Finding & Getting Data - Web Scraper

In this video I'm going to show you how to use a very neat extension in Google Chrome called Web Scraper. You can download Web Scraper if you go to Chrome Web Store and then you look for Web Scraper. You found a page like this.

Web Scraper allows you to scrape information out of websites so that you can start building your own data sets. So for example, take a look at the Billboard 200 web page which lists the top songs in any given week, the 200 top songs for Billboard. And if you wanted to get the data set out of this you probably would have to copy and paste all of the information here. You can't really copy all of this, and then paste somewhere else. You probably have to manually put all the information here. And this is not actually a table. So, importHTML on Google Sheets will not work here. So what we want to try to do is then capture all of the information that is here to making to a table with columns so that in the end it will look something like this. Have a song in a column, the artist, the position, you know, the URL for the image and even the week for the particular day that we scraped this information.

So coming back here what we want to do is to find patterns that we can identify so that we can scrape this information into, this information into a data set. And what we're going to do is just take a look at the web page and see that there are boxes right there like, white boxes here and scraping information is always trying to find patterns. So how you can find these patterns and then transform this into the data that the data set that you need. So in this case there are all of these white boxes. There are two hundred of them. They have all of the information that we need. And if we take a look at you know like the variables, like the name of the columns that we want they're all in here. The name of the song is here, the name of the artist. The position is here and also in the album, right, the image of the album. Right. And even the information about the week is here. So we're going to scrape all of this information using Web Scraper.

So you do that by accessing the Web Inspector menu. So right click anywhere on the web page and then click on "Inspect", and see here that on the rightmost part of the tab there is a new option called Web Scraper. Now if you, if you had your tab on the right side just click here on this. The three dots here and select here this option, "Dock to bottom", and then you can see Web Scraper here. Now Web Scraper starts here as blank. You have three options at the top, Sitemaps, that's how they call the robots that will start scraping and the routines that will start scraping things for you. We have a Sitemaps options and then here to "Create a new sitemap" or "Import a new sitemap". We're going to go ahead and select "Create sitemap" going to give a name, Billboard 200, and then the URL. Is this URL right here. And then we click on "Create sitemap".

And now what we're going to do is to add a new selector. Selector are, you know, the information that we can identify elements on the web page. So just click here and the blue button and we want to tell Web Scraper where these white boxes are. Right. We want to tell Web Scraper, "Web Scraper find 200 boxes on this page", and then after that we're going to tell Web Scraper where the information is in all of these boxes so that it can scrape it. But for now let's just call this "box" and it's an element on the web page. Right. So the type here is "Element". We click here on "Select", and this is, this is going to be multiple boxes. So we check this box here called multiple. And look when you start

hovering your mouse around you see that Web Scraper interact with the webpage showing all of these elements and how clickable they are and how selectable they are.

So what we want to do is to find a spot here in the right lower corner that highlights in green the full box at the top. We're going to click here and then it will become red, as soon as we click. And then we're going to do the same for the next one and see that Web Scraper tries to guess where all of the other boxes are. But then it stops at the 21st. So we're gonna do that again for the 21st, and see that it identified all of the 200 boxes here on the page. So this is where we want to be, right. We do this, and now notice that here, there is selector that identifies that describes all of these boxes you don't have to worry about this. Web Scraper automatically selects, identify the selector for you. Just click here on "Done selecting!" And notice that this will come here. And then once you do that this is the right signal so you can even preview this data. There's no data here that we're scraping but it can also do an "Element preview" and you see that all of the the boxes are selected. So we save this, and now we have the box. Now we have the routine to capture the box, but we need the routine to capture information in all of the 200 boxes.

So if you, if you hover your mouse here you see that it highlights in gray the box row here. And if you click on the box, notice that we're now inside of this like generic boxes that we selected here. We can come back to "_root" and we're going to come back there in a second, but then we would click on the box and now we're in this box that we just described and we want to tell Web Scraper, "look you know the location of the two hundred boxes, but now I want you to capture information in every single box and do the data set for me".

So, we're going to capture the name of the song, the name of the artist, the position, and the URL of the image of the album, and let's do that. So, you "Add new selector", remember inside here of the box. So we add a new selector, let's call this selector "song". And then, we click here on "Select", and then we select the name of the song, it identifies which selector it is and then click on "Done selecting!". We can do a data preview. So it captures all of the name of the songs here so it looks fine, and then we save the selector. Now we add a new one. Click on the blue button. This will be, this will be the artist.

This is also a text type because it's text here on the page. So we click here, this is the artist. It's an "a" selector. We click on "Done selecting!" And then we can do a data preview. Here it shows all the names of the artists that it captured on this page. We save the selector. Now we're going to add the position. Now this is also a text element. So we click here on "Select" and here's the position and then then "Done selecting!".

