# O Uso do Web Scraping

**Vitor T. Oliveira[1], Ayslan T. Possebom[1]**

[1]Instituto Federal do Paraná – Campus Paranavaí (IFPR)

– Paranavaí – PR – Brazil

vitortavares.o55@gmail.com, ayslan.possebom@ifpr.edu.br

## 1. Introdução

Quando se refere à *web scraping*, o primeiro pensamento que deve vir à mente é a de procura, captação e visualização de dados. Porém, de acordo Mouromtsev (2016), duas coisas costumam ser esquecidas na aplicação dessa garimpagem de informações: método de busca dos dados e a utilização dos dados. O método de busca poder ser de diferentes maneiras, já a usabilidade desses dados é a definição da finalidade de recolher centenas de dados. Devido a falta de difusão de como essa área do *Data Science* funciona e influencia nossas vidas, é necessário uma maior divulgação e explicação de tais funcionalidades.

## 2. Bases do *web scraping*

A identificação de padrões e tendências atuais é uma das áreas mais importantes na criação ou implementação de qualquer projeto. Conhecer o padrão mais comum de atitudes de usuários e a próxima tendência, permite a implementação do projeto de forma que o projeto se adeque a essas tendências e ao mesmo tempo se mantenha flexível a futuras mudanças.

### 2.1. Padrões

Utilizando do *web scraping*, é possível fazer um recolhimento de dados de diversas fontes de informação na Internet. Como explicado por Means (2012), ao transferir esses dados para tabelas e gráficos é possível identificar diversos padrões que antes estavam escondidos, como atrativos de atenção, cores que ajudam no destaque, interfaces mais intuitivas baseado nos padrões mais fortes, funcionalidades que podem ser incluídas no sistema, até mesmo formas de estruturação do sistema como um todo. O conhecimento desses padrões permite que ao invés de ficar se questionando ou realizando testes, você possa chegar a uma certeza de senso comum de funcionamento, aumentando muito as chances de sucesso.

### 2.2. Tendências

Tendência se trata de uma previsão de futuras mudanças e alteração de padrões baseada em múltiplos dados. Apesar de parecer simples de identificar, a tendência se trata de uma análise mais aprofundada nos dados que foram recolhidos e nos padrões que esses dados apresentam, realizando o web scraping diversas vezes e em diferentes períodos de tempo. Por exemplo, em uma base de dados que você recolhe dados de um site em 3

trimestres seguidos, você pode identificar qual dos padrões teve um aumento maior e qual teve uma redução expressiva. Isso pode apontar qual é a tendência de mudança.

## 3. Utilização do *web scraping*

Como identificado pelo tópico anterior, o web scraping pode fornecer milhares de dados e caso saiba trabalhar com eles, podem ser de grande ajuda na formulação de um projeto. Existem formas simples de recolher dados das pessoas, como a utilização de formulários, questionários e até mesmo recolhimento direto, buscando os dados em bancos abertos. Porém, essa maneira está sujeita a falhas e a um gasto excessivo de tempo. A forma mais rápida e precisa de recolher dados atualmente é por meio da utilização de scripts e bots, que vão recolher os dados de sites definidos em uma quantidade muito superior que qualquer pessoa pode fazer, assim entregando massas de informações que podem ser formatadas para uma visualização mais ampla, como demonstrado na Figura 1.

| conversation_id | created_at | favorite_count | full_text | hashtags/0 | hashtags/: |
|---|---|---|---|---|---|
| 1453014525805015044 | 2021-10-26T15:03:53.000Z | 17 | garotos do ifpr tem um histórico de inventar coisa sobre as mulheres de lá né inacreditável | | |
| 1452992036974518285 | 2021-10-26T13:34:31.000Z | 0 | IFPR vc me dá vontade de amarrar uma corda no pescoço | | |
| 1452614187302146061 | 2021-10-25T12:33:05.000Z | 45 | Ifpr aglomera alunos para falar sobre não aglomerar | | |
| 1452041980607877142 | 2021-10-23T22:39:20.000Z | 1 | Não a divisão do IFPR!!! https://t.co/iCrx6sqoLk | | |
| 1453028049587429378 | 2021-10-26T15:57:37.000Z | 43 | eu te amo ifpr https://t.co/PX6giZOGj1 | | |
| 1451322871255552004 | 2021-10-21T23:01:51.000Z | 15 | "vamos dividir o IFPR em 2?"<br><br>Comunidade: NÃO<br><br>"Tá bom, a gente divide em 3" | | |
| 1452451597808717826 | 2021-10-25T02:15:44.000Z | 43 | @monicabergamo Mesmo com a rejeição da comunidade do IFPR, o projeto de divisão continua a todo vapor, estamos em campanha contra, não aumenta vagas p estudantes, só cargos de chefia e sem | | |

