eliminación de las incluidas en la stoplist y las reagrupaciones necesarias, obtendremos los *glosarios*, tanto de unidades léxicas como de segmentos repetidos.

[2] *stematización*: neologismo formado a partir de *stem*, que en inglés significa raíz.

Las tablas generadas permiten estudiar la distribución de las palabras entre individuos[3] (*tabla léxica*: cruzando individuos y palabras) o entre categorías de individuos (*tabla léxica agregada*: cruzando categorías de individuos y palabras). En adelante, para una mejor adecuación de la terminología utilizada al caso de aplicación, se utilizará el término *documento* como equivalente a *individuos*.

2.2.1. Tabla léxica *Documentos* x *Palabras*

Para acometer el estudio de la distribución del vocabulario entre los documentos se codifica la variable textual en una tabla de frecuencias *Documentos* x *Palabras* que contiene el número de veces que en cada documento aparece cada una de las palabras.

2.2.2. Tabla léxica agregada *Categorías* x *Palabras*

En determinadas ocasiones puede resultar de interés la comparación de la distribución de palabras entre categorías de documentos, en función de una variable categórica. En este caso, para cada categoría se cuenta el número de veces que cada palabra aparece en los documentos de la categoría, es decir, se agregan los documentos por categorías, obteniendo una tabla léxica agregada *Categorías* x *Palabras*.

2.3. Sobre la dimensionalidad.

En la mayor parte de los casos el tratamiento previo del corpus textual da lugar a matrices de términos-documentos de una elevada dimensionalidad, lo que tradicionalmente ha constituido un serio inconveniente para el tratamiento de estos datos. Aunque la propia etapa de preprocesado incorpora técnicas para abordar este problema, ha sido principalmente el avance tecnológico de las últimas décadas lo que en mayor medida ha facilitado este tratamiento. En cualquier caso, el problema de la dimensionalidad sigue constituyendo un reto y su reducción es un objetivo principal en la mejora de los algoritmos existentes y en el desarrollo de nuevas metodologías.

[3]Entiéndase individuo en sentido amplio en el contexto adecuado, por ejemplo, en el estudio de encuestas con respuesta abierta será cada una de las respuestas individuales, en el caso de un corpus formado por varios documentos, cada uno de los documentos individuales, etc

Capítulo 3

Medidas de proximidad para datos multivariantes.

3.1. Distancias y similaridades.

A la hora de abordar el estudio de un conjunto de datos multivariantes, típicamente formado por n individuos sobre los que se han obtenido valores de p variables, es de interés poder establecer el grado de *similitud* entre diferentes individuos de cara, por ejemplo, a realizar agrupaciones o facilitar su representación gráfica y la interpretación de sus relaciones.

En estadística multivariante es habitual distinguir entre medidas de asociación para individuos y para variables aunque técnicamente estas medidas son válidas tanto para uno como para otras.

Existen muchas medidas de asociación multivariante. Entre las mas conocidas podemos citar el coseno del ángulo entre vectores, el coeficiente de correlación, la distancia euclídea, la distancia de Mahalanobis o distancias basadas en el estadístico χ^2.

Cada medida diferente refleja la asociación entre dos variables (o individuos) en un sentido particular por lo que es de gran importancia elegir, en cada situación, la medida apropiada para el problema concreto de que se trate.

En el caso del análisis de variables textuales es generalizado cuantificar el grado de similaridad entre dos textos mediante el coseno del ángulo que forman los vectores que los representan.

Se denomina *distancia* o *métrica* entre dos puntos, x_i y x_j, pertenecientes a $\mathbb{R}^p$ a la función $d : \mathbb{R}^p$ x $\mathbb{R}^p \to \mathbb{R}^+$ que verifica las siguientes propiedades:

- $d(x_i, x_j) \geq 0$
- $d(x_i, x_i) = 0 \ \ \forall i$
- $d(x_i, x_j) = d(x_j, x_i)$
- $d(x_i, x_j) \leq d(x_i, x_p) + d(x_p, x_j), \ \forall \ x_j \in \ \mathbb{R}^p$

De forma similar, se puede definir la *similaridad* entre dos puntos, x_i y x_j, pertenecientes a $\mathbb{R}^p$ como la función $d : \mathbb{R}^p$ x $\mathbb{R}^p \to \mathbb{R}^+$ que verifica las siguientes propiedades:

- $s(x_i, x_j) \leq s_0$
- $s(x_i, x_i) = s_0 \forall i$
- $s(x_i, x_j) = s_0 \Rightarrow x_i = x_j$
- $s(x_i, x_j) = s(x_j, x_i)$
- $|s(x_i, x_p) + s(x_p, x_j)|s(x_i, x_j) \geq s(x_i, x_p)s(x_p, x_j),\ \forall\ x_p \in\ \mathbb{R}^p$

3.2. Las métricas de Minkowski y la distancia euclídea.

Unas distancias empleadas muy frecuentemente en análisis multivariante son las denominadas distancias o métricas de *Minkowski*, definidas en función de un parámetro r de la siguiente forma:

$$d_{ij}^{(r)} = \left[\sum_{s=1}^{p}(x_{is} - x_{js})^r\right]^{1/r}$$

Es inmediato que la distancia de *Minkowski* de parámetro $r = 2$ es la distancia euclídea, que es la mas utilizada pero tiene el inconveniente de que depende de las unidades de medida de las variables, por lo que en su lugar se suelen utilizar las denominadas *métricas euclídeas ponderadas* en las que se divide cada variable por un término que elimine el efecto de la escala. Es decir

$$d_{ij} = [(x_i - x_j)'M^{-1}(x_i - x_j)]^{1/2}$$

siendo M una matriz diagonal empleada para estandarizar las variables y conseguir que la medida sea invariante ante cambios de escala.

La matriz M puede no ser diagonal pero siempre deberá ser singular y definida positiva de forma que se verifique la condición $d_{ij} \geq 0$.

De nuevo, en el caso $M = I$ tendremos la distancia euclídea.

3.2.1. La distancia de Mahalanobis.

Esta distancia se obtiene al tomar $M = S$ y definir la distancia entre un punto y su vector de medias:

$$d_i = [(x_i - \bar{x})'S^{-1}(x_i - \bar{x})]^{1/2}$$

Es frecuente referirse al valor de esta distancia al cuadrado, d_i^2, con la misma denominación: distancia de Mahalanobis y así se hace a lo largo de este trabajo en el apartado dedicado al *Análisis discriminante*.

3.3. El coseno del ángulo entre vectores.

Si $x_i = (x_{i1}, \ldots, x_{ip})'$ y $x_j = (x_{j1}, \ldots, x_{jp})'$ son dos vectores p-dimensionales, su producto escalar, o suma de sus productos cruzados, es $x_i'x_j = \sum_{s=1}^{p} x_{is}x_{js}$, y $||x_i||^2 = \sum_{s=1}^{p} x_{is}^2 = \sum_{s=1}^{p} x_{is}x_{is}$ es su norma al cuadrado o suma de cuadrados. Entonces, si β es el ángulo formado por los vectores x_i y x_j, su producto escalar se puede expresar como:

$$x_i'x_j = ||x_i||||x_j||cos(\beta)$$

de donde

$$cos(\beta) = \frac{x_i'x_j}{||x_i||||x_j||} = \frac{\sum_{s=1}^{p} x_{is}x_{js}}{\sqrt{\sum_{s=1}^{p} x_{is}^2 \sum_{s=1}^{p} x_{js}^2}}$$

Es fácil comprobar que cuando los vectores se encuentran centrados respecto de su media esta medida coincide con el coeficiente de correlación de Pearson.

El coseno del ángulo formado por dos vectores es por tanto una medida de similaridad entre los mismos que toma valores entre -1 y 1 siendo, además, la mejor medida para establecer el paralelismo entre dos vectores, al ser en este caso 1 en valor absoluto. Además, el coseno es invariante ante homotecias excepto un eventual cambio de signo y, por tanto, independiente de la longitud de los vectores considerados. Todas estas características hacen que sea una medida muy utilizada en el análisis textual a la hora de cuantificar el grado de similitud entre dos textos.

Capítulo 4

Técnicas exploratorias para datos multivariantes.

4.1. Consideraciones sobre el análisis de documentos de texto.

Antes de aplicar cualesquiera técnicas estadísticas multivariantes sobre una matriz de Documentos x Palabras es preciso realizar algunas operaciones adicionales.

Debe tenerse en cuenta que cada fila de la matriz representa un documento y, generalmente, los distintos documentos no contienen el mismo número de palabras, por lo que en un documento muy extenso la presencia varias veces de una misma palabra puede tener menos importancia relativa (en términos de información aportada acerca del documento) que la presencia de esa misma palabra un número menor de veces en un documento menos extenso. En definitiva, para eliminar la influencia de la extensión de los documentos es necesario convertir las frecuencias absolutas de la matriz original en frecuencias relativas por documentos, es decir, por filas.

4.2. Análisis de componentes principales.

Dadas n observaciones de p variables, el análisis de componentes principales persigue representar adecuadamente esta información, con la mínima pérdida de información, mediante un número menor de variables construidas como combinaciones lineales de las originales. La aplicación de esta técnica permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos. Al mismo tiempo, la representación de las observaciones en un espacio de menor dimensión propicia la identificación de las posibles variables latentes o no observadas que generan los datos.

En otras palabras podemos decir que el análisis de componentes principales consiste en encontrar un subespacio de dimensión menor que p de forma que la proyección sobre él de los n puntos conserve su estructura con la menor distorsión posible. Esta condición de distorsión mínima equivale a exigir que las distancias entre los puntos originales y sus proyecciones en el subespacio obtenido sean lo más pequeñas posible.

Así, dado un elemento x_i y una dirección definida por un vector de norma unidad, $a_1 = (a_{11}, \ldots, a_{1p})'$, la proyección de x_i sobre esta dirección será:

$$z_i = a_{11}x_{i1} + \cdots + a_{1p}x_{ip} = a_1'x_i$$

y el vector que representa esta dirección será $z_i a_1$. Si r_i es la distancia entre x_i y su proyección sobre a_1, entonces

$$minimizar(\sum_{i=1}^{n} r_i^2) = \sum_{i=1}^{n} |x_i - z_i a_1^2|$$

siendo $|u|$ la norma euclídea o módulo de u.

Ahora bien,

$$x_i'x_i = z_i^2 + r_i^2 \Rightarrow \sum_{i=1}^{n} x_i'x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2 \Rightarrow minimizar \sum_{i=1}^{n} r_i^2 = maximizar \sum_{i=1}^{n} z_i^2$$

y, puesto que las proyecciones z_i son variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza y, en definitiva, el criterio de minimizar la distorsión equivale a encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

4.2.1. Obtención de las componentes principales.

Supongamos ahora que la matriz X de dimensiones n x p contiene los valores de p variables observadas sobre n elementos, y supongamos que los valores de cada variable se encuentran centrados respecto de su media, de forma que las variables de la matriz X tienen media cero y la matriz de varianzas y covarianzas de X es $S = \frac{1}{n}X'X$.

La primera componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima y sus valores para los n elementos se representarán por el vector $z_1 = Xa_1$, que tendrá media cero puesto que las variables originales tienen media cero, y cuya varianza será:

$$Var(z_1) = \frac{1}{n}z_1'z_1 = \frac{1}{n}a_1'X'Xa_1 = a_1'Sa_1$$

Para que la maximización de la expresión anterior tenga solución se impone la restricción $a_1'a_1 = 1$ que se introduce mediante el multiplicador de Lagrange:

$$M = a_1'Sa_1 - \lambda(a_1'a_1 - 1)$$

Esta expresión se maximiza de la forma habitual derivando respecto de los componentes de a_1 e igualando a cero obteniéndose que

$$Sa_1 = \lambda a_1$$

lo que implica que a_1 es un vector propio de S y λ su correspondiente valor propio.

Ahora bien,

$$Sa_1 = \lambda a_1 \Rightarrow a_1' S a_1 = \lambda a_1' a1 = \lambda \Rightarrow \lambda = Var(z_1)$$

y, puesto que buscamos maximizar $Var(z_1)$, λ será el mayor valor propio de la matriz S y su vector asociado, a_1, definirá los coeficientes de las variables originales en la primera componente principal.

La segunda componente principal proporciona el mejor plano de proyección de X y se calcula estableciendo como función objetivo que la suma de las varianzas de las dos primeras componentes principales sea máxima, es decir, siendo a_1 y a_2 los vectores que definen el plano, la función objetivo será:

$$\phi = a_1' S a_1 + a_2' S a_2 - \lambda_1(a_1' a_1 - 1) - \lambda_2(a_2' a_2 - 1)$$

con las restricciones de que las direcciones tengan módulo unidad, $a_i' a_i = 1,\ i = 1, 2$. De nuevo, derivando e igualando a cero obtenemos la solución:

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

que indica que a_1 y a_2 son vectores propios de S. Tomando los vectores propios de norma uno y sustituyendo en la función objetivo se obtiene que, en el máximo, su valor es

$$\phi = \lambda_1 + \lambda_2$$

y por tanto, λ_1 y λ_2 deben ser los dos autovalores mayores de S, y a_1 y a_2 sus autovectores.

Nótese que z_1 y z_2 están incorreladas, puesto que:

$$a_1' a2 = 0 \Rightarrow Cov(z_1, z_2) = a_1' S a_2 = 0$$

Análogamente puede demostrarse que el espacio de dimensión r que mejor representa a los n puntos p-dimensionales viene definido por los vectores propios asociados a los r mayores valores propios de S. Estas direcciones se denominan *direcciones principales* de los datos y las nuevas variables que definen, componentes *principales*.

4.2.2. Relación entre las componentes principales y las variables originales.

Si X (y por tanto también S) tiene rango p, existirán tantas componentes principales como variables y se obtendrán calculando los valores propios de S mediante:

$$|S - \lambda I| = 0$$

y sus vectores propios asociados:

$$(S - \lambda_i I) a_i = 0$$

En este caso S será simétrica y definida positiva, y en consecuencia los términos λ_i serán reales y positivos., siendo ortogonales cualesquiera a_i y a_j vectores asociados a las raíces λ_i y λ_j, $\lambda_i \neq \lambda_j$.

Si S es semidefinida positiva de rango $r < p$, es decir $p - r$ variables son combinación lineal del resto, habrá únicamente r valores propios positivos en S y el resto serán nulos.

Si Z es la matriz cuyas columnas contienen los valores obtenidos para las p componentes en los n elementos, las nuevas variables se relacionan con las originales mediante $Z = XA$, con $A'A = I$. Obsérvese que calcular los componentes principales no es mas que calcular una transformación ortogonal A de las variables X para obtener unas nuevas variables Z incorreladas entre sí, lo que puede interpretarse como elegir unos nuevos ejes coordenados que coincidan con los *ejes naturales* de los datos.

4.2.3. Componentes principales estandarizadas.

Es preciso hacer notar que cuando las escalas de medida de las variables son muy distintas, la maximización de su variabilidad dependerá decisivamente de estas escalas de medida y las variables con valores mas grandes tendrán mas peso en el análisis. Así, si una variable tiene una varianza mucho mayor que las demás, la primera componente principal coincidirá muy aproximadamente con esa variable, situación no deseable si es debida a una diferencia en la escala de medida de esta variable respecto de las demás. Asimismo, aun cuando la escala de medida de todas las variables sea la misma, si las variabilidades son muy distintas, al calcular la primera componente principal tendrán mucha mas influencia las variables con varianzas mas elevadas. Para evitar estas situaciones es conveniente estandarizar las variables con carácter previo al cálculo de las componentes principales, transformando la ecuación a maximizar en:

$$M' = 1 + \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j r_{ij}$$

en la que r_{ij} es el coeficiente de correlación lineal entre las variables i y j, de forma que ahora la solución depende de las correlaciones.

Las componentes principales se obtendrán entonces calculando los vectores y valores propios de la matriz de coeficientes de correlación R y se denominarán *componentes principales estandarizadas*.

Por lo tanto, cuando las variables originales están medidas en unidades distintas es conveniente calcular las componentes principales estandarizadas, mientras que si están medidas en la misma unidad se podrá aplicar indistintamente cualquiera de las dos opciones pero, si las diferencias entre las varianzas de las variables son informativas y el analista pretende preservar esta información deberá tenerlas en cuenta y no proceder a la estandarización previa de las variables originales.

4.2.4. Interpretación.

Si existe una alta correlación entre todas las variables, la primera componente principal tendrá todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables, constituyendo un factor global de *forma*. Las demás componentes principales se interpretan como factores de forma y habitualmente incluirán coordenadas positivas y negativas contraponiendo grupos de variables.

4.2.5. Selección del número de componentes.

No existe una regla única para decidir el número de componentes principales idóneo, dependiendo de las características del problema concreto ante las que el analista deberá seguir uno u otro criterio. Las reglas mas utilizadas son las siguientes:

- Buscar un "codo" en el gráfico de λ_i frente a i, es decir localizar en dicho gráfico el punto a partir del que los autovalores λ_i son aproximadamente iguales y elegir un número de componentes que excluya los asociados a λ_i pequeños y aproximadamente de la misma magnitud.
- Seleccionar componentes hasta alcanzar una proporción de varianza determinada (usualmente ≥ 90
- Establecer una cota desechando las componentes asociadas a autovalores inferiores (suele utilizarse la varianza media: $\sum_{i=1}^{n} \frac{\lambda_i}{p}$). En el caso estandarizado, el valor medio de las componentes es 1 y esta regla equivale a seleccionar los valores propios mayores que 1. De nuevo, debe aplicarse con precaución ante la posibilidad de que una de las variables sea independiente de las demás, configurando por si misma una componente principal y podría ser seleccionada, pero si está incorrelada con el resto puede resultar poco relevante.

4.3. Escalado multidimensional.

El escalado multidimensional (MDS) es una técnica que representa las medidas de similaridad (o disimilaridad) entre pares de objetos como distancias entre puntos en un espacio de dimensionalidad menor. De este modo la representación MDS muestra los objetos como puntos en un plano n-dimensional en los que la distancia entre puntos es menor cuanto mas similares son los objetos lo que facilita al analista identificar comportamientos regulares que podrían permanecer ocultos al estudiar las matrices numéricas en su dimensión original.

