Extracción de datos para el desarrollo con herramientas digitales

Pablo Martín | PR3SSH

Índice

- Entornos no Open Data
- Importar datos a Google Sheets
- Extracción genérica en HTMLs (*scraping*)
- Extracción de tablas en PDFs

Entornos no Open Data

Entornos no Open Data

En la mayoría de portales de datos abiertos nos encontramos con datos que no está liberada basada en las especificaciones de los datos abiertos y que solo se encuentran incrustados dentro de páginas web.

Por desgracia esta práctica es más común de lo que se desearía, como mecanismo aglutinador de enlaces a otras webs en donde ya se encontraba dicha información que se pretende liberar.

Entornos no Open Data

Siempre podemos acceder a dichos sitios web y copiar y pegar la información que nos interese. En ese caso puede aparecer el problema del tiempo (que se puede resolver añadiendo más personas al proceso) pero para rescatar dicha información de una manera lo más automática posible se suele usar una técnica denominada *scraping*.

Entornos no Open Data

Todas las páginas web están construidas con código HTML, por lo que en resumen el *scraping* lo que hace como técnica es recuperar el código fuente de la página web de la que queremos extraer la información y procesar dicho contenido para extraer del mismo únicamente los apartados que nos interesen y de una forma lo más automatizada posible.

Entornos no Open Data

Con la aplicación del mecanismo de *scraping* reducimos drásticamente el tiempo que se tarda en recuperar los datos necesitados, añadiendo al problema una pizca de pensamiento computacional a la hora de encontrar la mejor solución.

Importar datos a Google Sheets

Importar datos a Google Sheets

- Google Sheets es la versión web de Google de la famosa categoría de software llamada hojas de cálculo
- Pertenece a la suite de Google Apps por lo que se le permite añadir comentarios, compartir con colegas, gestionar niveles de acceso y muchas otras opciones comunes
- Dispone de un gran número de herramientas y utilidades para trabajar con datos en el entorno Web

Importar tablas HTML a Google Sheets

```
IMPORTHTML(url, “table”, index)
```

- `url`: Dirección web de la que se desea extraer los datos
- “`table`”: Parámetro para indicar que se desea obtener una tabla
- `index`: Lugar en el que se encuentra la tabla dentro de la web

Importar listas HTML a Google Sheets

```
IMPORTHTML(url, “list”, index)
```

- `url`: Dirección web de la que se desea extraer los datos
- “`list`”: Parámetro para indicar que se desea obtener una lista
- `index`: Lugar en el que se encuentra la lista dentro de la web

Importar *feed* a Google Sheets

```
IMPORTFEED(url, query, headers, num_items)
```

- `url`: Dirección web del *feed*
- `query`: Elemento que se desea importar del *feed*
- `headers`: Se desea incluir cabeceras o no (TRUE o FALSE)
- `num_items`: Número de elementos a importar

Importar XML a Google Sheets

```
IMPORTXML(url, query)
```

- `url`: Dirección web de los datos estructurados (XML o xHTML)
- `query`: Criterio de búsqueda en el XML usando XPATH

Importar CSV a Google Sheets

```
IMPORTDATA(url)
```

- `url`: Dirección web en donde se encuentra el fichero CSV (o TSV) que se desea importar

PRÁCTICA

Mostraremos unos ejemplos y realizaremos una serie de ejercicios usando las funciones de obtención de datos de Google Sheets

Extracción Genérica en HTMLs (*scraping*)

Flujo general del *scraping*

1. Encontrar la web
2. Seleccionar los datos
3. Obtener los datos
4. Procesar los datos

Flujo del código del *scraping*

1. Encontrar la web
2. Obtener la url inicial
3. Obtener el HTML de la url
4. Procesar el contenido del HTML
5. Seleccionar datos
6. Procesar los datos
7. (volver a 3 si hay más urls)

Data-Miner.io

- Herramienta online para realizar *scraping*
- Se materializa en una extensión de navegador
- Facilidad de uso
- Requiere alta de usuario
- Plan gratuito y de pago
- Se puede instalar desde Google Chrome Store

Data-Miner.io

GRATUITO	DE PAGO
500 páginas/mes	100.000+ pages/month
50.000+ recetas públicas	50.000+ recetas públicas
Navegación “siguiente página”	Navegación “siguiente página”
Descarga de datos	Descarga de datos
	Automatización avanzada
	Scrapear sitios premium

Data-Miner.io: la extensión

La extensión de navegador permite generar recetas, las cuales se encargarán de obtener los datos seleccionados.

Consta de:

- Obtener datos similares (*get similar*)
- Gestor de recetas (*recipe manager*)
- Creador de recetas (*recipe creator*)
- Colecciones de datos (*data collections*)
- Trabajos (*jobs*)

Data-Miner.io: obtener datos similares

- Se activa pulsando el botón derecho del ratón en la sección de la web que deseemos obtener.
- Previamente hay que seleccionar el texto (o parte de él) que contiene los datos de dicha web.
- Funciona bastante bien con tablas y listas.
- Una vez pulsado se te abre el gestor de recetas para ver los resultados y editar la receta creada automáticamente.

