Coleta de Dados através de *Web Scraping*: uma aplicação com as taxas de incidência de Doenças Ocupacionais dos Estados Unidos

Thaline Ferreira Silva[1] **e Gecynalda Soares da Silva Gomes** [2]

[1] Departamento de Estatística, Universidade Federal da Bahia, Brasil

[2]Departamento de Estatística, Universidade Federal da Bahia, Brasil

Resumo

O web scraping é uma técnica que envolve a coleta automatizada para extração de informação de páginas web. Suas ferramentas possibilitam poupar tempo na etapa de coleta dos dados. É capaz de transformar informação não estruturada em informação estruturada que pode depois ser armazenada e analisada. O objetivo do artigo foi utilizar a ferramenta de extração de dados em ambiente web para auxiliar no estudo espacial da taxa de incidência de lesões ou doenças ocupacionais. Inicialmente, realizou-se uma fundamentação teórica sobre as ferramentas de extração de informações via web. Assim, foi utilizada uma biblioteca do software R para extração de dados do site da secretária de estatísticas trabalhistas dos Estados Unidos, buscando a taxa de incidência de doenças relacionadas ao trabalho por estado, transformando-as em um banco de dados estruturado. A metodologia apresentada poderá auxiliar as pessoas, de um modo geral, na extração de informações estratégicas, principalmente na esfera pública que estão disponibilizadas na web, com baixo custo, otimizando ações e garantindo uma melhoria no uso de recursos.

Keywords: *Ciência de Dados; Web Scraping; Doenças Ocupacionais.*

1. Introdução

WEBSITES são criados na maioria das vezes para auxiliar na visualização de informações e não para exposição de dados de forma estruturada. Mas extrair essas informações de forma manual pode consumir muito tempo e ser suscetível a erros de extração. Nesse contexto, a extração das informações de forma automatizada possui um papel fundamental para a análise dos dados.

2. A proposta

WEB scraping é o conjunto de técnicas usadas para obter automaticamente algumas informações de um site, em vez de copiá-las manualmente. O objetivo de um *web scraper* é procurar certos tipos de informações, extrair e agregá-las em novas páginas da *web*. Em particular, os *scraper* estão focados em transformar dados não estruturados e salvá-los em bancos de dados estruturados.

O processo de extração da *web* é bastante simples, embora a implementação possa ser complexa. Ela ocorre em 3 etapas:

1. O código usado para extrair as informações envia uma solicitação HTTP GET para um site específico.
2. Quando o site responde, o *scraper* analisa o documento HTML para um padrão específico de dados.
3. Depois que os dados são extraídos, eles são convertidos em qualquer formato específico criado pelo autor do *scraper*.

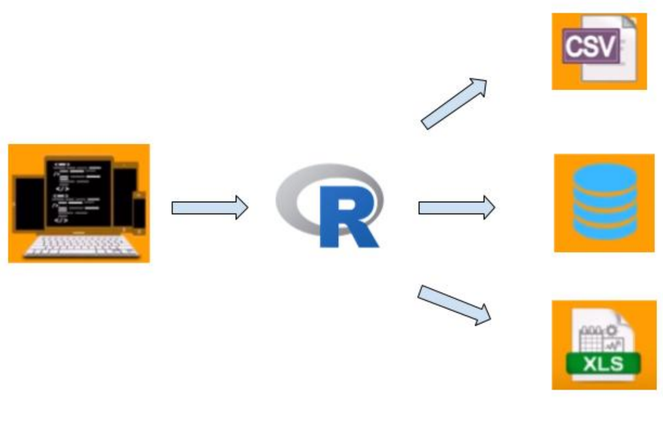

Figura 1: *O processo.*

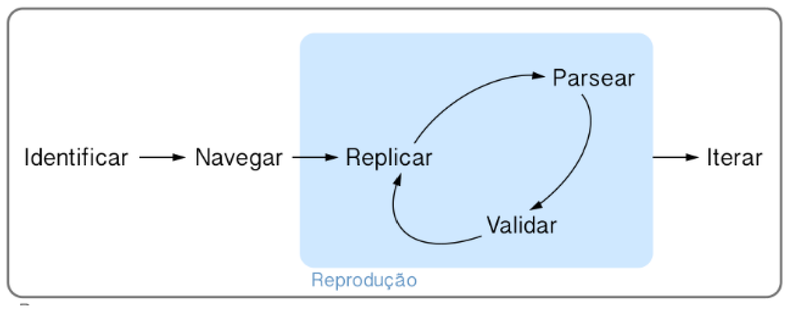

Figura 2: *Fluxo do* Web scraping.

