# Web Scraping con R

Jhoan Esteban Ruiz Borja

24 de mayo de 2018

## 1. Web Scraping

### 1.1. ¿Qué es Web Scraping?

Web Scraping se encarga de obtener los datos mediante un procesamiento del código HTML (HyperText Marckup Language) que forma la página. La práctica del Web Scraping es bastante útil para acceder, recopilar y organizar datos para la realización de reportajes de investigación, así como realizar cruce de datos con otra información.

### 1.2. Funcionamiento de Web Scraping

El funcionamiento del Web Scraping consiste en la transformación del contenido no estructurado, por lo general en formato HTML, en datos estructurados que pueden ser almacenados y analizados.

### 1.3. Técnicas usadas para el Web Scraping

Existe una gran variedad de técnicas que permiten el Web Scraping:

- **Recolección manual:** es quizás el más conocido y sencillo, consiste simplemente en copiar la información que nos interesa de un sitio web y pegarla.
- **Expresiones regulares:** son una secuencia de caracteres que forman un patrón de búsqueda. Normalmente es usado para la búsqueda o reconocimiento de cadenas de caracteres.
- **Protocolo HTTP:** se obtienen páginas web mediante peticiones HTTP a un servidor remoto mediante sockets.
- **Algoritmos de minería de datos:** la minería de datos se encarga de extraer información de un conjunto de datos, transformándola en una estructura comprensible para su posterior uso.
- **Parsers de HTML:** mediante el uso de lenguajes de programación se procesan documentos HTML, recuperando y transformando el contenido.
- **Reconocimiento de Web Semántica:** algunas páginas contienen metadatos o información semántica, como anotación o comentarios. Esta "Metainformación" puede ser usada para extraer la información deseada.

### 1.4. Herramientas para el Web Scraping

Existe una multitud de herramientas destinadas a la obtención de datos, cada herramienta se encuentra escrita en un lenguaje de programación, como puede ser Python, Ruby, PHP, R, Java, C++, entre otros.

## 2. Definiciones básicas

Entre las definiciones básicas se encuentran:

- **HTML**
- **XML**
- **JSON**
- **Xpath**

### 2.1. HTML

Existe un estándar oculto detrás de casi todo lo que vemos y hacemos cuando navegamos por la web, el Lenguaje de Marcado de Hipertexto, abreviado: HTML (HyperText Markup Language). Ya sea que busquemos información en Wikipedia, busquemos sitios en Google, revisemos nuestra cuenta bancaria o seamos sociales en Twitter, Facebook y YouTube, cuando usamos un navegador, usamos HTML.
HTML es un lenguaje para presentar contenido en la Web que fue propuesto por primera vez por Tim Berners-Lee (1989). El estándar ha evolucionado continuamente desde la introducción inicial, la encarnación más reciente es HTML5 que está siendo desarrollada por el World Wide Web Consortium (W3C) y el Grupo de Trabajo de Tecnología de Aplicaciones de Hipertexto Web (WHATWG). Aunque cada revisión de HTML ha establecido nuevas características y reestructurado las anteriores, la gramática básica de los documentos HTML no ha cambiado mucho a lo largo de los años y es probable que permanezca bastante estable en el futuro previsible, convirtiéndolo en uno de los estándares más importantes para trabajar con y sobre la web.

### 2.2. XML

XML, el eXtensible Markup Language, es uno de los formatos más populares para el intercambio de datos a través de la Web. Pero es más que eso. Es omnipresente en nuestra vida diaria. Como Harold y Means (2004, xiii) notan:

"XML se ha convertido en la sintaxis de elección para los formatos de documentos recientemente diseñados en casi todas las aplicaciones informáticas. Se usa en Linux, Windows, Macintosh y muchas otras plataformas informáticas. Los mainframes en Wall Street intercambian sus existencias intercambiando documentos XML. Los niños que juegan juegos en sus PC hogareños guardan sus documentos en XML. Los fanáticos de los deportes reciben puntajes de juegos en tiempo real en sus teléfonos celulares en XML. XML es simplemente la sintaxis de documentos más robusta, confiable y flexible que se haya inventado."

