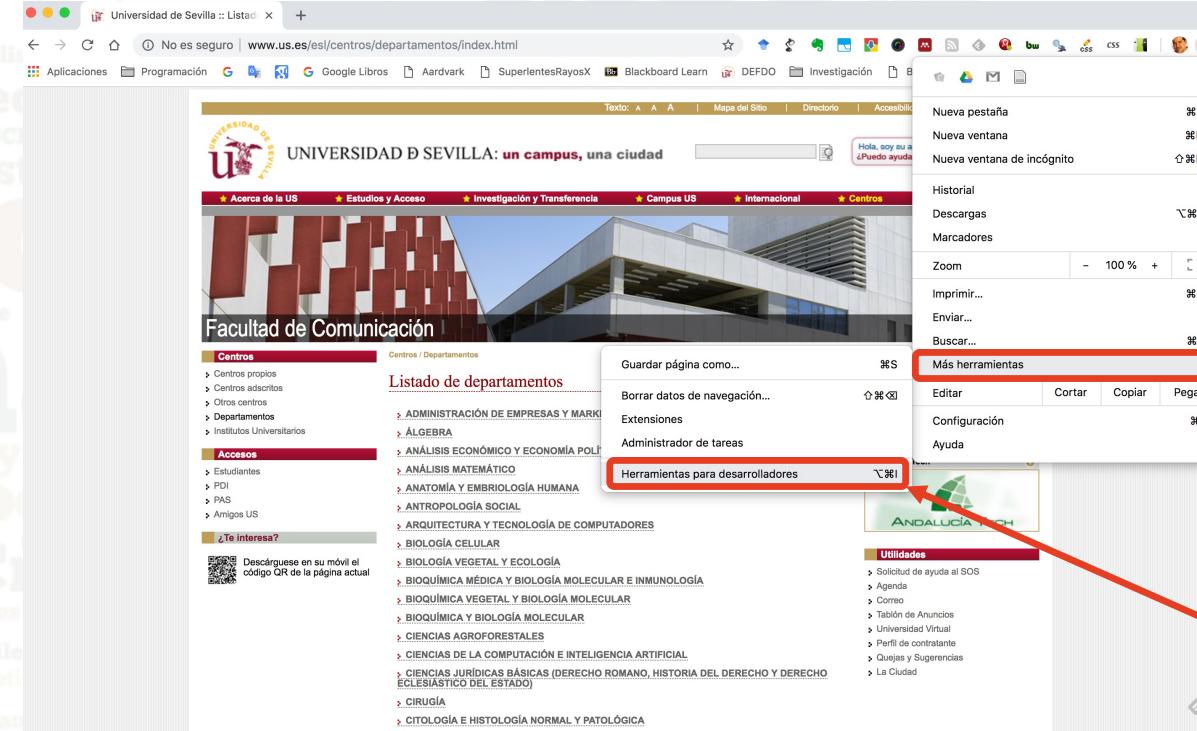


Cómo identificar contenidos en páginas web

Ejemplo captura desde consola navegador

Ejemplo de cómo extraer información directamente del HTML I

Listado de Departamentos US.ES : <https://www.us.es/centros/departamentos>



Abrimos en el **Navegador CHROME** las **Herramientas para Desarrolladores** (Developer Tools)

Abrir en : W10 = **F12**, Ctrl + Mayús + I Mac = Cmd + Opción + I

Ejemplo de cómo extraer información directamente del HTML II



XPath = XML Path Language.
Lenguaje que usa "path like" sintaxis para **identificar y seleccionar nodos** en un documento XML.

Con ella podemos acceder al elemento seleccionado y a todos sus atributos a través del **DOM**

Una vez terminado el paso 5, tenemos la ruta XPath del elemento seleccionado.
Con ella podemos acceder al elemento y a todos sus atributos a través del DOM

Ejemplo de cómo extraer información directamente del HTML III

```
//*[@id="contenedor_ficha"]/ul/li[1]/a
```

(1) Identificador XPath del elemento seleccionado en paso 2)

```
$x('//*[@id="contenedor_ficha"]/ul/li[1]/a')
```

(2) Instrucción que obtiene un conjunto de nodos del DOM a través de su identificador XPath. Se ejecuta en la Consola de las Dev Tools.

