Búsqueda y obtención de datos – Web Scraper

En este video, les mostraré cómo usar una extensión realmente chévere en Google Chrome llamada Web Scraper. Pueden descargar Web Scraper si van a Chrome Web Store y luego buscan Web Scraper. Encuentran una página como esta. Web Scraper les permite raspar información de sitios web para que puedan comenzar a construir sus propios conjuntos de datos.

Por ejemplo, echen un vistazo a la página de Billboard 200, que enumera las principales canciones éxitos en una semana determinada, las 200 canciones principales en la lista de Billboard. Y si desearan construir el conjunto de datos a partir de este sitio, probablemente tendrían que copiar y pegar toda la información aquí. No podrían copiarlo todo y luego pegarlo en otro lugar. Probablemente tendrían que ingresar manualmente toda la información aquí. Y esto no es una tabla de hecho. Por lo que el comando importHTML en Google Sheets no funcionará aquí.

Entonces, lo que queremos intentar hacer es capturar toda la información que está aquí para hacer una tabla con columnas para que al final se vea algo así. Tener una canción en una columna, el artista, la posición, la URL de la imagen e incluso la semana del día específico que raspamos esta información.

Devolviéndonos aquí, lo que queremos hacer es encontrar patrones que podamos identificar para poder raspar esta información y convertirla en un conjunto de datos. Y qué vamos a hacer: echen un vistazo a la página web y vean que hay casillas allí, casillas blancas aquí y raspar información es siempre intentar encontrar patrones. Entonces, ¿cómo pueden encontrar estos patrones y transformarlos en el conjunto de datos que necesitan?

En este caso están todas estas casillas blancas. Hay 200 de ellas. Tienen toda la información que necesitamos. Y si echamos un vistazo a las variables, como el nombre de las columnas que queremos, están todas aquí. El nombre de la canción está aquí, el nombre del artista. La posición está aquí y también el álbum, ¿cierto?, la imagen del álbum. E incluso la información sobre la semana está aquí. Así que vamos a raspar toda esta información usando Web Scraper.

Lo hacen accediendo al menú Web Inspector. Haga nclic derecho en cualquier lugar de la página y hagan clic en “Inspect”, y vean aquí que en la parte derecha de la pestaña hay una nueva opción llamada Web Scraper. Ahora, si tienen su pestaña en el lado derecho, simplemente hagan clic en los tres puntos aquí y seleccionen esta opción “Dock to bottom”, y luego pueden ver el Web Scraper aquí.

Ahora Web Scraper comienza aquí en blanco. Tienen tres opciones en la parte superior, Sitemaps, así es como ellos llaman a los robots que comenzarán el raspado y los procesos que empezarán a raspar cosas para ustedes. Tenemos la opción de Sitemaps y luego vamos a

“Create a new sitemap” o “Import a new sitemap”. Vamos a continuar y seleccionar “Create a new sitemap” y asígnenle un nombre, Billboard 200, y luego la URL. Esta es la URL aquí. Luego hacemos clic en “Create sitemap”.

Y ahora lo que vamos a hacer es agregar un nuevo selector. Selector es la información que podemos identificar como elementos en la página web. Así que solo hagan clic en el botón azul aquí y queremos decirle a Web Scraper dónde están estas casillas blancas, ¿correcto? Queremos decirle a Web Scraper: “Web Scraper, encuentra 200 casillas en esta página”, y luego le diremos a Web Scraper dónde está la información en todas estas casillas para que pueda rasparla.

Pero por ahora, llamemos a esto una “casilla” y este es un elemento en la página web, ¿verdad? Entonces el tipo aquí es “Element”. Hacemos clic aquí en “Select”, y estas serán múltiples casillas. Luego seleccionamos esta casilla aquí llamada “múltiple”. Y tengan en cuenta que cuando comiencen a mover el mouse, verán que Web Scraper interactúa con la página que muestra todos estos elementos y qué tan *clicables* y seleccionables son. Entonces, lo que queremos hacer es encontrar un lugar aquí en la esquina inferior derecha que resalte en verde la casilla completa desde la parte superior. Hagamos clic aquí y se volverá rojo tan pronto como hagamos clic. Y luego haremos lo mismo para el próximo y veremos que Web Scraper intenta adivinar dónde están las demás casillas. Pero luego se detiene en la 21. Así que haremos esto nuevamente para la 21 y veremos que identificó todas las 200 casillas aquí en la página. Entonces aquí es donde queremos llegar, ¿verdad?

