# WebScraping Ripley.cl y Falabella.com

**Descripción**

Diseñar la automatización de un proceso que compare una vez al día los precios de los televisores LED que se encuentran en www.falabella.cl con los que aparecen en www.ripley.cl.

**Planteamiento del problema y tecnologías**

En un escenario ideal en el que las APIs de ambas empresas fueran públicas la extracción sería mucho más sencilla. En este caso no nos queda más alternativa que extraer la información directamente del HTML con una técnica de Web Scraping. La gran contrapartida de esta técnica es que un cambio en la web echa a perder parte del trabajo realizado.

Para realizar esto vamos a poner a prueba tecnologías más novedosas. Node nos permite escribir aplicaciones en JavaScript y contiene una infinidad de módulos que podemos aprovechar para realizar esta tarea.

Módulos que usaremos

- **fs** nos permitirá escribir en un fichero de texto.
- **request** descargaremos el HTML necesario para el *parser*.
- **cheerio** Nos crea un objeto JQuery a partir del HTML descargado.

Los resultados obtenidos para cada web las vamos a guardar en formato JSON en diferentes ficheros dentro de la carpeta OUTPUT. Cada objeto contiene un par con el nombre del modelo y el precio.

El resultado del código creado puede ser ejecutado en cualquier máquina con independencia del sistema operativo. Se debería programar una tarea para ejecutar la herramienta preferiblemente todos los días a la misma hora cuando no haya tanto tráfico.

**Detallando el problema y dándole solución**

En este problema además nos encontramos con que la extracción de ambas webs se tiene que hacer de forma diferente. A continuación se detalla el *modus operandi* para cada caso:

**Ripley.cl**

En el caso de Ripley.cl toda la información que necesitamos puede ser extraída de la misma página, aunque esta se muestra los productos de 24 en 24. Para ello podemos engañar al navegador para forzar que muestre por ejemplo de 5000 en 5000 y así tener todos los objetivos en el mismo HTML. La URL que fuerza esto es:

http://www.ripley.cl/ripley-chile/tecnologia/tv/SearchDisplay?urlRequestType=Base&storeId=10151&catalogId=10051&langId=-5&categoryId=12184&urlLangId=-5&beginIndex=0&pageSize=5000

Con esto ya solo tenemos que buscar los tags HTML donde se encuentra el modelo y el precio de cada televisor. Dentro del contenedor donde se muestran todos los aparatos se repite el patrón para mostrar cada producto. El modelo se encuentra dentro de un tag que lleva por clase “product_name” y el precio dentro de otro tag que lleva por clase “price”.

**Falabella.com**

En Falabella la cosa se complica ya que el en la lista principal no podemos encontrar el nombre del modelo. Para atacar el problema obraremos como en el caso anterior forzando que la página principal muestre todos los televisores de golpe. La URL que fuerza esto es:

http://www.falabella.com/falabella-cl/category/cat70043/Televisores?No=0&[redacted]&userSelectedFormat=list

Para cada producto que encontramos en la lista buscaremos la URL de detalles técnicos de cada televisor. Cuando saquemos la URL hacemos un *request* de los detalles y sacamos el nombre del modelo. En el código fuente se puede ver mejor las etiquetas HTML que se buscan para este caso.