XML parece familiar para alguien con conocimientos básicos sobre HTML, ya que comparte las mismas características de un lenguaje de marcado. Sin embargo, HTML y XML sirven para

sus propios propósitos específicos. Si bien HTML se utiliza para dar forma a la visualización de la información, el objetivo principal de XML es almacenar datos. Por lo tanto, el contenido de un documento XML no se vuelve mucho más agradable cuando se abre con un navegador: los datos XML están envueltos en etiquetas definidas por el usuario. Las etiquetas definidas por el usuario hacen que XML sea mucho más flexible para almacenar datos que HTML

### 2.3. JSON

En esta sección, nos familiarizaremos con los beneficios del estándar de intercambio de datos JSON. El acrónimo (pronunciado "JSON") significa JavaScript Object Notation. JSON fue diseñado para las mismas tareas que XML a menudo se utiliza para el almacenamiento y el intercambio de datos legibles. Muchas API de aplicaciones web populares proporcionan datos en formato JSON.
Como su nombre lo sugiere, JSON es un formato de datos que tiene su origen en el lenguaje de programación JavaScript. Sin embargo, JSON es independiente del lenguaje y se puede analizar con muchos lenguajes de programación existentes, incluido R. JSON se ha convertido en uno de los formatos más populares para el suministro de datos web. Por lo tanto, vale la pena estudiarlo para nuestros propósitos.

### 2.4. Xpath

En un flujo de trabajo de análisis de datos típico, estos son importantes, pero solo pasos intermedios en el proceso de ensamblar conjuntos de datos bien estructurados y limpios de páginas web. Antes de que podamos aprovechar al máximo la Web como fuente de datos casi infinita, una vez que se hayan identificado y descargado los documentos web relevantes, se seguirán una serie de pasos de filtrado y extracción. El objetivo principal de estos pasos es refundir la información codificada en formatos que utilizan el lenguaje de marcado en formatos que son adecuados para su posterior procesamiento y análisis con software estadístico. Inicialmente, esta tarea comprende preguntar en qué información estamos interesados e identificar dónde se encuentra la información en un documento específico. Una vez que sabemos esto, podemos adaptar una consulta al documento y obtener la información deseada. Además, a menudo es necesario cambiar la forma de los datos y manejar las excepciones para convertir los valores extraídos en un formato que facilite un análisis posterior.

## 3. Web Scraping con R

Existe una gran cantidad de información valiosa que está disponible públicamente en línea, pero parece estar encerrada en páginas web que no son susceptibles de análisis de datos. Si bien muchas organizaciones facilitan sus datos a los investigadores, esta práctica sigue siendo la excepción y no la regla, y muchas veces los datos que buscamos pueden no estar centralizados o los creadores pueden no haber tenido la intención de crear una base de datos. Pero los investigadores recurren cada vez más a estas útiles fuentes de datos.
Por ejemplo, Grimmer (2013) analiza los comunicados de prensa de los miembros del Congreso para estudiar cómo los representantes se comunican con sus electores. Del mismo modo, Nielsen y Simmons (2015) utilizan los comunicados de prensa de la Unión Europea para medir si la ratificación del tratado lleva a los elogios de la comunidad internacional. En ambos casos, estos documentos estaban disponibles en línea, pero no estaban centralizados y sus autores no tenían la intención de crear una base de datos de comunicados de prensa para que los investigadores la analizaran. No obstante, representan datos valiosos para preguntas importantes de ciencias sociales, si tenemos una forma de ponerlos en un formato más útil.

Afortunadamente, hay muchas herramientas disponibles para traducir HTML ingobernable en bases de datos más estructuradas. El objetivo de este tutorial es proporcionar una introducción a la filosofía y la implementación básica de "web scrapingütilizando el lenguaje de programación estadística de código abierto R. Creo que la mejor manera de aprender web scraping es hacerlo, por lo que después de una breve descripción de las herramientas, la mayor parte de este documento se dedicará a trabajar a través de ejemplos.

### 3.1. Descripción general de alto nivel: el proceso de Web Scraping

En esencia, hay seis pasos para extraer datos basados en texto de un sitio web:

1 Identifica la información en internet que deseas usar.

2 Si esta información se almacena en más de una página web, descubra cómo navegar automáticamente a las páginas web. En el mejor de los casos, tendrá una página de directorio o la URL tendrá un patrón consistente que podrá recrear, por ejemplo, www.pelisplus.tv.

3 Ubique las características en el sitio web que marcan la información que desea extraer. Esto significa mirar el HTML subyacente para encontrar los elementos que desea y/o identificar algún tipo de patrón en el texto del sitio web que pueda explotar.

4 Escriba una secuencia de comandos para extraer, formatear y guardar la información que desee utilizando los indicadores que identificó.

5 Recorra todos los sitios web desde el paso 2 y aplique el guion a cada uno de ellos.

6 ¡Haz algunos análisis increíbles sobre tus datos recién desbloqueados!