MDS trabaja sobre una matriz de distancias (o proximidades) $D = [d_{ij}]_{n_x n}$. En el caso general estas distancias serán euclídeas, es decir, supuestas dos observaciones, x_i y x_j en un espacio p-dimensional: $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$.

Estas distancias no se ven alteradas si se centran las variables respecto de su media, es decir:

$$d_{ij}^2 = \sum s = 1^p (x_{is} - x_{js})^2 = \sum_{s=1}^{p} [(x_{is} - \bar{x}_s) - (x_{js} - \bar{x}_s)]^2$$

y por lo tanto no hay pérdida de generalidad en suponer que las variables tienen media cero.

Se trata de encontrar, a partir de la matriz de distancias D, una matriz $\tilde{X}$ n x p con variables de media cero. Su matriz de covarianzas tendrá la forma:

$$S = \frac{\tilde{X}'\tilde{X}}{n}$$

siendo $Q = \tilde{X}\tilde{X}'$ la correspondiente matriz de productos cruzados que puede a su vez interpretarse como una matriz de similitudes entre los n elementos. Sus términos serán de la forma:

$$q_{ij} = \tilde{x}_i'\tilde{x}_j = \sum_{s=1}^{p} x_{is}x_{js}$$

con lo que las distancias pueden ser deducidas inmediatamente a partir de Q:

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js} = q_{ii} + q_{jj} - 2q_{ij}$$

Dado que $\tilde{X}'1 = 0$, también $Q1 = 0$, es decir, $\sum_{i=1}^{n} q_{ij} = 0$ (y por tanto, al ser Q simétrica, $\sum_{j=1}^{n} q_{ij} = 0$). Imponiendo estas restricciones resulta:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + nq_{jj} = t + nq_{jj}$$

$$\sum_{j=1}^{n} d_{ij}^2 = \sum_{j=1}^{n} q_{jj} + nq_{ii} = t + nq_{ii}$$

con $t = traza(Q) = \sum_{i=1}^{n} q_{ii}$, y entonces

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nt \;\Rightarrow\; d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij} = d_{i.}^2 + d_{.j}^2 - d_{..}^2 - 2q_{ij}$$

y por lo tanto

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)$$

donde $d_{i.}^2$, $d_{.j}^2$ y $d_{..}^2$ son, respectivamente, la media por filas, por columnas y total de los elementos de D, es decir:

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$$

$$d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$$

Hemos obtenido entonces la matriz Q de similitud a partir de la matriz de distancias D.

Si suponemos que Q es definida positiva de rango p podemos expresarla como $Q = V\Lambda V'$, donde V contiene los vectores propios que corresponden a los valores propios no nulos de Q y Λ es diagonal y contiene los valores propios de Q. Entonces, podemos escribir:

$$Q = (V\Lambda^{1/2})(\Lambda^{1/2}V')$$

obteniendo una matriz $V\Lambda^{1/2} = Y$, n x p, con p variables incorreladas que reproducen la métrica inicial.

Se debe observar que la matriz obtenida no estará formada por las variables originales sino por sus componentes principales pues existe una indeterminación en el problema al partir únicamente de la matriz de distancias D que es función de la matriz de similitud, Q, y esta es invariante ante rotaciones de las variables originales.

Es frecuente que la matriz de distancias D no sea compatible con una métrica euclídea, pero también lo es que la matriz de similitud Q obtenida a partir de ella tenga p valores propios positivos y mayores que el resto y, si estos restantes $n - p$ valores propios no nulos son mucho menores, podemos obtener una representación aproximada de los puntos originales mediante los p vectores propios asociados a los primeros p valores propios positivos de Q, en cuyo caso las representaciones conservarán la distancia entre los puntos de manera aproximada.

El procedimiento para obtener las *coordenadas principales* de los puntos originales a partir de una matriz de distancias D es el siguiente:

1. Se construye $Q = -\frac{1}{2}PDP$, con $P = I - \frac{1}{n}11'$ (se puede comprobar que en estas condiciones Q es semidefinida positiva y D compatible con una métrica euclídea)
2. Se obtienen los valores propios de Q y se toman los r mayores de forma que los restantes $n - r$ valores sean próximos a 0.
3. Se considera $Q \approx (V_r\Lambda_r^{1/2})(\Lambda_r^{1/2}V_r')$ lo que implica tomar como coordenadas de los puntos originales $Y_r = V_r\Lambda_r^{1/2}$ y por tanto $y_i = v_i\sqrt{\lambda_i}$ donde λ_i es un valor propio de Q y v_i su vector propio asociado.

Además, es posible calcular la precisión obtenida mediante la aproximación realizada a partir de los p valores propios positivos de Q. Por ejemplo, mediante el coeficiente propuesto por *Mardia*:

$$m_{1,p} = 100\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{p}|\lambda_i|}$$

4.3.1. Relación entre coordenadas y componentes principales.

El escalado multidimensional está muy relacionado con el análisis de componentes principales. Ambos persiguen la reducción de la dimensionalidad de los datos originales. En el caso del análisis de componentes principales se obtienen los valores propios de la matriz $X'X$ y se proyectan las variables sobre las direcciones obtenidas para obtener los valores de las componentes principales, mientras que en el caso del escalado multidimensional las coordenadas principales se obtienen directamente como vectores propios de la matriz XX', pero si la matriz de similaridades Q proviene de una métrica euclídea los dos métodos conducen al mismo resultado.

No obstante es necesario precisar que el escalado multidimensional tiene aplicación en una gama de problemas mayor, puesto que siempre es posible obtener las coordenadas principales, aun cuando la matriz de distancias D no provenga exactamente de un conjunto de variables originales, como en el caso del escalado multidimensional *no métrico,* en los que la matriz de partida esta compuesta por las diferencias o *disimilitudes* entre objetos, obtenidas en general por las opiniones de un conjunto de *jueces* o por procedimientos de *ordenación*.

4.3.2. Escalado no métrico.

En estos casos se parte de la premisa de que la matriz de disimilaridades se relaciona con una matriz de distancias de una manera compleja, es decir, las variables explicativas de las similitudes entre los elementos comparados determinan unas distancia euclídeas, d_{ij}, entre los mismos relacionadas con las similitudes, δ_{ij}, mediante una función desconocida:

$$\delta_{ij} = f(d_{ij})$$

imponiendo la única condición de que f sea monótona, es decir, $\delta_{ij} > \delta_{ih} \Leftrightarrow d_{ij} > d_{ih}$, con lo que se pretende encontrar unas coordenadas capaces de reproducir las distancias originales a partir únicamente de la condición de monotonía. Para ello será preciso definir un criterio de bondad del ajuste que sea invariante ante transformaciones monótonas de los datos y un algoritmo para obtener las coordenadas optimizando el criterio anterior.

Aunque para este tipo de problemas no existe solución única y se han propuesto diversos procedimientos, uno de los mas utilizados consiste en minimizar las diferencias entre las distancias derivadas de las coordenadas principales, $\hat{d}_{ij}$, y las similitudes de partida, δ_{ij}, para todos los términos de la matriz, es decir: minimizar $\sum_{i<j}(\delta_{ij} - \hat{d}_{ij})^2$. Al estandarizar esta cantidad se obtiene un criterio de ajuste denominado STRESS, S^2, dado por la expresión:

$$S^2 = \frac{\sum_{i<j}(\delta_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}\delta_{ij}^2}$$

Otro criterio habitual es el conocido como S-STRESS basado en minimizar las distancias al cuadrado, $\hat{d}_{ij}$, que se determinarán obteniendo p coordenadas principales que se emplean como variables implícitas, y_{ij}, $i = 1, \ldots, n, \quad j = 1, \ldots, p$, de forma que las distancias euclídeas entre ellas se expresan en la forma:

$$\hat{d}_{ij}^2 = \sum_{s=1}^{p}(y_{is} - y_{js})^2$$

Se suele considerar como valor inicial de estas variables la solución proporcionada por las *coordenadas principales* y se itera para mejorar la solución minimizando el criterio S^2. El número de dimensiones para obtener una buena representación, p, se suele estimar probando distintos valores y estudiando la evolución de S^2. Una vez fijado p se plantea el problema de minimizar S^2 para las distancias entre las variables y_{ij}. Para ello, se deriva S^2 respecto a cada término y_{ip} y se iguala a cero, obteniendo:

$$\frac{\partial S^2}{\partial y_{ip}} = 2\sum_{j=1}^{n}(\delta_{ij} - \hat{d_{ij}})\frac{\partial \hat{d_{ij}}}{\partial y_{ip}} = 0$$

Dado que $\frac{\partial \hat{d_{ij}}}{\partial y_{ip}} = \frac{(y_{ip}-y_{jp})}{\hat{d_{ij}}}$, sustituyendo en las ecuaciones anteriores llegamos a las ecuaciones:

$$y_{ip}\sum_{j=1}^{n}\frac{(\delta_{ij}-\hat{d}_{ij})}{\hat{d}_{ij}} - \sum_{j=1}^{n}\frac{(\delta_{ij}-\hat{d}_{ij})}{\hat{d}_{ij}}y_{jp} = 0$$

Entonces, el sistema de ecuaciones resultante derivando para los np valores de las coordenadas principales puede expresarse como

$$FX = 0$$

siendo F una matriz cuadrada y simétrica de orden n cuyos coeficientes son:

$$f_{ij} = -\frac{\delta_{ij}-\hat{d}_{ij}}{\hat{d}_{ij}}, \quad i \neq j$$

$$f_{ii} = \sum_{j=1, j\neq i}^{n} f_{ij}, \quad i = j$$

4.4. Análisis de correspondencias.

El análisis de correspondencias es una técnica descriptiva enfocada al estudio y representación de tablas de contingencia.

Una tabla de contingencia consiste, en general, en un conjunto de números positivos dispuestos en forma matricial de forma que el valor de cada casilla representa la frecuencia absoluta observada para la combinación de las dos variables correspondientes a la fila y columna en cuestión, es decir, se tiene una matriz de valores numéricos k_{ij} que representan el número de individuos pertenecientes a la clase i de la característica I y a la clase j de la característica J, donde tanto I como J clasifican o particionan la población objeto de estudio. En este tipo de tablas no tendrá sentido, por tanto, distinguir entre variables e individuos, jugando ambos un papel simétrico.

El análisis de correspondencias es una técnica equivalente al análisis de componentes principales específicamente desarrollada para el estudio de variables cualitativas, siendo la información de partida la tabla de contingencia que recoge las frecuencias absolutas observadas de dos variables cualitativas (representadas, respectivamente, en las filas y en las columnas) en n elementos.

Se trata de un procedimiento que permite resumir la información de una tabla de contingencia obteniendo una representación de las variables en un espacio de dimensión menor, como en el análisis de componentes principales, pero realizando unas determinadas transformaciones sobre la tabla de contingencia inicial y utilizando la distancia χ^2 en lugar de la distancia euclídea que no es apropiada para la interpretación con este tipo de datos.

La tabla de contingencia inicial necesita ser transformada de acuerdo con su naturaleza y características particulares. Para ello, se trabaja con la matriz F de frecuencias relativas en la que cada casilla de la tabla inicial se divide por el número total de elementos observados, n, con lo que las casillas de F verifican:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1$$

Es decir, dada una tabla inicial de contingencia:

	1	...	j	...	J	
1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...
I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

consideraremos la tabla transformada conocida como tabla o matriz de correspondencias:

	1	...	j	...	J
1	$f_{11} = \frac{n_{11}}{n}$	...	$f_{1j} = \frac{n_{1j}}{n}$	...	$f_{1J} = \frac{n_{1J}}{n}$
...	...	...	...	...	...
i	$f_{i1} = \frac{n_{i1}}{n}$	...	$f_{ij} = \frac{n_{ij}}{n}$	...	$f_{iJ} = \frac{n_{iJ}}{n}$
...	...	...	...	...	...
I	$f_{I1} = \frac{n_{I1}}{n}$	...	$f_{Ij} = \frac{n_{Ij}}{n}$	...	$f_{IJ} = \frac{n_{IJ}}{n}$

Se realiza esta transformación para eliminar el efecto que, sobre el cálculo de distancias entre puntos-fila (o puntos-columna) tiene el efectivo total de cada fila (o columna).

4.4.1. Proyección de las filas.

Las I filas pueden tomarse como I puntos en el espacio $\mathbb{R}^J$ y se busca una representación en un espacio de dimensión menor que permita apreciar las distancias relativas entre ellos. Para ello es necesario tener en cuenta que, por las características intrínsecas de las tablas de contingencia, las filas no tienen el mismo peso puesto que el número de datos de cada una puede ser diferente, es decir, cada fila de F tiene una frecuencia relativa específica $f_{i.} = \sum_{j=1}^{J} f_{ij}$. En esta situación la distancia euclídea no es una buena medida de la proximidad puesto que aunque las frecuencias relativas de dos filas sean muy distintas esto puede ser debido únicamente al distinto número de elementos *contados* en cada fila siendo una de ellas simplemente producto de la otra por un determinado escalar. La distancia euclídea entre dos filas de este tipo arrojará un valor alto que no será reflejo de diferencias en la estructura de las filas sino únicamente de su distinta frecuencia relativa. Para evitar esta distorsión se divide cada casilla de la matriz por la frecuencia relativa de su fila, $f_{i.}$, obteniendo una nueva matriz transformada en la que los valores de cada casilla representan ahora la frecuencia relativa de la variable columna condicionada a la variable fila. De esta forma, las dos filas *linealmente dependientes* anteriores resultarán ahora idénticas y la distancia euclídea entre ellas será 0 reflejando que no existen diferencias en la estructura de ambas aun cuando sus frecuencias relativas sean muy distintas.

Si llamamos R a esta matriz de frecuencias relativas condicionadas a los totales por filas y D_fa la matriz diagonal de dimensiones I x I con las frecuencias relativas de las filas, $f_{i.}$, podemos escribir:

$$R = D_f^{-1} F$$

Ahora cada fila de R representa la distribución de la variable en columnas condicionada al atributo correspondiente a la fila.

Las filas, r_i' de R pueden considerarse puntos de $\mathbb{R}^J$ y puesto que $\sum_{j=1}^{J} r_{ij}' = 1, \;\; i = 1, \ldots, I,$ todos los puntos r_i' se encuentran en un espacio de dimensión $J - 1$.

Se trata ahora de proyectar estos puntos en un espacio de dimensión menor que preserve las distancias relativas entre filas en el sentido de que las filas que tengan estructuras similares estén próximas y las que tengan estructuras diferentes se encuentren alejadas. Para ello se debe definir una medida de distancia entre los puntos-fila pero la distancia euclídea no resulta apropiada puesto que considera de la misma forma todos los componentes de cada punto-fila y es necesario tener en cuenta la magnitud de las frecuencias de los atributos representados por las columnas en los cambios relativos entre estos atributos, para lo que se ponderan las diferencias en frecuencia relativa entre dos atributos de manera inversamente proporcional a la frecuencia de cada atributo, es decir, en lugar de calcular la distancia euclídea entre dos puntos r_a y r_b:

$$d_{ab} = \sum_{j=1}^{J} (r_{aj} - r_{bj})^2$$

se calculará la distancia χ^2 definida como:

$$D^2(r_a, r_b) = \sum_{j=1}^{J} \frac{(r_{aj} - r_{bj})^2}{f_{.j}}$$

o, en expresión matricial, $D^2(r_a, r_b) = (r_a - r_b)' D_c^{-1} (r_a - r_b)$, siendo D_c la matriz diagonal con términos $f_{.j}$.

Esta distancia equivale a la distancia euclídea entre los vectores transformados $y_i = D_c^{-1/2} r_i$ lo que permite simplificar el problema definiendo una nueva matriz de datos transformada de forma que tenga sentido aplicar la distancia euclídea:

$$Y = R D_c^{-1/2} = D_f^{-1} F D_c - 1/2$$

cuyos términos representan las frecuencias relativas condicionadas por filas y estandarizadas por su variabilidad, que serán directamente comparables entre sí y que serán de la forma:

$$y_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}, \quad i = 1, \ldots, I, \;\; j = i, \ldots, J$$

La distancia χ^2 tiene una propiedad importante conocida como *principio de equivalencia distribucional* que viene a decir que si dos filas (o análogamente dos columnas) tiene la misma estructura relativa,

$$\frac{f_{ij}}{f_{i.}} = \frac{f_{kj}}{f_{k.}}, \quad j = 1, \ldots, J$$

y se agregan en una nueva fila única, las distancias entre las restantes filas permanecen invariables.

En el Análisis de Correspondencias esta propiedad permite que si existe una suficiente proximidad entre los perfiles de dos filas (o columnas) puedan sustituirse por una única fila (o columna) como agregación de las anteriores sin que los resultados se vean alterados sustancialmente.

Ahora bien, para calcular la proyección de Y no será del todo correcto considerarla como una matriz de variables continuas tal y como se hace para encontrar las *componentes principales* puesto que, una vez mas, hay que tener en cuenta que al tratarse de una tabla de contingencia las filas tienen una distinta frecuencia relativa y por tanto deben tener distinto peso. Es decir, es necesario que las filas con mayor número de elementos estén bien representadas aunque ello suponga una peor representación de las filas con pocos elementos. Para ello se otorga a cada fila un peso proporcional al número de elementos que contiene, lo cual se lleva a cabo maximizando la suma de cuadrados ponderada:

$$m = a'Y'D_fYa$$

sujeto a $a'a = 1$, lo que equivale a: $m = a'D_c^{-1/2}F'D_f^{-1}FD_c^{-1/2}a$.

Alternativamente, se puede construir la matriz que estandariza las frecuencias relativas en cada casilla por el producto de las raíces cuadradas de las frecuencias relativas totales de la fila y columna:

$$Z = D_f^{-1/2}FD_c^{-1/2}$$

que tendrá sus componentes de la forma

$$z_{ij} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}}, \quad i = 1, \ldots, I, \ j = i, \ldots, J$$

Se tratará entonces de encontrar el vector a que maximice $m = a'Z'Za$ sujeto a $a'a = 1$, es decir, encontrar las *componentes principales* de Z y, por tanto, su solución será:

$$D_c^{-1/2}F'D_f^{-1}FD_c^{-1/2}a = \lambda a$$

donde a es un vector propio de $Z'Z$ y λ su valor propio asociado.