CUIDADO con las comillas. Vease como la expresión (1) aparece dentro de unas comillas simples. Ej: '(1)'

```
$x('//*[@id="contenedor_ficha"]/ul/li/a')[18].textContent
```

Modificamos la instrucción anterior para obtener el texto del elemento número 19.
RECUERDE que el contador empieza en 0.

```
$x('//*[@id="contenedor_ficha"]/ul/li/a')[1].href
```

En este caso queremos obtener el "link" del elemento número 2

XPath:

Expresión	Descripción
nodename	Selecciona todos los nodos con el nombre "nodename" podría ser : div, table, span, tr, td, h1 ... etc
/	Selecciona desde el nodo raiz
//	Selecciona nodos en el documento desde el actual, que coincidan con la selección no importa donde esten
.	Selecciona el nodo actual
..	Selecciona el nodo padre de el actual
@	Selecciona atributos

Ejemplo: `//*[@id='titleCast']//span[@class='itemprop']`

Esta expresión significa : Buscamos cualquier etiqueta “span” con el atributo `class='itemprop'` no
importa lo que sea PERO localizada bajo una etiqueta `div[@id='titleCast']`

Ejemplo de cómo extraer información directamente del HTML IV

Snippet :

Seleccione pestaña : Sources > sni

<https://www.us.es/centros/departamentos>

The screenshot shows the Chrome DevTools interface with the Sources tab selected. A red arrow labeled '1)' points to the 'Sources' tab in the top navigation bar. Another red arrow labeled '2)' points to the 'Snippets' button in the left sidebar. A third red arrow labeled '3)' points to the '+ New snippet' button. The main pane displays two snippets of code:

Snippet 1:

```
/* Obtenemos nombres de Dpto.
cntnd = $x('//*[@id="block-views-block-dej-1"]/div/div/div/div[2]/div');
for(var i=0; i < cntnd.length; i++){
    console.log(cntnd[i].textContent);
}
```

Snippet 2:

```
var cont = "";
cntnd = $x('//*[@id="block-views-block-departamentos-block-1"]/div/div/div/div[2]/div//a');
for(var i=0; i < cntnd.length; i++){
    cont = cont + "|" + cntnd[i].href;
}
```

Explicación del Javascript utilizado :

```
for(var i=0; i < f; i++){ <instrucciones> }
```

Bucle que hace tomar a la variable ‘i’ valores desde 0 hasta ‘f’, repitiendo sucesivamente las <instrucciones> que contiene.

```
console.log( variable );
```

Muestra en la consola una determinada “variable”

```
cont = cont + " | " + cntnd[i].href;
```

Carga en la variable “cont” una expresión (marcada en verde).

Efecto del Javascript en una página web :

Puede verse ese efecto en algunas páginas como estas : <https://www.rtve.es/> , Google.com o Elpais.com

Para deshabilitar Javascript de forma permanente en Chrome, tendremos que dirigirnos a :

Menú -> Configuración > Privacidad y seguridad > Configuración de sitios -> Contenido >
Javascript

... y lo desabilitamos.

Podemos llegar a este panel de opciones rápidamente si escribimos en la barra de direcciones: **chrome://settings/content/javascript**

Bloque 1 – Parte 3

Utilidad del WebCrawling y WebScraping
estático y dinámico. Ejemplos.



“The open web is by far the greatest global repository for human knowledge, there is almost no information that you can't find through extracting web data. This list of tools will help you take advantage of this information for your own projects and businesses.

Happy scraping!