Este tutorial se centrará en los pasos 3 y 4, que son la parte más difícil del web scraping.

También hay otra forma más sencilla de hacer web scraping de la que mostraré un ejemplo: a saber, el uso de interfaces de programación de aplicaciones (API) que algunos sitios web ponen a disposición. Las API básicamente le dan una forma simple de consultar una base de datos y devolver los datos que solicita en un formato agradable (generalmente JSON o XML).

Las API son geniales, pero generalmente no están disponibles, así que no las enfatizo aquí.

### 3.2. Conceptos básicos de HTML e identificación de la información que se quiere

El lenguaje de marcado de hipertexto - o HTML - es un sistema estandarizado para escribir páginas web. Su estructura es bastante simple, y comprender sus fundamentos es importante para lograr un raspado web exitoso.

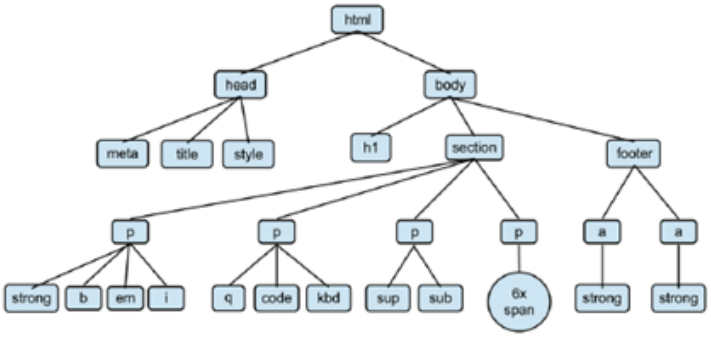


Figure 1: HTML document tree. Source: http://www.openbookproject.net/tutorials/getdown/css/images/lesson4/HTMLDOMTree.png

Esto es básicamente lo que parece un sitio web:

```
<!DOCTYPE html>
<html>

  <head>
    <title>This is the title of the webpage</title>
  </head>

  <body>
    <h1>This is a heading</h1>
    <p class="notThisOne">This is a paragraph</p>
    <p class="thisOne">This is another paragraph with a different class!</p>

    <div id="myDivID">
      <p class="divGraf">
        This is a paragraph inside a division, along with a
        <a href="http://stanford.edu">a link</a>.
      </p>
    </div>
  </body>

</html>
```

La Figura 1 también proporciona una representación visual de un árbol HTML.

Hay varias cosas que destacar sobre esta estructura. En primer lugar, los elementos siempre están rodeados de código que les dice a los navegadores web qué son. Estas etiquetas se abren con corchetes triangulares <tag>y se cierran con una barra dentro de más corchetes triangulares </tag>. En segundo lugar, estas etiquetas a menudo tienen información adicional, como información sobre la clase. En tercer lugar, estos elementos siempre están anidados dentro de otros elementos. Juntos, podemos utilizar estas características para extraer los datos que queremos.

Es fácil ver el HTML subyacente para cualquier página web: en Chrome, haz clic en Ver. ¡Desarrollador! Ver fuente. Esto es lo primero que debe hacer cuando desea extraer datos de una página web. También hay un complemento excelente de Chrome llamado SelectorGadget

que le permite señalar y hacer clic en las partes del sitio web que desea extraer. Le indicará automáticamente cuáles son las etiquetas subyacentes, y puede copiarlo y pegarlo en su secuencia de comandos.

### 3.3. Paquete rvest

En R existen muchos paquetes para realizar Web Scraping, como rvest.

¿Cómo se pueden seleccionar elementos de un sitio web en R? El paquete **rvest** es el kit de herramientas. El flujo de trabajo normalmente es el siguiente:

1 Lea una página web usando la función **read_html()**. Esta función descargará el HTML y lo almacenará para que **rvest** pueda navegarlo.

2 Seleccione los elementos que desee utilizando la función **html_nodes()**. Esta función tomará un objeto HTML (de **read_html**) junto con un selector CSS o Xpath (por ejemplo, p o span) y guardará todos los elementos que coincidan con el selector. Aquí es donde **SelectorGadget** puede ser útil.

3 Extraiga los componentes de los nodos que ha seleccionado usando funciones como **html_tag()** (el nombre de la etiqueta), **html_text()** (todo el texto dentro de las etiquetas), **html_attr()** (contenido de un solo atributo) y **html_attrs()** (todos los atributos).

