

Procedimiento del script “Procesamiento en Python” – Web scraping

La siguiente nota tiene como motivo instruir en la formación de *web scraping*, entendido este –en pocas palabras– en una forma de obtener información ágil y velozmente de páginas web sin una intervención directa sobre la misma.

El producto final que se obtiene es una lista de inmuebles en venta dentro de la ciudad de Santa Fe que son provistos por un medio web. En particular, se obtienen datos del título, precio y fecha de publicación.

Para ejecutar el *script* es necesario tener algunas librerías –o bibliotecas– y funciones específicas. Tal como se visualiza en las celdas de código uno y tres del *script*, se requiere de:

- BeautifulSoup: biblioteca de Python que se utiliza para analizar documentos HTML o XML.
- requests: biblioteca de Python que facilita la realización de solicitudes HTTP.
- datetime.date: es parte del módulo¹ datetime de Python que permite trabajar con fechas
- pandas: es una poderosa biblioteca de Python para el análisis y manipulación de datos.

Las primeras dos bibliotecas se utilizan para realizar el *scrapeo* de la página web, mientras que las dos segundas se utilizan para manipular los datos traídos desde la página web.

La segunda celda de código es la que requiere mayor atención dado que allí es donde se realiza el *scrapeo*. En primer lugar, se definen elementos vacíos (*precios_def*, *titulo_def* y *fecha_pub_def*). En segundo lugar, se realiza un *loop* que tendrá por objeto completar los elementos vacíos que fueron creados. A decir verdad, si bien se ha especificado que se ha utilizado un medio web, cuando uno realiza una búsqueda de inmuebles, obtiene una cierta cantidad de resultados por página. Lo que interesa es entrar en cada una de estas páginas.

En particular, este loop entra en cada una de las páginas que se especifica. La cantidad de páginas a especificar dependerá de la cantidad de resultados que uno quiera obtener. Para ello, se recomienda visualizar la cantidad máxima de resultados por páginas, y la cantidad máxima de resultados en todas las páginas.

Especial atención deberá ponerse sobre tres elementos de este bloque de código:

```
titulo_divs = pagina_filt.find_all("div", {"class": "titulo"})
precios_divs = pagina_filt.find_all("div", {"class": "precio"})
fecha_pub_divs = pagina_filt.find_all("div", {"class": "fecha_pub"})
```

Los elementos distintivos que se encuentran entre llaves, {}, contienen el elemento que en definitiva se *scrapea*, en caso de buscar replicar para otros medios web, se requiere modificar específicamente estos elementos.

Por último, una vez que se completan los [elementos vacíos previamente creados](#), se almacena todo en una base de datos. Por último, se ejecuta un código que permite visualizar el resultado y, seguidamente, otro código que permite descargar los mismos en una planilla de Excel.

¹ Una biblioteca puede contener más de un módulo. En este sentido, si bien los módulos tienen funciones, son más pequeños que las bibliotecas.