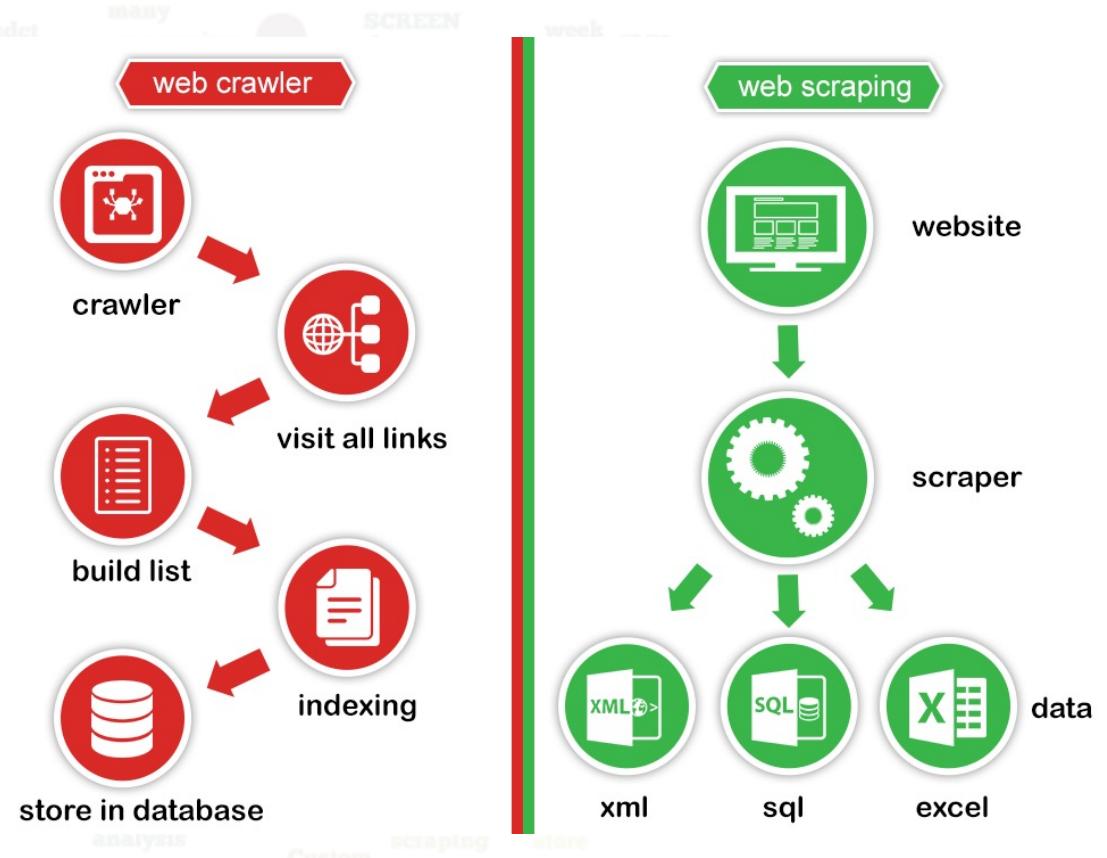
Fuente: <https://www.scrapingapi.com/blog/the-10-best-web-scraping-tools>

Extracción semiautomática de datos de Internet

Es necesario utilizar un **web crawler** cuando los contenidos que necesitamos extraer (**web scraping**) están distribuidos en una multitud de páginas web.

Normalmente se utiliza una base de datos con las URLs (direcciones en Internet) como **índice**. Esa base de datos podemos crearla desde una página donde aparezca una lista.

El crawler recorrerá el índice obteniendo las páginas contenidas en él. Una vez obtenidas esas páginas se les aplicará un algoritmo de scraping para extraer los datos que se pretenden conseguir.



Nota: En el **snippet 2** se puede ver una forma de crear un índice.

Dificultades scraping:

- Detección uso abusivo y bloqueo IP desde servidor.
- Utilización masiva de Javascript
- Cambios frecuentes de estructura/estilo website
- Limitación de acceso por CAPTCHA
- Limitación acceso a determinados dispositivos (agents)
- Carga de contenidos a partir de acciones del usuario (pulsar botón, o desplazarse al final de la página)

WebScraping estático vs dinámico

Caso 1. Podemos aplicar lo que llamamos **WebScraping estático**, cuando la página sobre la que queremos realizarlo NO utiliza javascript (AJAX) para cargar nuevos contenidos.