Now you might be asking, why not check the box multiple? Right. Because we check that for the multiple boxes that we were capturing. Since we're capturing only one position here, there are not multiple positions inside of this yellow box, we're not going to check the multiple boxes. Only when there are occasions that you have to select multiple elements that repeat in positions or in different cases, that you would select multiple. But in this case is just one position, one name of song, one artist, so you don't check the multiple box.

All right. So we save the selector too. Now we're going to add the image, right. So this is an image, and then we'll call this "image". We select here, the image of the album and then "Done selecting!", and then we save. Now if we go back to "_root" and do "Data preview" here, you will see that it already almost has everything that we need right.

We have the song, the artist, the position and then the URL of the song. But it's not quite everything that we want, because we also want the date, right? And date is here at the top. So what we want to do is to apply this date for every single row here. And to do that, we go back to "_root" where we are. So we were in a box, now we go back to "_root" and then we add a new selector for the date. And this is also a text selector, and then we're going to highlight here all of this, and then we're gonna do "Done selecting!" But notice when you preview this, it gets all of this information here that I don't want, I just want this "August 31st" and then 2019. So how do how do you can extract just this?

I'm not gonna go into details because I'm going to use regular expressions to do that. Feel free to google regular expressions or great tutorials, great classes about regular expression. They're very powerful, especially in programming. But thankfully the Web Scraper also has a "Regex" or a regular expression field here that you can apply regular expressions. So what we're gonna do is just I'm going to get all of this, and I want to find a pattern to extract just this part here. This "August 31st, 2019" and I'm gonna do it in a way that it extracts every time there is a text here that shows a word in a number comma, a year it's going to extract just that.

So, I'm going to go here to this online regular expression tester, and see that I already have here my string. So the week of August 31st, 2019 and then last week, next week, current week, date search. What I'm doing here is using regular expression and I have three elements. The first one is a "\w+" and it means this backslash is like a standard when you want to use like in something of a token, that's called this "w" here. So "w" means any word character. So any a, b, c, d, or any number. So it matches anything that happens here and the "+" means any character to just one character, or an infinite number of characters until it hits the space. But it's not only a space, it's a space that comes before like a "\d" which means digits. Right. So it matches a digit equals to 0-9, and the plus means matches between 1 and unlimited time, so any number of digits. So, it's a word with any number of characters, a space that comes before any number of digits, and then there is a comma, and then a space, and then four digits. That's what this means here.

And if you copy this here to the regular expression field in a web scraper and you do a data preview you see that it extracts everything out and then it relieves only the August 31st, 2019. And that's exactly what we want. I'm gonna save this selector here and then we're all done. So we're gonna go here. Here you can select other options, you can see the selector graph where you see the "_root" and then all of the other selectors and then the relation between them.

It can get fairly complicated depending on the complexity of the page. You can edit the metadata, the name of the site, or the URL. You can also scrape. You can browse the scrape data and you can even export this site map. So this is like a Jason that you can string, that you can export and use in other computers, or send to a friend that will load the same scraping routine in other computers. Or you can tweak it a bit to change the website. So there is a way to export.

And then you can import as well and then you can export the data that you scraped as "CSV". So we're gonna go ahead and scrape here on this option. So it gives you two, two options. The "request interval" which is like the amount of time that you will wait until it does a request to the website. And two ms, which is two seconds is good practice. So you don't wanna be hammering the site with requests, so we might look suspicious. The

webmaster might think that you're trying to take this website, down so we don't want it dead. You want to use this option responsibly.

And then there's the page load delay. The page load delay is the amount of time that web scraper waits for the page to load and then to scrape the data. So you might want to give the Web site some time to load the data and then to scrape it to finish, so that you can guarantee that all elements load before you start capturing information with that.

So two and two seconds are good numbers to start but you might tweak it depending on your case. So you click here on "start scraping", it opens a window. It waits two seconds to load the page and then you wait two seconds to do the request, and then it scrapes the data. And if you click here on "refresh" it will load all of the data that it just scraped, and voila, we have here the same metadata that Web Scraper adds. This is the web scraper I.D. for each one of the records, you have to start URL. Here we have the data that we really want to scrape which song, the artist, the position, the image URL, and also the date that was applied to every record here.

So now we can go ahead and click here and export this site at CSV. And when you click here you're going to download the CSV file to your computer, and then you can import the CSV file to any other spreadsheet application to start analyzing or cleaning, or editing, or building your dataset. So that's it for Web Scraper. So, go to the Chrome Web Store, download the extension and start scraping.