Fuentes de Datos - 1

Introducción

Los problemas de minería de datos relacionan distintas disciplinas/herramientas para su abordaje (Figura 1). Dentro de éstas, las bases de datos van a ser el principal insumo para resolver el problema.

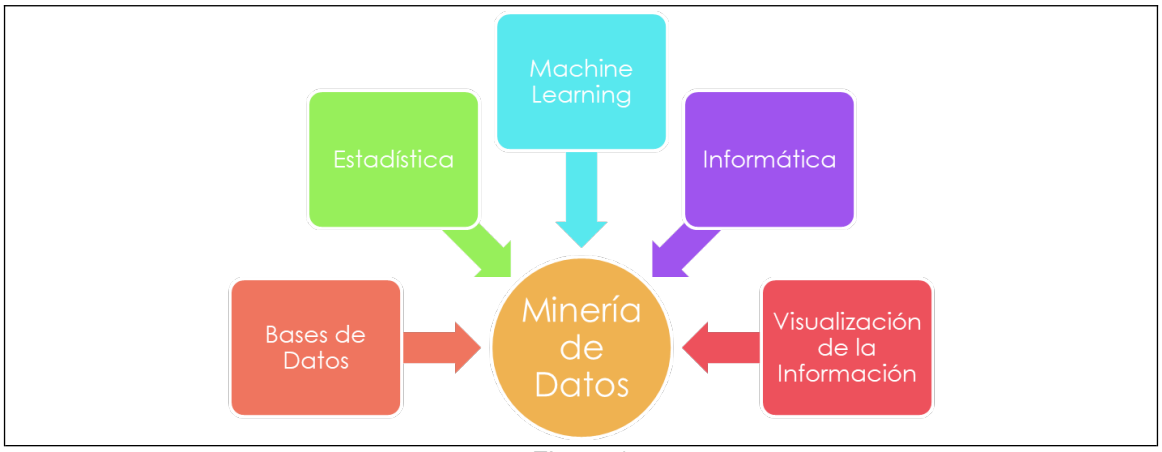

Figura 1

Aquí nos referimos como “base de datos” al conjunto de datos que tenemos disponible, independientemente del formato. Muy frecuentemente los datos disponibles van a estar en papel y para poder utilizarlos se deberá pasar a un formato digital.

Las bases de datos pueden ser una base de datos literalmente hablando (SQL o no SQL), a la que tendremos que acceder para poder obtener los registros que necesitemos; o tablas en formato de archivo (valores separados por comas, tabulaciones, de texto, Excel, etc.) que serán las fuentes de datos que se van a utilizar en el cursado de la materia.

Pero además de estos tipos de datos “clásicos”, muchas veces es necesario obtenerlos de otras fuentes, como puede ser de la web (un archivo o información de una página) o de otros formatos de archivo que requieren un tratamiento previo para poder utilizarlos (imágenes o archivos PDF).

Acceso clásico a datos en R

La instalación base de R tiene conjuntos de datos de prueba que pueden utilizarse, del paquete *Datasets*. Ejecutando la función *data()*, sin argumento, podemos acceder al listado completo.

Para cargar el conjunto de datos, usamos la función mencionada con el nombre:

data(nombre)

Algunos de los más utilizados son: iris, cars, mtcars, airpassengers, entre otros.

También podemos utilizar datos externos, cómo archivos de texto (TXT o CSV) o archivos de Excel, que son los más comúnmente utilizados.

Para leer archivos de texto vamos a utilizar la función *read.csv()* y para archivos de excel, la función *read_excel()* del paquete **readxl** [1].

Para el caso de archivos grandes (incluso de varios gigabytes), es recomendable utilizar la función *fread()* del paquete **data.table** [2] ya que es significativamente más rápida que *read.csv()* del paquete base de R.

Otras fuentes de datos: web

Para obtener datos de páginas web, desde R lo vamos a poder hacer de dos formas, descargando directamente el archivo o capturando la información de interés directamente de la página web.

