Localizando e obtendo os dados - Web Scraper

Neste vídeo eu vou mostrar como usar uma extensão muito legal no Google Chrome chamada Web Scraper. Você pode baixar o Web Scraper se você for à Chrome Web Store e então você busca Web Scraper. Você encontra uma página como esta. O Web Scraper permite que você raspe informações dos sites para que você possa começar a construir seus próprios conjuntos de dados.

Por exemplo, dê uma olhada na página web Billboard 200, que lista as canções no topo das paradas em qualquer semana, as 200 canções na parada da Billboard. E se você quisesse extrair o conjunto de dados deste site, você provavelmente teria que copiar e colar todas as informações aqui. Você não poderia copiar tudo isso, e depois colar em outro lugar. Você provavelmente teria que colocar manualmente todas as informações aqui. E isso não é uma tabela, na verdade. Então o comando importHTML no Google Sheets não vai funcionar aqui.

Então, o que nós queremos tentar fazer é capturar toda a informação que está aqui para fazer uma tabela com colunas de modo que, no final, será algo parecido com isso. Ter uma canção em uma coluna, o artista, a posição, a URL para a imagem e até mesmo a semana do dia específico em que raspamos esta informação.

Então, voltando aqui, o que queremos fazer é encontrar padrões que podemos identificar para que possamos raspar estas informações e transformá-las em um conjunto de dados. E o que vamos fazer: dê uma olhada na página web e veja que há caixas ali, caixas brancas aqui e raspar informações é sempre tentar encontrar padrões. Então como você pode encontrar esses padrões e transformar isso no conjunto de dados de que você precisa.

Neste caso, há todas estas caixas brancas. Há 200 delas. Elas têm todas as informações de que precisamos. E se damos uma olhada nas variáveis, como o nome das colunas que queremos, elas estão todas aqui. O nome da canção está aqui, o nome do artista. A posição está aqui e também o álbum, certo, a imagem do álbum. E até as informações sobre a semana estão aqui. Então vamos raspar toda esta informação usando Web Scraper.

Você faz isso acessando o menu do Web Inspector. Clique com o botão direito em qualquer lugar na página e clique em Inspecionar, e veja aqui que na parte mais à direita da guia há uma nova opção chamada Web Scraper. Agora, se você tiver sua guia no lado direito basta clicar sobre os três pontos aqui e selecionar esta opção "Dock to bottom", e então você pode ver o Web Scraper aqui.

Agora Web Scraper começa aqui em branco. Você tem três opções na parte superior, Sitemaps, assim são chamados os robôs que vão começar a raspagem e os processos que vão começar a raspar as coisas para você. Temos a opção Sitemaps e depois vamos aqui em "Create a new sitemap" ou "Import a new sitemap". Vamos selecionar "Create sitemap" e dar um nome, Billboard 200, e depois a URL. É esta URL aqui. Então clicamos em "Create sitemap".

E agora o que vamos fazer é adicionar um novo seletor. Seletor são as informações que podemos identificar como elementos na página web. Então basta clicar aqui no botão azul e queremos dizer ao Web Scraper onde essas caixas brancas estão, certo? Queremos dizer ao Web Scraper: "Web Scraper, encontre 200 caixas nesta página", e depois disso vamos dizer ao Web Scraper onde está a informação em todas estas caixas, para que ele possa raspá-la.

Mas, por enquanto, vamos apenas chamar isso de "caixa" e este é um elemento na página web, certo? Então, o tipo aqui é "Element". Clicamos aqui em "Select", e essas serão múltiplas caixas. Então selecionamos esta caixa aqui chamada "multiple". E observe que quando você começar a passear com o mouse, você vai ver que o Web Scraper interage com a página mostrando todos esses elementos e como eles são clicáveis e selecionáveis. Então, o que queremos fazer é encontrar um lugar aqui no canto inferior direito que destaca em verde a caixa cheia no topo. Vamos clicar aqui e ela vai ficar vermelha assim que nós clicarmos. E então nós vamos fazer o mesmo para a próxima e veja que o Web Scraper tenta adivinhar onde estão todas as outras caixas. Mas aí ele pára na 21ª. Então nós vamos fazer isso de novo para a 21ª, e veja que ele identificou todas as 200 caixas aqui na página. Portanto, é aqui que queremos chegar, certo?

