Minería de datos. Introducción y guía de estudio

PID_00237681

Patricia Carracedo
Manuel Terrádez

Universitat Oberta de Catalunya

Índice

1. Definición de minería de datos

Se podría discutir eternamente sobre el significado de la minería de datos sin llegar a un acuerdo.

Basta pasarse por la página que Wikipedia dedica a este concepto, tanto en su versión en inglés (https://en.wikipedia.org/wiki/Data_mining) como en castellano (https://es.wikipedia.org/wiki/Minería_de_datos), para constatarlo: en la versión en inglés se habla de que es un *misnomer* (término que induce a error) y un *buzzword* (término de moda durante un periodo de tiempo), se dice que es un subcampo multidisciplinario, se comienza dando otra serie de términos con los que no se debe confundir, etc.

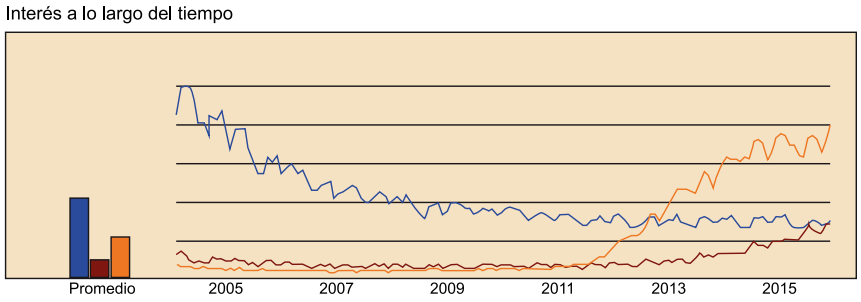
Para empezar, el propio término –que proviene de la traducción literal del término en inglés *data mining*– resulta curioso, pues combina dos conceptos con aparente poca relación entre sí, como son la minería y los datos.

Es evidente que la metáfora se basa en el paralelismo entre la actividad de la tradicional minería industrial, que consiste en explotar una mina para extraer de ella un mineral que tiene cierto valor y utilidad, y el proceso de extraer conocimiento (muy útil y valioso, sin duda) partiendo de información (datos que pueden ser desestructurados o de poca calidad).

Afortunado o no, lo cierto es que este término gozó de un gran éxito a finales del siglo pasado y comienzos de este, si bien recientemente ha sido algo desplazado por otros que han surgido con mucha fuerza, como por ejemplo *data science,* o la gran estrella dentro de este ámbito, *big data*: sirva de ejemplo que hasta el presidente de los EE. UU., Barack Obama, ha hablado recientemente (febrero de 2015) sobre estos términos en ocasión de la presentación del Dr. D. J. Patil como el primer *chief data scientist and deputy chief technology officer for datapolicy* del Gobierno de EUA.

El siguiente gráfico muestra la evolución en los últimos años de las búsquedas en Google relacionadas con *data mining* (azul), *data science* (rojo) *y big data* (amarillo).

Figura 1. Evolución de las búsquedas en Google relacionadas con *data mining* (azul), *data science* (rojo) y *big data* (amarillo) en el período 2005-2015



Fuente: elaboración propia.

Por el momento, podemos quedarnos con una definición simple que podríamos considerar de consenso, y que asimilaría la minería de datos al proceso de extraer patrones y relaciones útiles de grandes volúmenes de datos con el objetivo de obtener una estructura comprensible que permita transformar la información en conocimiento.

A lo largo de los próximos apartados profundizaremos en esta definición.

2. Relación entre Estadística, *Machine Learning, Data Mining, Data Science, Big Data* y otras disciplinas asociadas

En el principio existía la Estadística...

La práctica totalidad de los campos de conocimiento de los que vamos a hablar en este apartado tienen su origen común en la Estadística, disciplina cuyos inicios formales se remontan al siglo XVIII (http://www.statslife.org.uk/images/pdf/timeline-of-statistics.pdf) –aunque hay algunas referencias mucho más antiguas–, y que tradicionalmente se ha considerado una rama de las Matemáticas (de hecho, hasta la aparición de las primeras titulaciones específicas de Estadística a finales del siglo XX, solo se ofrecía por las universidades españolas como especialidad dentro de la licenciatura en Matemáticas), pese a que esta relación es hasta cierto punto discutible, porque si bien la mayoría de las técnicas estadísticas se basan en un rigor y unos fundamentos que provienen de otras ramas de las Matemáticas (Álgebra, por ejemplo), no es menos cierto que la Probabilidad y la Estadística no comparten la famosa «exactitud» que siempre se asocia a las Matemáticas (recordemos que la titulación se llamó en el pasado Ciencias Exactas), y por tanto, muchas veces dan lugar a confusión y a malinterpretaciones.

Es frecuente, de hecho, escuchar que se asocia la palabra *estadística* a otras como *manipulación* («Solo me creo las estadísticas que he manipulado». Winston Churchill) o incluso *mentira* («Hay tres tipos de mentiras: las mentiras, las malditas mentiras y las estadísticas». De origen incierto, aunque se suele atribuir a Mark Twain).

En la segunda mitad del siglo XX, la confluencia de diversos factores influyó en la progresiva consideración de la Estadística como una materia con entidad propia y relativamente independiente de las Matemáticas, y en la eclosión de otros campos multidisciplinares asociados.

Por un lado, la ampliación de los ámbitos de aplicación de las técnicas estadísticas propició el rápido desarrollo de la Estadística aplicada, que además permitió desarrollar técnicas diferenciadas según el tipo de uso que para cada uno de los campos resultaba de mayor interés. El empleo de técnicas estadísticas empezó a ser frecuente en disciplinas como la Psicología, la Economía, la Medicina, todo tipo de industrias, etc., generando una serie de nuevas ramas en ellas con el sufijo *-metría* (Biometría, Econometría, Psicometría...), o algunas más independientes, como el *Business Intelligence*, que englobaría la aplicación de técnicas estadísticas para analizar y sacar partido a la información en diversos tipos de negocios.