### 3.4. Ejemplo simple de Web Scraping

Trabajando sobre el html, expuesto anteriormente se desarrolla un ejemplo simpre de extracción de información básica.

```
## Primero se instalan los paquetes que se requieren para Web Scraping
intall.packages("rvest")
intall.packages("magrittr")
intall.packages("httr")
intall.packages("XML")
intall.packages("stringr")
```

```
## Segundo se cargan las librerias instaladas
library(xml2)
library(rvest)
library(magrittr)
library(httr)
library(stringr)
```

```
## Se lee la página html con read_html(), llamada ejemplo.html
pagina_web = read_html("C:/Analitica/Books/Web Mining/Web Scraping/ejemplo.html")

## Para obtener los nodos ("p")
pagina_web %>% html_nodes("p")
```

```
## {xml_nodeset (3)}
## [1] <p class="notThisOne">This is a paragraph</p>
## [2] <p class="thisOne">This is another paragraph with a different class! ...
## [3] <p class="divGraf">\r\n\t\t\tThis is a paragraph inside a division,  ...
```

```
## Para obtener la clase (class "thisOne")
## se usa un punto para denotar la clase
pagina_web %>% html_nodes(".thisOne")
```

```
## {xml_nodeset (1)}
## [1] <p class="thisOne">This is another paragraph with a different class! ...
```

```
## Para obtener elementos con id "myDivID"
## se usa el hashtad para denotar el id
pagina_web %>% html_nodes("#myDivID")
```

```
## {xml_nodeset (1)}
## [1] <div id="myDivID">\r\n\t\t<p class="divGraf">\r\n\t\t\tThis is a par ...
```

```
## Para obtener el texto del elemento con id "myDivID"
## se utiliza la siguiente sentencia
pagina_web %>% html_nodes("#myDivID") %>% html_text()
```

```
## [1] "\r\n\t\t\r\n\t\t\tThis is a paragraph inside a division, along with a\r\n\t\t\t
```

```
## Para extraer el links de "myDivID".
## Primero se extrae todos los nodos "a", (como un href = "sitioweb.com")
## y se extrae un atributo "href" de estos nodos, de la siguiente forma
pagina_web %>% html_nodes("a") %>% html_attr("href")
```

```
## [1] "http://stanford.edu"
```

Aquí también se puede utilizar un complemento que tiene Google Crhome que se llama **SelectorGadget** que es muy útil para seleccionar texto html.