Caso 2. Por otra parte, cuando la página web utiliza javascript para obtener datos mediante AJAX, consideramos que estamos ante **WebScraping dinámico**. Si se utilizan las mismas librerías de R que en el caso anterior, **NO se obtendrá la información deseada**, ya que mediante javascript, el navegador está generando nuevos contenidos en tiempo real.

¿Por qué necesitamos dejar clara esta clasificación?

En cada caso se emplearán unas librerías de R distintas. Y además en el caso 2 (dinámico) se va a necesitar de un "conector" entre R y el navegador, que por suerte también instala R automáticamente.

Concretamente la librerías serán (entre otras) :
library("rvest") para el caso estático; y library("Rselenium") para el caso dinámico.

Efecto del Javascript en una página web :

Puede verse ese efecto en algunas páginas como estas : <https://www.rtve.es/> , Google.com o Elpais.com

Para deshabilitar Javascript de forma permanente en Chrome, tendremos que dirigirnos a :

Menú -> Configuración > Privacidad y seguridad > Configuración de sitios -> Contenido >
Javascript

... y lo desabilitamos.

Podemos llegar a este panel de opciones rápidamente si escribimos en la barra de direcciones: **chrome://settings/content/javascript**

Extensiones Chrome para editar páginas web

- Live CSS Editor
 - Stylebot
 - Code Cola
 - CSS Brush

<https://chrome.google.com/webstore/detail/css-brush-live-css-editor/mamnhhinmmdpcipjnhflilmhnogoobj>

Bloque 2

Iniciación a diversas herramientas para la obtención de datos en páginas web.
Aplicación práctica.

Bloque 3

¿Por qué R como herramienta?
Cómo configurar R para la obtención de bases de datos de páginas web.

If we go to the R web site in order to discover what R is all about, the first sentence we see is **R is a free software environment for statistical computing and graphics.**

I haven't been to the R web site in quite some time, but it struck me that the word "data" does not appear in that first sentence.

<https://simplystatistics.org/2018/07/12/use-r-keynote-2018/>

Is it an interactive system for data analysis or is it a sophisticated programming language for software developers? Or is system for developing reproducible workflows in data analysis? Or is it a platform for developing interactive graphics, dashboards, and web apps? Or is it a language for doing complex data wrangling and data management? Or...

<https://simplystatistics.org/2018/07/12/use-r-keynote-2018/>

¿Merece la pena "pedir" tiempo con R?

- Si sólo vas a utilizar R para extraer unos pocos datos concretos, mejor búscate alguien que lo haga por ti. O contrata un servicio de pago.
- Curva de aprendizaje de R **puede ser** algo mayor que otras herramientas estadísticas. **Aunque no necesariamente**. Existen multitud de GUI's (Graphical User Interface)