Algo más tarde surgieron también la Cienciometría, la Informetría y la Bibliometría en el ámbito profesional de la documentación, si bien también hay referencias más antiguas sobre los primeros estudios que introdujeron técnicas matemáticas y estadísticas para la medición y el análisis de la información. De estas disciplinas nos ocuparemos en el próximo apartado.

Por otro lado, la rápida evolución de la informática tuvo tres efectos decisivos: en primer lugar, permitió automatizar y evaluar cálculos a una velocidad antes impensable, de tal forma que análisis que pocos años antes resultaban inviables se convertían en factibles e incluso sencillos al utilizar programas informáticos; en segundo lugar, y relacionado con lo anterior, se empezaron a popularizar los paquetes de software estadístico especializado (SPSS, uno de los pioneros, data de 1968); y en tercer lugar, obviamente también relacionado con los dos aspectos anteriores, confluyeron los ámbitos de estudio de los estadísticos y los informáticos, dando lugar a campos multidisciplinares como el *machine learning*. Los primeros algoritmos de minería de datos, precursores de las redes neuronales o de técnicas de proximidad como el *nearest neighbour*, datan de esta época.

La International Association for Statistical Computing se fundó en 1977 con el objetivo de «vincular la metodología estadística tradicional, la tecnología informática moderna, y el conocimiento de expertos para convertir los datos en información y conocimiento».

En 1989 tuvo lugar el primer congreso de *Knowledge Discovery in Databases* (KDD), que posteriormente se convirtió en la *Conference on Knowledge Discovery and Data Mining*.

A mediados de los años noventa del siglo pasado apareció por primera vez el término *data science*, aunque no se popularizó hasta una década después, con la eclosión de las grandes compañías tecnológicas y las redes sociales: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/.

En cuanto al *big data*, también se popularizó a partir de los años 2010-2011-2012, cuando empezaron a proliferar datos que demostraban que en los últimos años se genera más información que en toda la historia anterior: http://www.mediapost.com/publications/article/173345/ibm-92-of-worlds-data-created-in-last-two-years.html.

Otros enlaces que pueden aportar algo de luz a esta discusión de términos:

http://101.datascience.community/2015/11/12/the-data-science-industry-who-does-what/.

http://www.datasciencecentral.com/profiles/blogs/66-job-interview-questions-for-data-scientists.

3. Importancia de la minería de datos en las ciencias sociales y de la información/documentación

Pero ¿por qué es importante también la minería de datos en las ciencias sociales, que siempre han tenido una consideración más «de letras» o, cuando menos, mixta?

Realmente, ¿puede aportar algo el estudio de las técnicas cuantitativas en el ámbito de las ciencias sociales, donde generalmente la cantidad de datos disponibles es mucho menor que en la de otros campos de los que hemos hablado anteriormente? Por supuesto que sí.

El primer motivo, y más obvio, es el amplio e importante uso que se les pueden dar:

En cualquier disciplina es importante medir cosas, extraer patrones de comportamiento, predecir situaciones futuras, etc., y también en casi cualquier campo existen de hecho dichos patrones, relaciones, distribuciones y regularidades, que se prestan a ser analizados mediante técnicas cuantitativas. Hay una línea de opinión algo discutible que defiende que todo es medible (http://www.cio.com/article/2438921/it-organization/everything-is-measurable.html); sin necesidad de llegar a ello, lo que sí parece claro es que, como dijo Edward Deming, «Sin datos solo eres otra persona más con una opinión».

Centrándonos en las materias que nos ocupan especialmente, las ciencias de la información y la documentación, las aplicaciones son múltiples: medición de factores de impacto de las diversas publicaciones, de las audiencias televisivas, búsqueda de patrones en diversos eventos (préstamos en bibliotecas, visitas a museos...), análisis textual para valorar autorías o similitudes entre autores, análisis de redes de relaciones (entre enlaces en páginas web, citas entre autores...), etc.

Y el segundo, que es común a aquellas disciplinas no específicamente cuantitativas pero que hacen uso de las técnicas estadísticas, es que muchas veces la validez de ciertos estudios queda en entredicho por carencias en el uso de la estadística: escasa formación de los investigadores, ausencia de una correcta planificación del análisis, uso de técnicas demasiado elementales, datos de dudosa calidad, tamaños de muestra escasos, etc.

Por tanto, es fundamental un sólido conocimiento de este tipo de técnicas, para dotar del necesario rigor a los análisis y evitar así que su valor pueda ser puesto en duda.

Hoy en día está ampliamente aceptado entre la comunidad científica que el llamado método científico es la forma de distinguir las ciencias de las seudociencias, y por tanto, es conveniente que cualquier estudio cuantitativo cumpla los principios de reproducibilidad y refutabilidad.

4. *Web mining, text mining* y otras disciplinas asociadas

4.1. *Web mining*

Con la eclosión de internet y el crecimiento exponencial de páginas web, pronto surgió la necesidad de analizar la información que aparecía en ellas, pero los primeros intentos de análisis toparon con ciertas dificultades inherentes a la estructura y contenido tan particulares de un sitio web: información de diversos tipos (texto, imágenes estáticas, imágenes dinámicas, vídeo...); estructura en niveles, con una página principal y varias subpáginas; multitud de enlaces e hipervínculos, algunos internos y otros externos, etc.

De hecho, la información de una página web se suele considerar como un tipo de información llamada no estructurada, que se caracteriza por un gran volumen de datos, de los cuales pocos son relevantes porque hay mucho ruido e información superflua. Por tanto, adquiere una gran importancia la fase de extracción de la información como paso previo a las típicas fases de depuración de la misma y posterior análisis.