### 3.5. Ejemplo complejo de Web Scraping

Se desea extraer datos de la página, pelisplus y poder estructurar los datos de una forma que permita ser almacenada como una tabla de una base de datos. Por propósito de mostrar como extraer los datos se van a estraer los datos de 5 películas de acción.

```
## Primero se lee la página con la función read_html()
pelisplus = read_html("https://www.pelisplus.tv/")
```

```
## Luego se extraen los links de la página principal y
## se guardan en lista_links
```

```r
lista_links = pelisplus %>% html_nodes("a") %>% html_attr("href")
lista_links
```

```
##  [1] NA
##  [2] "https://www.pelisplus.tv"
##  [3] "https://www.pelisplus.tv/peliculas/"
##  [4] "https://www.pelisplus.tv/series/"
##  [5] "https://www.pelisplus.tv/documentales/"
##  [6] "https://www.animeshd.tv/"
##  [7] "https://www.pelispedia.tv/"
##  [8] "https://www.pelisplus.tv/kids/"
##  [9] "/login/"
## [10] "https://www.pelisplus.tv/registro/"
## [11] "/login/"
## [12] "/registro/"
## [13] "https://www.pelisplus.tv/"
## [14] "https://www.pelisplus.tv/peliculas/"
## [15] "https://www.pelisplus.tv/series/"
## [16] "https://www.pelisplus.tv/documentales/"
## [17] "https://www.pelisplus.tv/kids/"
## [18] "https://www.pelisplus.tv"
## [19] "#"
## [20] ""
## [21] "#car-hero"
## [22] "#car-hero"
## [23] "https://www.facebook.com/PelisplusOficial/?ref=bookmarks"
## [24] "https://www.pelisplus.tv/peliculas/ultimas-peliculas/"
## [25] ""
## [26] "#car-5"
## [27] "#car-5"
## [28] "https://www.pelisplus.tv/peliculas/actualizadas/"
## [29] ""
## [30] "#car-5"
## [31] "#car-5"
## [32] "https://www.pelisplus.tv/series/ultimas-series/"
## [33] ""
## [34] "#car-5"
## [35] "#car-5"
## [36] "https://www.pelisplus.tv/peliculas/popular-peliculas/"
## [37] ""
## [38] "#car-5"
## [39] "#car-5"
## [40] "https://www.pelisplus.tv/series/popular-series/"
## [41] ""
## [42] "#car-5"
## [43] "#car-5"
## [44] "https://www.pelisplus.tv/peliculas/accion/"
## [45] ""
## [46] "#car-5"
```

```
## [47] "#car-5"
## [48] "https://www.pelisplus.tv/peliculas/romance/"
## [49] ""
## [50] "#car-5"
## [51] "#car-5"
## [52] "https://www.pelisplus.tv/peliculas/comedia/"
## [53] ""
## [54] "#car-5"
## [55] "#car-5"
## [56] "https://www.pelisplus.tv/peliculas/terror/"
## [57] ""
## [58] "#car-5"
## [59] "#car-5"
## [60] "https://www.pelisplus.tv/es/pelisplus"
## [61] "https://www.pelisplus.tv/es/faq"
## [62] "https://www.pelisplus.tv/es/contact"
## [63] "https://www.pelisplus.tv/es/termsandconditions"
## [64] "https://www.pelisplus.tv/es/privacypolicy"
## [65] "http://pelispblog.com/"
## [66] "https://www.facebook.com/pelisplus"
## [67] "https://www.facebook.com/pelisplus"
## [68] "https://www.facebook.com/pelisplus"
## [69] "https://www.facebook.com/pelisplus"
## [70] "#"
## [71] "#"


## Luego se busca solo el link del que se desea extraer los datos y
## se guarda en link_peli_accion
link_peli_accion = grep("https://www.pelisplus.tv/peliculas/accion/", lista_links)
link_peli_accion


## [1] 44


## Luego se extrae es link de la peliculas de acción y se guarda en pelis_accion
pelis_accion = lista_links[link_peli_accion]
pelis_accion


## [1] "https://www.pelisplus.tv/peliculas/accion/"


## Luego vuelve y se lee esa página con read_html()
pelisplus_accion = read_html(pelis_accion)
pelisplus_accion


## {xml_document}
## <html lang="es">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body class="u-portal u-notLogged ">\n<header id="header" class="Hea ...
```

```r
## Luego se extraen los links de la página de películas de acción
lista_links_accion = pelisplus_accion %>% html_nodes("a") %>% html_attr("href")
lista_links_accion
```

```