R console (Rgui), Rstudio,

R commander, Rattle,

Eclipse/StatET,

RapidMiner R extensión, RapidAnalytics,

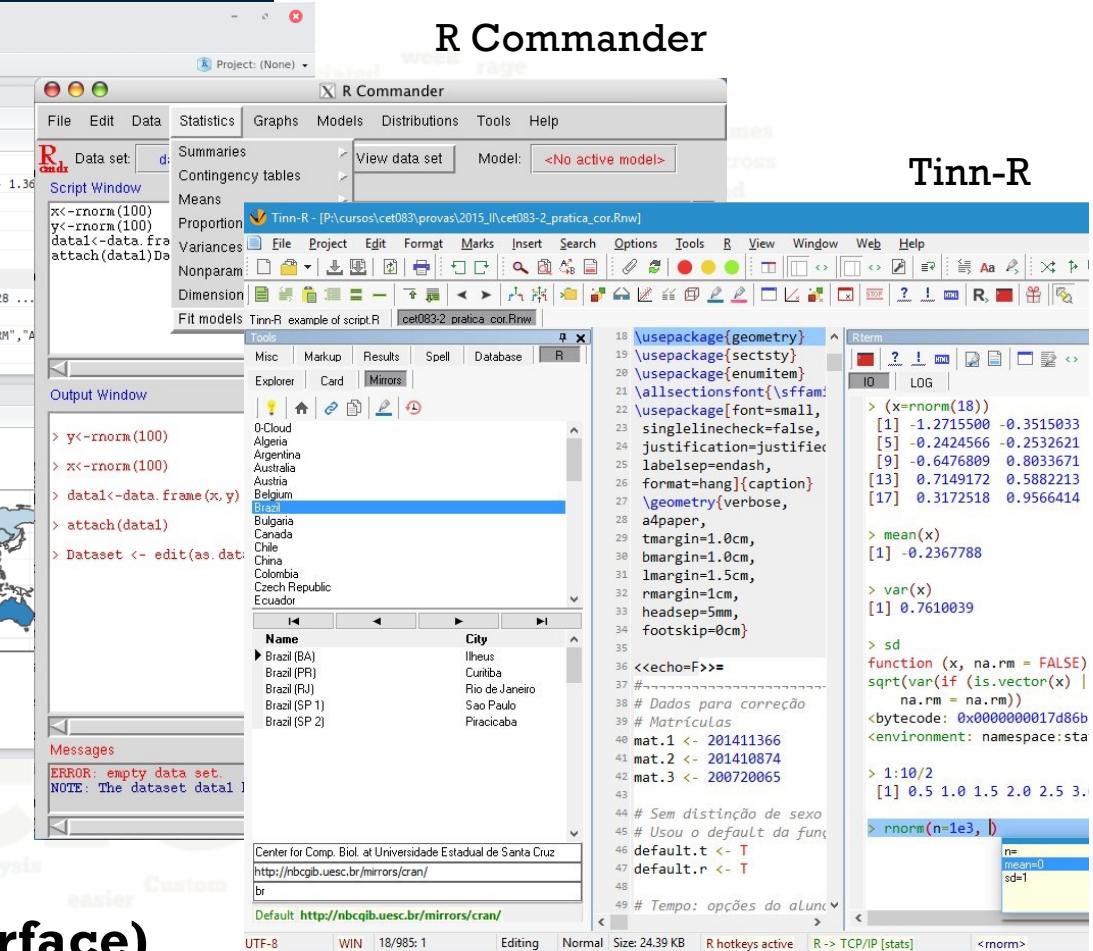
Tinn-R, ...

(<https://www.kdnuggets.com/polls/2011/r-gui-used.html>)

```
7 #Loading the rvest package
8 library('rvest')
9 #Specifying the url for desired website to be scrapped
10 url <- "http://www.wine-geography.com/wine-area.html"
11 # Reading the webpage
12 webpage <- read_html(url)
13
14 #Using CSS selector
15 Busqueda <- html_nodes(webpage, "div#content")
16 Bq <- html_node(Busqueda, "table")
17 html_attr(Bq, "border", 1)
18 html_attr(Bq, "width", "100%")
19
20 html_text(htm
21
rgui
```

The screenshot shows the RStudio interface with several windows open:

- Code Editor:** Displays the R script with code for scraping wine area data from a website.
- Environment:** Shows the global environment with objects like `geocep`, `intlabels`, `word.df`, `word.point`, and `Ws`.
- Plots:** A world map showing wine regions colored by wine area percentage, with labels for major wine-producing countries.
- Console:** Displays the R command used to generate the map.



Algunas R GUI's (Graphical User Interface)

¿Merece la pena "pedir" tiempo con R?

- Sin embargo, una vez aprendidos los fundamentos, las posibilidades son infinitas.



Además oportunidades de Investigación desarrollando en R

Revistas :

R Journal

ISSN: 2073-4859

JCR Q2 - SCIMAGO Q2 - H index 23

<https://www.scimagojr.com/journalsearch.php?q=21100255423&tip=sid&clean=0>

The Journal of Open Source Software

A developer friendly journal for research software packages.