Dado que la matriz $Z'Z$ tiene como mayor valor propio siempre 1, se descarta esta solución trivial que no aporta información acerca de la estructura de las filas y se toma el mayor valor propio menor que 1 y su vector propio asociado, proyectando Y sobre la dirección encontrada para obtener la mejor representación, en una dimensión, de las filas de la tabla de contingencia original:

$$y_f(a) = Ya = D_f^{-1}FD_c^{-1/2}a$$

Análogamente, al extraer el vector propio ligado al siguiente mayor valor propio de $Z'Z$ se obtiene la segunda coordenada para cada fila de su mejor representación en un espacio de dimensión dos. De esta forma, las coordenadas de la mejor representación bidimensional de las filas vendrán dadas por las filas de la matriz:

$$C_f = YA_2 = D_f^{-1}FD_c^{-1/2}A_2$$

donde $A_2 = [a_1 a_2]$ es la matriz que contiene en columnas los dos vectores propios de $Z'Z$ asociados a los dos mayores valores propios menores que la unidad.

Este procedimiento se puede extender para obtener la mejor representación de las filas en mas dimensiones mediante el cálculo de los vectores propios asociados a los siguientes valores propios de $Z'Z$ en orden decreciente.

4.4.2. Proyección de las columnas.

Dado que en las tablas de contingencia, en general, no tiene sentido diferenciar entre individuos y variables, es decir, entre filas y columnas, debido a que cada una de estas dimensiones representa los atributos de una determinada variable, es posible aplicar a las columnas un análisis equivalente al descrito en la sección anterior para las filas, considerando ahora las J columnas como J puntos en $\mathbb{R}^I$.

Si llamamos $c = F'1$ al vector de frecuencias relativas de las columnas y D_c a la matriz diagonal que las contiene, la búsqueda de la mejor representación de los puntos-columna en un espacio de dimensión menor conducirá, aplicando la distancia χ^2 a estudiar la matriz $D_c^{-1}F'D_f^{-1/2}$, problema idéntico al de la sección anterior intercambiando el papel de las matrices D_c y D_f por lo que las direcciones de proyección serán ahora los vectores propios de la matriz:

$$ZZ' = D_f^{-1/2}FD_c^{-1}F'D_f^{-1/2}$$

siendo Z la matriz definida en la sección anterior.

Dado que $Z'Z$ y ZZ' tienen los mismos valores propios no nulos, esta última matriz tendrá también un valor propio unidad ligado al vector propio $\mathbf{1}$ y, llamando b al mayor valor propio menor que 1 de ZZ', la mejor representación unidimensional de los puntos-columna vendrá dada por:

$$y_c(b) = Y'b = D_c^{-1}F'D_f^{-1/2}b$$

Análogamente, la mejor representación bidimensional de las columnas viene dada por las coordenadas definidas por las filas de la matriz:

$$C_c = Y'B_2 = D_c^{-1}F'D_f^{-1/2}B_2$$

siendo $B_2 = [b_1 b_2]$ la matriz que contiene en columnas los vectores propios ligados a los dos valores propios mayores de ZZ' menores que la unidad.

4.4.3. Análisis conjunto.

El carácter simétrico del problema estudiado con el análisis de correspondencias hace que resulte de especial interés la representación conjunta de las filas y columnas de la matriz.

Las matrices $Z'Z$ y ZZ' tienen los mismos valores propios no nulos y, si a_i es un vector propio de $Z'Z$ ligado al valor propio λ_i: $Z'Za_i = \lambda_i a_i \Rightarrow ZZ'(Za_i) = \lambda_i(Za_i) \Rightarrow b_i = Za_i$, donde b_i es un vector propio de ZZ' ligado a λ_i.

En consecuencia, un método para obtener los vectores propios es calcular directamente los correspondientes a la matriz de dimensión menor, $Z'Z$ o ZZ', y a partir de estos obtener los restantes como Za_i o $Z'b_i$

En el análisis de correspondencias, como en el caso del análisis de componentes principales, la proporción de variabilidad explicada por cada dimensión se calcula descartando el valor propio igual a 1 y tomando la proporción que representa cada valor propio en relación a los restantes.

En definitiva, el análisis de correspondencias de una tabla de contingencia de dimensiones I x J se lleva a cabo en los siguientes pasos:

1. Se obtiene la matriz de frecuencias relativas, F y se transforma en la matriz estandarizada de frecuencias relativas en la que cada celda se divide por la raíz de los totales de su fila y columna, $z_{ij} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J$
2. Se calculan los h vectores propios ligados a valores propios mayores distintos de la unidad de la matriz de menor dimensión entre $Z'Z$ y ZZ', y se obtienen los restantes de la siguiente forma: si $Z'Z$ es la matriz de menor dimensión se calculan directamente sus vectores propios a_i y a continuación los restantes aplicando $b_i = Za_i$, mientras que si la dimensión de ZZ' es menor se calcular sus vectores propios b_i y posteriormente los vectores propios de $Z'Z$ como $a_i = Z'b_i$

Las I filas y las J columnas se representarán como puntos en $\mathbb{R}^h$ con coordenadas dadas, respectivamente, por

$$C_f = D_f^{-1/2} Z A_h$$

$$C_c = D_c^{-1/2} Z' B_h$$

donde A_h contiene en columnas los h vectores propios de $Z'Z$ correspondientes a los h valores propios distintos a la unidad en orden decreciente, y B_h los correspondientes h vectores propios asociados de ZZ'.

En general se suele considerar $h = 2$ para obtener una representación bidimensional aunque se puede establecer otro valor de h en función de las particularidades del problema concreto para lo que se suele fijar una cota inferior para la proporción acumulada de variabilidad explicada y considerar el valor de h que la supere.

Capítulo 5

Técnicas estadístcas de clasificación. Análisis discriminante.

El problema de la clasificación (o discriminación) es común en muchas áreas de las ciencias, tanto experimentales como sociales, desde la biología o la medicina hasta la sociología o la economía. En ingeniería ha sido ampliamente estudiado para diseñar máquinas capaces de clasificar de forma automática, por ejemplo billetes y monedas, sonidos, imágenes, etc. También en el ámbito financiero tiene aplicación desde hace mucho tiempo para la clasificación de riesgo crediticio donde, a partir de una serie de variables conocidas sobre la persona que solicita un crédito (ingresos, profesión, miembros de la unidad familiar, patrimonio, edad, etc) se aplica para decidir, de manera automatizada, acerca de su concesión. Existen otros muchos ejemplos de aplicación: en el diagnóstico de enfermedades, en procesos de control de calidad para procesos de fabricación o, como en el caso de este trabajo, para asignar un texto a uno de varios autores a partir del análisis de las frecuencias de utilización de las palabras.

En términos generales el planteamiento estadístico de este tipo de problemas es el siguiente: se cuenta con un conjunto de elementos pertenecientes a dos o mas poblaciones distintas sobre los que se ha observado una variable aleatoria p-dimensional, x, y se desea clasificar un nuevo elemento, a partir de sus valores conocidos para la variable x, en una de las poblaciones.

En las situaciones en las que se cuenta a priori con una serie de clases o categorías predefinidas que particionan la población las técnicas para construir un modelo que asigne los elementos a la clase correspondiente se conocen también como técnicas de *clasificación automática supervisada.* En contraposición, las técnicas en las que el objetivo es construir un modelo para particionar la población en grupos homogéneos pero sin partir de ningún tipo de información previa acerca del número y características de estos, y en las que se aplica la metodología del *análisis cluster* o *análisis de conglomerados, se denominan técnicas* de *clasificación automática no supervisada* .

Entre las técnicas de clasificación supervisada mas utilizadas encontramos algoritmos probabilísticos (entre los que destaca el conocido algoritmo de Naive-Bayes que consiste en estimar la probabilidad de que un documento pertenezca a una categoría en función de la probabilidad de poseer una serie de características conocida para cada uno de los elementos que pertenecen a la categoría en cuestión), el algoritmo del vecino mas próximo, extensivo a los k vecinos mas próximos (en el que se calcula la similitud entre el elemento a clasificar y cada uno de los elementos del conjunto de entrenamiento, asumiendo que la categoría del elemento coincide con la del mas similar entre estos últimos), algoritmos basados en redes neuronales (una de las aplicaciones mas extendidas de las redes neuronales es precisamente el reconocimiento de patrones) o algoritmos basados en árboles de clasificación.

Aunque existen muchos algoritmos de clasificación supervisada la idea básica es la misma en todos ellos: construir un patrón para cada una de las clases que, mediante la aplicación de alguna función, permita estimar el parecido o similitud entre cada elemento a clasificar y los patrones de las categorías. Para construir los patrones se utiliza un conjunto de individuos de los que se conoce previamente su clase y que se denomina conjunto de entrenamiento, conociéndose como *entrenamiento* o aprendizaje el proceso de formación de los patrones de cada clase a partir de estos individuos conocidos.

Los conceptos y técnicas de clasificación utilizados con variables de tipo numérico son aplicables a variables textuales sin mas que tener en cuenta que en este último caso debe realizarse una tarea previa de procesamiento de los datos textuales no estructurados que proporcione una matriz de datos estructurados sobre la que sea posible su aplicación. Por lo tanto, las dos etapas de las técnicas de clasificación: construcción del clasificador y clasificación de nuevos documentos, vendrán precedidas en el caso de la variable textual del preprocesamiento de los datos textuales.

Así, la clasificación supervisada de textos se puede definir como la tarea de aproximar una función de asignación de categoría desconocida $F : D\ x\ C\ \rightarrow\ \{0, 1\}$, donde D es el conjunto de documentos de texto y C es el conjunto de categorías predefinidas. El valor de $F(d, c)$ es 1 si el documento d pertenece a la categoría c mientras que de otra manera el valor es 0.

En este trabajo se aplicará el análisis discriminante clásico a un problema de clasificación de textos para la identificación de los autores. Se pretende obtener un modelo discriminante que clasifique los textos en función de su autor. Para reducir la dimensionalidad de los datos y, al mismo tiempo, contar con un conjunto de variables continuas bajo la hipótesis de normalidad multivariante (necesaria para la aplicación del análisis discriminante), se aplicarán técnicas del escalado multidimensional a la matriz de *Documentos* x *Términos*.

5.1. Análisis Discriminante Lineal (LDA).

El enfoque para la resolución del problema de clasificación propuesto por Fisher bajo la denominación de *análisis discriminante* está basado en la normalidad multivariante de los datos considerados y es óptimo bajo este supuesto. En los casos en que todas las variables son continuas, aun cuando los datos originales no estén normalmente distribuidos es posible transformarlos para obtener normalidad y con ello posibilitar la aplicación de esta técnica, pero cuando en el conjunto de variables exista alguna de tipo discreto, la aceptación de la hipótesis de normalidad es poco realista y será preferible aplicar otro tipo de técnicas.

Se denomina *regla de decisión* a cualquier partición del espacio muestral E_x en regiones: $A_1, \ldots, A_n$ tales que $\bigcap_{i \neq j} A_i A_j = 0,\ i, j \in (1, \ldots, n)$ y $E_x = \bigcup_{i=1}^{n} A_i$, y de manera que si $x_0 \in A_i \Rightarrow d_i,\ i = (1, \ldots, n)$, donde d_i representa la decisión de clasificar x_0 en la población P_i

Consideremos ahora una variable x p-variante y absolutamente continua y dos poblaciones, P_1 y P_2 en las que x tiene funciones de densidad conocidas f_1 y f_2, y planteemos el problema de clasificar una nueva observación x_0.

Si conocemos las probabilidades *a priori* de que x_0 provenga de cada una de las poblaciones, $P[x_0 \in P_1] = \pi_1$ y $P[x_0 \in P_2] = \pi_2$, con $\pi_1 + \pi_2 = 1$, entonces $f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$, y observado x_0 será posible calcular, por el teorema de Bayes, las probabilidades *a posteriori* de que haya sido generado por cada una de las poblaciones:

$$P[P_1|x_0] = \frac{\pi_1 P[x_0|P_1]}{\pi_1 P[x_0|P_1] + \pi_2 P[x_0|P_2]} = \frac{\pi_1 f_1(x_0)}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

$$P[P_2|x_0] = \frac{\pi_2 P[x_0|P_2]}{\pi_1 P[x_0|P_1] + \pi_2 P[x_0|P_2]} = \frac{\pi_2 f_2(x_0)}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

Entonces, clasificaremos x_0 en la población mas probable *a posteriori*, es decir, clasificaremos x_0 en P_2 si $\pi_2 f_2(x_0) > \pi_1 f_1(x_0)$. Nótese que en el caso en que las probabilidades a priori fueran iguales, $\pi_1 = \pi_2$, la regla de clasificación anterior se reduce a clasificar x_0 en P_2 si $f_2(x_0) > f_1(x_0)$.

En muchas ocasiones, los errores de clasificación tienen distintas consecuencias que se pueden cuantificar, en cuyo caso, planteándolo como un problema bayesiano de decisión, podemos incluir estas consecuencias en la solución. Sea $c(i|j)$ el coste de clasificar en P_i un elemento que pertenece a P_j, con $c(i|i) = 0$ y $c(i|j)$ conocido para todo $i \neq j$. Entonces, se trata ahora de minimizar el coste esperado.

En el caso de dos poblaciones, los costes esperados de las decisiones d_1 y d_2 serán:

$$E(d_1) = c(1|1)P[P_1|x_0] + c(1|2)P[P_2|x_0] = c(1|2)P[P_2|x_0] = c(1|2)\pi_2 f_2(x_0)$$

$$E(d_2) = c(2|1)P[P_1|x_0] + c(2|2)P[P_2|x_0] = c(2|1)P[P_1|x_0] = c(2|1)\pi_1 f_1(x_0)$$

Entonces, clasificaremos x_0 en P_1 si su coste esperado es menor, es decir, si $c(1|2)\pi_2 f_2(x_0) < c(2|1)\pi_1 f_1(x_0)$ o, equivalentemente, si:

$$\frac{\pi_2 f_2(x_0)}{c(2|1)} < \frac{\pi_1 f_1(x_0)}{c(1|2)}$$

y puede comprobarse que este criterio es equivalente minimizar la probabilidad total de error en la clasificación.

5.1.1. Función lineal discriminante.

Supongamos que f_1 y f_2 son distribuciones normales con distintos vectores de medias, μ_1 y μ_2, pero igual matriz de covarianzas, V.

La partición óptima, como acabamos de ver, es la que clasificará los nuevos elementos en P_1 si

$$\frac{\pi_2 f_2(x_0)}{c(2|1)} < \frac{\pi_1 f_1(x_0)}{c(1|2)}$$

Tomando logaritmos y sustituyendo f_i por sus expresiones, esto equivale a:

$$-\frac{1}{2}(x_0-\mu_2)'V^{-1}(x_0-\mu_2)+log\frac{\pi_2}{c(2|1}>-\frac{1}{2}(x_0-\mu_1)'V^{-1}(x_0-\mu_1)+log\frac{\pi_1}{c(1|2}$$

que se puede reescribir a su vez como:

$$D_1^2 - log\frac{\pi_1}{c(1|2} > D_2^2 - log\frac{\pi_2}{c(2|1}$$

siendo D_i^2 la distancia de Mahalanobis entre el elemento x_0 y la media de P_i, μ_i: $D_i^2 = (x_0-\mu_i)'V^{-1}(x_0-\mu_i)$

En el caso de que los costes y las probabilidades *a priori* sean iguales, ($c(1|2) = c(2|1)$ y $\pi_1 = \pi_2$) la regla anterior se reduce a clasificar x_0 en P_1 si $D_1^2 < D_2^2$. Nótese además que si las variables son incorreladas, $V = I\sigma^2$, D_i^2 coincide con la distancia euclídea.

La generalización de lo anterior para G poblaciones es inmediata. Supongamos los costes de clasificación constantes e independientes de la población en que se clasifique la población, en cuyo caso A_g será la región definida por los puntos con probabilidad máxima de ser generados por P_g, es decir, la región en la que el producto de la probabilidad *a priori* y la verosimilitud sea máximo:

$$A_g = \{x \in E_x | \pi_g f_g(x) > \pi_i f_i(x);\ \forall i \neq g\}$$

Cuando las probabilidades *a priori* son iguales ($\pi_i = \frac{1}{G},\ i = (1,\ldots,G)$ y $f_i(x) \rightsquigarrow N(\mu_i, V)$, la definición anterior equivale a calcular la distancia de Mahalanobis desde el elemento al centro de cada población y clasificarlo en la población que la haga mínima.

5.1.2. Poblaciones desconocidas.

Consideremos la matriz de datos X, n x p, particionada en G matrices correspondientes a las G poblaciones. Entonces los elementos de X serán de la forma x_{ijg}, donde i representa la fila, j la columna y g la submatriz. Si n_g es el número de elementos que pertenecen a la submatriz g, entonces $n = \sum_{g=1}^{G} n_g$, y si $x'_{ig} = (x_{i1g},\ldots,x_{ipg})$, entonces el vector de medias de cada población y la correspondiente matriz de varianzas y covarianzas serán

$$\bar{x}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} x_{ig}$$

$$\hat{S}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(x_{ig}-\bar{x}_g)(x_{ig}-\bar{x}_g)'$$

Si suponemos que todas las poblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de las matrices de varianzas y covarianzas en cada población con peso proporcional a su precisión:

$$\hat{S}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \hat{S}_g$$

Llamaremos W a la matriz de sumas de cuadrados dentro de las clases

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)' = (n - G)\hat{S}_w$$

Ahora, para obtener las funciones discriminantes se utiliza $\bar{x}_g$ como estimador de μ_g y $\hat{S}_w$ como estimación de V. Así, suponiendo iguales las probabilidades *a priori* y los costes de clasificación ($c(i|j) = k,\ \pi_i = 1/G,\ \forall i \in \{1, \ldots, G\},\ i \neq j$) y llamando $\hat{w}_g = \hat{S}_w^{-1}\bar{x}_g$ se clasificará x_0 en la población g que verifique:

$$min[(x_0 - \bar{x}_g)'\hat{S}_w^{-1}(x_0 - \bar{x}_g)] = min[w_g'(\bar{x}_g - x_0)]$$

lo que equivale a calcular las variables indicadoras

$$z_{g,g+1} = \hat{w}_{g,g+1}' x_0,\ g = 1, \ldots, G - 1$$

donde $\hat{w}_{g,g+1} = \hat{S}_w^{-1}(\bar{x}_g - \bar{x}_{g+1}) = \hat{w}_g - \hat{w}_{g+1}$ y la regla de decisión será clasificar en g frente a $g + 1$ si

$$|z_{g,g+1} - \hat{m}_g| < |z_{g,g+1} - \hat{m}_{g+1}|$$

siendo $\hat{m}_g = \hat{w}_{g,g+1}'\bar{x}_g$

5.1.2.1. Probabilidades de error y validación cruzada.

El método mas inmediato para aproximar el error asociado a una regla de clasificación consiste en aplicar la función discriminante a los n elementos observados para obtener su clasificación, llamando n_{ij} al número de elementos de P_i clasificados en P_j, de forma que el error aparente se calculará como el cociente entre el número de elementos mal clasificados y el número de elementos clasificados correctamente, es decir:

$$\epsilon = \frac{\sum_{i \neq j}^{G} n_{ij}}{\sum_{i=1} n_{ii}}$$

El método anterior tiende a subestimar las probabilidades de error al emplear los mismos datos en la estimación de los parámetros y en la evaluación de la regla de clasificación resultante. Por este motivo es preferible aplicar la técnica denominada *validación cruzada* consistente en clasificar cada elemento con una regla construida prescindiendo de dicho elemento. Así, se construyen n funciones discriminantes con los n conjuntos de $n - 1$ elementos cada uno resultantes de eliminar uno a uno cada elemento del conjunto original, para a continuación clasificar cada uno de los elementos con la regla construida sin él.