##  [1] NA
##  [2] "https://www.pelisplus.tv"
##  [3] "https://www.pelisplus.tv/peliculas/"
##  [4] "https://www.pelisplus.tv/series/"
##  [5] "https://www.pelisplus.tv/documentales/"
##  [6] "https://www.animeshd.tv/"
##  [7] "https://www.pelispedia.tv/"
##  [8] "https://www.pelisplus.tv/kids/"
##  [9] "/login/"
## [10] "https://www.pelisplus.tv/registro/"
## [11] "/login/"
## [12] "/registro/"
## [13] "https://www.pelisplus.tv/"
## [14] "https://www.pelisplus.tv/peliculas/"
## [15] "https://www.pelisplus.tv/series/"
## [16] "https://www.pelisplus.tv/documentales/"
## [17] "https://www.pelisplus.tv/kids/"
## [18] "https://www.pelisplus.tv/pelicula/el-agente/"
## [19] "https://www.pelisplus.tv/pelicula/batman-ninja/"
## [20] "https://www.pelisplus.tv/pelicula/27-el-club-de-los-malditos/"
## [21] "https://www.pelisplus.tv/pelicula/el-diario-de-los-muertos/"
## [22] "https://www.pelisplus.tv/pelicula/sanson/"
## [23] "https://www.pelisplus.tv/pelicula/black-panther/"
## [24] "https://www.pelisplus.tv/pelicula/psychokinesis/"
## [25] "https://www.pelisplus.tv/pelicula/braven/"
## [26] "https://www.pelisplus.tv/pelicula/vigilancia-extrema/"
## [27] "https://www.pelisplus.tv/pelicula/gun-shy/"
## [28] "https://www.pelisplus.tv/pelicula/bushwick/"
## [29] "https://www.pelisplus.tv/pelicula/tremors-a-cold-day-in-hell/"
## [30] "https://www.pelisplus.tv/pelicula/deep-blue-sea-2/"
## [31] "https://www.pelisplus.tv/pelicula/mayhem/"
## [32] "https://www.pelisplus.tv/pelicula/la-4-ordf-compania/"
## [33] "https://www.pelisplus.tv/pelicula/el-corredor-del-laberinto-la-cura-
mortal/"
## [34] "https://www.pelisplus.tv/pelicula/league-of-gods-feng-shen-bang/"
## [35] "https://www.pelisplus.tv/pelicula/revolt/"
## [36] "https://www.pelisplus.tv/pelicula/just-getting-started/"
## [37] "https://www.pelisplus.tv/pelicula/condorito-la-pelicula/"
## [38] "https://www.pelisplus.tv/pelicula/el-pasajero/"
## [39] "https://www.pelisplus.tv/pelicula/day-of-the-dead-bloodline/"
## [40] "https://www.pelisplus.tv/pelicula/lara-croft-tomb-raider-la-cuna-de-
la-vida/"
## [41] "https://www.pelisplus.tv/pelicula/lara-croft-tomb-raider/"
## [42] "https://www.pelisplus.tv/pelicula/proud-mary/"
## [43] "https://www.pelisplus.tv/pelicula/suicide-squad-hell-to-pay/"
```

```
## [44] "https://www.pelisplus.tv/pelicula/game-over-man/"
## [45] "https://www.pelisplus.tv/pelicula/codigo-abierto/"
## [46] "https://www.pelisplus.tv/pelicula/el-nacimiento-del-dragon/"
## [47] "https://www.pelisplus.tv/pelicula/talento-de-barrio/"
## [48] "https://www.pelisplus.tv/pelicula/las-mil-caras-de-dunjia/"
## [49] "https://www.pelisplus.tv/pelicula/cold-war-ii/"
## [50] "https://www.pelisplus.tv/pelicula/steel-rain/"
## [51] "https://www.pelisplus.tv/pelicula/o-matador/"
## [52] "https://www.pelisplus.tv/pelicula/el-implacable/"
## [53] "https://www.pelisplus.tv/pelicula/kickboxer-venganza/"
## [54] "https://www.pelisplus.tv/pelicula/star-wars-episodio-viii-los-ultimos-
jedi/"
## [55] "https://www.pelisplus.tv/pelicula/jumanji-bienvenidos-a-la-jungla/"
## [56] "https://www.pelisplus.tv/pelicula/beyond-skyline/"
## [57] "https://www.pelisplus.tv/pelicula/thor-ragnarok/"
## [58] "https://www.pelisplus.tv/pelicula/el-ultimo-disparo/"
## [59] "https://www.pelisplus.tv/pelicula/overdrive/"
## [60] "https://www.pelisplus.tv/pelicula/liga-de-la-justicia/"
## [61] "https://www.pelisplus.tv/pelicula/asalto-al-convoy/"
## [62] "https://www.pelisplus.tv/pelicula/actos-de-venganza/"
## [63] "https://www.pelisplus.tv/pelicula/plan-b/"
## [64] "https://www.pelisplus.tv/pelicula/enemigo-publico/"
## [65] "https://www.pelisplus.tv/pelicula/the-villainess/"
## [66] "https://www.pelisplus.tv/pelicula/suburbicon/"
## [67] "https://www.pelisplus.tv/pelicula/batman-gotham-luz-de-gas/"
## [68] "javascript:void(0)"
## [69] "/peliculas/accion/"
## [70] "/peliculas/accion/?page=2"
## [71] "/peliculas/accion/?page=3"
## [72] "/peliculas/accion/?page=4"
## [73] "/peliculas/accion/?page=5"
## [74] "/peliculas/accion/?page=6"
## [75] "/peliculas/accion/?page=7"
## [76] "/peliculas/accion/?page=8"
## [77] "/peliculas/accion/?page=9"
## [78] "/peliculas/accion/?page=10"
## [79] "/peliculas/accion/?page=23"
## [80] "/peliculas/accion/?page=2"
## [81] "https://www.pelisplus.tv/es/pelisplus"
## [82] "https://www.pelisplus.tv/es/faq"
## [83] "https://www.pelisplus.tv/es/contact"
## [84] "https://www.pelisplus.tv/es/termsandconditions"
## [85] "https://www.pelisplus.tv/es/privacypolicy"
## [86] "http://pelispblog.com/"
## [87] "https://www.facebook.com/pelisplus"
## [88] "https://www.facebook.com/pelisplus"
## [89] "https://www.facebook.com/pelisplus"
## [90] "https://www.facebook.com/pelisplus"
## [91] "#"
```

```