Los modelos que se suelen utilizar en este campo están entre los llamados relacionales (grafos o redes, por ejemplo).

Un aspecto diferenciador de los modelos que tratan de explicar los patrones de este tipo de datos es que la interpretación del modelo, que en otros campos es de suma importancia (hasta el punto de que en ocasiones se descartan modelos con muy buena capacidad predictiva por su difícil interpretabilidad), aquí pasa a un segundo plano.

Se pueden clasificar los diversos tipos de *web mining* según la dimensión de la página web que se pretende analizar:

- ***Web content mining***: Se centra en el contenido de la página (texto, enlaces, imágenes...).

- ***Web structure mining***: el objetivo principal es analizar la estructura del sitio (relaciones entre las diversas páginas, los enlaces entrantes y salientes, etc.).

- ***Web usage mining***: Se centra en el uso, para tratar de establecer patrones de comportamiento de los usuarios que visitan una página e interactúan con los diversos elementos de la misma.

4.2. Text mining

El *text mining* es una disciplina más transversal y de creciente interés, que tiene relación y coincidencias con el *web mining*, el *sentiment analysis*, etc.

Posiblemente tiene su origen en la lingüística computacional, disciplina surgida a finales del siglo XX, y que si bien algunas líneas engloban en ella todas las aplicaciones y aportaciones de la informática al análisis del lenguaje, otras son más específicas y restringen su ámbito a los sistemas informáticos que procesen estructura lingüística, y cuyo objetivo sea la simulación parcial de la capacidad lingüística de los hablantes de una lengua.

No obstante, parece claro que, como ocurre con la mayor parte de disciplinas de las que nos hemos ocupado en este y los anteriores apartados, la delimitación de su ámbito de estudio no está claramente definida, pues se trata de campos híbridos y multidisciplinares.

Puede consultarse una discusión profunda sobre estos temas en el siguiente artículo: http://www.elprofesionaldelainformacion.com/contenidos/2004/enero/2.pdf.

Se incluirían entre sus aplicaciones algunas tan diversas como las siguientes: indexación de documentos, traducción automática, resumen automático de textos, reconocimiento de voz, identificación de la autoría de textos...

Esta última es la que suele tener mayor repercusión en los medios de comunicación: http://cultura.elpais.com/cultura/2016/04/11/actualidad/1460388427_850730.html.

Pero en internet se pueden encontrar múltiples usos que dan lugar a interesantes aplicaciones; sirvan de ejemplo las siguientes:

- Palabras más frecuentes (del inglés): http://www.wordcount.org/index2.html.
- Visualización de redes de asociación de palabras: http://visuwords.com/.
- Visualización de búsquedas relacionadas con palabras: http://answerthepublic.com/.

Uno de los ámbitos de mayor interés en los últimos tiempos es el del *sentiment analysis*, que busca determinar el cariz (positivo *vs*. negativo, esencialmente) de un mensaje de texto (un tuit, un post de un blog, etc.) a partir de la identificación y clasificación de las palabras que lo componen. Se puede ver una interesante aplicación de este campo en http://www.austinwehrwein.com/digital-humanities/comparing-twitter-sentiment-analysis-scores/.

5. Guía de estudio

Los materiales que tenemos a nuestra disposición consisten en unos módulos elaborados por los profesores Enric Mor, Ramón Sangüesa y Luis C. Molina.

Dado que se trata de unos documentos con contenido esencialmente técnico, mediante este apartado pretendemos introducir al lector en los principales conceptos tratados en cada módulo, y de relacionar los mismos, con el objetivo de facilitar su lectura y, a partir de esta, el aprendizaje de la materia de la que se ocupan.

En el módulo 1 («El proceso de descubrimiento de conocimiento a partir de datos») se sientan las bases de la minería de datos mediante una detallada explicación del proceso de extracción de conocimiento y sus fases, que, como se puede comprobar, no difiere apenas del clásico proceso seguido en estadística, como disciplinas intrínsecamente relacionadas que son (tal y como hemos comentado en apartados anteriores de este documento): es decir, en primer lugar se debe definir claramente el **objetivo** buscado (lo cual determinará la/s técnica/s que será más adecuado utilizar); posteriormente hay una fase muy importante (y generalmente muy tediosa) de **preparación de la información**, para pasar a la **construcción del modelo** propiamente dicho y la **evaluación** de su corrección y su idoneidad (validación), y finalizar con su **interpretación** y la obtención de **resultados** a partir del mismo.

Todos estos conceptos se irán desarrollando con mayor profundidad en los sucesivos módulos.

En el módulo 2 («Preparación de datos») se detalla la primera fase del proceso antes citado, la preparación de los datos. Con mucha frecuencia, y por diversos motivos, los datos que queremos utilizar en nuestro estudio no cumplen los requisitos para poder ser analizados directamente, y por tanto, es necesario llevar a cabo un proceso previo de análisis de su calidad y preparación (limpieza, transformación, síntesis) para poder extraer la mayor y mejor cantidad de información.

Los problemas principales con los que se enfrenta el investigador en esta fase son dos: por un lado, es una tarea algo ingrata y tediosa, que además con frecuencia ocupa mucho tiempo (se suele decir que cerca del 80 % de un proyecto de análisis de datos se dedica a esta fase), y en consecuencia reduce el dedicado a la construcción y validación del modelo; y por otro, es una tarea difícilmente sistematizable o automatizable, pues cada base de datos requiere su propio banco de pruebas y acciones de corrección de los posibles problemas encontrados.

No obstante, las principales tareas que se acometen en esta fase son la identificación de valores perdidos y de valores atípicos (pues su presencia suele afectar al análisis y puede ser debida a factores externos), la transformación de datos para facilitar su posterior tratamiento, la discretización de variables continuas (dado que en ocasiones es más sencillo su tratamiento de forma agrupada) y, en el caso de disponer de excesiva información, la reducción de la dimensionalidad (que trata de reducir el número de variables disponibles sin perder mucha información, aprovechando las correlaciones entre las mismas).