<https://joss.theoj.org/>

Congresos:

useR! 2019 - @UseR2019_Conf

<http://user2019.r-project.org/dates/>

- **Reproducibility, Reporting, and Automation.** With the development of `knitr` and its combination with `R Markdown`, the writing of reproducible reports was made infinitely easier. (Markdown itself, probably deserves its own discussion, but it's not specifically R-related.)
- **Graphics.** R still has the ability to make great data graphics and with the introduction of `ggplot2`, it has become easier to make and extend good graphics.
- **R Packages and Community.** With over 13,000 packages on CRAN alone, there's pretty much a package to do anything. More importantly, the people contributing those packages and the greater R community have expanded tremendously over time, bringing in new users and pushing R to be useful in more applications. **Every year now there are probably hundreds if not thousands of meetups, conferences, seminars, tutorials, and workshops all around the world, all related to R.**
- **RStudio.** The development of the RStudio IDE has made **getting started with R much easier**. RStudio has significantly simplified the development of R packages via `devtools` and `roxygen2`. While it's not yet perfect, these tools have changed what used to be a labor-intensive and finicky process into a more manageable and easier to learn work flow. In addition, RStudio has funded the development of many critical R packages, including the members of the `tidyverse`.

<https://simplystatistics.org/2018/07/12/use-r-keynote-2018/>

Cómo configurar R para la obtención de bases de datos de páginas web.



R versión 4.1.1 for Windows

86 MB, 32/64 bit (2018-07-02) -- "Feather Spray"

<https://cran.r-project.org/bin/windows/base/R-4.1.1-win.exe>

<https://cran.r-project.org/>

Lanzamiento nueva versión : 10-08-2021



RStudio Desktop 2021.09.0+351 - Windows 10

156.88 MB

<https://www.rstudio.com/>

<https://download1.rstudio.org/desktop/windows/RStudio-2021.09.0%2B351.exe>

Existen también versiones de ambos para sistemas operativos LINUX y Mac

Packages:



[CRAN
Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)

<https://cran.rstudio.com/web/packages/index.html>

<https://cran.rstudio.com/web/views/>

<https://www.r-project.org/>

The screenshot shows the CRAN Packages index page. At the top, there is a banner with various words like "proud", "many", "SCREEN", "week", and "rage". Below the banner, the title "Contributed Packages" is centered. Underneath it, the heading "Available Packages" is followed by the text: "Currently, the CRAN package repository features **18287** available packages." A large red arrow points from the text "18287 available packages." to the right. Below this, there are two links: "Table of available packages, sorted by date of publication" and "Table of available packages, sorted by name". Further down, the heading "Installation of Packages" is shown, followed by a note about installing packages from the repository. At the bottom, it mentions "CRAN Task Views" and the number of available views.