Fazemos isso, e agora perceba que aqui, há um seletor que identifica, que descreve todas estas caixas. Você não precisa se preocupar com isso. O Web Scraper seleciona automaticamente, identifica o seletor para você. Basta clicar aqui em "Done selecting!" ("Seleção concluída!") e note que isso vai vir para cá. E uma vez que você faz isso, este é o sinal certo para que você possa até pré-visualizar esses dados. Não há dados aqui que estamos raspando, mas ele também pode fazer um "Element preview" ("Pré-visualização de elemento") e você vê que todas as caixas estão selecionadas. Então nós salvamos isso, e agora temos a caixa. Agora temos o processo para capturar a caixa, mas precisamos do processo para capturar informações em todas as 200 caixas.

Então, se você passar o mouse aqui, você vê que ele destaca em cinza a linha da caixa. E se você clicar na caixa, perceba que estamos agora dentro destas caixas genéricas que selecionamos aqui. Podemos voltar a "_root" e vamos voltar lá em um segundo, mas então clicaríamos na caixa e agora estamos nesta caixa que acabamos de descrever e queremos dizer ao Web Scraper: "olha, você sabe a localização das 200 caixas, mas agora eu quero que você capture informações em cada uma das caixas e faça o conjunto de dados para mim."

Então, nós vamos capturar o nome da música, o nome do artista, a posição, e a URL da imagem do álbum, e vamos fazer isso. Você clica em "Add new selector" ("Adicionar novo seletor"), lembre-se: aqui dentro da caixa. Então nós adicionamos um novo seletor, vamos chamar este seletor de "canção". Nós clicamos aqui em "Select", e então selecionamos o nome da canção, ele identifica qual é o seletor, e em seguida clique em "Done selecting!" ("Seleção concluída!"). Podemos fazer uma pré-visualização dos dados. Ele captura todos os nomes das canções aqui e parece que está tudo bem, e então nós salvamos o seletor.

Agora vamos adicionar um novo. Clique no botão azul. Este será o artista. Este também é um tipo de texto porque é texto aqui na página. Clicamos aqui, este é o artista. É um seletor "a". Clicamos em "Done selecting!" e então podemos fazer uma pré-visualização dos dados. Aqui ele mostra todos os nomes dos artistas que ele capturou nesta página. Salvamos o seletor.

Agora vamos adicionar a posição. Este também é um elemento de texto. Clicamos aqui em "Select" e aqui está a posição e então em "Done selecting!". Você pode estar se perguntando: por que não selecionar a caixa "multiple"? Já que a selecionamos para as múltiplas caixas que estávamos capturando. Como só estamos capturando uma posição aqui, não há múltiplas posições dentro desta caixa amarela, não vamos selecionar as caixas múltiplas. Somente quando há ocasiões em que você tem que selecionar vários elementos que se repetem em posições ou em diferentes casos, você seleciona "multiple". Mas neste caso é apenas uma posição, um nome da canção, um artista, então você não seleciona a caixa "multiple". Tudo bem. Então, salvamos o seletor também.

Agora vamos adicionar a imagem. Isso é uma imagem, então vamos chamar isso de "imagem". Selecionamos aqui a imagem do álbum e então "Done selecting!", e então nós salvamos. Agora, se nós voltamos para "_root" e vamos a "Data preview" aqui, você vai ver que ele já tem quase tudo de que precisamos, certo? Temos a música, o artista, a posição e a URL da canção. Mas não é exatamente tudo o que queremos, porque também queremos a data, certo? E a data está aqui no topo.

Então, o que nós queremos fazer é aplicar esta data para cada linha aqui. E para fazer isso, vamos voltar a "_root", onde estamos. Estávamos em uma caixa, agora vamos voltar para "_root" e adicionar um novo seletor para a data. E isso também é um seletor de texto, e então nós vamos destacar aqui tudo isso, e depois clicamos em "Done selecting!". Mas perceba quando você visualiza isso, ele pega toda esta informação aqui que eu não quero, eu só quero este "August 31" e 2019. Então como você pode extrair apenas isso? Eu não vou entrar em detalhes porque eu vou usar expressões regulares para fazer isso. Sinta-se livre para googlar expressões regulares. Há ótimos tutoriais, ótimas aulas sobre expressões regulares. Elas são muito poderosas, especialmente em programação. Mas, felizmente, o Web Scraper também tem um "Regex" ou um campo de expressão regular aqui onde você pode aplicar expressões regulares. Então o que nós vamos fazer é: eu vou pegar tudo isso, e eu vou encontrar um padrão para extrair apenas esta parte aqui, este "August 31, 2019" e eu vou fazer isso de modo que ele extraia cada vez que haja um texto aqui que mostre uma palavra e um número, vírgula, um ano. Ele vai extrair apenas isso.