El módulo 3 («Clasificación: árboles de decisión») se ocupa de una de las técnicas que goza de mayor aceptación por su versatilidad y su simplicidad: los árboles de clasificación o decisión.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de manera que la decisión final se puede determinar siguiendo las reglas que se cumplen desde la raíz del árbol hasta alguno de sus nodos finales. Una de las grandes ventajas de esta técnica es que las opciones posibles a partir de una determinada condición son excluyentes, lo cual permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar (Hernández y otros, 2004).

La tarea de aprendizaje para la cual los árboles de decisión se adecuan mejor es la clasificación, para lo cual utilizan la técnica de partición (es decir, el algoritmo va construyendo el árbol añadiendo particiones o divisiones sucesivas de cada nodo, que pueden ser nominales o por intervalos, según la naturaleza del atributo sea nominal o continua), siendo el criterio de selección de las particiones lo que diferencia los principales algoritmos utilizados.

En el módulo se introducen primero dos de los algoritmos más antiguos pero que son la base de la mayoría de versiones posteriores: el ID3 y el C4.5.

Posteriormente, se explican las características de dos de los métodos más extendidos actualmente: CART o CRT (*classification and regression tree*) y CHAID (*chi-squared automatic interaction detector*). Este tipo de métodos se adaptan bien a otras tareas distintas de la clasificación, como la regresión o la estimación de probabilidades, para lo que utilizan una función de dominio real en lugar de discreto, y etiquetan los nodos del árbol con valores reales.

Por último, se introduce también una técnica de construcción de árboles multivariantes: LMDT (*linear machine decision trees*).

Un problema relativamente frecuente de los árboles de decisión es el del sobreajuste, y para evitarlo surge el concepto de poda, que consiste en eliminar algunos nodos que se consideren demasiado específicos.

Los algoritmos de este tipo, por su carácter voraz y su estructura «divide y vencerás», se comportan bien con grandes volúmenes de datos.

En el módulo 5 se detalla ampliamente otra técnica muy importante de la minería de datos, denominada métodos de agregación (*clustering*).

Así pues, *clustering* es una técnica de *data mining* que identifica clústeres o agrupaciones de observaciones de acuerdo a una medida de similitud entre ellos. El objetivo fundamental de esta técnica es conseguir la máxima homogeneidad dentro cada grupo, y a su vez, la mayor diferencia entre los diversos grupos. Por tanto, se trata de obtener una descripción inicial que separe grupos de observaciones con características similares. Esta agrupación nos permitirá identificar las características (variables) comunes de las observaciones que pertenecen a un determinado clúster y no a otro, detectando a su vez por qué son similares, y también qué los hace diferentes de los otros clústeres.

La identificación de clústeres o grupos de observaciones se basa en una medida de similitud. Según la medida de similitud que se adopte, la aplicación del método podrá dar lugar a diferentes clústeres o agrupaciones.

Por ello, el módulo se inicia definiendo qué son las medidas de similitud, en términos de distancia, y a continuación detalla alguno de los diferentes métodos de agrupación existentes: método de los centroides (*k-means*), método de los vecinos más cercanos (*k-nearest neighbours*), métodos incrementales o aglomerados, métodos de agregación probabilistas, métodos probabilistas de construcción de agregaciones cuando el número de clases es conocido *a priori* y métodos de construcción de agregaciones cuando el número de clases es desconocido *a priori*.

A continuación, se detallan los pasos que hay que seguir para saber interpretar los modelos o clústeres obtenidos y cómo realizar predicciones con dichos modelos. Además de explicar cómo interpretar y predecir con los modelos obtenidos, se muestran ejemplos para ayudar a mejorar su comprensión.

Seguidamente, se muestran dos medidas de bondad de ajuste de los modelos obtenidos, es decir, medidas que evalúan el rendimiento del modelo. Estas medidas son: el principio de mínima longitud de descripción y la medida de Kullback-Leibler. Igual que antes, aquí también se muestran ejemplos de las medidas.

Por último, se comentan las principales ventajas y desventajas de los métodos de agrupación.

En el módulo 8 («Evaluación de modelos») se trata un tema muy importante, y al que no siempre se le concede la debida atención: la evaluación o validación del modelo.

Utilizar un método de validación es necesario para evitar que la precisión del modelo esté sobreestimada por el hecho de proporcionar mucho mejores resultados en la muestra empleada para construirlo que en otras a las que se aplique con posterioridad (problema conocido como sobreajuste).

El método de validación más básico y tradicional reserva un porcentaje de la base de datos como conjunto de prueba (o validación o test). El resto de los datos forman el conjunto de entrenamiento que se usa para construir el modelo. La división de los datos en estos dos grupos debe ser aleatoria para que la estimación sea correcta.

Un método alternativo de validación, que se utiliza sobre todo cuando no se dispone de muestra suficiente, es el conocido como validación cruzada, que se suele implementar mediante el método de los k pliegues (*k-fold crossvalidation*), el cual divide aleatoriamente los datos en k grupos (frecuentemente k=10) de tamaño similar. Un grupo se reserva como conjunto de prueba, y con los k-1 restantes se construye un modelo, y se utiliza para predecir el resultado de los datos del grupo reservado. Este proceso se repite k veces, dejando cada vez un grupo diferente para la prueba. Finalmente, se construye un modelo con todos los datos y se obtienen sus ratios de error y/o precisión, promediando las k ratios disponibles.

Existen otras formas de validación externa, como la validación *out-of-time*, que consiste en aplicar los resultados del modelo obtenido con datos de un periodo temporal, a datos que hacen referencia a otro periodo temporal distinto.

