# TRABALHOS RELACIONADOS

Visando o entendimento e análise da importância da implementação de técnicas de coletas de dados, foram ponderados artigos cujo o conceito principal é análogo ao do presente trabalho.

O estudo [1] de autoria de Carolina Franco Zanon, apresenta a criação de um sistema, utilizando a técnica Web Scraping, para auxiliar na definição do perfil epidemiológico dos pacientes da Clínica de Estomatologia da FO-UFRJ. Os dados foram coletados, de um sistema já existente denominado Estomato Web, um software incapaz de fornecer estatísticas ou exportar grande quantidade de dados de uma única vez. A técnica foi implementada fazendo a coleta dos conjuntos de dados e transferindo-os para planilhas do Excel. Foi possível agrupar os dados por categorias e fazer uma análise detalhada do perfil dos pacientes. Concluiu-se que o uso do novo sistema, chamado Estomato Web Scraper, facilitou a extração e qualidade dos dados, e, assim sendo, poderá ser utilizado para diversos estudos na área de estomatologia.

Um segundo artigo [2] que tem Dhaniel Nunes Mazini e Renato Cesar Sato como autores e aborda uma análise da rentabilidade dos dividendos das empresas que compõem o IBrX 50. Foi implementado um Web Scraper a partir da linguagem Python. As informações foram extraídas do site Yahoo!Finanças [[4](#REFERENCIASBIBLIOGRÁFICAS_4)], onde os dados não estão disponíveis para uma coleta em massa. As análises feitas para a área do mercado financeiro têm como característica uma grande massa de dados, fazendo-se necessária extração em diferentes fontes. Utilizando a tecnologia proposta, foi possível a automatização da coleta em uma escala de tempo reduzida. Os elementos obtidos foram armazenados em um arquivo que favorece a manipulação, possibilitando a realização de uma análise de regressão múltipla para chegar no objetivo proposto pelo estudo.

Por fim o trabalho [3] de autoria de Thiago da Cunha Borges e Zeus Olenchuk Ganimi discorre sobre a implementação de um Web Scraper para capturar dados de valores de automóveis, extraídos do site da FIPE (Fundação Instituto de Pesquisas Econômicas). A ferramenta permite a coleta automatizada dos dados semiestruturados do módulo de consulta da página da FIPE e persistência das informações estruturadas em banco de dados PostgreSQL. A análise do conteúdo estruturado pode ser utilizada, por exemplo, para otimização de recursos de empresas adquirentes de veículos, e sua visualização foi possibilitada através do software de Tableau Desktop.

## DISCUSSÃO

A partir dos trabalhos supracitados é possível analisar que existem diversas aplicações da tecnologia de Web Scraping para a obtenção de dados. Constatou-se que a implementação da técnica é capaz de auxiliar a automação, assiduidade e estudo dos dados obtidos, além de dar suporte para distintas áreas de estudo.

Uma das áreas na qual foi identificada a oportunidade de automatização é a farmacêutica. Parte do procedimento de proposta de um novo fármaco envolve a consulta aos ingredientes mais utilizados pela comunidade para se compor a receita, tendo como parâmetro de busca o princípio ativo do medicamento. Não há um veículo de comunicação oficial que permita aos pesquisadores acompanharem essas tendências de utilização de ingredientes eficientemente, de forma que as informações são extraídas de sites especializados da área.

A aplicação de Web Scraping nesse contexto de extração das informações de sites farmacêuticos especializados através da busca por elementos-chave que compõe a página web proverá agilidade em uma etapa importante da criação de novos medicamentos, principalmente nessa época na qual a nossa sociedade demanda por soluções rápidas e eficientes que combatam os agentes infecciosos recém descobertos.

# FUNDAMENTAÇÃO TEÓRICA

Com objetivo de disponibilizar embasamento teórico para uma acertada análise e interpretação do presente estudo, neste capítulo serão abordados os principais conceitos e definições encontrados na literatura sobre tipos de dados, técnicas de coleta e de análise e armazenamento.