Cuando el número de elementos es muy elevado, el método de validación cruzada es computacionalmente muy costoso y se suele dividir el conjunto de datos en k subconjuntos de igual tamaño con los que se aplica la *validación cruzada* pero prescindiendo de cada uno de los subconjuntos para obtener las reglas de clasificación.

5.2. Discriminación cuadrática (QDA).

Si aun presumiendo la normalidad de las variables observadas no fuera posible admitir la hipótesis de igualdad de varianzas, el problema se resolverá clasificando cada nueva observación en la población con máxima probabilidad *a posteriori*, es decir, clasificar x_0 en la población que minimice la siguiente función

$$\frac{1}{2}log|V_g| + \frac{1}{2}(x_0 - \mu_g)'V_g^{-1}(x_0 - \mu_g) - log(C_g\pi_g), \quad g = 1, \ldots, G$$

Si V_g y μ_g no son conocidas se estimarán de la forma habitual mediante S_g y $\bar{x}_g$.

En este caso el término $x_0'V_g^{-1}x_0$ no se puede anular puesto que depende de la población, y las funciones discriminantes no serán lineales al contar con un término de segundo grado.

El número de parámetros a estimar en este caso cuadrático se mucho mayor que en el lineal lo que hace que, excepto en el caso de muestras muy grandes, la discriminación cuadrática sea bastante inestable y, aun con matrices de covarianzas muy diferentes, es frecuente obtener mejores resultados con la clasificación lineal. Un problema adicional es la extremada sensibilidad de la función discriminante cuadrática a desviaciones de la normalidad de las variables observadas.

En general, la evidencia indica que la discriminación lineal es mas robusta que la discriminación cuadrática.

5.3. Métricas de evaluación..

Existen diversas medidas para evaluar la "bondad" de los modelos de clasificación obtenidos que, en general, están basadas en el concepto de *matriz de confusión.*

Esta matriz es una tabla:

	C_1'	$\ldots$	C_p'	
C_1	n_{11}	$\ldots$	n_{1p}	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
C_p	n_{p1}	$\ldots$	n_{pp}	$n_{p.}$
	$n_{.1}$	$\ldots$	$n_{.p}$	

Las filas, C_i, representan las clases o categorías reales y las columnas, C_i', las clases o categorías predichas por el modelo. Los valores n_{ij} corresponden con el número de elementos de la clase real i que han sido predichos por el modelo en la clase j.

Por lo tanto el tamaño del conjunto de test es igual a la suma de todos los elementos de la matriz es, $n_{..} = \sum_{i=1}^{p}\sum_{j=1}^{p} n_{ij}$, y el número total de elementos bien clasificados vendrá determinado por la suma de la diagonal principal,$\sum_{i=1}^{p} n_{ii}$.

5.3.1. Exactitud, precisión y recuperación.

La proporción de elementos clasificados correctamente por el modelo corresponderá entonces con el conciente entre la suma de la diagonal principal y la suma de todos los elementos de la matriz de confusión. Este valor, en forma de porcentaje, se conoce como *exactitud* (o *accuracy*) y se utiliza como medida global del "acierto" del modelo pues representa la proporción total de elementos de todas las categorías clasificados correctamente.

$$Exactitud = 100\frac{\sum_{1=1}^{p} n_{ii}}{\sum_{i=1}^{p}\sum_{j=1}^{p} n_{ij}} = 100\frac{\sum_{1=1}^{p} n_{ii}}{n_{..}}$$

Además se definen algunas otras medidas individuales para evaluar el comportamiento del modelo respecto de cada una de las categorías. Las mas habituales son la *precisión* y la *recuperación* (o *recall*).

La *precisión* representa la proporción de elementos de una categoría que el modelo clasifica correctamente y se calcula como el cociente de cada elemento de la diagonal principal entre la suma de la fila a la que pertence.

$$Precisión_i = 100\frac{n_{ii}}{\sum_{j=1}^{p} n_{ij}} = \frac{n_{ii}}{n_{i.}}$$

La *recuperación* representa la proporción de elementos clasificado por el modelo en una determinada categoría que realmente pertenecen a la misma y se obtiene calculando el cociente entre cada elemento de la diagonal principal y la suma de la columna a la que pertenece.

$$Recuperación_i = 100\frac{n_{jj}}{\sum_{i=1}^{p} n_{ij}} = 100\frac{n_{jj}}{n_{.j}}$$

5.3.2. Adjusted Rand Index (ARI)

Se trata de una medida definida en el área del análisis cluster para comparar dos particiones de entre el conjunto finito de particiones posibles de una población determinada. Fue propuesta por Hubert y Arabie en 1985 a partir del índice de Rand, una de las medidas mas populares para comparar particiones en aquel momento.

Supongamos una población de n objetos, $S = \{o_1, ..., o_n\}$ y dos particiones de S, $U = \{u_1, ..., u_p\}$ y $V = \{v_1, ..., v_q\}$. En el análisis clúster "clásico" el número de conjuntos que forman cada una de las particiones puede no coincidir ($p \neq q$) pero en la utilización de estos indices para evaluar modelos de clasificación supervisada este número representa el de las clases y enfrentaremos la partición correspondiente a las clases reales de los elementos del conjunto de test frente a las clases predichas por el modelo, por lo que siempre se verificará $p = q$.

Consideramos ahora todos los posibles pares de objetos de S, (o_i, o_j), $i = 1, \ldots, n$; $i \neq j$ y definimos:

- a: número de pares (o_i, o_j) que se encuentran en el mismo subconjunto de U y en el mismo subconjunto de V, es decir, $o_i, o_j \in U_h$; $o_i, o_j \in V_k$

- b: número de pares (o_i, o_j) que se encuentran en distintos subconjuntos de U y en distintos subconjuntos de V, es decir, $o_i \in U_h, o_j \in U_k \neq U_h \ ; \ o_i \in V_m, o_j \in V_n \neq V_m$
- c: número de pares (o_i, o_j) que se encuentran en el mismo subconjunto de U y en distintos subconjuntos de V, es decir, $o_i, o_j \in U_h \ ; \ o_i \in V_m, o_j \in V_n \neq V_m$
- d: número de pares (o_i, o_j) que se encuentran en distintos subconjuntos de U y en el mismo subconjunto de V, es decir, $o_i \in U_h, o_j \in U_k \neq U_h \ ; \ o_i, o_j \in V_k$

De forma que diremos que se produce un *acuerdo* entre U y V cuando el par de objetos (o_i, o_j) se encuentra en los casos a o b, es decir, si ambas particiones incluyen a los dos objetos en un mismo subconjunto o ambas particiones incluyen a cada uno de los objetos en distintos subconjuntos.

Al contratio diremos que se produce un *desacuerdo* entre U y V cuando el par de objetos (o_i, o_j) se encuentra en los casos c o d, es decir, si una partición incluye a los dos objetos en un mismo subconjunto y la otra partición incluye a cada uno de los objetos en distintos subconjuntos.

La suma de posibles *acuerdos* y *desacuerdos* $(a + b + c + d)$ será entonces el número total de pares posibles, $\binom{n}{2}$.

El índice de Rand representa la proporción de acuerdos entre U y V:

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

El *índice de Rand ajustado* propuesto por Hubert y Arabie es la versión corregida por el azar del índice de Rand. Su expresión a partir de una matriz de confusión es la siguiente:

$$ARI = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\binom{n_{ij}}{2} - \frac{\sum_{i=1}^{p}\binom{n_{i\cdot}}{2}\sum_{j=1}^{q}\binom{n_{\cdot j}}{2}}{\binom{n_{\cdot\cdot}}{2}}}{\frac{\sum_{i=1}^{p}\binom{n_{i\cdot}}{2}+\sum_{j=1}^{q}\binom{n_{\cdot j}}{2}}{2} - \frac{\sum_{i=1}^{p}\binom{n_{i\cdot}}{2}\sum_{j=1}^{q}\binom{n_{\cdot j}}{2}}{\binom{n_{\cdot\cdot}}{2}}}$$

En los problemas de clasificación supervisada, en los que se utiliza este índice para comparar la "partición" obtenida (clases predicas) con la "partición" perfecta (clases reales), ARI=0 cuando la clasificación obtenida se corresponde con la esperada para un modelo puramente aleatorio, mientras que ARI=1 cuando la clasificación del modelo es perfecta, es decir coincide con la clasificación real del conjunto de test. Se da la circunstancia de que este índice puede tomar valores negativos cuando la clasificación obtenida es peor que la esperada para un modelo aleatorio puro.

Capítulo 6

Análisis del corpus *Reuters Corpus Volume I.*

La aplicación de los métodos y técnicas estadísticas presentadas en este trabajo se reailizará sobre un subconjunto del *dataset Reuters Corpus Volume I* (Lewis, Yang, Rose, & and Li, 2004). Este conjunto de datos contiene más de 800.000 noticias periodísticas de la agencia Reuters clasificadas por autores. Se han seleccionado 100 noticias de cada uno de 4 autores.

```
##
##  AlexanderSmith BenjaminKangLim     BradDorfman     DavidLawder
##             100              100             100             100
```

Como hemos visto es necesario realizar un determinado procesamiento previo del corpus para lo que se aplican las siguientes acciones:

- Convertir todas las palabras a minúsculas
- Eliminar los números
- Eliminar espacios en blanco innecesarios
- Eliminar los signos de puntuación
- Eliminar las *stopWords* del idioma inglés
- Eliminar las palabras de 1 y 2 caracteres
- *Lematizar* todas las palabras resultantes

6.1. Matriz *Documentos* x *Palabras*.

Con este corpus transformado se construye la matriz de Documentos x Palabras, obteniendo así una primera aproximación a la información de interés, estadísticamente hablando, de los textos objeto de análisis.