## [92] "#"

## Luego se extraen los links de las películas y se guarda en peliculas_accion
peliculas_accion = grep("https://www.pelisplus.tv/pelicula/.", lista_links_accion)

## Luego se seleccionan solo esas películas y se guarda en links_peliculas_accion
links_peliculas_accion = lista_links_accion[peliculas_accion]
links_peliculas_accion

##  [1] "https://www.pelisplus.tv/pelicula/el-agente/"
##  [2] "https://www.pelisplus.tv/pelicula/batman-ninja/"
##  [3] "https://www.pelisplus.tv/pelicula/27-el-club-de-los-malditos/"
##  [4] "https://www.pelisplus.tv/pelicula/el-diario-de-los-muertos/"
##  [5] "https://www.pelisplus.tv/pelicula/sanson/"
##  [6] "https://www.pelisplus.tv/pelicula/black-panther/"
##  [7] "https://www.pelisplus.tv/pelicula/psychokinesis/"
##  [8] "https://www.pelisplus.tv/pelicula/braven/"
##  [9] "https://www.pelisplus.tv/pelicula/vigilancia-extrema/"
## [10] "https://www.pelisplus.tv/pelicula/gun-shy/"
## [11] "https://www.pelisplus.tv/pelicula/bushwick/"
## [12] "https://www.pelisplus.tv/pelicula/tremors-a-cold-day-in-hell/"
## [13] "https://www.pelisplus.tv/pelicula/deep-blue-sea-2/"
## [14] "https://www.pelisplus.tv/pelicula/mayhem/"
## [15] "https://www.pelisplus.tv/pelicula/la-4-ordf-compania/"
## [16] "https://www.pelisplus.tv/pelicula/el-corredor-del-laberinto-la-cura-
mortal/"
## [17] "https://www.pelisplus.tv/pelicula/league-of-gods-feng-shen-bang/"
## [18] "https://www.pelisplus.tv/pelicula/revolt/"
## [19] "https://www.pelisplus.tv/pelicula/just-getting-started/"
## [20] "https://www.pelisplus.tv/pelicula/condorito-la-pelicula/"
## [21] "https://www.pelisplus.tv/pelicula/el-pasajero/"
## [22] "https://www.pelisplus.tv/pelicula/day-of-the-dead-bloodline/"
## [23] "https://www.pelisplus.tv/pelicula/lara-croft-tomb-raider-la-cuna-de-
la-vida/"
## [24] "https://www.pelisplus.tv/pelicula/lara-croft-tomb-raider/"
## [25] "https://www.pelisplus.tv/pelicula/proud-mary/"
## [26] "https://www.pelisplus.tv/pelicula/suicide-squad-hell-to-pay/"
## [27] "https://www.pelisplus.tv/pelicula/game-over-man/"
## [28] "https://www.pelisplus.tv/pelicula/codigo-abierto/"
## [29] "https://www.pelisplus.tv/pelicula/el-nacimiento-del-dragon/"
## [30] "https://www.pelisplus.tv/pelicula/talento-de-barrio/"
## [31] "https://www.pelisplus.tv/pelicula/las-mil-caras-de-dunjia/"
## [32] "https://www.pelisplus.tv/pelicula/cold-war-ii/"
## [33] "https://www.pelisplus.tv/pelicula/steel-rain/"
## [34] "https://www.pelisplus.tv/pelicula/o-matador/"
## [35] "https://www.pelisplus.tv/pelicula/el-implacable/"
## [36] "https://www.pelisplus.tv/pelicula/kickboxer-venganza/"
## [37] "https://www.pelisplus.tv/pelicula/star-wars-episodio-viii-los-ultimos-
jedi/"
```

```
## [38] "https://www.pelisplus.tv/pelicula/jumanji-bienvenidos-a-la-jungla/"
## [39] "https://www.pelisplus.tv/pelicula/beyond-skyline/"
## [40] "https://www.pelisplus.tv/pelicula/thor-ragnarok/"
## [41] "https://www.pelisplus.tv/pelicula/el-ultimo-disparo/"
## [42] "https://www.pelisplus.tv/pelicula/overdrive/"
## [43] "https://www.pelisplus.tv/pelicula/liga-de-la-justicia/"
## [44] "https://www.pelisplus.tv/pelicula/asalto-al-convoy/"
## [45] "https://www.pelisplus.tv/pelicula/actos-de-venganza/"
## [46] "https://www.pelisplus.tv/pelicula/plan-b/"
## [47] "https://www.pelisplus.tv/pelicula/enemigo-publico/"
## [48] "https://www.pelisplus.tv/pelicula/the-villainess/"
## [49] "https://www.pelisplus.tv/pelicula/suburbicon/"
## [50] "https://www.pelisplus.tv/pelicula/batman-gotham-luz-de-gas/"
```

```
## Luego se sacan los datos que se desean estructurar y
## se guardan en datos_pelisplus_accion
datos_pelisplus_accion = data.frame(
                  Titulo = rep(NA_character_, length(links_peliculas_accion)),
                   Anno = NA_character_,
                   Duracion = NA_character_,
                   Categoria = NA_character_,
                   stringsAsFactors = F)

## Se crea un ciclo para recorrer la lista de películas
for(i in 1:length(links_peliculas_accion)){

    # Titulo de la pelicula
    datos_pelisplus_accion$Titulo[i] = read_html(links_peliculas_accion[i]) %>% html_no
title") %>% html_text()
    # Año de la pelicula
    datos_pelisplus_accion$Anno[i] = (read_html(links_peliculas_accion[i]) %>% html_noc