## COLETA DE DADOS

Ao iniciar uma pesquisa é necessário um estudo intensivo do conjunto de dados que gere informações relevantes para a análise a ser feita. A obtenção desses dados possibilita um posterior exame pormenorizado de diferentes fontes para o estudo da área de interesse.[18]

De acordo com a finalidade da pesquisa e tipo de dado interessante a mesma, a coleta poderá ser realizada através de questionários, entrevistas, observações, experimentos ou estudo e revisões de documentos já disponíveis, além de poder ser necessária a utilização de mais de um método para explorar com profundidade as informações desejadas. No mesmo sentido, cabe realçar que independentemente do tipo de coleta utilizado, a prioridade desta será a garantia da precisão, coerência e confiabilidade nas informações obtidas.

## TIPOS DE DADOS

Para compreendermos melhor o desenvolvimento do trabalho, que será discorrido no capítulo a seguir, é importante definirmos com clareza os tipos de dados que existem. Estes são divididos em dois grupos, quanto a sua característica: qualitativos e quantitativos, e em dois outros grupos, quanto a sua origem: primários e secundários.

### QUALITATIVOS

Os dados qualitativos possuem a função de rotular e atribuir uma identificação ao objeto analisado. Como exemplos de dados qualitativos, temos: nível de escolaridade, estado civil ou nome. [7]

### QUANTITATIVOS

Os dados quantitativos são numéricos e, como o nome sugere, possuem o objetivo de quantificar as amostras, além de serem amplamente utilizados em modelos estatísticos. São considerados discretos, quando assumem valores dentro de um intervalo enumerável (por exemplo, minutos inteiros em uma hora), e contínuos, quando assumem valores reais que não podem ser enumerados. [7]

### PRIMÁRIOS

Os dados denominados primários são aqueles coletados diretamente de sua fonte. A obtenção de dados primários é planejada especialmente para o estudo de interesse e, portanto, são mais adequados aos seus objetivos. Em contrapartida, há a desvantagem de o esforço de coleta ser alto. São exemplos de dados primários as respostas de entrevistas e questionários dadas diretamente pelo entrevistado. [7]

### SECUNDÁRIOS

Dados secundários são obtidos a partir de coletas previamente realizadas. Por já terem sido publicados, comumente os dados secundários não possuem o formato ou a integridade mais adequada ao estudo de interesse. Desta forma, apesar do menor esforço de obtenção, em comparação aos dados primários, os dados secundários requerem maiores esforços na definição de fontes e na metodologia empregada em seu manejo. [7]

## ORGANIZAÇÃO E ARMAZENAMENTO

Posterior a coleta dos dados, os mesmos têm de ser armazenados de acordo com sua organização que pode ser estruturada, não estruturada ou semiestruturada. É priorizada formas acessíveis e a escolha leva em consideração a análise e nível de controle necessário dos dados.

### ESTRUTURADOS

São dados que possuem um formato específico e bem definido. Nesse tipo de organização, cada registro possui características obrigatórias que devem ser respeitadas (figura 1). [7]



Figura 1 – Dados não estruturados, semiestruturados e estruturados

### NÃO ESTRUTURADOS

São dados que não possuem estrutura fixa, como arquivos de texto comuns, cujo conteúdo pode conter inúmeras variações em formato, idioma, formatação, entre outros (figura 1). [7]

### SEMIESTRUTURADOS

Os dados semiestruturados apresentam um avanço com relação aos dados não estruturados, pois são dotados de regras que orientam a organização do seu conteúdo, mesmo ainda possuindo considerável liberdade com relação à composição como um todo (figura 1). [7]

## FUNCIONAMENTO DAS PÁGINAS WEB

A técnica de coleta de dados proposta no presente trabalho terá como base fundamental a navegação através da estrutura de sites para a captura dos dados desejados. Portanto, é de suma importância a compreensão dos conceitos envolvendo as páginas web e os navegadores que conhecemos.

### HISTÓRIA DA INTERNET

As motivações que culminaram na estruturação da rede que hoje conhecemos como internet datam-se nas primeiras décadas da Guerra Fria. No ano de 1962, os militares americanos, com o apoio do governo, concluíram o desenvolvimento da rede que ficou conhecida como ARPANET. Concebida como uma resposta ao Projeto Sputnik da URSS, lançado alguns anos antes, a rede era limitada foi utilizada principalmente para a comunicação entre universidades de ponta e institutos de pesquisa da época [5].