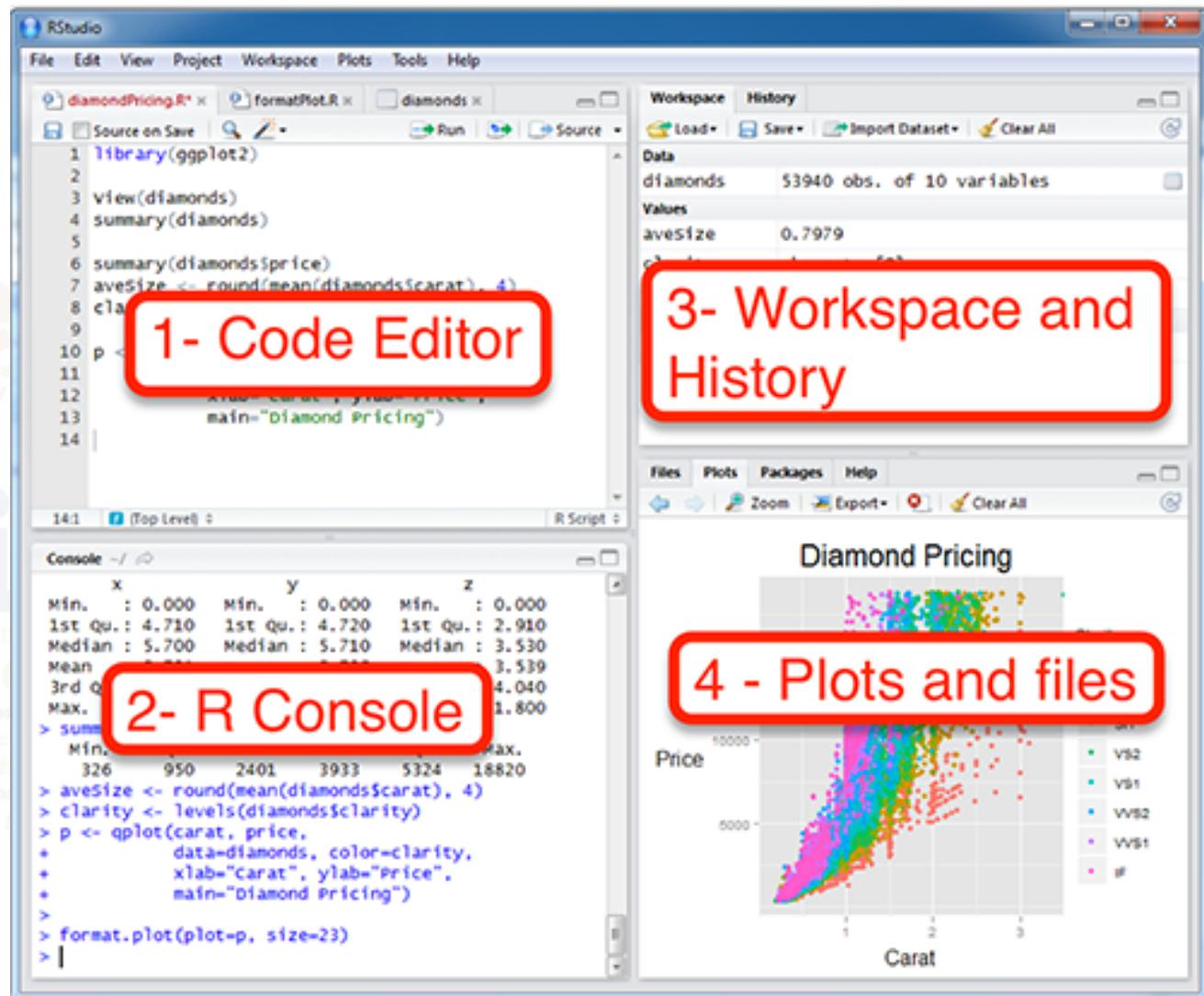
Instalación y configuración

R es muy fácil de instalar. Basta con descargar el fichero de instalación de la página web y ejecutarlo como administrador en nuestro sistema.

Para facilitar nuestro trabajo en R instalaremos también Rstudio, de la misma forma, y seleccionamos las opciones de instalación por defecto.