**Figura 1. Exemplo de tabulação de dados obtidos por um script de *web scraping* aplicado em postagens no twitter envolvendo o termo IFPR.**

## 4. Uso do *web scraping* nos dias atuais

Apesar do *web scraping* estar ganhando popularidade atualmente, ele vem sendo utilizado desde o início da década de 90, obviamente em menor escala. Aos poucos, o *web scraping* começou a se tornar uma ferramenta útil para quase todas as áreas do mercado global, partindo desde a criação de softwares, até engenharias, como civil e biomédica, e a escrita. Segundo Feng (2012), essa ocorrência se dá pelo fato de que o acesso a tão diversos dados de diversas pessoas diferentes ajuda em todas as áreas, sendo quase como um recolhimento de opiniões sem as pessoas saberem suas próprias opiniões.

A utilização mais característica do *web scraping* é para pesquisas. De acordo com Mouromtsev (2016), essa disponibilidade gigantesca de dados ajuda em criações, validações e mitificações, principalmente em áreas que necessitam de diversas pesquisas sociais, na criação de artigos, na disponibilização de dados, formas de ensinamento eficientes e outros pontos. A utilização desse método é questão de tempo para se tornar algo comum, chegando a nível onde professores estão podendo ter acesso a padrões e comportamentos dos alunos em individual, assim aprendendo a lidar com o ensino de cada estudante.

Cada vez mais podemos ver a utilização do *web scraping* integrada a outras tecnologias, como *machine learning* e *deep learning*, que vem incrementando cada vez mais funções e atrativos no mundo. Um dos principais casos de implementação são os algoritmos de redes sociais, entre eles o Instagram e o Tik Tok. Como explicado por Orgaz (2020), estas empresas ficaram conhecidas por possuírem algoritmos de captação e recomendação extremamente avançados, identificando o padrão de cada indivíduo e somando isso ao conjunto, aumentando a divulgação desses conteúdos específicos.

Apesar de ser uma forma de melhoria tecnológica a nível global, essa capacidade de compreender padrões e tendências é extremamente perigosa quando utilizado incorretamente. O caso mais famoso foi o da empresa Cambridge Analytica (CONFESSORE, 2018), que foi contratada pelo organizador da campanha do ex-presidente Donald Trump, realizando coleta de dados em massa nas redes sociais, como Facebook e Instagram, se utilizando desses dados para criar o discurso e a entrega de diferentes materiais às pessoas. Basicamente, a estratégia era a adaptação da campanha do candidato às vontades mais comuns das pessoas, manipulando suas redes e utilizando seus dados contra eles mesmos.

Devido aos fatos citados nos tópicos anteriores, a utilização do web scraping já foi questionada diversas vezes. Apesar de todas críticas recebidas, *o web scraping* continua crescendo exponencialmente em todas as áreas possíveis, pois um método que apresenta soluções e atitudes mais certeiras é sempre algo importante para criação e expansão de qualquer ideia.

## Referências

MOUROMTSEV, Dimitry and D'AQUIN, Mathieu (2016), Open Data for Education Linked, Shared, and Reusable Data for Teaching and Learning, Springer.

BIENKOWSKI, Marie and FENG, Mingyu and MEANS, Barbara (2012), Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief, Center for Technology in Learning, U.S Department of Education.

ORGAZ, Cristina (2020), 'TikTok foi feito para ser viciante': o homem que investigou as entranhas do aplicativo. Disponível em: https://www.bbc.com/portuguese/geral-55173900. Acesso em: 30/08/2021.

CONFESSORE, Nicholas (2018), 'Cambridge Analytica and Facebook: The Scandal and the Fallout So Far'. Disponível em: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html Acesso em: 30/08/2021.