Hacemos esto, y ahora nos damos cuenta de que aquí, hay un selector que identifica, que describe todas estas casillas. No necesitan preocuparse por eso. Web Scraper selecciona automáticamente, identifica el selector por ustedes. Simplemente hagan clic aquí en “Done selecting!” (¡Selección terminada!). Y tengan en cuenta que esto vendrá aquí. Y una vez que hacen eso, esta es la señal correcta para que incluso puedan obtener una vista previa de estos datos. No hay datos aquí que estemos raspando, pero también pueden hacer una “Element preview” (Vista previa del elemento) y verán que todas las casillas están marcadas. Así que guardamos eso, y ahora tenemos la casilla. Ahora tenemos el proceso para capturar la casilla, pero necesitamos el proceso para capturar información en todas las 200 casillas.

Entonces, si pasan el mouse aquí, verán que resalta en gris la línea de la casilla. Y si hacen clic en la casilla, noten que ahora estamos dentro de estas casillas genéricas que hemos seleccionado aquí. Podemos volver a “_root” y volveremos en un segundo, pero luego haremos clic en la casilla y ahora estamos en esta casilla que acabamos de describir y queremos decirle al Web Scraper: “mira, sabes la ubicación de las 200 casillas, pero ahora quiero que captures la información en cada casilla y crees el conjunto de datos para mí”.

Entonces, capturaremos el nombre de la canción, el nombre del artista, la posición y la URL de la imagen del álbum, y hagámoslo. Hagan clic en “Add new selector” (agregar nuevo selector), recuerden: aquí dentro de la casilla. Así que agregamos un nuevo selector, llamemos a este

selector “song” (canción). Hacemos clic en “Select” aquí, luego seleccionamos el nombre de la canción, este identifica el selector, y luego hacemos clic en “Done selecting!” (¡Selección terminada!). Podemos previsualizar los datos. Captura todos los nombres de las canciones aquí y parece que todo está bien, y luego guardamos el selector.

Ahora agreguemos uno nuevo. Hagan clic en el botón azul. Este será el artista. Este también es un tipo de texto porque es texto aquí en la página. Hacemos clic aquí, este es el artista. Es un selector “a”. Hacemos clic en “Done selecting!” (¡Selección terminada!) y luego podemos obtener una vista previa de los datos. Aquí muestra todos los nombres de los artistas que capturó en esta página. Guardamos el selector.

Ahora agreguemos la posición. Este también es un elemento de texto. Hacemos clic aquí en “Select” y aquí está la posición y luego en “Done selecting!” (¡Selección terminada!). Quizás se pregunten por qué no marca la casilla “múltiple”. Como lo seleccionamos para las múltiples casillas que estábamos capturando. Como solo estamos capturando una posición aquí, no hay varias posiciones dentro de esta casilla amarilla, no seleccionaremos las casillas múltiples. Solo en las ocasiones en las que tienen que seleccionar múltiples elementos que se repiten en posiciones o en diferentes casos, seleccionen “múltiples”. Pero en este caso es solo una posición, un nombre de canción, un artista, entonces no marcamos la casilla “múltiple”. Todo bien. Así que también guardamos el selector.

Ahora vamos a agregar la imagen. Esta es una imagen, así que llamemos a esto una “imagen”. Seleccionamos la imagen del álbum aquí y luego “Done selecting!”, y luego guardamos. Ahora, si volvemos a “_root” y vamos a “Data preview” aquí, verán que ya tienen casi todo lo que necesitamos, ¿verdad? Tenemos la canción, el artista, la posición y la URL de la canción. Pero eso no es exactamente lo que queremos, porque también queremos la fecha, ¿verdad? Y la fecha está aquí en la parte superior.

Entonces, lo que queremos hacer es aplicar esta fecha a cada fila aquí. Y para hacer eso, volvemos a “_root” donde estamos. Estábamos en una casilla, ahora volvamos a “_root” y agreguemos un nuevo selector para la fecha. Y este también es un selector de texto, por lo que resaltaremos todo esto aquí y luego haremos clic en “Done selecting!”. Pero noten que cuando van a la vista previa, este toma toda esta información aquí que no quiero, sólo quiero este “August 31” y 2019. Entonces, ¿cómo pueden extraer sólo esto? No entraré en detalles porque usaré expresiones regulares para hacer esto. Siéntanse libre para buscar en Google expresiones regulares. Hay excelentes tutoriales, excelentes lecciones de expresiones regulares. Hay unas muy buenas, especialmente en programación. Pero afortunadamente, Web Scraper también tiene un “Regex” o campo de expresión regular aquí donde pueden usar expresiones regulares. Entonces, lo que vamos a hacer es: voy a copiar todo, y encontraré un patrón para extraer solo esta parte aquí, este “August 31, 2019” y lo haré para que se extraiga cada vez que haya un texto aquí que muestra una palabra y un número, una coma, un año. Él extraerá sólo eso.