    # Duracion de la pelicula
    datos_pelisplus_accion$Duracion[i] = read_html(links_peliculas_accion[i]) %>% html_
data-item--duration") %>% html_text()
    # categoria
    datos_pelisplus_accion$Categoria[i] = (read_html(links_peliculas_accion[i]) %>% htm

}
datos_pelisplus_accion
```

```
##                                      Titulo Anno Duracion Categoria
## 1                                 El agente  2017   95 min    Acción
## 2                               Batman Ninja  2018   85 min    Acción
## 3                 27: El club de los malditos  2018      N/A    Acción
## 4                    El diario de los muertos  2007   95 min    Acción
## 5                                    Sansón  2018  110 min    Acción
## 6                              Black Panther  2018  134 min    Acción
## 7                              Psychokinesis  2018  101 min    Acción
```

```
## 8                                      Braven  2018   94 min    Acción
## 9                            Vigilancia Extrema  2013  119 min    Acción
## 10                                     Gun Shy  2017   92 min    Acción
## 11                                   Bushwick  2017   94 min    Acción
## 12                  Tremors: A Cold Day in Hell  2018   98 min    Acción
## 13                               Deep Blue Sea 2  2018   94 min    Acción
## 14                                       Mayhem  2017   86 min    Acción
## 15                                La 4ª compañía  2016  109 min    Acción
## 16   El corredor del laberinto: La cura mortal  2018  141 min    Acción
## 17              League of Gods (Feng shen bang)  2016  109 min    Acción
## 18                                       Revolt  2017   87 min    Acción
## 19                          Just Getting Started  2017   91 min    Acción
## 20                        Condorito: la película  2017   88 min    Acción
## 21                                  El pasajero  2018  105 min    Acción
## 22                    Day of the Dead: Bloodline  2018   90 min    Acción
## 23  Lara Croft Tomb Raider: La cuna de la vida  2003  117 min    Acción
## 24                       Lara Croft: Tomb Raider  2001  100 min    Acción
## 25                                   Proud Mary  2018   89 min    Acción
## 26                  Suicide Squad: Hell to Pay  2018   86 min    Acción
## 27                               ¡Game Over, Man!  2018     N/A    Acción
## 28                                 Código abierto  2017   98 min    Acción
## 29                       El Nacimiento del Dragón  2016   95 min    Acción
## 30                              Talento de Barrio  2008  107 min    Acción
## 31                        Las mil caras de Dunjia  2017  113 min    Acción
## 32                                    Cold War II  2016  110 min    Acción
## 33                                     Steel Rain  2017  139 min    Acción
## 34                                      O Matador  2017   99 min    Acción
## 35                                  El Implacable  2017  113 min    Acción
## 36                             Kickboxer: Venganza  2016   90 min    Acción
## 37 Star Wars: Episodio VIII - Los últimos Jedi  2017  152 min    Acción
## 38              Jumanji: Bienvenidos a la jungla  2017  119 min    Acción
## 39                                 Beyond Skyline  2017  106 min    Acción
## 40                                 Thor: Ragnarok  2017  130 min    Acción
## 41                              El último disparo  2017   97 min    Acción
## 42                                     Overdrive  2017   93 min    Acción
## 43                             Liga de la Justicia  2017  120 min    Acción
## 44                                Asalto al convoy  2016  102 min    Acción
## 45                              Actos de venganza  2017   87 min    Acción
## 46                                        Plan B  2016   98 min    Acción
## 47                                Enemigo público  1998  132 min    Acción
## 48                                 The Villainess  2017  129 min    Acción
## 49                                     Suburbicon  2017  105 min    Acción
## 50                    Batman: Gotham: Luz de gas  2018   78 min    Acción
```