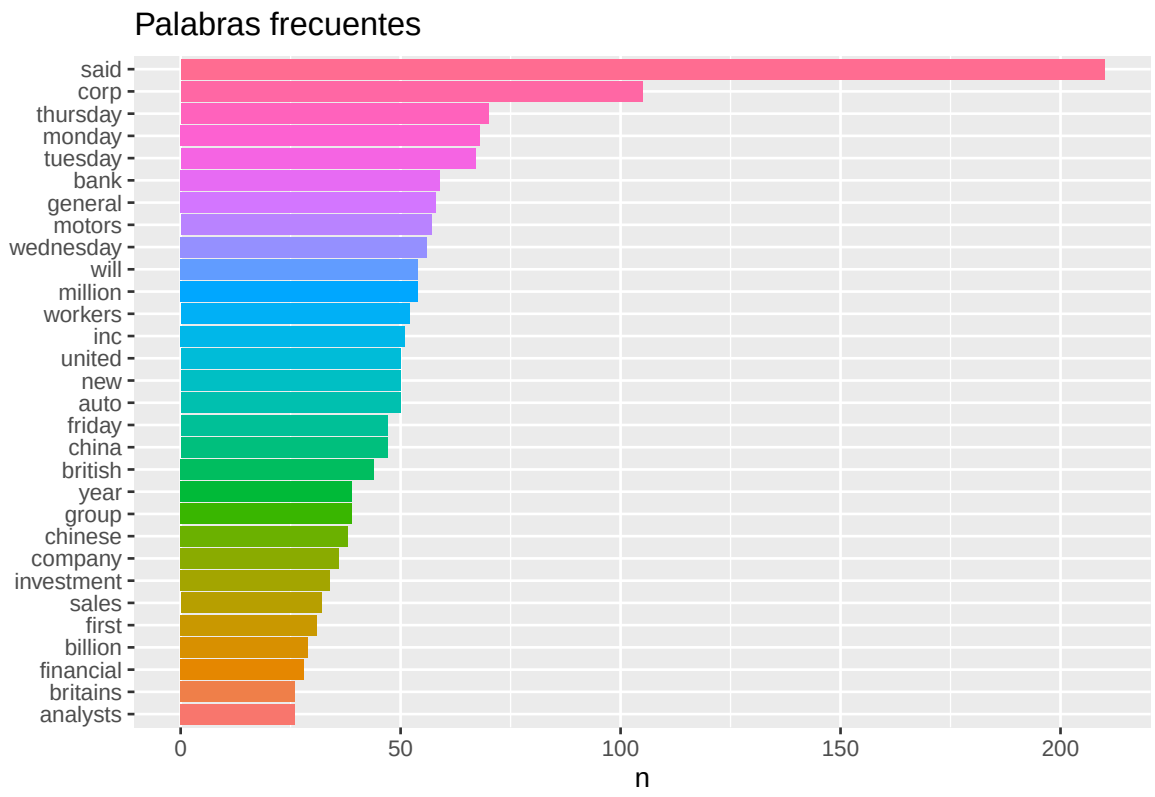
```
## <<DocumentTermMatrix (documents: 400, terms: 1990)>>
## Non-/sparse entries: 7781/788219
## Sparsity           : 99%
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

La simple transformación en una matriz numérica proporciona una primera información sobre el corpus objeto de estudio. Esta matriz está compuesta por 400 filas (que corresponden a las noticias que forman el corpus) y 1990 columnas (que corresponden a las palabras conservadas tras el procesamiento). Se trata de una matriz muy dispersa en la que el 99 % de los elementos son ceros (788219 elementos iguales a 0 frente a 7781 elementos mayores que cero) y la palabra de mayor longitud tiene 21 caracteres.

6.2. Análisis exploratorio

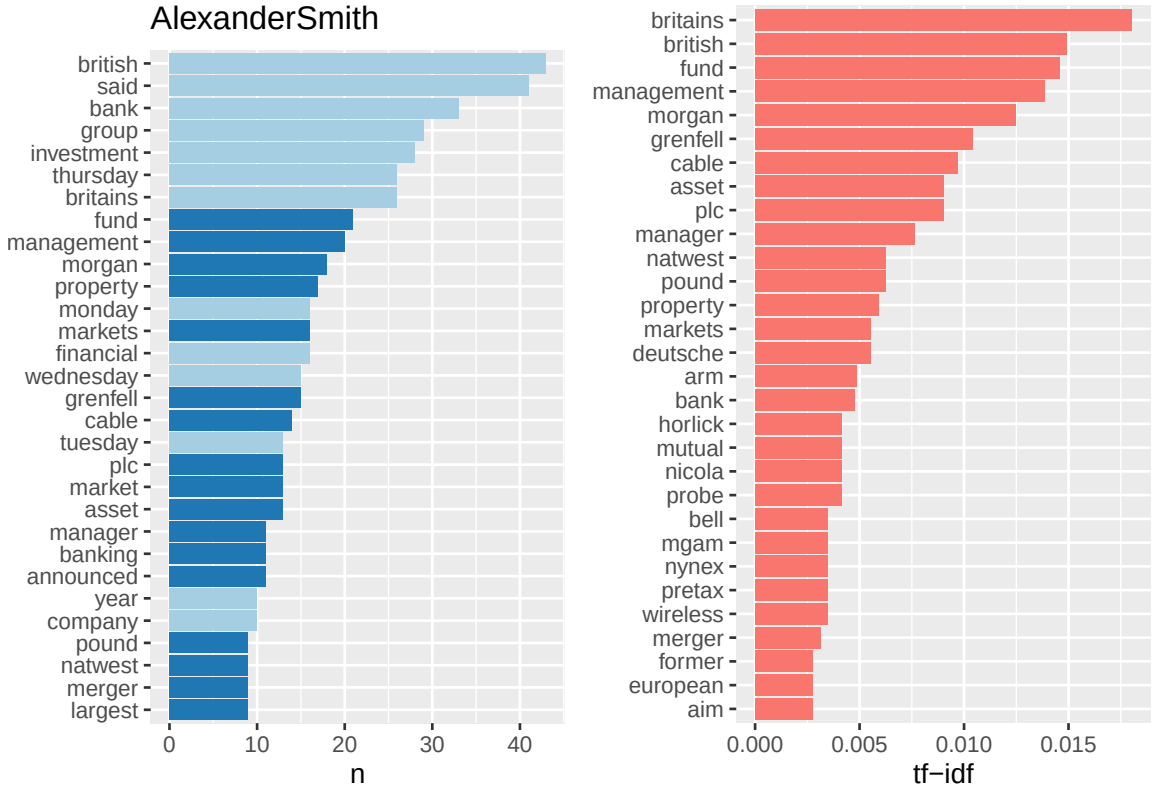
En primer lugar es habitual identificar las palabras mas frecuentes a nivel de *corpus* y, en su caso, a nivel de cada una de las categorías que clasifican los documentos del corpus (en nuestro caso los autores) así como aplicar la función tfIdf a fin de identificar los términos mas representativos de cada categoría.

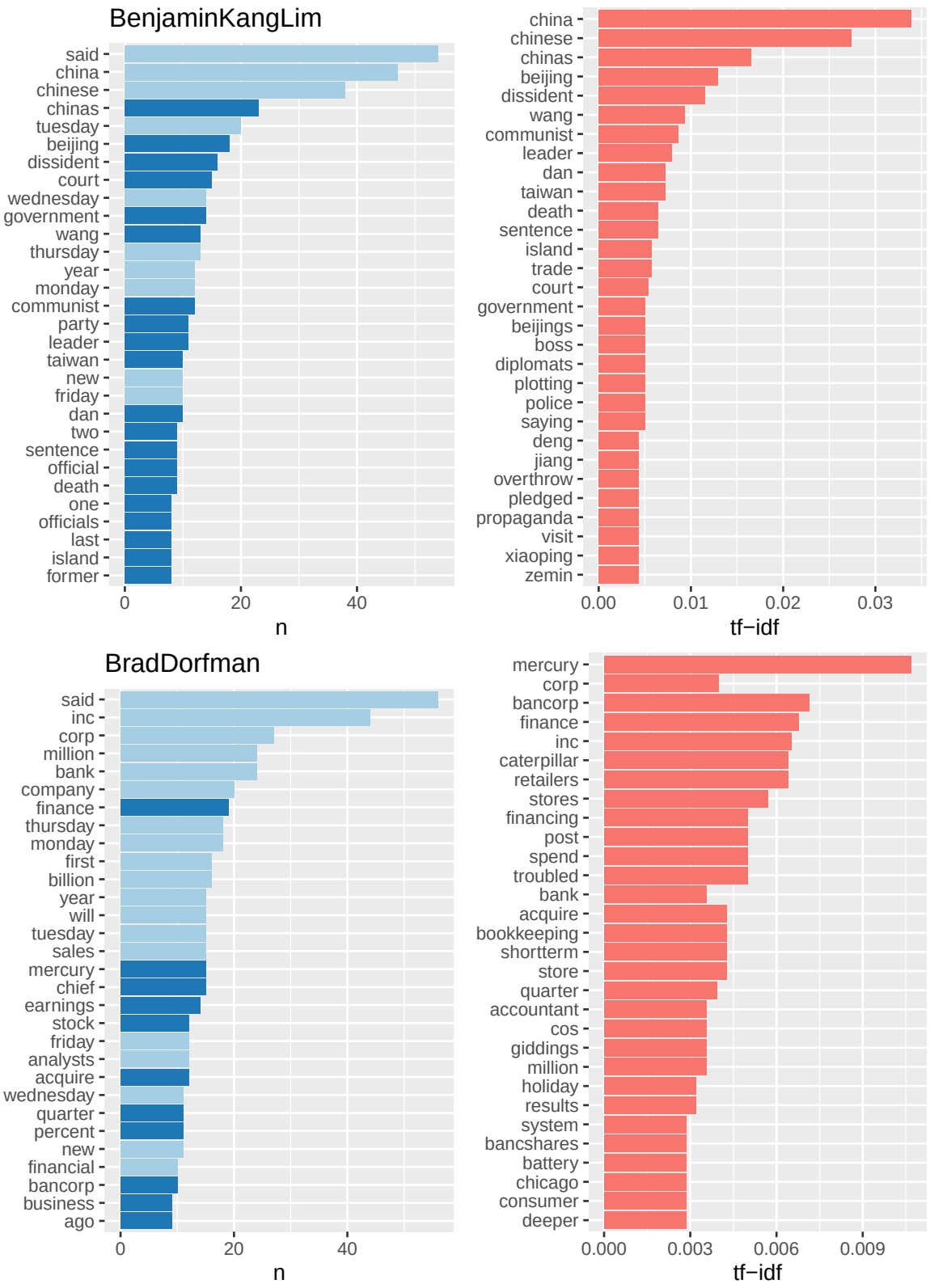
6.2.1. Palabras frecuentes, *tfIdf* y distribución del vocabulario.

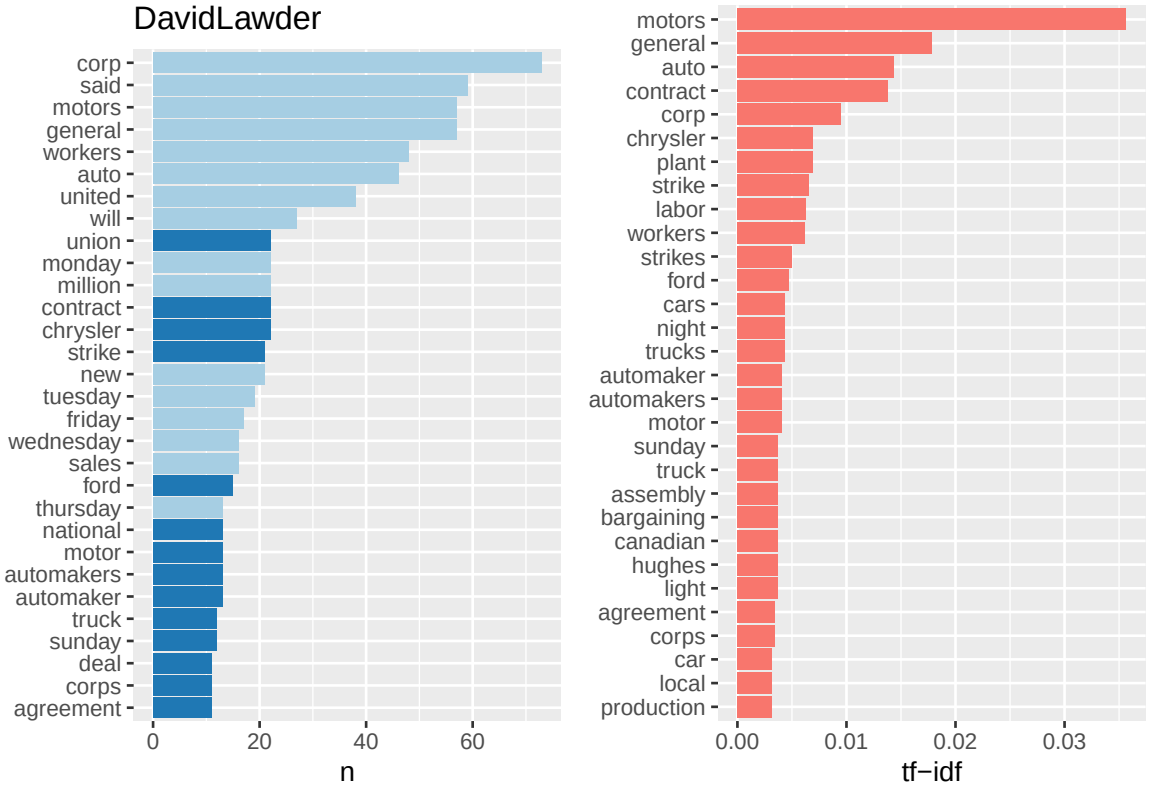


Este gráfico nos permite detectar las temáticas predominantes en las noticias que forman el corpus: temas empresariales, financieros, banca, sector automovilístico, china, etc.

Observar las palabras mas frecuentes de cada autor facilita la identificación de posibles diferencias entre ellos en cuanto a vocabulario o temáticas subyacentes en sus documentos. Por este motivo estudiaremos a continuación, para cada uno de los cuatro autores, las palabras mas frecuentes y las que mejor lo caracterizan (las de mayor valor tfIdf). En los gráficos de palabras mas frecuentes se muestran en una tonalidad mas clara aquellas que también se encuentran presentes entre las mas frecuentes a nivel de corpus, frente a una tonalidad mas oscura para las palabras mas frecuentes en cada autor que no encuentran correspondencia a este nivel.




BenjaminKangLim
said
china
chinese
chinas
tuesday
beijing
dissident
court
wednesday
government
wang
thursday
year
monday
communist
party
leader
taiwan
new
friday
dan
two
sentence
official
death
one
officials
last
island
former
0
20
40
n
china
chinese
chinas
beijing
dissident
wang
communist
leader
dan
taiwan
death
sentence
island
trade
court
government
beijings
boss
diplomats
plotting
police
saying
deng
jiang
overthrow
pledged
propaganda
visit
xiaoping
zemin
0.00
0.01
0.02
0.03
tf−idf
BradDorfman
said
inc
corp
million
bank
company
finance
thursday
monday
first
billion
year
will
tuesday
sales
mercury
chief
earnings
stock
friday
analysts
acquire
wednesday
quarter
percent
new
financial
bancorp
business
ago
0
20
40
n
mercury
corp
bancorp
finance
inc
caterpillar
retailers
stores
financing
post
spend
troubled
bank
acquire
bookkeeping
shortterm
store
quarter
accountant
cos
giddings
million
holiday
results
system
bancshares
battery
chicago
consumer
deeper
0.000
0.003
0.006
0.009
tf−idf

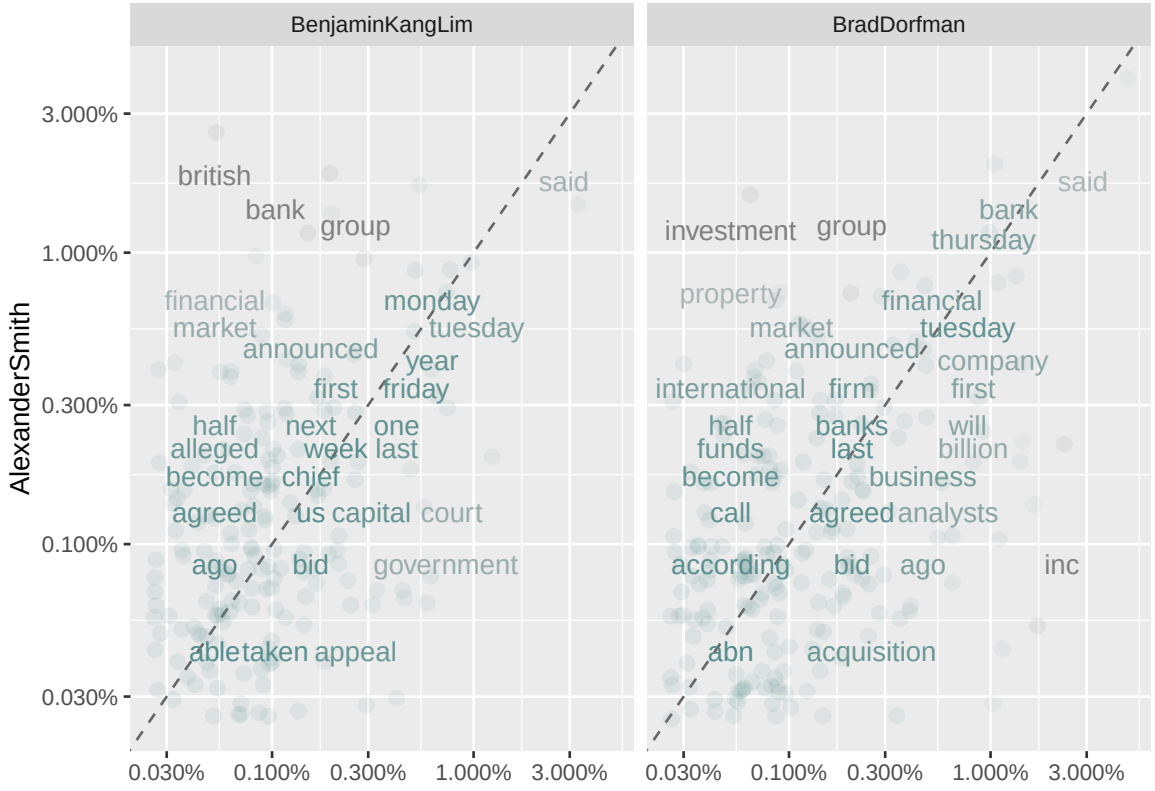


La observación detallada de estos gráficos permite profundizar en el vocabulario empleado y en la temática propia en cada uno de los autores. Así, en Alexander Smith destaca la palabra *british*, siendo además la segunda con mayor valor tfIdf lo que indica que es un término, en comparación, muy poco empleado por el resto de autores; en Benjamin Kang Lim predomina un vocabulario muy relacionado con la política de China; en Brad Dorfman observamos una mayoría de términos relacionados con temática bancaria y financiera así como la palabra *chicago* entre las de valores tfIdf altos; y, por último, las palabras que mas destacan para David Lawder tienen que ver con la industria automovilística y, entre las palabras de valores tfIdf altos, encontramos *canadian*.

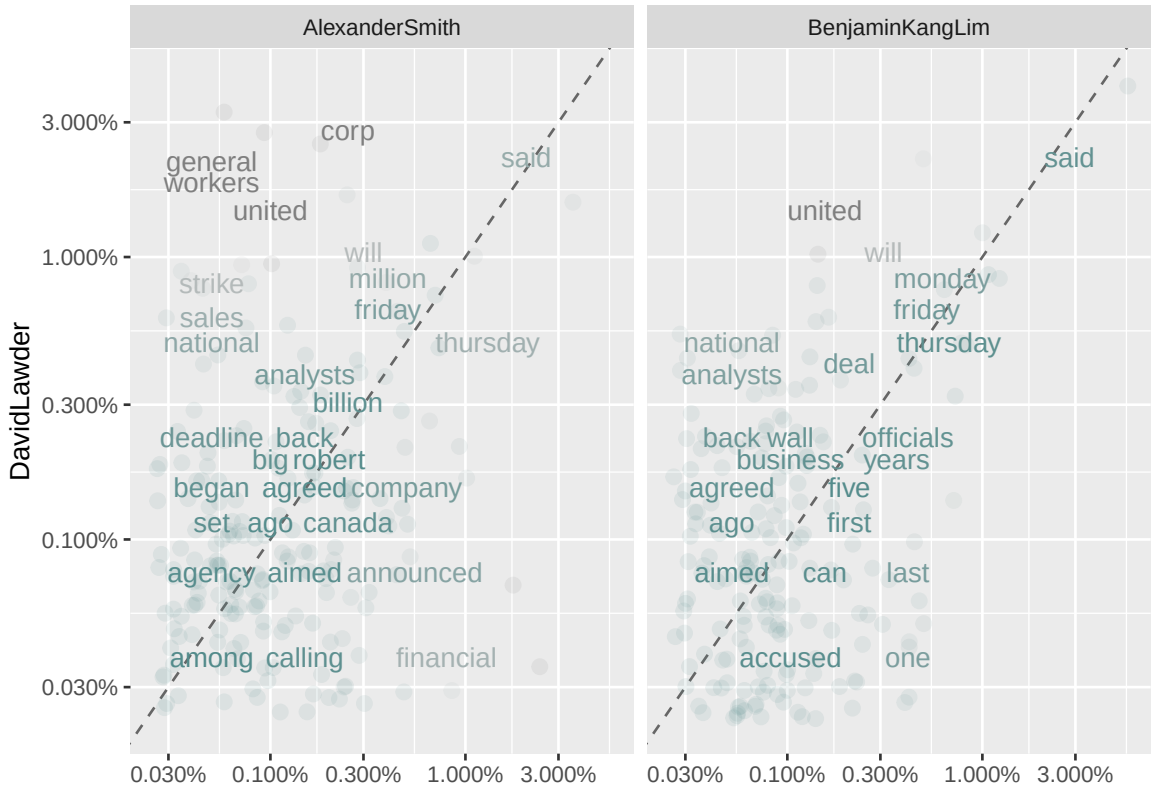
El análisis de la frecuencia de las palabras tanto a nivel del corpus completo como a nivel de cada categoría aporta mucha información sobre el vocabulario. Esta información se completa y enriquece con la observación de las palabras que arrojan valores altos de tfIdf. Un ejemplo de ello es el término *british*. Se trata de una palabra muy frecuente a nivel de corpus (en la posición 19 de las mas frecuentes) lo que podría llevar a pensar que es un término transversal utilizado frecuentemente por varios autores. Al individualizar el análisis de las frecuencias para los autores resulta ser la palabra mas utilizada por *Alexander Smith* y, observando el gráfico del valor tfIdf para este autor, resulta ser la segunda palabra con mayor tfIdf (estando además la primera, *britains*, muy relacionada), lo que implica que, comparativamente, es mucho mas utilizada por Alexander Smith que por el resto de autores, además se da la circunstancia de que no aparece entre las 30 palabras mas frecuentes ni con mayores valores tfIdf para ningún otro autor, por lo que, en definitiva, aun siendo una palabra muy frecuente a nivel de corpus, puede resultar muy indicativa para identificar los textos de *Alexander Smith*.

Tras esta primera aproximación tenemos una primera idea acerca de la temática propia de las noticias de cada uno de los autores así como del vocabulario mas representativo de cada uno. También resultará muy interesante realizar una comparativa de la distribución del vocabulario compartido entre autores, es decir, analizar las diferencias y similitudes en las frecuencias de uso de las palabras que aparecen en varios autores.

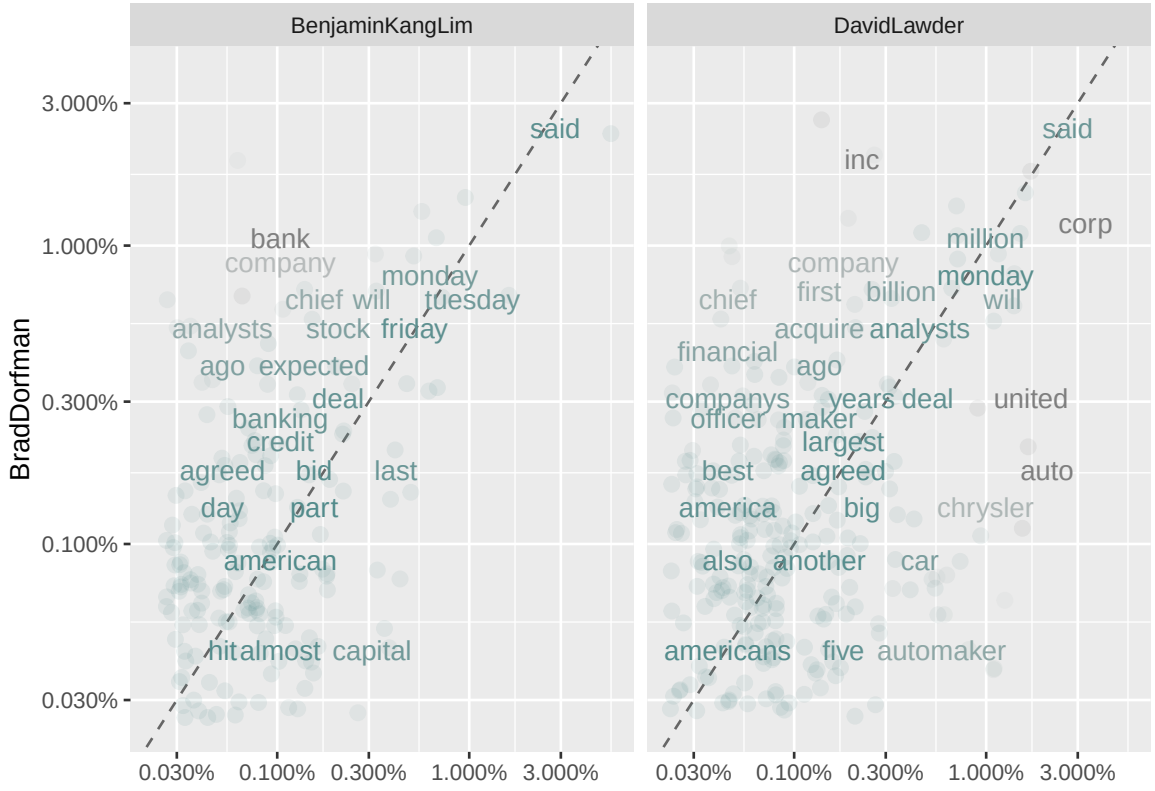
En los siguientes gráficos se realiza la comparación, dos a dos, de esta distribución. La escala corresponde a la frecuencia relativa de los términos en cada uno de los autores comparados, que están representados en cada uno de los ejes. De esta forma las palabras cercanas a la línea de puntos diagonal tienen una frecuencia similar en ambos autores, mientras que si el término se encuentra alejado de la diagonal será mas utilizada, proporcionalmente, por uno de ellos frente al otro. Asimismo, la mayor o menor dispersión de las palabras respecto de la línea de puntos diagonal nos da una idea acerca del grado de diferencia en la frecuencia de uso del conjunto del vocabulario compartido entre los dos autores.



De la observación de estas representaciones gráficas podemos extraer algunas conclusiones que pueden aportar información adicional. En el primer gráfico observamos, por ejemplo, que la palabra *government* es utilizada por Alexander Smith con una frecuencia que no alcanza el 0,1 % mientras que en las noticias de Benjamin Kang Lim aparece con una frecuencia que supera el 0.3 %, mas de tres veces superior. Por otra parte las palabras *chief*, *weed*, *monday* o *tuesday* se encuentran prácticamente sobre la diagonal lo que indica que son empleadas emplean con frecuencias muy similares por ambos autores. En el segundo gráfico destaca como Brad Dorfman emplea la palabra *inc* con una frecuencia muy superior a Alexander Smith, mientras que se da el caso contrario respecto de *investment*, *group* o *property*.



En este caso observamos en la parte superior izquierda del primer gráfico las palabras *general*, *corp*, *workers* y *united* lo que indica que son mucho mas empleadas por Alexander Smith que por David Lawder mientras la palabra *financial* es mucho mas empleada por este último que por el primero. Respecto de la comparativa entre Benjamin Kang Lim y David Lawder este último emplea las palabras *united*, *national* y *analysts* con mayor frecuencia que el primero.



Por último observamos que Benjamin Kang Lim emplea *capital* con mayor frecuencia que Brad Dorfman, mientras que este último emplea mas frecuentemente *bank*, *company* o *analysts* y, por otra parte, Brad Dorfman hace un uso mas frecuente de *inc* que David Lawder que lo supera en cuanto al empleo de *united* o *auto*.

En definitiva, hemos comprobado que el vocabulario mas empleado en el conjunto del corpus difiere sustancialmente del vocabulario mas frecuente restringido a cada uno de los autores, es decir, cada autor utiliza un conjunto de palabras mas frecuentes diferente, en térmminos generales, al empleado por los demás autores. Además hemos visto como el empleo de la frecuencia de aparición de los términos ponderada por la frecuencia inversa de los documentos (aplicando la función *TfIdf* que minora considerablemente las frecuencias asociadas a los términos que aparecen con una frecuencia elevada de manera transversal en todo el corpus) facilita la identificación de los términos que, teniendo o no frecuencias elevadas, son representativos de cada uno de los autores.

6.2.2. Análisis de correspondencias.

El análisis de las frecuencias, los valores tfIdf y la comparativa de las frecuencias de utilización del vocabulario nos ha permitido detectar las palabras mas relevantes, tanto a nivel de corpus como para cada uno de los autores, así como la presencia de una temática subyacente en cada autor. Por lo tanto, ahora resultará de interés profundizar en el análisis de las diferencias entre autores. Para ello, agruparemos los documentos en función de su autor y construiremos la correspondiente tabla léxica agregada que da lugar, una vez aplicadas las tareas de preprocesado, a la siguiente matriz de *Categorías* x *Palabras*:

```
## <<DocumentTermMatrix (documents: 4, terms: 1990)>>
## Non-/sparse entries: 2914/5046
## Sparsity           : 63%
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

El análisis de correspondencias de esta matriz nos permitirá la exploración y visualización de la relación entre sus filas y columnas, es decir, entre los autores y el vocabulario. La siguiente salida proporciona los valores propios y porcentajes de inercia calculados al aplicar el análisis de correspondencias.

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.6372633              38.58889                           38.58889
## dim 2  0.5583585              33.81089                           72.39978
## dim 3  0.4557946              27.60022                          100.00000
```

La segunda columna indica el porcentaje de la inercia que conserva cada uno de los ejes obtenidos. El primer eje conserva un 38.59 % de la inercia total, el segundo conserva el 33.81 % y el tercero el 27.6 %. El primer eje acumula un porcentaje de variabilidad superior al que le correspondería proporcionalmente en detrimento del tercer eje.

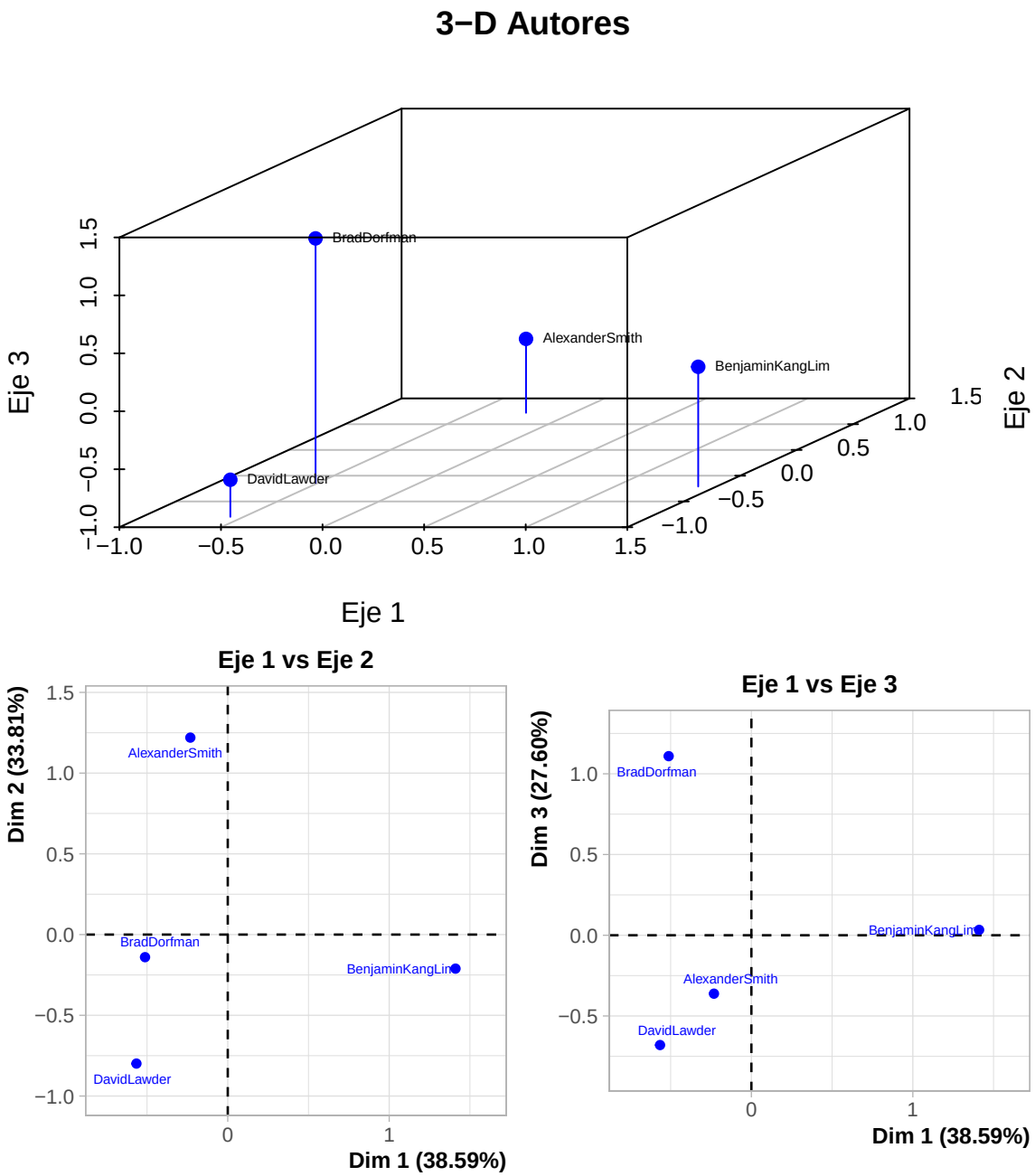
A continuación analizaremos la contribución de los autores en la construcción de estos tres ejes:

```
##                      Dim 1     Dim 2       Dim 3
## AlexanderSmith    2.087152 66.023678  7.11512102
## BenjaminKangLim 74.264525  1.894197  0.05769461
## BradDorfman      9.928563  0.845777 65.10779492
## DavidLawder     13.719761 31.236348 27.71938946
```

Benjamin Kang Lim es el autor que contribuye mayoritariamente a la construcción del primer eje con un 74,26 %, *Brad Dorfman* es el autor que contribuye mayoritariamente a la construcción del tercer eje con un 65,11 % y *Alexander Smith* es el que aporta mas en la construcción del segundo eje con un 66,02 %, mientras que *David Lawder* es el autor que contribuye en segunda posición a la construcción del primer eje, aunque tan solo en un 13,72 %, pero tiene una aportación importante en la construcción de los ejes segundo y tercero ocupando en ambos la segunda posición con un 31,24 % y un 27,72 % respectivamente. En otras palabras, el primer eje se ha construido mayoritariamente con la información de *Benjamin Kang Lim*, el segundo eje con la información de *Alexander Smith* y, en menor medida, de *David Lawder*, y el tercer eje principalmente con la información de *Brad Dorfman* y, en menor medida, de *David Lawder*.

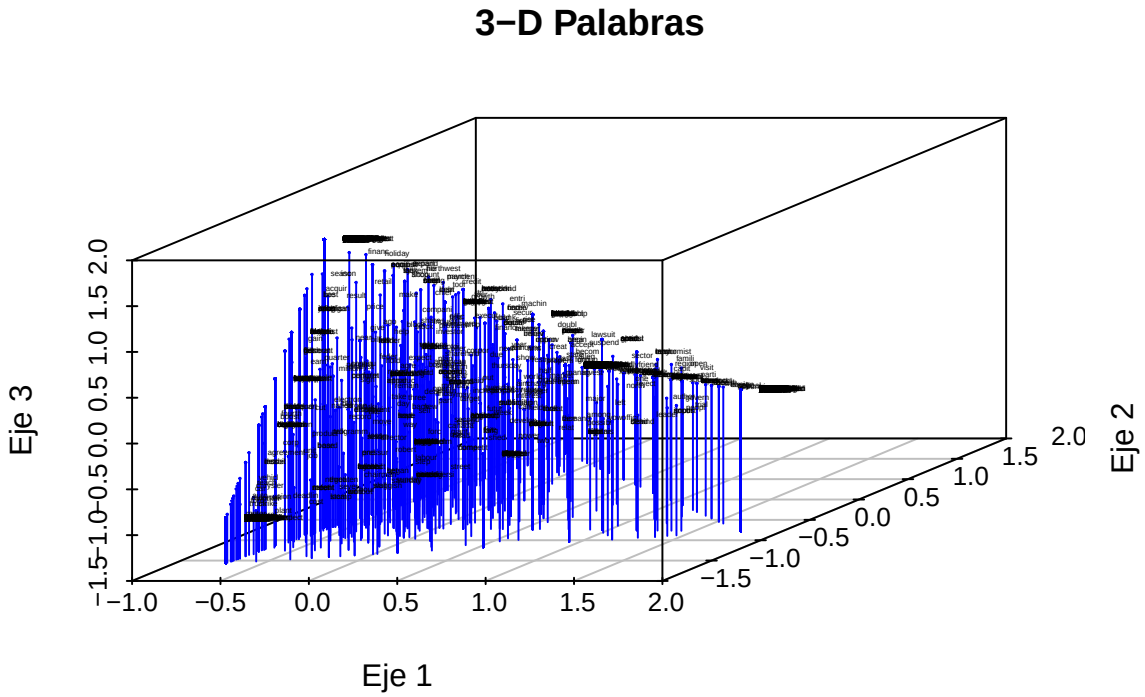
Por tanto, el gráfico bidimensional de los autores sobre los ejes 1 y 2 proporcionará una buena representación de Benjamin Kang Lim respecto del primer eje y de Alexander Smith y David Lawder sobre el segundo, mientras que en el gráfico sobre los ejes 1 y 3 obtenemos una buena representación de Brad Dorfman y David Lawder sobre el tercer eje. En la representación bidimensional sobre los ejes segundo y tercero tendremos una buena representación de Alexander Smith, Brad Dorfman y David Lawder. En otras palabras, el primer eje proporciona la mejor representación posible de Benjamin Kang Lim, el segundo eje la de Alexander Smith y David Lawder y el tercer eje la de Brad Dorfman.

Se muestran a continuación la representación tridimensional seguida de los gráficos bidimensionales sobre los ejes 1-2 y 1-3 (suficientes para representar de manera óptima a los cuatro autores).



Vemos como el primer eje situa a *Benjamin Kang Lim* en un extremo frente al resto de autores que ocupan una posición similar cercana al extremo contrario. El segundo eje contrapone a *Alexander Smith* y a *David Lawder* mientras que situa a *Benjamin Kang Lim* y a *Brad Dorfman* en una posición cercana a valores nulos. Por último el tercer eje reproduce una forma similar al segundo eje contraponiendo en este caso a *Brad Dorfman* y a *David Lawder*.

La representación tridimensional de las columnas (palabras) sobre el subespacio generado por las tres direcciones obtenidas es la siguiente:



En las primeras fases de exploración del corpus hemos identificado las palabras mas relevantes. Vamos a aprovechar este conocimiento para intentar reducir el número de palabras representadas y al mismo tiempo facilitar la exploración de los datos y su interpretación.

En primer lugar procederemos a identificar las palabras que mas han contribuido a la construcción de los ejes y, una vez identificadas, determinar si se encuentran entre las mas representativas (atendiendo al valor tfIdf y a la frecuencia) para cada uno de los autores.

Las palabras que cuentan con una aportación $\geq 0{,}25\,\%$ en la construcción de alguno de los tres ejes son las siguientes:

```
##   [1] "acquir"      "acquisit"     "agreement"    "alleg"        "appeal"
##   [6] "arm"         "asset"        "auto"         "automak"      "bancorp"
##  [11] "bancshar"    "bank"         "batteri"      "beij"         "bell"
##  [16] "bookkeep"    "boss"         "britain"      "british"      "cabl"
##  [21] "car"         "caterpillar"  "chen"         "chicago"      "chief"
##  [26] "china"       "chines"       "chrysler"     "communist"    "compani"
##  [31] "consum"      "contract"     "corp"         "cos"          "countri"
##  [36] "court"       "creat"        "credit"       "dan"          "death"
##  [41] "deeper"      "deng"         "depart"       "detain"       "deutsch"
##  [46] "diplomat"    "dissid"       "door"         "european"     "expand"
##  [51] "export"      "financ"       "financi"      "firm"         "first"
##  [56] "ford"        "fund"         "general"      "gid"          "govern"
##  [61] "grenfel"     "group"        "harnischfeg"  "holiday"      "horlick"
##  [66] "human"       "inc"          "invest"       "island"       "jail"
##  [71] "jiang"       "korean"       "labor"        "leader"       "least"
##  [76] "lewi"        "london"       "make"         "manag"        "market"
##  [81] "mercantil"   "mercuri"      "merger"       "mgam"         "million"
```

```
##  [86] "morgan"       "mother"       "motor"        "mutual"       "natwest"
##  [91] "newspap"      "nicola"       "northwestern" "nynex"        "offici"
##  [96] "overthrow"    "paramount"    "parti"        "pension"      "peopl"
## [101] "plant"        "plc"          "pledg"        "plot"         "polic"
## [106] "post"         "pound"        "pretax"       "probe"        "propaganda"
## [111] "properti"     "result"       "retail"       "revamp"       "riot"
## [116] "roebuck"      "sale"         "say"          "sear"         "season"
## [121] "sentenc"      "shortterm"    "son"          "spend"        "stock"
## [126] "store"        "strike"       "sunday"       "system"       "taiwan"
## [131] "tcf"          "textil"       "tibet"        "tire"         "trade"
## [136] "trial"        "troubl"       "truck"        "trust"        "union"
## [141] "unit"         "vehicl"       "visit"        "wang"         "wireless"
## [146] "worker"       "worth"        "xiaop"        "xinjiang"     "year"
## [151] "zemin"
```

Veamos las diez que mas han contribuido a la construcción de cada uno de los ejes:

```
##  [1] "china"     "chines"    "beij"      "dissid"    "corp"      "sentenc"
##  [7] "taiwan"    "wang"      "communist" "court"
```

En el primer eje destacan palabras relacionadas con China, muy representativas de las noticias de *Benjamin Kang Lim*.

```
##  [1] "british" "manag"   "motor"   "fund"    "britain" "corp"    "bank"
##  [8] "general" "invest"  "auto"
```

En el segundo eje la palabra mas importante es *british*, fuertemente asociada a *Alexander Smith*. Entre las demás se encuentran palabras características de este mismo autor (como *fund* o *britain*) así como de *David Lawder* (como *motor* o *auto*).