En el primer caso, vamos a utilizar la función *download.file()* de la instalación base de R. Los argumentos que necesitamos son la URL del archivo y la ubicación y nombre del archivo de salida:

download.file(url = "dirección del archivo",
destfile = "path archivo de salida")

El segundo caso es un poco más complejo, ya que vamos a tener en cuenta el código fuente de la página en cuestión. Este código es HTML (Hyper Text Markup Language), que define el contenido y la estructura de una página web (Figura 2).

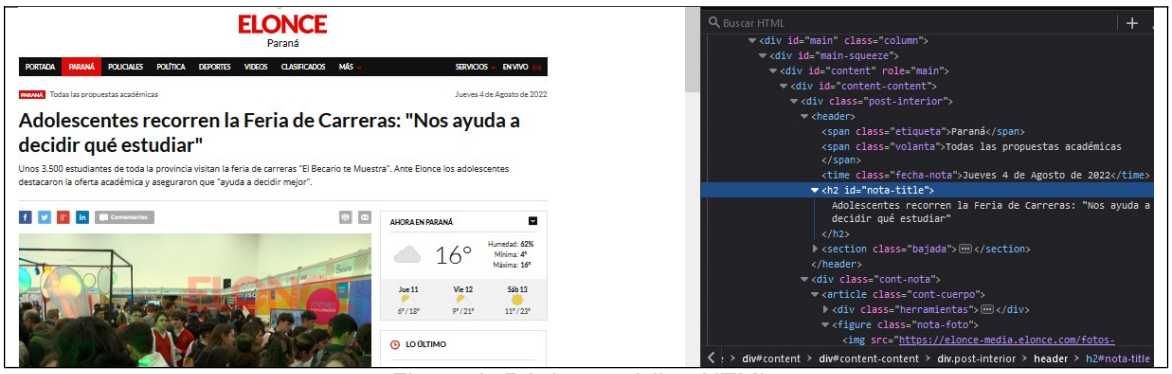

Figura 2. Página y código HTML.

Para poder recorrer el código desde el navegador y ver entre que etiquetas está el contenido que nos interesa, hacemos click con el botón derecho en la página que nos interesa y elegimos "Inspeccionar" o "Ver código fuente" en Firefox o Chrome.

Para extraer información de una página web vamos a utilizar el paquete **rvest** [3] de R. En particular las funciones *read_html()* para cargar el código de la página; *htlm_nodes()/html_elements()* para obtener un elemento determinado

del código de la página web; *html_table()* para obtener elementos de tipo tabla directamente y *html_text()* para obtener texto directamente.

Para *read_html()*, el único parámetro que necesitamos es la URL de la página que vamos a utilizar.

Las otras funciones van a tener como argumento el código de la página que obtenemos con *read_html()* y luego, vamos a utilizar el parámetro **css**.

CSS (Cascading Style Sheets) define la apariencia de los elementos HTML. Los selectores CSS suelen utilizarse para dar estilo a determinados subconjuntos de elementos, pero también se pueden utilizar para extraer elementos de una página web. Para encontrar el selector CSS que nos interesa, hacemos click derecho sobre el código y elegimos en “Copiar”, la opción “Selector CSS”.

Actividades

1 – Descargue el archivo TetuanCityPowerConsumption.csv desde la URL:

https://archive.ics.uci.edu/ml/machine-learning-databases/00616/Tetuan%20City%20power%20consumption.csv

2 – Obtenga la tabla de ganadores del Super Bowl de la NFL a partir de la página: http://www.espn.com/nfl/superbowl/history/winners

Guarde los datos en un archivo de valores separados por comas.

3 – Obtenga los datos de la tabla de graduados de la Universidad de Stanford, desde el año 1980 al 2007:

https://statistics.stanford.edu/people/alumni

Guarde el resultado en un archivo de valores separados por comas.

4 – Leer el artículo sobre ética en el rastreo web de James Densmore:

https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

Referencias

1. Package readxl. Disponible en: https://cran.r-project.org/web/packages/readxl/readxl.pdf
2. Package data.table. Disponible en: https://cran.r-project.org/web/packages/data.table/data.table.pdf
3. Package rvest. Disponible en: https://cran.r-project.org/web/packages/rvest/rvest.pdf