

Conversas com dados

O texto começa com a definição do que é "raspagem" de dados: "o processo que permite extrair conteúdo de uma página da Web usando uma ferramenta especializada ou com a escrita de um código”.

As limitações da extração de dados são:

- sites com HTML mal formatado com muito pouca informação estrutural.
- sistemas de autenticação que impedem o acesso automatizado.
- mudanças nos marcadores de uma página da web.

Faça uma re-verificação do seu código para não perder nenhum dado

Esta seção foi escrita por Peter Aldhous, repórter de ciência do BuzzFeed News.

Ele usou o Python Requests and Beautiful Soup para extrair de um site registros de ações disciplinares contra médicos no estado de Nova York. Com esses dados, fez uma reportagem sobre como médicos viciados estavam conduzindo os testes de drogas para os pacientes.

Com as mesmas ferramentas, ele extraiu os nomes dos autores, a disciplina científica, o número de citações e outros metadados dos trabalhos publicados por 10 anos na Revista da Academia Nacional de Ciências. Atualmente, ele usa o pacote rvest do tidyverse (uma coleção de bibliotecas R). Alguns exemplos das reportagens que ele fez com o rvest R:

- "[Por que as estrelas do atletismo não estabelecem recordes mundiais como costumavam fazer \(mas os nadadores, sim\)](#)", revelada coletando dados das 100 melhores performances ao ar livre em muitos eventos de atletismo no site da Associação Internacional das Federações de Atletismo.
- Aldhous extraiu do site “The American Presidency Project” os textos completos de todos os discursos do “estado da União” (State of the Union Address) e outros discursos presidenciais dados ao Congresso para [comparar as palavras usadas por Trump](#), com as dos outros presidentes, desde que George Washington fez o primeiro discurso desse tipo na história dos Estados Unidos em 1790.

Uma advertência do autor: é essencial garantir que todos os dados estejam sendo extraídos. Variações sutis ou falhas na maneira como os sites são codificados podem causar problemas e podem fazer com que algumas lacunas tenham que ser preenchidas manualmente. Ele sugere a utilização do inspetor do navegador da web e estudo cuidadoso do código do site para saber como escrever o raspador de dados. Para o Chrome, você pode usar a extensão Selectgadget.

Extração de dados de boa qualidade leva tempo: comunique-se eficazmente e busque as APIs existentes. Esta seção é escrita pelo jornalista Mikołaj Mierzejewski, do maior jornal da Polônia -Gazeta Wyborcza- que explica que os dados a serem extraídos podem estar em três situações :

- 1.- Os dados estão em HTML simples e podem ser facilmente extraídos com ferramentas como o Portia.
- 2.- A segunda situação é mais difícil, porque precisa de cookies ou manter uma sessão do usuário; ou os dados podem ser carregados, mas são necessárias ferramentas especiais para baixá-los.
- 3.- No terceiro nível de dificuldade, os dados são carregados apenas à medida que você interage com o site, então você precisa de um programador que desenvolva um pequeno “raspador” que atue como um navegador para baixar os dados.

Um dos grandes desafios é poder comunicar as necessidades aos membros não técnicos da equipe, entre outras coisas, porque eles precisam entender que uma boa extração de dados leva tempo.

Mierzejewski oferece exemplos do que eles fizeram extraindo dados do Instagram.

Por fim, ele sugere sempre pesquisar a API primeiro, antes de extrair dados, bem como recomenda o Postman e Insomnia como ferramentas para explorá-las.

9 coisas para se lembrar sobre raspagem de dados

Esta seção é de Paul Bradshaw, que realiza o curso de mestrado em Jornalismo de Dados na Birmingham City University e é autor de "Scraping for Journalists".

- 1.- Nem sempre precisa de código. Seu colega da BBC usou as planilhas do Google para uma história sobre reclamações de ruído.
- 2.- Pense na política dos Termos e Condições para o uso da informação. Ele diz que queria extrair dados de um site de propriedades e os termos de serviço os proibiam. No fim, eles convenceram os proprietários do site. O departamento jurídico deve ser consultado quando os termos e condições do site proíbem a extração de informações.
- 3.- Use a extração de dados como uma segunda opção. Você pode fazer solicitações via Lei de Acesso à Informação para não ter que raspar um determinado site governamental, por exemplo.
- 4.- Lembre-se de que, se houver uma estrutura ou um padrão, você provavelmente poderá extrair os dados. Em uma história sobre violações, eles extraíram os principais dados dos relatórios policiais porque as informações estavam no mesmo formato.
- 5.- Verifique se os dados não estão disponíveis de antemão, antes de extraí-los. Às vezes, o dado desejado já está disponível no formato JSON para download ou você pode encontrá-lo em outro site.
- 6.- Faça um controle aleatório dos seus dados extraídos e compare-os com os dados no site original.
- 7.- Use filtros e tabelas dinâmicas para encontrar resultados incomuns. Quando a extração não funciona bem, ela não funciona bem sistematicamente. Esses desvios podem ajudá-lo a verificar.
- 8.- Extraia a maior quantidade de dados primeiro e depois filtre e limpe. São dois processos diferentes.
- 9.- Extraia as informações mais de uma vez e identifique as informações adicionadas ou excluídas.

Use raspadores como uma ferramenta de monitoramento

Seção por Maggie Lee, uma repórter freelancer de Atlanta.

A extração de dados também serve para monitorar assuntos com cobertura recorrente. A extração de dados de forma contínua e sistemática pode automatizar o processo para identificar alterações relevantes ou economizar a necessidade de ir constantemente ao site para verificar se há novas informações publicadas, pois é possível receber um email, por exemplo, quando um novo documento for incluído. Assim, o repórter economiza tempo.

Certifique-se de que o raspador seja resistente e tenha sempre um plano B

Esta seção foi escrita por Gianna-Carina Grün, chefe do DW Data Journalism.

Se os dados são extraídos regularmente com um “raspador”, eles devem ser projetados de modo que pequenas alterações na escrita de uma página ou dos dados, por exemplo, não quebrem o código. O extrator deve ser o mais resistente possível.

Você pode prever alternativas para solucionar possíveis problemas futuros, por exemplo, pensando em quais outras fontes de informação podem ter os dados.

Não se esqueça de fazer backup dos dados na raspagem

Seção de Erika Panuccio, assistente de comunicação da ALTIS.

É importante fazer backup dos dados extraídos, pois eles podem ser excluídos. Ele dá o exemplo da análise das contas do Twitter de pessoas a favor do ISIS, com as quais reuniu um banco de dados de 30.000 tweets de 100 usuários. A raspagem foi feita com uma plataforma automatizada (IFTTT) que guardava os tweets em uma planilha, assim que eles eram publicados. Isso foi fundamental, pois havia contas excluídas pela política do Twitter sobre publicidade terrorista ou tweets excluídos pelos usuários das contas.