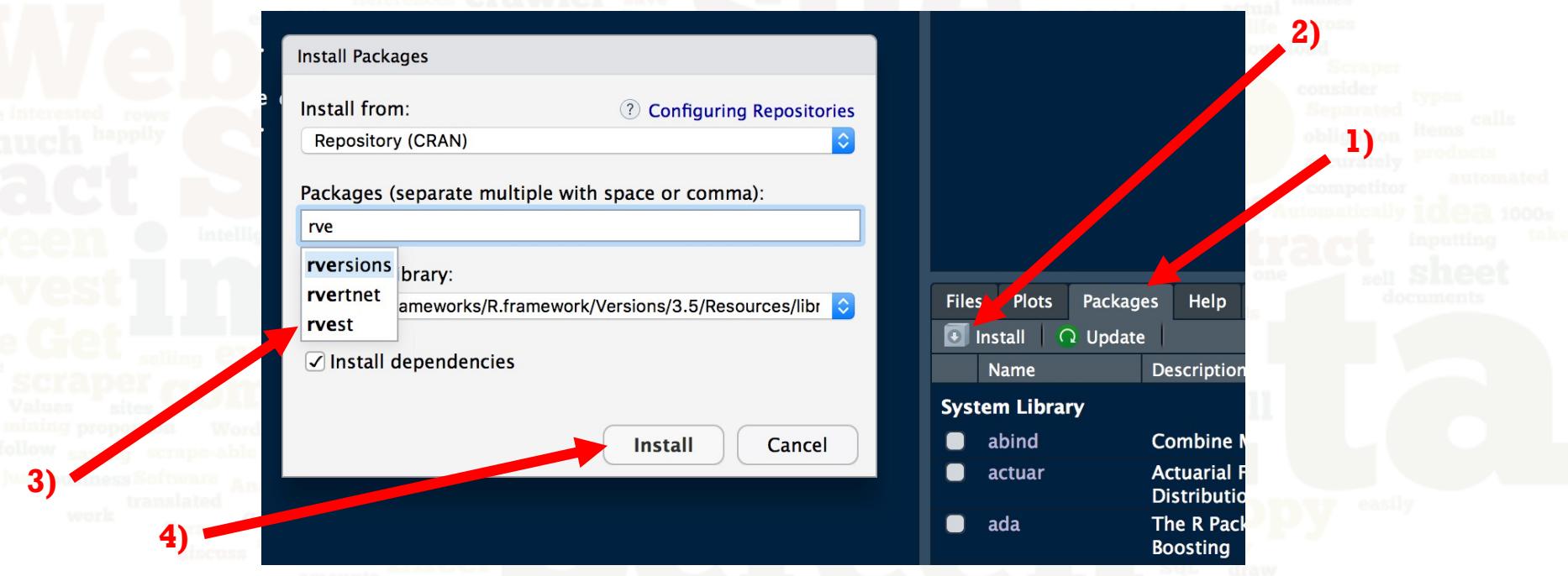
Una vez instalados, ejecutamos Rstudio.

Y ya podemos empezar a trabajar.



<http://www.sthda.com/english/wiki/running-rstudio-and-setting-up-your-working-directory-easy-r-programming>

Para realizar la extracción de datos de páginas web, existen varias posibilidades, como casi siempre en R. En nuestro caso utilizaremos el **paquete “rvest”**. Para instalarlo siga los pasos indicados en el gráfico o bien, introduzca el comando `install.packages("rvest")` en la Consola:



Una vez instalado el **paquete “rvest”**, tenemos que cargarlo en memoria. Para ello lo buscamos en la lista de Packages (paso 1) y lo seleccionamos. También podría hacerse introduciendo el siguiente comando en la Consola: `library("rvest")`

Bloque 4

Ejemplos prácticos de creación de bases de datos.

Ejemplo captura de datos de Internet con R

Práctica 1

- Descargar datos de contacto de Dptos. de la US desde R
Ejemplo de aplicación de Scraping para páginas web con un índice

Consultar código en :

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practical.R

Práctica 2

- Obtenga un fichero con los datos de los alumnos contenidos en las fichas que proporciona el sistema de la Universidad de Sevilla en formato PDF

Ejemplo de aplicación de Scraping sobre **fichero PDF**

Consultar código en :

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practica2.R

Práctica 3

- Obtenga un fichero con los nombres de los paquetes de R disponibles en la web

Ejemplo de aplicación de web scraping sobre una **TABLA**

Consultar código en :

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practica3.R

Práctica 4, 5 y 6

- Google Scholar
- TwitterR
- EXCEL ejemplo SEPE (tb puede verse web turismo alojamientos juntadeandalucia)

Consultar código en :

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practica4.R

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practica5.R

https://github.com/tonimoreno-us/Rscrap/blob/main/Sesion_2de3/R_practica6.R

Bloque 5

Otras posibilidades de R en el tratamiento automático de la información.

Otras posibilidades de R en el tratamiento automático de la información

Por su versatilidad y por la gran cantidad de paquetes disponibles, R puede ser utilizado para el tratamiento de datos desde otros formatos, no necesariamente desde la web.

Para ello será útil familiarizarse con otros paquetes como:

- rio
- datapasta
- clipr

Consúltese entre otros:

<https://cran.rstudio.com/web/views/WebTechnologies.html> (Sobre todo el apartado **Social Media Clients**)

Para poder hacer consultas en Facebook y Twitter se necesita registrarse y obtener unos tokens de acceso.