3. Resultados

Table 6. Incidence rates[1] of nonfatal occupational injuries and illnesses, by industry and case type, 1996

Alabama

Industry [2]	SIC code [3]	1996 Annual average employment [4] (000's)	Injuries and illnesses				Injuries			
			Total cases	Lost workday cases		Cases without lost workdays	Total cases	Lost workday cases		Cases without lost workdays
				Total [5]	With days away from work [6]			Total [5]	With days away from work [6]	
Private industry [7]		1,455.1	8.9	4.0	2.5	4.9	8.4	3.7	2.4	4.7
Agriculture, forestry, and fishing [7]		18.2	8.5	4.4	3.7	4.1	7.9	4.1	3.4	3.8
Agricultural production [7]	01-02	6.0	15.0	6.5	5.1	8.5	13.5	5.4	4.1	8.1
Mining [8]		10.7	9.2	6.5	6.4	2.6	9.1	6.4	6.2	2.6
Construction		93.3	10.3	4.6	3.6	5.7	10.2	4.6	3.6	5.6
General building contractors	15	25.1	9.5	3.5	2.6	6.0	9.5	3.5	2.6	6.0
Residential building construction	152	8.6	6.9	2.7	2.4	4.2	6.9	2.7	2.4	4.2
Nonresidential building construction	154	16.3	10.7	3.9	2.8	6.8	10.7	3.9	2.8	6.8
Heavy construction, except building	16	14.5	10.1	5.5	4.4	4.6	10.1	5.5	4.4	4.6
Highway and street construction	161	5.0	8.2	4.3	1.9	4.0	8.2	4.3	1.9	4.0
Heavy construction, except highway	162	9.5	11.1	6.1	5.8	5.0	11.1	6.1	5.8	5.0
Special trade contractors	17	53.7	10.7	4.8	3.8	5.9	10.5	4.8	3.8	5.7
Plumbing, heating, air-conditioning	171	14.0	12.0	5.3	3.6	6.7	11.5	5.2	3.4	6.3
Painting and paper hanging	172	3.2	10.7	4.3	2.9	6.4	10.4	4.3	2.9	6.1
Electrical work	173	9.5	13.5	4.5	4.3	9.0	13.5	4.5	4.3	9.0
Masonry, stonework, and plastering	174	7.3	7.5	4.7	4.7	2.7	7.4	4.7	4.7	2.7
Roofing, siding, and sheet metal work	176	3.3	16.3	7.0	5.6	9.3	16.1	7.0	5.6	9.1
Miscellaneous special trade contractors	179	10.3	9.2	4.4	2.9	4.8	9.0	4.4	2.9	4.6

Figura 3: *Exemplo de uma tabela do estado do Alabama, 1996.*

Figura 4: *Inspecionando o código-fonte.*

”https://www.bls.gov/iif/oshwc/osh/os/pr156al.pdf”Ano 2015 Estado Alabama
”https://www.bls.gov/iif/oshwc/osh/os/pr146tx.pdf”Ano 2014 Estado Texas
”https://www.bls.gov/iif/oshwc/osh/os/pr036ny.pdf”Ano 2003 Estado New York

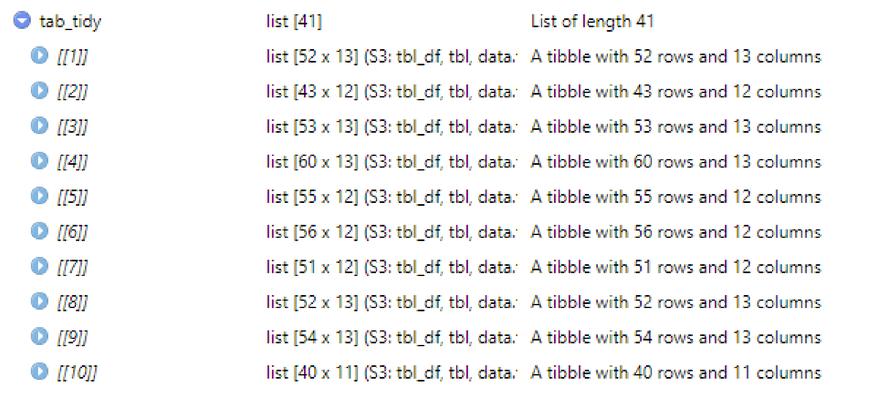

Figura 5: *Descrição das estruturas de cada tibble, 1996.*

Conclusões

FAZER um *scraper* não é uma tarefa fácil, entretanto, se toda vez que o pesquisador seguir um método consistente e robusto, pode-se melhorar um pouco o trabalho. Em um contexto no qual grande parte das informações publicadas pelas organizações estão na *web*, faz-se necessário o desenvolvimento de técnicas computacionais para a organização destas bases e a exploração do conhecimento nelas contido.

Referências

DEVIKA, K.; SURENDRAN, S (2013). *An overview of web data extraction techniques*. International Journal of Scientific Engineering and Technology. **2**, 278,–287.