Data-Miner.io: gestor de recetas

- Se activa pulsando en el icono de Data Miner, el cual se puede encontrar en la parte superior derecha del navegador.
- Se abre una interfaz flotante en la que se puede acceder a:
  - Mis recetas
  - Recetas favoritas
  - Recetas públicas
  - Crear nueva receta (con el creador de recetas)
  - Importar datos (CSV o plantilla)
  - Acceso a las colecciones de datos
  - Acceso a los trabajos

Data-Miner.io: creador de recetas

En esta opción se te permitirá crear, de forma más personalizada, recetas para obtener datos de distintas páginas web. Consta de 7 pasos:

1. Inicio
2. Filas
3. Columnas
4. Navegación
5. Acciones
6. Javascript
7. Guardar

Data-Miner.io: creador de recetas (1)

En el paso 1 (inicio) se dará la bienvenida a la herramienta de creación de recetas y se solicitará el tipo de página de la que se quiere obtener la información. Los dos tipos de los que dispone son listados (más de un resultado) y detalle (un solo resultado).

Data-Miner.io: creador de recetas (2)

En el paso 2 (filas) se le permitirá al usuario seleccionar los datos que definan cada una de las filas del fichero resultante. Se permite realizar mediante el botón “Find” (usando el ratón y la tecla Shift para seleccionar) o si no usando directamente selectores CSS o XPATH (búsquedas en XML).

Data-Miner.io: creador de recetas (3)

En el paso 3 (columnas) se podrá seleccionar cada una de las columnas con las que contará el fichero de datos resultante. Dichas columnas deben ser extraídas de cada una de las filas del paso anterior.

Para seleccionar cada una de las columnas podemos hacer uso de la herramienta “Find”, que también la vimos en el paso 2 (filas).

Data-Miner.io: creador de recetas (4)

En el paso 4 (navegación) lo que se va a definir es el botón o enlace que se debe pulsar para avanzar en la página de la web sobre la que estamos obteniendo los datos. La forma de definir dicho elemento es similar a las realizadas en el paso 2 y 3.

Además, en este mismo paso, y una vez seleccionado el botón siguiente, podemos simular una prueba para comprobar que la navegación está correctamente configurada.

Data-Miner.io: creador de recetas (5)

En el paso 5 (acciones) se permite añadir una serie de acciones previas a la obtención de los datos, de forma que éstas aseguren que toda la información a obtener está cargada satisfactoriamente en el HTML de la página.

Algunas de estas acciones *pre-scraping* son:

- Pulsar botón para que se muestre información
- Hacer scroll hasta el final de la página
- Hacer scroll infinito

Data-Miner.io: creador de recetas (6)

En el paso 6 (Javascript) se permite añadir código de programación para realizar modificaciones sobre los datos obtenidos en el proceso de obtención de los mismos.

Algunas de las modificaciones más comunes están relacionadas con la limpieza de datos, conversiones de formatos, etc.

Data-Miner.io: creador de recetas (7)

En el paso 7 (guardar) se permite guardar la receta o guardarla como nueva, además de ejecutarla para probarla.

PRÁCTICA

Mostraremos unos ejemplos y realizaremos una serie de ejercicios usando la herramienta Data-Miner.io

Extracción de tablas en PDFs

Datos en formato PDF

Como se comentaba al principio, no siempre nos encontramos los datos en formatos Open Data. Nada es perfecto (aunque sería fácil que así fuera). Es muy común que los datos/información sean liberados en PDF, por el carácter inmutable de dicho formato.

En este caso, es muy común que dentro de los PDFs hayan tablas cargadas de datos. Esto último le da sentido a este apartado.

Advertencia

Los PDFs pueden venir en dos formatos:

- Texto, si fue correctamente generado
- Imagen, si fue generado de otra forma

En este apartado nos centraremos en los PDF de tipo texto, ya que para la extracción de datos de los PDF de tipo imagen deberíamos de aplicar técnicas algo más complejas del tipo OCR (*Optical Character Recognition* o Reconocimiento Óptico de Caracteres).

Tabula.technology

- Tabula es una herramienta que nos permite extraer de forma sencilla tablas incrustadas dentro de ficheros de tipo PDF.
- Disponible para Windows, Mac y GNU/Linux.
- Herramienta de software libre (licencia MIT)

Tabula.technology: el proceso básico (1)

- Tenemos un PDF con tablas de datos
- Abrimos Tabula y se nos carga una página web
- Pulsamos en *Browse* y seleccionamos el PDF
- Pulsamos en *Import* para cargar el PDF
- Vemos el PDF en una lista y pulsamos en *Extract Data*
- Se nos carga el PDF a pantalla completa

Tabula.technology: el proceso básico (2)

- Ahora podemos indicarle a Tabula que *Autodetect Tables*
- O si no seleccionarla nosotros mismos con el ratón
- Además podemos buscar tablas similares a la seleccionada en el PDF con la opción *Repeat this Selection*.
- Luego pulsamos en *Preview & Export Extracted Data*
- En esta pantalla final solo nos queda seleccionar el formato en el que deseemos guardar los datos y pulsar *Export*.

PRÁCTICA

Mostraremos unos ejemplos y realizaremos una serie de ejercicios usando la herramienta Tabula

URLs para los ejercicios en

https://gist.github.com/pr3ssh/13cc6dee48c783b8e9f87a06381dfe8e