Por otro lado, si los posibles errores cometidos por el modelo no son igualmente importantes a la hora de evaluarlo (circunstancia habitual, por ejemplo, en los diagnósticos médicos), se puede incorporar una función de costes para ponderarlos.

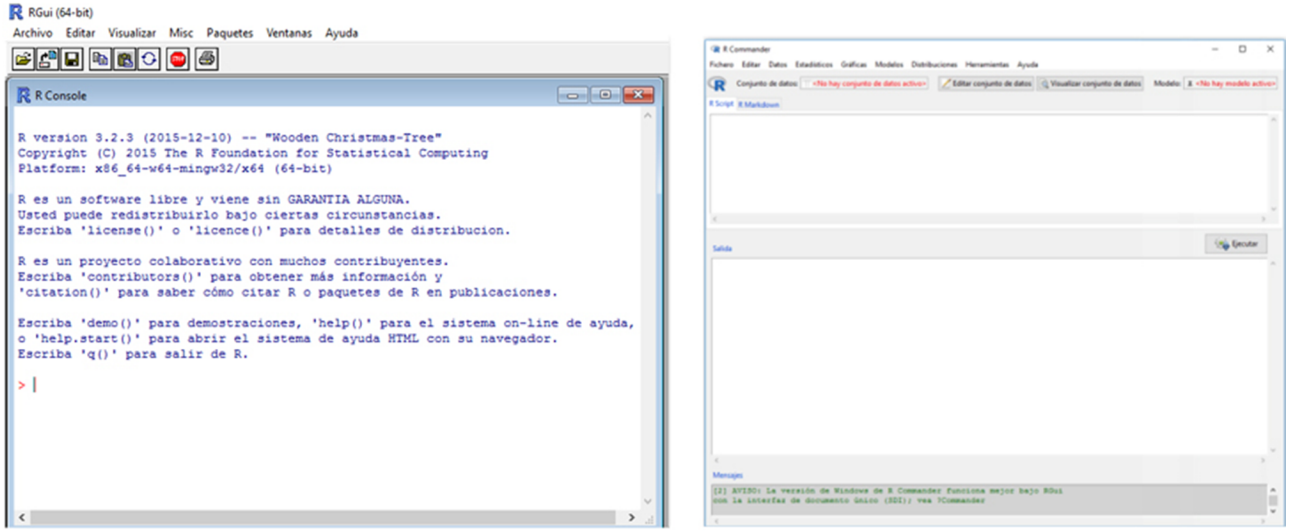
6. Uso de R como software especializado

El software estadístico utilizado en la asignatura *Minería de datos* del grado de Información y Documentación es R. Un software para el análisis estadístico de datos y generación de gráficos de excelente calidad. Está considerado como uno de los más importantes por su precisión a la hora de obtener resultados. Cada vez con mayor frecuencia está aumentando su uso, sobre todo en investigación, por parte de la comunidad estadística, siendo además muy popular en materias como la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras.

Su mayor ventaja es que es un software libre, es decir, gratuito. Se puede descargar para distintas plataformas: Mac, Windows o Linux. Además, el programa base de R contiene funciones básicas para un gran número de procedimientos estadísticos. Para funciones más específicas, existen las llamadas librerías o *packages*, las cuales se instalan fácilmente. Las librerías son contribuciones de diferentes autores. Continuamente aparecen nuevos *packages* gratuitos que expanden la capacidad de R para estimar o solucionar diferentes problemas de diferentes áreas. Al ser un software al alcance de todos, es fácil y rápido encontrar cualquier información sobre su instalación, uso, descarga, librerías... en la red.

El mayor inconveniente que presenta es que, como es un lenguaje de programación, es necesario saber programar. Para solucionar esta desventaja, existe una librería o paquete llamado R Commander, el cual permite utilizar R sin necesidad de saber programar. Es una interfaz gráfica de usuario, que permite realizar cálculos estadísticos solo pulsando un botón.

Figura 2. A la izquierda, una imagen de programa R, a la derecha, una de la librería Rcmdr



Fuente: elaboración propia.

Para obtener toda la información necesaria sobre el software R, podéis entrar en https://www.r-project.org/.

Dependiendo del sistema operativo que tenga el usuario, para su descarga hay que pulsar en:

- Windows: https://cran.r-project.org/bin/windows/base.
- MAC: https://cran.r-project.org/bin/macosx/.
- Linux: https://cran.r-project.org/bin/linux/.

Una vez instalado R, se pueden instalar las librerías o *packages* que se deseen. Para ello, se debe tener conexión a internet y entrar en *Paquetes> Instalar paquete(s)> HTTPS CRAN correspondiente > Seleccionar la librería que se desea instalar*.

A continuación, se muestran algunos enlaces donde se pueden descargar manuales tanto para la instalación como para profundizar en el manejo del software estadístico R.

- https://cran.r-project.org/doc/contrib/Chicana-Introduccion_al_uso_de_R.pdf.
- http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf.
- https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf.

Para facilitar el trabajo con R se creó RStudio. Este es un entorno de desarrollo integrado en R. Se puede descargar también gratuitamente en https://www.rstudio.com/products/rstudio/download/ para Windows, MAC y Linux. Cuando entras en él, se observa que la pantalla se ha dividido en cuatro ventanas:

1) La ventana superior izquierda sirve para abrir y editar ficheros con código R.

2) En la ventana inferior izquierda hay una consola de R, en la cual aparecen los resultados de lanzar el código seleccionado en la ventana anterior.