```
##  [1] "inc"     "motor"   "financ"  "general" "mercuri" "store"   "worker"
##  [8] "auto"    "retail"  "bancorp"
```

En el tercer eje la palabra mas importante es *inc*, una de las palabras mas frecuente y de mayor tfIdf de *Brad Dorfman*. El resto son palabras igualmente representativas de este autor (como *financ* o *mercuri*) y de *David Lawder* (como *motor* o *general*).

El estudio de las frecuencias y valores *tfIdf* de las palabras por autores es una buena manera de identificar las palabras mas características o representativas de cada uno puesto que observamos como las mayores aportaciones de las palabras en la construcción de cada uno de los ejes se corresponde, en lineas generales, con las mas representativas de los autores que mas contribuyen a la construcción de cada uno de los ejes.

Veamos ahora cuantas de estas palabras que mas contribuyen a la construcción de los ejes se encuentran entre las mas representativas de cada uno de los autores. La siguiente tabla resume numéricamente el resultado de cruzar las palabras que cuentan con una aportación $\geq 0{,}25\,\%$ a cualquiera de los ejes con el mismo número de palabras (151) con mayor frecuencia y mayor valor *tfIdf* de cada autor:

```
##
##                    1  2  3
##   AlexanderSmith   2 11  1
##   BenjaminKangLim  8  2  1
##   BradDorfman      1  5  9
##   DavidLawder      3  7  7
```

```
## El  37.75  % ( 57  palabras) de las que mas aportan a la construcción de
##  los ejes se encuentra entre las 151 palabras de mayor frecuencia o de
##  mayores valores tfIdf de alguno de los cuatro autores
```

Esta tabla muestra como se reproduce la influencia de los autores sobre cada uno de los ejes en la contribución de las palabras. El autor que mas contribuye a la construcción del primer eje es *Benjamin Kang Lim* y 8 palabras de las que mayor influencia tienen en la construcción de este eje se encuentran entre las mas destacadas, por su valor TfIdf, para este autor frente a 3 de *David Lawder*, 2 de *Alexander Smith* y 1 de *Brad Dorfman*. En cuanto al segundo eje, los autores que mas contribuyen a su construcción son *Alexander Smith* y *David Lawder* y observamos como, respectivamente, 11 y 7 palabras de las que mas contribuyen se encuentran entre las mas representativas de estos autores frente a 5 de *Brad Dorfman* y 2 de *Benjamin Kang Lim*. Por último, en la construcción del tercer eje la mayor contribución corresponde a *Brad Dorfman* y *David Lawder* y en la contribución de las palabras observamos, respectivamente, 9 y 7 palabras representativas de estos autores frente a 1 de cada uno de los restantes.

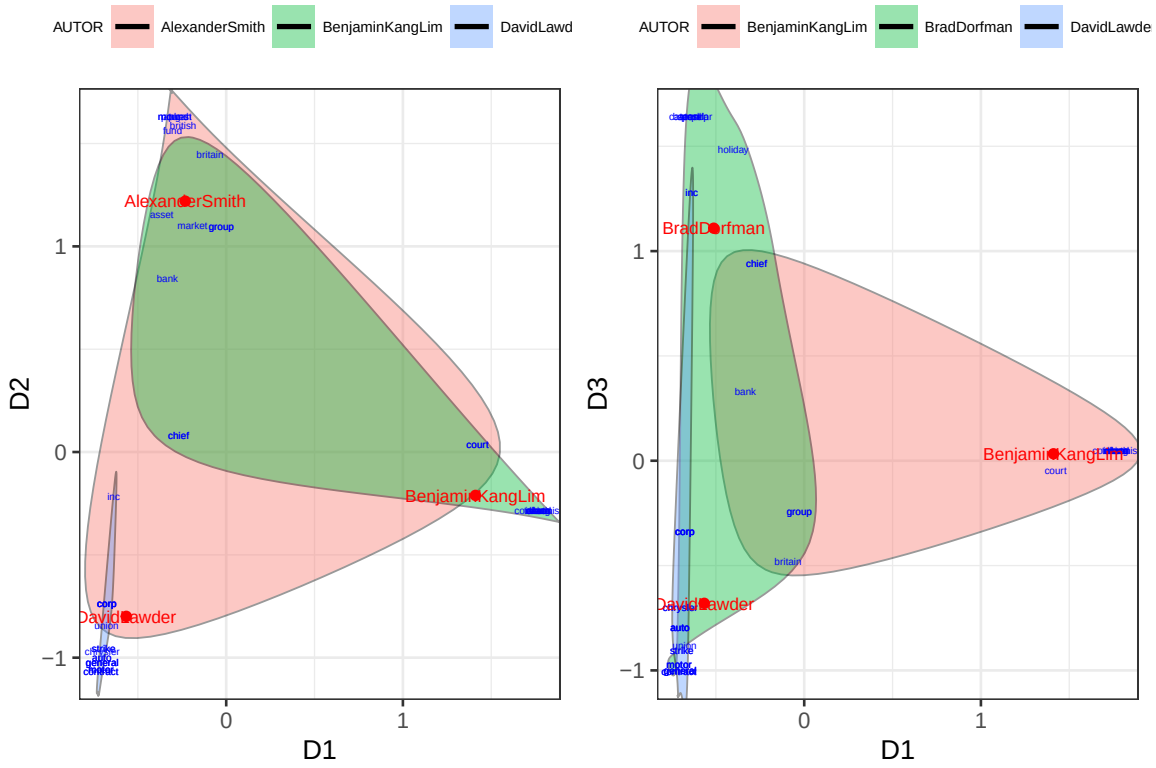
Estas palabras pueden contribuir a la construcción de uno o mas ejes y hacerlo por parte de uno o mas autores, con lo que estas 57 se reducen a las 43 siguientes:

```
##  [1] "china"       "inc"         "british"     "motor"       "motor"
##  [6] "general"     "fund"        "britain"     "corp"        "bank"
## [11] "general"     "auto"        "group"       "morgan"      "corp"
## [16] "store"       "market"      "auto"        "taiwan"      "wang"
## [21] "plc"         "bancorp"     "communist"   "court"       "strike"
## [26] "motor"       "caterpillar" "contract"    "strike"      "dan"
## [31] "island"      "holiday"     "pound"       "union"       "post"
## [36] "contract"    "general"     "chief"       "death"       "chrysler"
## [41] "natwest"     "spend"       "asset"
```

```
##    PALABRA DIMENSION CONTRIBUCION
## 15   china         1     4.246520
## 37     inc         3     2.497152
## 5  british         2     2.463342
```

La palabra que mas ha aportado a la construcción del primer eje ha sido *china* con un $4{,}25\,\%$. La de mayor contribución al segundo eje ha sido *british* con un $2{,}46\,\%$ y respecto del tercer eje la primera posición la ha ocupado *inc* con un $2{,}5\,\%$.

Las representaciones bidimensionales conjuntas de los autores y sus palabras mas representativas sobre los ejes obtenidos 1-2 y 1-3 son las siguientes:



6.2.3. Escalado multidimensional.

Para aplicar las técnicas de escalado multidimensional es necesario partir de una matriz de similaridades o disimilaridades. En el caso de la tabla léxica agregada, aplicando la distancia del coseno, la matriz de distancias obtenida es la siguiente:

```
##                 AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
## AlexanderSmith        0.000000        0.696476    0.529259    0.739267
## BenjaminKangLim       0.696476        0.000000    0.673665    0.755474
## BradDorfman           0.529259        0.673665    0.000000    0.521640
## DavidLawder           0.739267        0.755474    0.521640    0.000000
```

En cuanto al número idóneo de dimensiones a considerar lo podemos determinar en función de los valores propios positivos de la matriz anterior teniendo en cuenta que su suma representa la explicación del 100 % de la varianza.

```
## [1]  47.15314  85.76762 100.00000 100.00000
```

Por tanto, la configuración bidimensional conservará el 85,77 % de la variabilidad de la matriz.

Vamos ahora aplicar las modalidades métrica y no métrica del escalado multidimensional y comparar los resultados obtenidos.

```
##  AlexanderSmith     DavidLawder     BradDorfman BenjaminKangLim
##        30.30229        29.24228        28.70841        11.74701
```

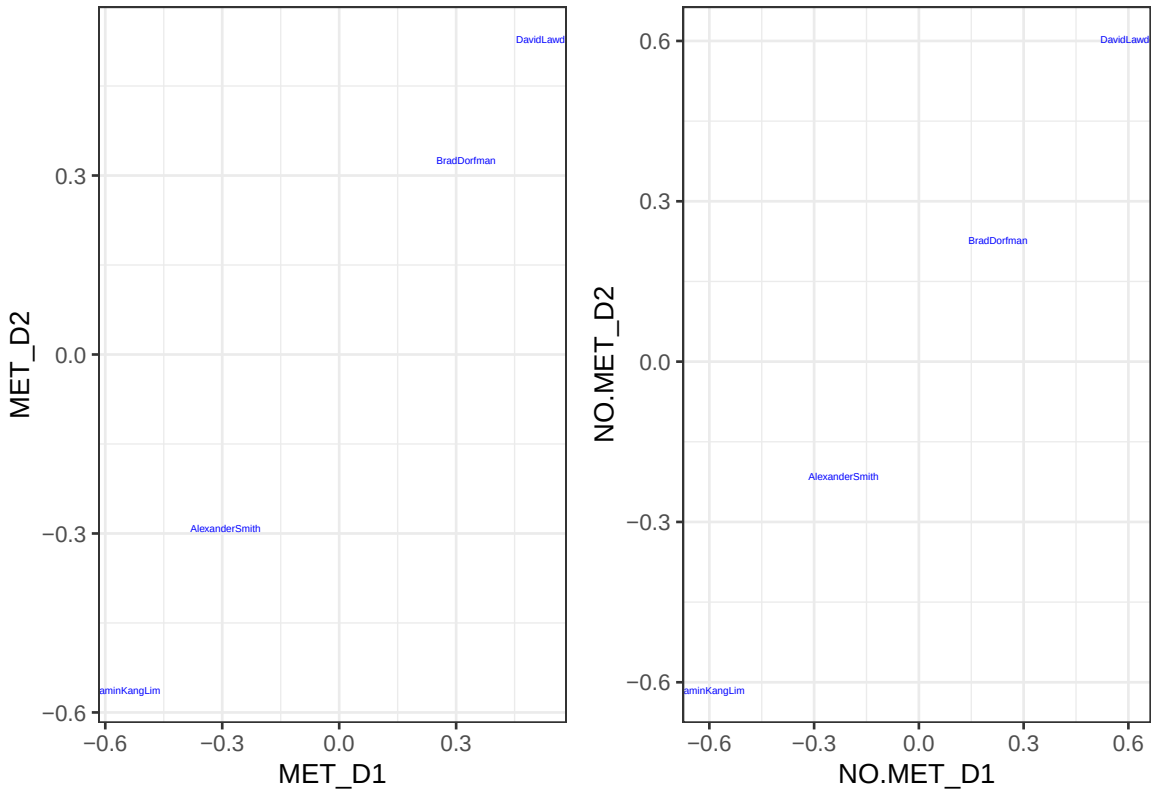
```
##     DavidLawder BenjaminKangLim  AlexanderSmith     BradDorfman
##    4.999998e+01    2.503706e+01    2.496295e+01    1.657732e-05
```

```
##     TIPO    BONDAD ITERACIONES
## 1    MET 0.9023460          22
## 2 NO.MET 0.9991735           8
```

Con el método métrico el algoritmo necesita 22 iteraciones para alcanzar el criterio de convergencia y el modelo ajustado conserva un 90,23 % de la variabilidad de la matriz de distancias mientras que el método no métrico alcanza el criterio de convergencia con tan solo 8 iteraciones conservando el 99,92 % de la variabilidad.

Veamos una comparativa de ambas representaciones:

```
##                     MET_D1     MET_D2  NO.MET_D1  NO.MET_D2
## AlexanderSmith  -0.2914495 -0.2914495 -0.2152174 -0.2152174
## BenjaminKangLim -0.5624844 -0.5624844 -0.6146073 -0.6146073
## BradDorfman      0.3257546  0.3257546  0.2268681  0.2268681
## DavidLawder      0.5281793  0.5281793  0.6029566  0.6029566
```



Al aplicar el análisis de correspondencias habíamos identificado el conjunto de palabras cuya aportación a la construcción de cualquiera de los ejes era $\geq 0{,}25\,\%$. Se trata de 151 palabras con las que vamos a construir la correspondiente matriz de distancias y a continuación determinaremos el número de dimensiones idóneo para aplicar el escalado multidimensional.

```
##   [,1]            [,2]            [,3]            [,4]
## h "46.51"         "79.42"         "99.9"          "100"
##   "..."       ..." "..."       ..." "..."       ..." "..."       ..."
## t "100"           "100"           "100"           "100"
```

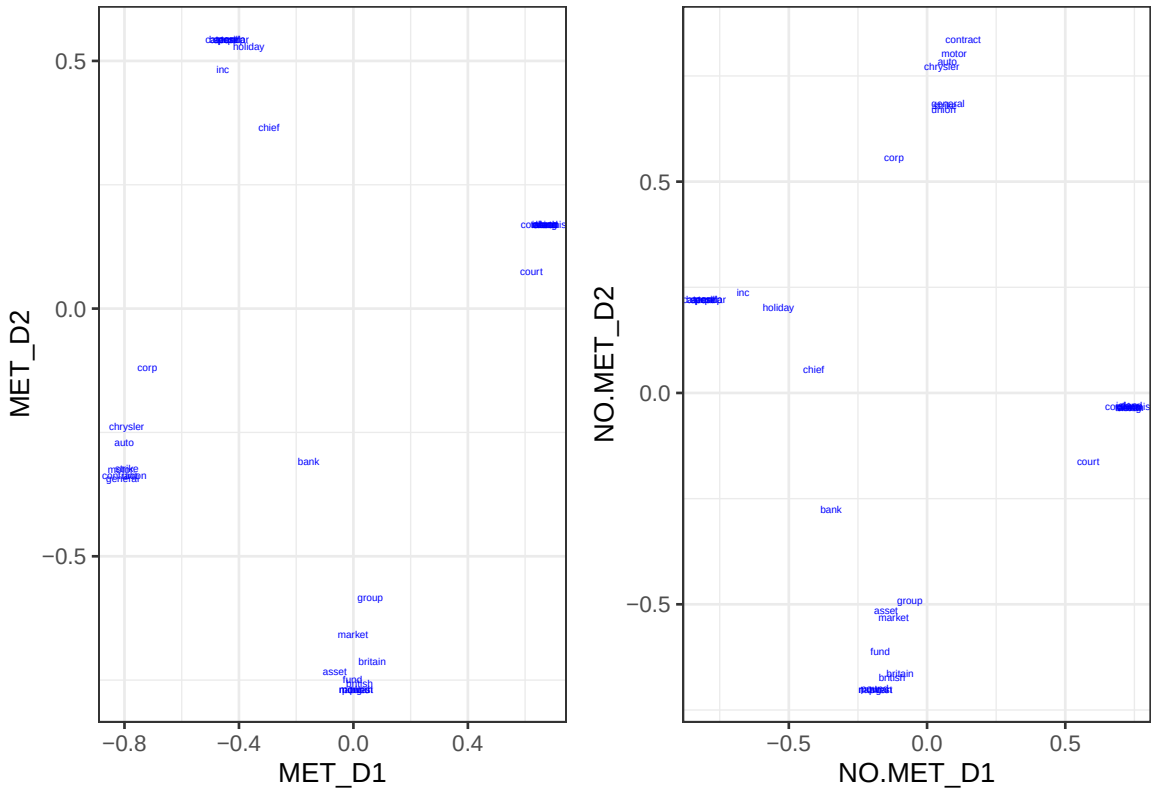
La representación bidimensional conservará el 79,42 % de la variabilidad, alcanzándose el 99,9 % en el caso tridimensional.

Aplicamos de nuevo los métodos métrico y no métrico y comparamos los resultados.

```
##     TIPO    BONDAD ITERACIONES
## 1    MET 0.8277487          66
## 2 NO.MET 0.9159765          90
```

Con el método métrico el algoritmo alcanza el criterio de convergencia tras 66 iteraciones conservando un 82,77 % de la variabilidad de la matriz de distancias mientras que el método no métrico necesita 90 iteraciones para alcanzar el criterio de convergencia pero conserva el 91,61 % de la variabilidad.

Veamos una comparativa de las representaciones para las palabras seleccionadas:



6.3. Clasificación mediante Análisis Discriminante Lineal.

Como ya sabemos, el análisis discriminante lineal resulta óptimo bajo el supuesto de normalidad multivariante de las variables aunque en la práctica son habituales las situaciones en que no cumpliéndose está condición proporciona buenos resultados.

Veamos en primer lugar los resultados de aplicar el análisis discriminante directamente sobre la matriz original con las frecuencias absolutas de las palabras por documento. Obtenemos la siguiente matriz de confusión:

```
##
##                   AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   AlexanderSmith              23              19          33          25
##   BenjaminKangLim             16              38          20          26
##   BradDorfman                 24              28          26          22
##   DavidLawder                 22              30          23          25
```

con los siguientes valores sobre la bondad del ajuste realizado

```
## [1] "Precision: 0.28"
```

```
## [1] "Adjusted Rand Index: 0.00330537515882275"
```

Teniendo en cuenta que contamos con cuatro autores y el mismo número de documentos de cada uno, un modelo de clasificación aleatorio puro tendría una precisión del 25 %. Se trata por tanto de un ajuste muy pobre. El índice ARI arroja un valor muy cercano a cero indicando también que el modelo es casi equivalente al azar.

A partir de esta matriz de confusión podemos obtener los siguientes indicadores para cada autor:

```
##                  Precis Recup
## AlexanderSmith     0.27  0.23
## BenjaminKangLim    0.33  0.38
## BradDorfman        0.25  0.26
## DavidLawder        0.26  0.25
```

Estos indicadores nos permiten interpretar la bondad del clasificador construido para cada uno de los autores. La primera columna contiene el índice de *precisión* (proporción de documentos de cada autor correctamente clasificados), mientras que en la segunda columna se encuentra el índice de *recuperación* (proporción de documentos clasificados en un autor que pertenecen al mismo).

Se trata ahora de encontrar, mediante escalado multidimensional, una configuración basada en la matriz original de documentos-términos (no agregada) sobre la que aplicar la técnica del análisis discriminante para construir un modelo que sirva para identificar al autor de cada documento.

Para ello, calculamos en primer lugar las similaridades entre documentos con la distancia del coseno y aplicamos la transformación $1 - cos(x)$ para obtener una matriz de disimilaridades. A partir de los valores propios de esta matriz determinamos el número de dimensiones apropiado que conserve un porcentaje suficiente de la variabilidad.