Entonces, voy aquí a este probador en línea de expresiones regulares, y veo que ya tengo mi cadena de caracteres aquí. Entonces es la semana del 31 de agosto de 2019 y luego la semana pasada, la próxima semana, la semana actual, buscar por fecha. Lo que estoy haciendo aquí es usar expresiones regulares y tengo tres elementos. El primero es un "\w+" y lo que significa es que esta barra invertida (\) es como un patrón cuando quieran usar una especie de símbolo, que es esta "w" aquí. Entonces "w" significa cualquier carácter de palabra. Cualquier a, b, c, d, o cualquier número. Por lo tanto, se une a todo lo que sucede aquí y el "+" significa cualquier carácter para un sólo carácter, o un número infinito de caracteres hasta que llegue al espacio. Pero no es solo un espacio, es un espacio anterior, como un "\d", que significa dígitos, ¿verdad? Luego corresponde a un dígito igual a 0-9, y el "+" significa coincidencias entre 1 y tiempo ilimitado, por lo tanto, cualquier número de dígitos. Entonces, es una palabra con cualquier número de caracteres, un espacio que precede a cualquier número de dígitos, y luego hay una coma, y luego un espacio, y luego cuatro dígitos. Es lo que significa eso.

Y si copian esto aquí en el campo de expresión regular en un Web Scraper y obtienen una vista previa de los datos, verán que extrae todo y luego deja "August 31, 2019". Y eso es exactamente lo que queremos. Guardaré este selector aquí y listo, terminamos.

Entonces vamos a ir a aquí. Aquí pueden seleccionar otras opciones, pueden ver el gráfico selector donde ven "_root" y todos los demás selectores y la relación entre ellos. Esto puede ser bastante complicado dependiendo de la complejidad de la página. Pueden editar los metadatos, el nombre del sitio web o la URL. También pueden raspar. Pueden examinar los datos raspados y pueden exportar este mapa del sitio. Así que esto es como un JSON, una cadena de caracteres, que pueden exportar y usar en otros computadores, o enviarla a un amigo que cargará el mismo proceso de raspado en otros computadores. O pueden modificarlo un poco para cambiar el sitio. Entonces hay una manera de exportar. Y también pueden importar. Y pueden exportar los datos que rasparon como archivo CSV.

Avancemos y raspemos aquí en esta opción. Luego les dan dos opciones. El "request interval", que es la cantidad de tiempo que esperarán mientras realizan una solicitud al sitio. Y dos MS, que son dos segundos, es una buena práctica. No desean colapsar el sitio con solicitudes, ya que puede parecer sospechoso. El webmaster puede pensar que están intentando derribar el sitio, y no queremos eso. Ustedes querrán usar esta opción de manera responsable. Y luego está el retraso de carga de la página. El retraso de carga de la página es la cantidad de tiempo que Web Scraper espera a que se cargue la página para luego raspar los datos. Por lo tanto, es posible que deseen darles al sitio algo de tiempo para cargar los datos y luego rasparlos para asegurarse de que todos los elementos se carguen antes de comenzar a capturar información. Entonces dos y dos segundos son buenos números para comenzar, pero pueden ajustarlos dependiendo de sus casos.

Luego hagan clic en "start scraping" (comenzar a raspar) aquí, se abre una ventana. Este espera dos segundos para cargar la página y espera dos segundos para realizar la solicitud, y

luego raspa los datos. Y si hacen clic en “refresh” (actualizar) aquí, cargará todos los datos que acaba de raspar, y *voilá*, tenemos aquí los metadatos que agregó Web Scraper. Esta es la identificación de Web Scraper para cada registro, tienen la URL de inicio. Aquí tenemos los datos que realmente queremos raspar: la canción, el artista, la posición, la URL de la imagen y también la fecha que se aplicó a cada registro aquí. Así que ahora podemos seguir y hacer clic aquí y exportar este sitio como CSV. Y cuando hagan clic aquí, descargarán el archivo CSV a sus computadores, y luego podrán importar el archivo CSV a cualquier otra aplicación de hoja de cálculo para comenzar a analizar, limpiar, editar o construir su conjunto de datos.

Así que está listo para Web Scraper. Entonces vayan a Chrome Web Store, descarguen la extensión y comiencen a raspar.