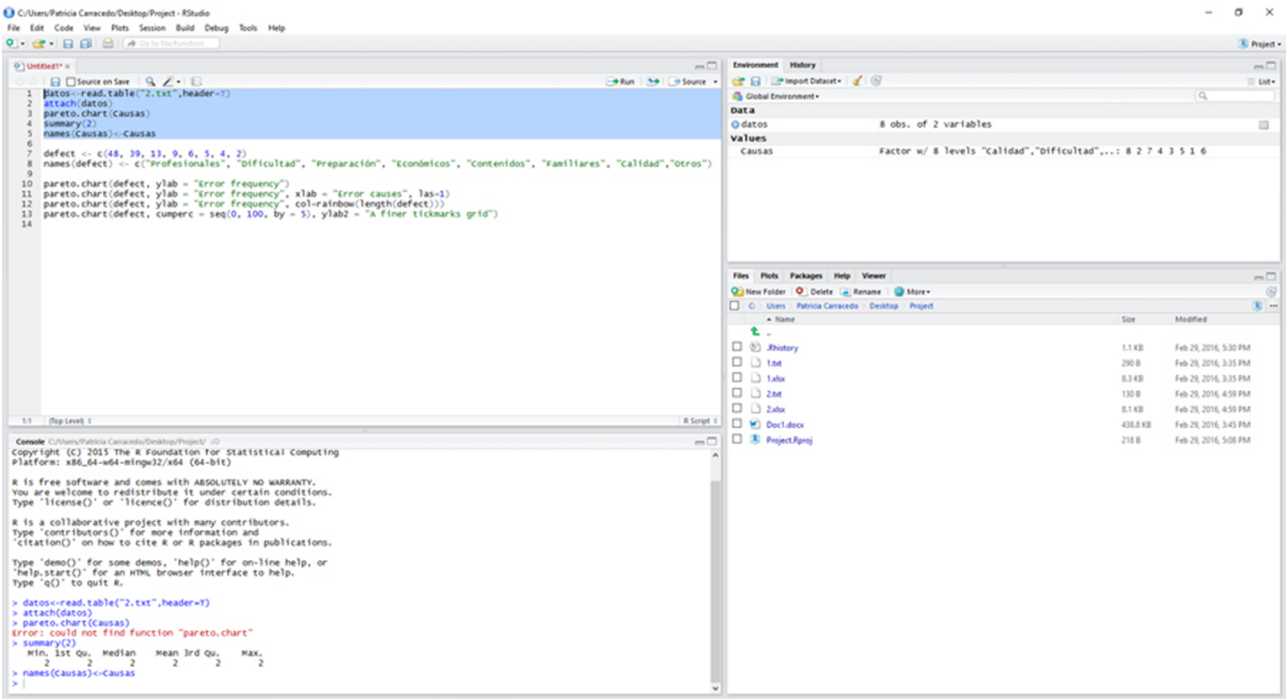
3) La ventana superior derecha está integrada por dos pestañas:

- ***Workspace***, en la cual aparece la lista de los objetos que se han creado en memoria.
- ***History***, contiene el histórico de las líneas de código que han sido ejecutadas en R.

4) La ventana inferior derecha está formada por cuatro pestañas:

- ***Files***, que da acceso al directorio y ficheros del disco duro.
- ***Plots***, donde aparecen los gráficos que se crean en la consola de R.
- ***Packages***, te permite tanto la instalación de paquetes como la visualización de los que ya están instalados en la máquina. Pulsando sobre ellos, los abres.
- ***Help***, página de ayuda.

Figura 3. Imagen del programa RStudio



Fuente: elaboración propia.

Por tanto, RStudio te permite trabajar de una forma más organizada al tener todos los archivos necesarios para trabajar con R en una misma ventana.

7. Un ejemplo de minería de datos con R

Una vez visto qué es R, sus ventajas e inconvenientes, el siguiente paso es realizar un ejercicio con él, en concreto de *data mining*. El ejercicio trata sobre el estudio de los dos únicos discursos en navidades que ha realizado el rey Felipe VI. Vamos a estudiar qué palabras son las más frecuentes, veremos cómo mostrarlo en un gráfico y además, con qué palabras tiene una alta correlación. Para ello, cada paso que se va a realizar se explica mediante (#), se adjunta el código (en courier) y se muestra el resultado obtenido (cursiva y negrita).

El primer paso será cargar las siguientes librerías:

```
library(NLP)
library(RColorBrewer)
library(wordcloud)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(ggplot2)
```

Disponemos de dos textos a analizar, ambos en formato .txt (Discurso Rey 2014.txt, Discurso Rey 2015.txt), los cuales se encuentran en una carpeta denominada «discursos». Para proceder a la carga de ambos ficheros, primero hay que indicar la ruta donde se encuentra la carpeta «discursos» y a continuación cargar los ficheros de la siguiente manera:

```
setwd("C:/Users/Documentos Data Mining/")
cname<- file.path(".","discursos")
dir(cname)
```

«Discurso Rey 2014.txt» «Discurso Rey 2015.txt»

Una vez cargados los dos ficheros de texto, se debe crear el corpus. Corpus es el cuerpo textual, es decir, la estructura principal para la gestión de documentos. A continuación se detalla el código que se debe utilizar para su creación e inspección.

Creamos el corpus:

```
corpus<- Corpus(DirSource(cname))
```

Para inspeccionar cada documento escribimos:

```
inspect(corpus[1])
```

<<VCorpus>>

Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 1

<<PlainTextDocument>>
Metadata: 7
Content: chars: 10264

```
inspect(corpus[2])
```

<<VCorpus>>
Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 1

<<PlainTextDocument>>
Metadata: 7
Content: chars: 10199

El siguiente paso es realizar las siguientes transformaciones en nuestro corpus:

- Convertir a corpus nuestro archivo de texto mediante la función *Corpus*.
- Convertir las letras de los textos en minúsculas mediante el argumento «*tolower*».
- Eliminar los números, mediante el argumento «*removeNumbers*».
- Eliminar los signos de puntuación, mediante el argumento «*removePunctuation*».
- Eliminarlas palabras comunes del español mediante el comando stopwords («spanish»).
- Eliminamos los espacios en blanco que se han generado con el comando «*stripWhitespace*».
- Por último, tratamos los documentos que se procesan como documentos de texto mediante el argumento «PlainTextDocument».

```
corpus<- Corpus(VectorSource(texto))
corpus<- tm_map(corpus, tolower)
corpus<- tm_map(corpus, removeNumbers)
corpus<- tm_map(corpus, removePunctuation)
corpus<- tm_map(corpus, removeWords, stopwords("spanish"))
corpus<- tm_map(corpus, stripWhitespace)
corpus<- tm_map(corpus, PlainTextDocument)
```

A continuación realizamos el *stemmingel*, que es un proceso de normalización lingüística en donde las diferentes formas que puede adoptar una palabra son reducidas a una única forma común, a la cual se denomina *stem*, por ejemplo *florero*, *florista*, *floral*, *flora* y *florido* se reducen al *stem flor*.