Então, eu vou aqui neste testador online de expressão regular, e veja que eu já tenho aqui a minha cadeia de caracteres. Então é a semana de 31 de agosto de 2019 e depois a semana passada, a próxima semana, a semana atual, pesquisa por data. O que eu estou fazendo aqui é usar expressão regular e eu tenho três elementos. O primeiro é um "\w+" e o que significa é que esta barra invertida é como um padrão quando você quer usar em uma espécie de símbolo, que é este "w" aqui. Assim, "w" significa qualquer caractere de palavra. Qualquer a, b, c, d, ou qualquer número. Então ele une qualquer coisa que acontece aqui e o "+" significa qualquer caractere para apenas um caractere, ou um número infinito de

caracteres até chegar ao espaço. Mas não é só um espaço, é um espaço que vem antes, como um "\d", que significa dígitos, certo? Então ele corresponde a um dígito igual a 0-9, e o "+" significa correspondências entre 1 e tempo ilimitado, portanto, qualquer número de dígitos. Então, é uma palavra com qualquer número de caracteres, um espaço que vem antes de qualquer número de dígitos, e depois há uma vírgula, e depois um espaço, e depois quatro dígitos. É o que isso significa.

E se você copiar isso aqui para o campo de expressão regular em um Web Scraper e fizer uma pré-visualização de dados, você vê que ele extrai tudo e depois deixa apenas "August 31, 2019". E isso é exatamente o que queremos. Eu vou salvar este seletor aqui e pronto, terminamos.

Então nós vamos aqui. Aqui você pode selecionar outras opções, você pode ver o gráfico seletor onde você vê o "_root" e todos os outros seletores e a relação entre eles. Isso pode se tornar bastante complicado, dependendo da complexidade da página. Você pode editar os metadados, o nome do o site ou a URL. Você também pode raspar. Você pode navegar pelos dados raspados e você pode exportar este mapa do site. Portanto, este é como um JSON, uma cadeia de caracteres, que você pode exportar e usar em outros computadores, ou enviar a um amigo que vai carregar o mesmo processo de raspagem em outros computadores. Ou você pode ajustá-lo um pouco para mudar o site. Então há uma maneira de exportar. E você pode importar também. E você pode exportar os dados que você raspou como arquivo CSV.

Vamos em frente e raspar aqui nesta opção. Então ele te dá duas opções. O "request interval", que é a quantidade de tempo que você vai esperar enquanto ele faz uma solicitação ao site. E dois MS, que é dois segundos, é uma boa prática. Você não quer martelar o site com solicitações, porque isso pode parecer suspeito. O webmaster pode pensar que você está tentando derrubar o site, e nós não queremos isso. Queremos usar essa opção de forma responsável. E depois há o delay do carregamento da página. O delay do carregamento da página é a quantidade de tempo que o Web Scraper espera a página carregar para, em seguida, raspar os dados. Então você pode querer dar ao site algum tempo para carregar os dados para depois raspar, de modo que você possa garantir que todos os elementos carreguem antes de começar a capturar informações. Dois e dois segundos são bons números para começar, mas você pode ajustar isso de acordo com seu caso.

Então você clica aqui em "start scraping" ("começar a raspagem"), ele abre uma janela. Ele aguarda dois segundos para carregar a página e espera dois segundos para fazer a solicitação, e então raspa os dados. E se você clicar aqui em "atualizar", ele vai carregar todos os dados que acabou de raspar, e voilá, temos aqui os metadados que o Web Scraper adicionou. Esta é a identificação do Web Scraper para cada um dos registros, você tem a URL de partida. Aqui temos os dados que realmente queremos raspar: a canção, o artista, a posição, a URL da imagem, e também a data que foi aplicada a cada registro aqui. Então agora podemos ir em frente e clicar aqui e exportar este site em CSV. E quando você clicar aqui você vai baixar o arquivo CSV em seu computador, e então você pode importar o

arquivo CSV para qualquer outro aplicativo de planilhas para começar a analisar ou limpar, ou editar, ou construir seu conjunto de dados.

Então é isto para o Web Scraper. Então vá à Chrome Web Store, baixe a extensão e comece a raspar.