```
##   [1]  6.29  9.94 13.31 16.18 18.45 20.70 22.66 24.56 26.39 28.15 29.82 31.46
##  [13] 33.04 34.55 35.95 37.30 38.59 39.86 41.06 42.25 43.35 44.40 45.45 46.44
##  [25] 47.41 48.35 49.27 50.17 51.05 51.91 52.72 53.53 54.32 55.11 55.86 56.61
##  [37] 57.34 58.07 58.78 59.49 60.15 60.79 61.42 62.04 62.64 63.23 63.82 64.40
##  [49] 64.97 65.54 66.10 66.65 67.19 67.72 68.24 68.76 69.25 69.73 70.21 70.68
##  [61] 71.14 71.60 72.05 72.50 72.94 73.37 73.79 74.22 74.63 75.04 75.45 75.85
##  [73] 76.24 76.62 77.00 77.37 77.74 78.10 78.46 78.81 79.16 79.49 79.83 80.16
##  [85] 80.49 80.81 81.13 81.45 81.76 82.06 82.37 82.66 82.96 83.24 83.53 83.81
##  [97] 84.09 84.36 84.63 84.90 85.16 85.42 85.68 85.94 86.19 86.44 86.69 86.93
## [109] 87.17 87.41 87.65 87.88 88.12 88.35 88.57 88.79 89.01 89.23 89.44 89.65
## [121] 89.86 90.07 90.27 90.46 90.66 90.85 91.04 91.23 91.41 91.59 91.77 91.95
## [133] 92.13 92.30 92.48 92.65 92.81 92.97 93.13 93.29 93.45 93.60 93.75 93.90
## [145] 94.05 94.19 94.34 94.48 94.61 94.75 94.88 95.01 95.14 95.27 95.39 95.51
## [157] 95.63 95.75 95.87 95.98 96.10 96.21 96.32 96.43 96.53 96.63 96.73 96.83
## [169] 96.93 97.03 97.12 97.21 97.30 97.39 97.48 97.56 97.65 97.73 97.81 97.89
## [181] 97.97 98.04 98.12 98.19 98.26 98.33 98.40 98.46 98.53 98.59 98.65 98.71
## [193] 98.77 98.82 98.88 98.93 98.99 99.04 99.08 99.13
```

Con 198 dimensiones alcanzamos el 99 % de la variabilidad de la matriz. Por lo tanto, la matriz resultante tendrá menos del 10 % de columnas que la original (recordemos que contaba con 1990 variables correspondientes a las palabras conservadas tras el preprocesamiento del corpus).

Realizaremos a continuación el escalado multidimensional métrico mediante el algoritmo Smacof y aplicamos el análisis discriminante lineal sobre la configuración formada por las 198 primeras coordenadas. Obtenemos la siguiente matriz de confusión y los siguientes resultados sobre la evaluación del modelo resultante.

```
##
##                   AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   AlexanderSmith              94               2           2           2
##   BenjaminKangLim              5              86           1           8
##   BradDorfman                  1               0          91           8
##   DavidLawder                  3               0           7          90
```

```
## [1] "Precision: 0.9025"
```

```
## [1] "Adjusted Rand Index: 0.756246749996156"
```

La precisión ha aumentado hasta el 90,25 % y el índice ARI hasta 0,76.

Los indicadores individuales por autores resultan:

```
##                 Precis Recup
## AlexanderSmith    0.91  0.94
## BenjaminKangLim   0.98  0.86
## BradDorfman       0.90  0.91
## DavidLawder       0.83  0.90
```

En cuanto a la aplicación del escalado multidimensional en su versión no métrica obtenemos los siguientes resultados

```
##
##                   AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   AlexanderSmith              95               0           3           2
##   BenjaminKangLim              7              85           1           7
##   BradDorfman                  1               1          90           8
##   DavidLawder                  1               0           7          92
```

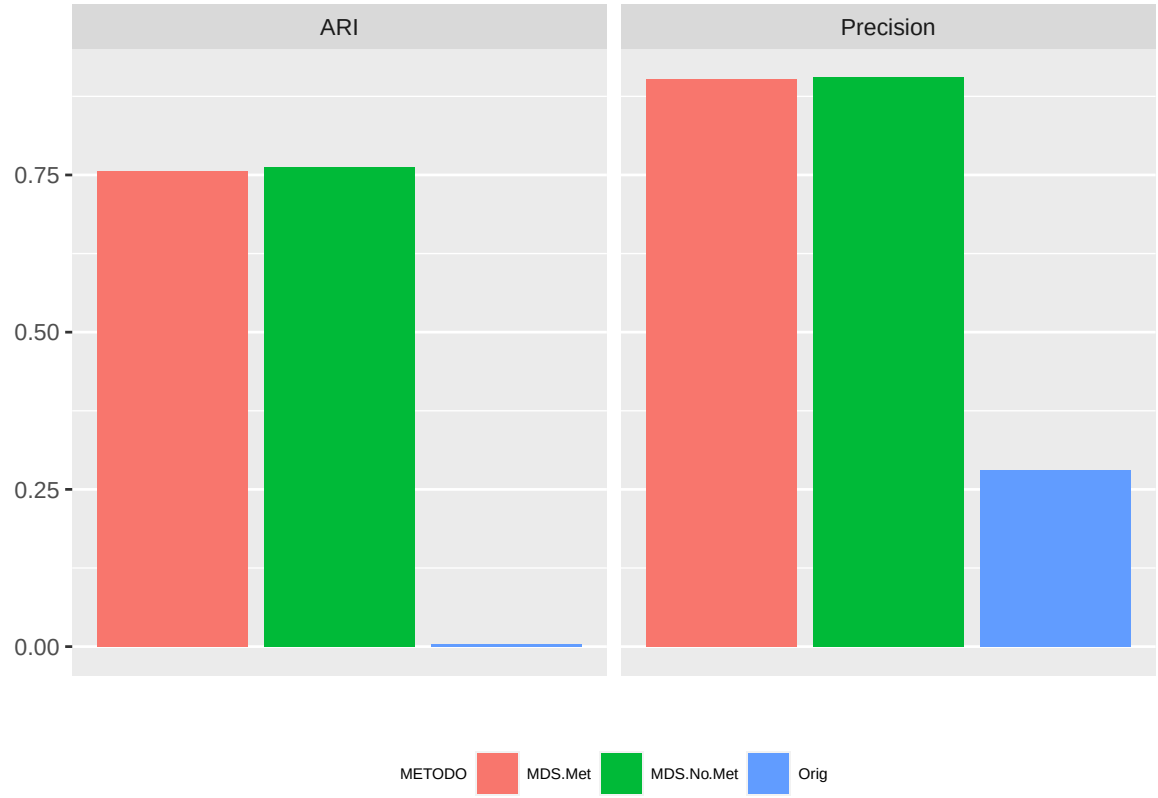
```
## [1] "Precision: 0.905"
```

```
## [1] "Adjusted Rand Index: 0.761964497894001"
```

```
##                 Precis Recup
## AlexanderSmith    0.91  0.95
## BenjaminKangLim   0.99  0.85
## BradDorfman       0.89  0.90
## DavidLawder       0.84  0.92
```

6.3.1. Conclusiones.

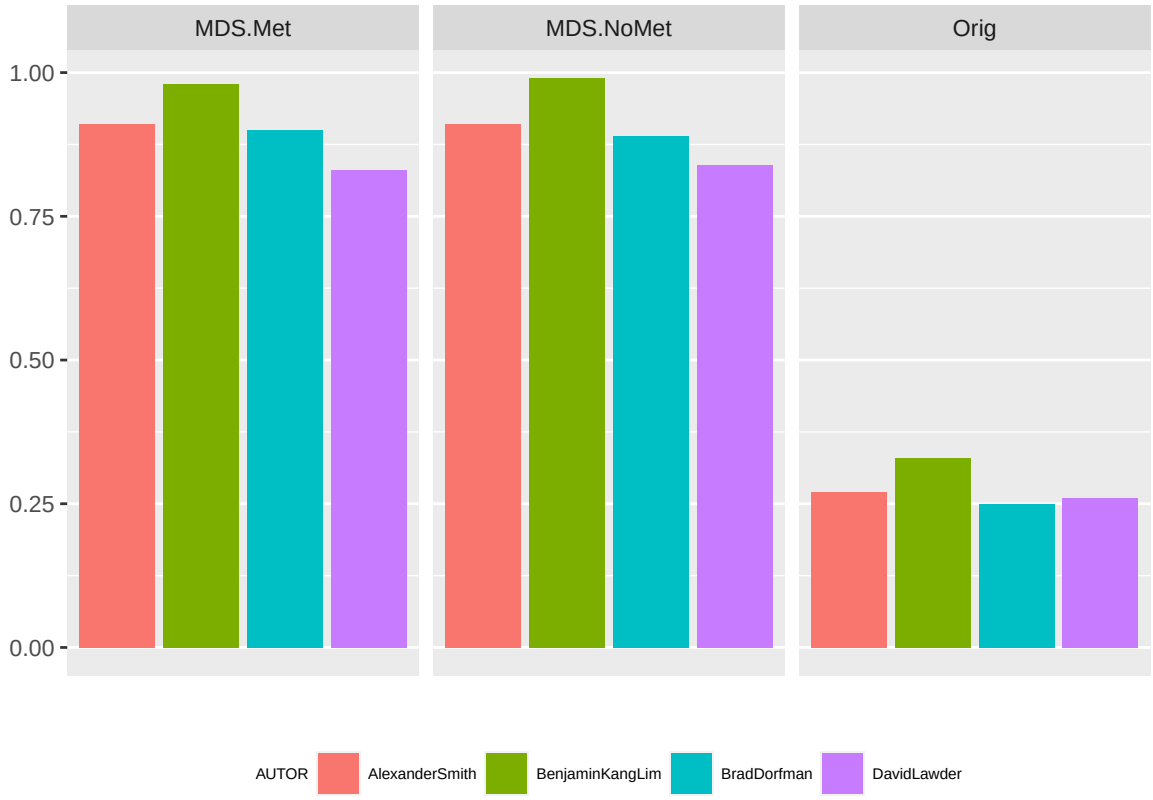
6.3.1.1. Resultados globales del ajuste.




```
##        MODELO PRECIS         ARI
## 1        Orig 0.2800 0.003305375
## 2    MDS.Met 0.9025 0.756246750
## 3 MDS.No.Met 0.9050 0.761964498
```

A nivel global se aprecia claramente la mejora producida al aplicar el análisis discriminante sobre la configuración resultante de un previo escalado multidimensional frente a su aplicación directa sobre la matriz de frecuencias original. En este último caso el índice ARI muestra que el modelo obtenido apenas difiere de un modelo aleatorio puro, contando con tan solo un 28 % de documentos correctamente clasificados, mientras que mediante el escalado multidimensional previo este índice supera el valor 0,75 tanto con el método métrico como con el no métrico con un porcentaje de documentos clasificados correctamente superior al 90 % en ambos casos.

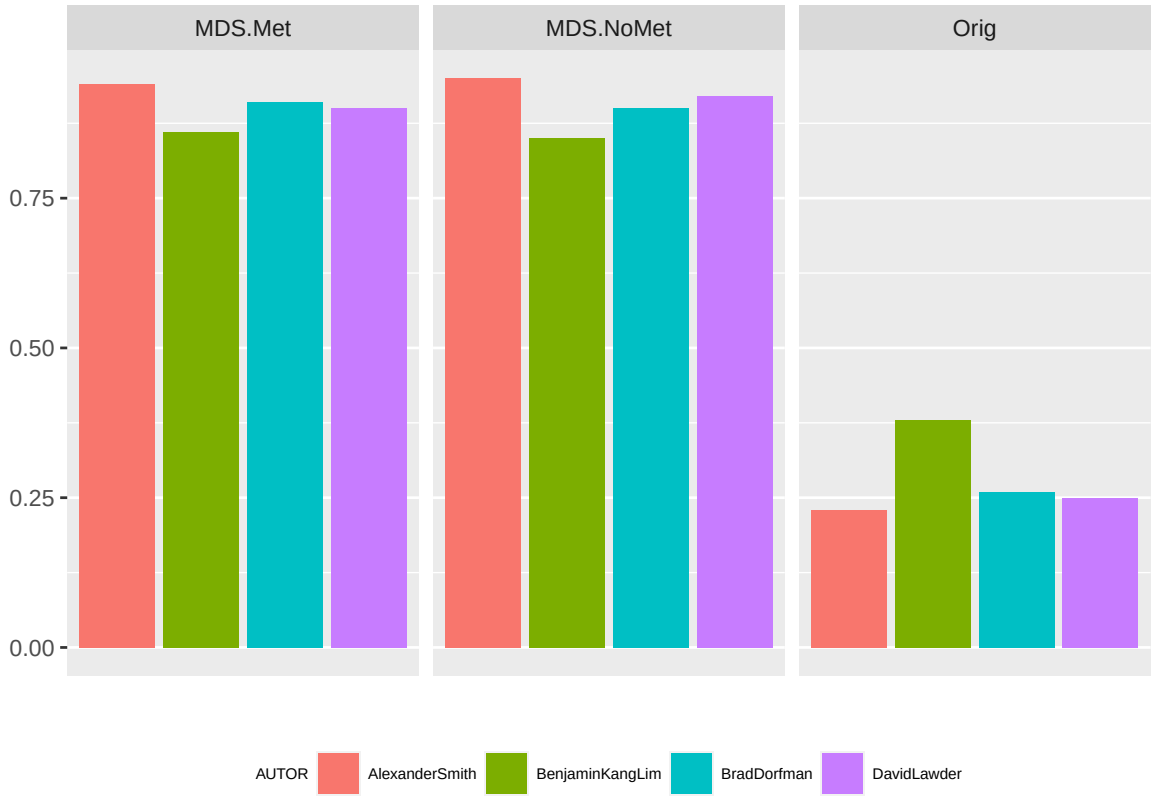
6.3.1.2. Resultados del ìndice de precisión por autores.




```
##                  Orig MDS.Met MDS.No.Met
## AlexanderSmith  0.27    0.91       0.91
## BenjaminKangLim 0.33    0.98       0.99
## BradDorfman     0.25    0.90       0.89
## DavidLawder     0.26    0.83       0.84
```

En todos los casos el autor con mayor proporción de sus documentos clasificados correctamente es *Benjamin Kang Lim* mientras que al que corresponde la menor proporción es *David Lawder* (salvo para la aplicación sobre los datos originales que es *Brad Dorfman*), con unos porcentajes de documentos clasificados correctamente del 33 y el 25 % respectivamente en el caso de la aplicación del análisis discriminante directo frente al 98 y 83 % en el caso del escalado métrico y el 99 y 84 % del no métrico.

6.3.1.3. Resultados del ìndice de recuperación por autores.




```
##                  Orig MDS.Met RECUP.MDS.No.Met
## AlexanderSmith   0.23    0.94             0.95
## BenjaminKangLim 0.38    0.86             0.85
## BradDorfman      0.26    0.91             0.90
## DavidLawder      0.25    0.90             0.92
```

Por contra, observamos como el autor con mayor y con menor proporción de documentos clasificados correctamente entre los que le son atribuidos se encuentra invertido entre la discriminación directa y la discriminación previo escalado de la matriz. En el caso de disccriminación lineal "directa" el autor con mayor proporción es *Benjamin Kang Lim* con un 38 % y el autor con menor proporción es *Alexander Smith* con un 23 % mientras que, previo escalado multidimensional, al que corresponde la mayor proporción es *Alexander Smith* con el 94 % en la modalidad métrica y el 95 % en la no métrica, y al que corresponde la menor proporción es *Benjamin Kang Lim* con el 86 % mediante escalado métrico y el 85 % mediante escalado no métrico.

En definitiva podemos concluir que la aplicación directa de la metodología del análisis discriminante clásico para obtener un modelo que permita clasificar los textos del corpus objeto de estudio en función de sus autores es poco útil cuando se aplica directamente sobre la matriz de documentos-términos directamente, con unos resultados que apenas superan los de un modelo aleatorio puro, mientras que la apliación a esta matriz de las técnicas de escalado multidimensional proporciona una configuración muy reducida (inferior al 10 % de las dimensiones de la matriz original) que produce una mejora sustancial de los resultados, superándose el 90 % de documentos clasificados correctamente con un índice de Rand ajustado mayor que 0,75.

Apéndice A

Apéndice: Implementación con R

La implementación con el lenguaje de programación R correspondiente al tratamiento y resultados mostrados en el capítulo 6 se encuentra en los siguientes enlaces de mi repositorio en gitHub:

- librerias Yfunciones.R
- codigo.R

Bibliografía

Karmele Fernández Aguirre. Análisis textual: generación y aplicaciones. *Metodología de encuestas*, 5(1):55–66, 2003.

Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling. Theory and Applications*. Springer Series in Statistics, 2005.

Mönica Bécue-Bertaut. *Minería de textos. Aplicación a preguntas abiertas en encuestas.* Cuadernos de Estadística, 2010.

Daniel Peña Sánchez de Rivera. *Estadística. Modelos y métodos. (1. Fundamentos).* Alianza Universidad Textos, 1997.

Daniel Peña Sánchez de Rivera. *Estadística. Modelos y métodos. (2. Modelos lineales y series temporales).* Alianza Universidad Textos, 1999.

Daniel Peña Sánchez de Rivera. *Análisis de Datos Multivariantes*. McGraw Hill, 2002.

José Manuel Ramírez Hurtado Flor María Guerrero Casas. El análisis de escalamiento multidimensional: una alternativa y un complemento a otras técnicas multivariantes. 03 2019.

Marcial Terrádez Gurrea. Frecuencias léxicas del español coloquial: Análisis cuantitativo y cualitativo. *Facultat de Filología, Universitat de València*, pages 37–42, 2001.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classifications*, (2):193–218, 1985.

F. Torres R. Gutiérrez, A. González. *Técnicas de Análisis de datos multivariable. Tratamiento computacional*. Alianza Universidad Textos, 1997.

Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. *Artificial Neural Networks - ICANN 2009*, pages 175–184, 2009.

Kari Torkkola. Linear discriminant analysis in document classification. *IEEE TextDM 2001*, 12 2001.