Para ello, utilizamos la librería *SnowballC,* que borra los sufijos de las palabras para dejarlas en su «raíz». Después volveremos a eliminar los espacios. Para ello el código es el siguiente:

```
corpus<- tm_map(corpus, stemDocument)
corpus<- tm_map(corpus, stripWhitespace)
```

Inspeccionamos nuevamente los documentos, así comprobamos que se deben haber reducido los caracteres:

```
inspect(corpus[1])
```

<<VCorpus>>
Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 1

<<PlainTextDocument>>
Metadata: 7
Content: chars: 6416

```
inspect(corpus[2])
```

<<VCorpus>>
Metadata: corpus specific: 0, document level (indexed): 0
Content: documents: 1

<<PlainTextDocument>>
Metadata: 7
Content: chars: 6140
Content: chars: 10264

Efectivamente, los caracteres se han reducido, en concreto, en el documento 1 se ha reducido de 10.264 a 6.416 caracteres, y en el documento 2, de 10.199 a 6140 caracteres.

A continuación se crea la matriz de términos del documento de la siguiente forma:

```
mdt<- DocumentTermMatrix(corpus)
mdt
```

Donde se obtiene:

<<DocumentTermMatrix (documents: 2, terms: 830)>>

Non-/sparse entries: 1021/639
Sparsity: 38 %
Maximal term length: 15
Weighting: term frequency (tf)

Nuestra matriz está compuesta por dos documentos de texto con 830 palabras en total, de las cuales, el 38 % están dispersas. Más adelante las trataremos.

Ahora estudiamos brevemente nuestra matriz. Para obtener una lista con los términos más frecuentes escribir:

```
freq<- colSums(as.matrix(dtm))
ord<- order(freq)
freq[tail(ord)]
```

ciudadano, hoy, futuro, debemo, español, españa
12, 12, 14, 16, 21, 23

Para conseguir los términos menos frecuentes:

```
freq[head(ord)]
```

abdicación, abierta, abierto, abordar, abrir, acentuada
1, 1, 1, 1, 1, 1

A continuación creamos una tabla de frecuencias con 15 términos:

```
head(table(freq), 15)
```

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 21
558, 128, 53, 30, 20, 13, 6, 4, 7, 1, 4, 2, 1, 1, 1

```
tail(table(freq), 15)
```

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 21, 23
128, 53, 30, 20, 13, 6, 4, 7, 1, 4, 2, 1, 1, 1, 1

Se puede observar, que tenemos 558 palabras que aparecen solo una vez en el texto, mientras que hay una palabra que se repite 23 veces.

El siguiente paso es obtener la dimensión de la matriz de términos del documento y a continuación eliminar los términos con poca frecuencia.

Obtención de la dimensión de la matriz de términos del documento.

```
dim(mdt)
```

2, 830

Eliminamos los términos menos frecuentes.

```
mdts<- removeSparseTerms(mdt, .1)
dim(mdts)
```

Inspeccionamos la matriz y vemos ahora que no tenemos ninguna palabra dispersa «Sparsity 0 %».

```
inspect(dtms)
dtms
```

DocumentTermMatrix (documents: 1, terms: 496)>>
Non-/sparse entries: 496/0
Sparsity: 0 %
Maximal term length: 15
Weighting: termfrequency (tf)

Una vez eliminada la dispersión:

Volvemos a obtener las palabras con que más frecuencia se repiten:

```
inspect(dtms)
freq<- colSums(as.matrix(mdts))
findFreqTerms(dtms, lowfreq=10)
```

ciudadano, debemo, españa, español, futuro, historia, hoy, mundo, política, tiempo, vida

Estudiamos las relaciones de las palabras más frecuentes.

```
findAssocs(mdts,"ciudadado", corlimit=0.7)
```

afecto, afrontar, ahora, buena, cada, confianza…

```
findAssocs(mdts,"debemo", corlimit=0.7)
```

afecto, afrontar, ahora, buena, cada, ciudadano, confianza, constitución…

```
findAssocs(mdts,"españa", corlimit=0.7)
```

común, comunidad, conjunto, conseguido, constitución, convivencia, crecimiento…

```
findAssocs(mdts,"español", corlimit=0.7)
```

común, comunidad, conjunto, conseguido constitución, convivencia, crecimiento, dado, deben…

Gracias a las librerías Wordcloud y RColorBrewer podemos obtener los siguientes gráficos:

Gráfico con las 80 palabras que aparecen con más frecuencia.

```
dark2 <- brewer.pal(6, "Dark2")
wordcloud(names(freq), freq, max.words=80, min.freq=5,rot.per=0.2, scale=c(3, .1),colors=dark2)
```

Figura 4. Gráfico con las 80 palabras que aparecen con más frecuencia

Fuente: elaboración propia.

Gráfico con las palabras que como mínimo aparecen 10 veces.

```
dark2 <- brewer.pal(6, "Dark2")
wordcloud(names(freq), freq, min.freq=10, scale=c(5, .1), colors=brewer.pal(6, "Dark2"))
```

Figura 5. Gráfico con las palabras que como mínimo aparecen 10 veces

Fuente: elaboración propia.

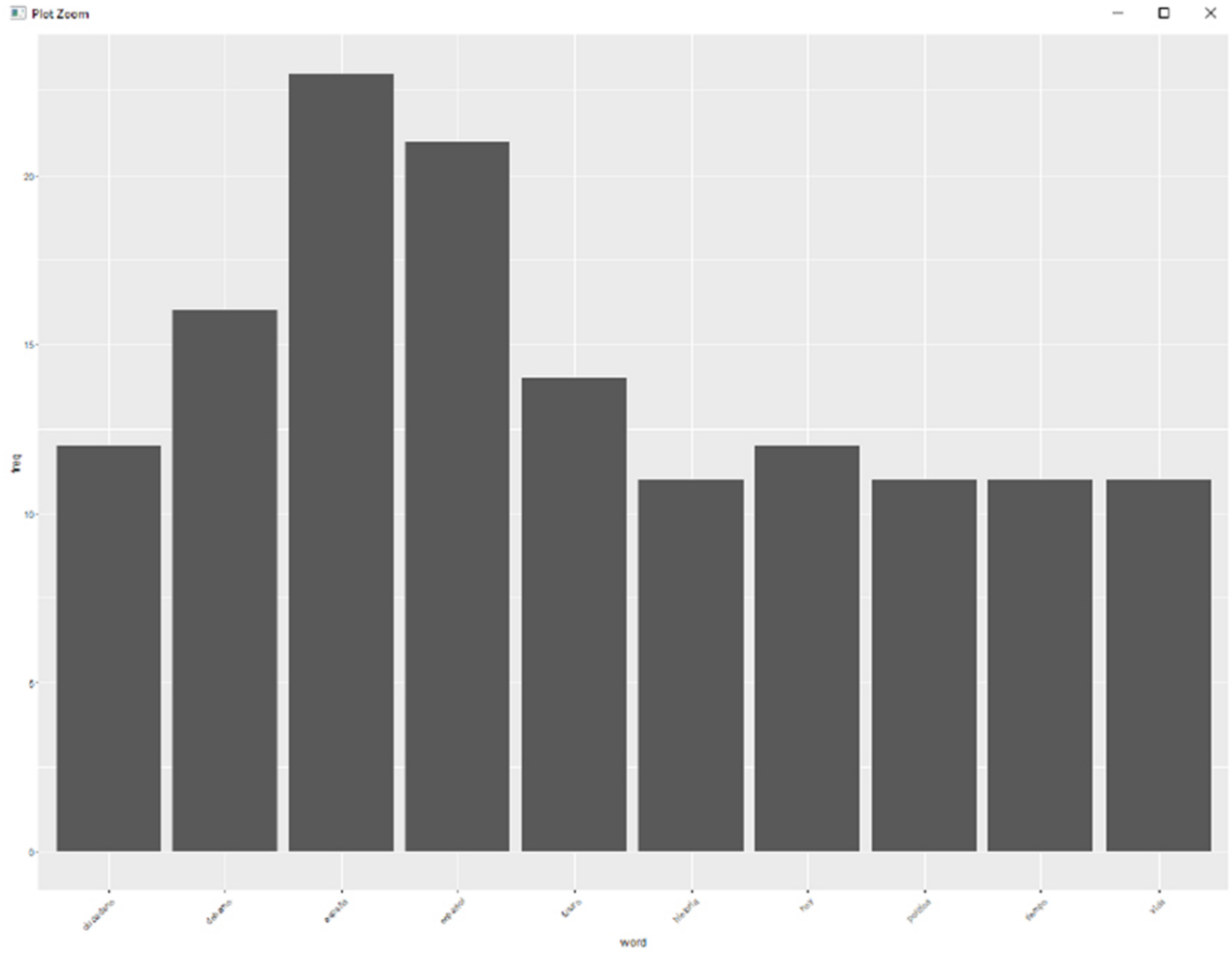
Por último, creamos una tabla de frecuencias y dibujamos las palabras con una frecuencia mayor o igual a 10 con ayuda de la librería ggplot2.

```
freq<- sort(colSums(as.matrix(mdts)), decreasing=TRUE)
head(freq, 10)
wf<- data.frame(word=names(freq), freq=freq)
head(wf)
```

freq
españa: 23
español: 21
debemo: 16
futuro: 14
ciudadano: 12
hoy: 12

```
plot<- ggplot(subset(wf, freq>10), aes(word, freq))
plot<- plot + geom_bar(stat="identity")
plot<- plot + theme(axis.text.x=element_text(angle=45, hjust=1))
plot
```

Figura 6. Tabla de frecuencias dibujando las palabras con una frecuencia mayor o igual a 10



Fuente: elaboración propia.

En conclusión a este ejercicio, hemos visto cómo crear y tratar un corpus. Una vez eliminada la dispersión de palabras en dicho corpus, hemos observado que las palabras más frecuentes en los dos únicos discursos en navidades que ha realizado el rey Felipe VI son: *ciudadano, debemo, españa, español, futuro, historia, hoy, mundo, política*… y las hemos mostrado en varios gráficos con las librerías *wordcloud, RColorBrewer y ggplot2,* y en una tabla de frecuencias.

Otro ejercicio similar al realizado se puede encontrar en (http://epub.wu.ac.at/3978/), donde se encuentra el artículo titulado «Text mining infrastructure in R».

Bibliografía

Referencias bibliográficas

Hernández, J.; Ramírez, M. J.; Ferri, C. (2004). *Introducción a la minería de datos*. Pearson Prentice Hall.

Meyer, D.; Hornik, K.; Feinerer, I. (2008). «Text mining infrastructure in R». *Journal of statistical software* (vol 5, núm. 25, pág. 1-54).