Até o final da década de 80, a Net foi majoritariamente um meio de comunicação entre militares e acadêmicos, porém já não mais para fins necessariamente relacionados à Defesa dos EUA [5]. Em 1979, concretizou-se a visão comercial da rede através do primeiro vendedor de serviços on-line americano. Dez anos depois, o pesquisador Tim Berners-Lee, que posteriormente ficou conhecido como o pai da internet, imaginou o que chamou de World Wide Web (ou apenas web). Os alicerces da proposta de Berners-Lee são os mesmos da rede de ampla utilização que conhecemos hoje. Segundo KRAEMER [5], ao final do século XX, estabeleceu-se a tríade informação, educação e entretenimento na web, porém, novas possibilidades de uso seguem sendo exploradas até os dias de hoje.

### NAVEGADORES E PÁGINAS WEB

É de suma importância a compreensão da forma básica das páginas estruturadoras dos websites que compõe a web, além da responsabilidade dos navegadores na interpretação do conteúdo das mesmas. Todo site, independente da natureza de seu conteúdo, é composto de uma ou mais páginas que são relacionadas entre si através de *links* (referências) [11]. Estejam aplicadas as devidas práticas de desenvolvimento web, frequentemente a navegação entre as páginas é fluida o suficiente para que os usuários percebam um único contexto em todas as páginas que compõe o mesmo site [11].

Toda as páginas web são arquivos compostos por dados semiestruturados, mais precisamente através de uma linguagem de marcação denominada HyperText Markup Language – HTML [8]. A estrutura do HTML é reconhecível através das marcações realizadas através das *tags* que, por sua vez, podem agrupar outras *tags* diversas. Os dois principais agrupamentos representam o cabeçalho e o corpo do documento (figura 2).



Figura 2 – Estrutura básica de um documento HTML

Dadas as características das páginas que compõe a internet, entra em evidência outro elemento também essencial ao funcionamento da web: os navegadores. Os navegadores são interpretadores dos diversos elementos que podem compor as páginas, como imagens, folhas de estilo CSS [13], scripts em Javascript [14] e a estrutura HTML previamente citada. Dentre os navegadores mais conhecidos estão o Google Chrome [15], Mozilla Firefox [16] e Internet Explorer [17].

Na figura 3, podemos visualizar o resultado da página estruturada na figura 2, quando interpretada por um navegador. Vale observar que os comentários adicionados no documento foram ignorados pelo navegador, conforme citado anteriormente.



Figura 3 – Visualização de um documento HTML em um navegador

Apesar de os navegadores serem a forma mais conhecida e utilizada pelos usuários para acessar páginas web, outros programas de computador também podem requisitar e interpretar as mesmas páginas, como é o caso de algumas bibliotecas e *frameworks* de linguagens de programação. A partir destas implementações, é possível extrair o código HTML de uma página web através de poucas linhas de programação.

Dadas as características de páginas web previamente citadas, além das possibilidades relacionadas ao seu consumo, seja através de navegadores ou outros meios, existem diversas técnicas empregadas na análise, interpretação, busca e recuperação de dados contidos em páginas web.

## web scraping

# IMPLEMENTAÇÃO

Conforme definições realizadas em reunião que contou com a presença de um farmacêutico pesquisador na área de interesse do presente trabalho (detalhada no Anexo A), foram levantados os requisitos necessários para a implementação da técnica de *web scraping*, visando a automatização da busca e coleta de dados, em fontes específicas, relacionados aos princípios ativos de determinados medicamentos e seus demais constituintes.

O assunto deste capítulo serão os detalhes técnicos da implementação aplicada, incluindo: diagrama e descrição dos casos de uso observados, informações sobre a tecnologia optada, um diagrama de classes provendo uma visão geral da arquitetura implementada e a definição das fontes de dados a serem exploradas, conforme orientação do especialista farmacêutico consultado.

## CASOS DE USO

REFERÊNCIAS BIBLIOGRÁFICAS

1. ZANON, Carolina Franco. **Aplicação de um programa de extração de dados através da técnica de web scraping para determinação do perfil epidemiológico dos pacientes da clínica de estomatologia da faculdade de odontologia da ufrj cadastrados no software estomato web**, 2017. Dissertação (Mestrado em Odontologia) – Curso de Pós-graduação em Odontologia, UFRJ, Rio de Janeiro, RJ.
2. MAZINI, Dhaniel Nunes; SATO, Renato Cesar. **Extração de dados financeiros com um web scraper: um estudo sobre a rentabilidade dos dividendos**, 2019. Dissertaação (Artigo na WAIAF) - Workshop of Artificial Intelligence Applied to Finance, ITA, São José dos Campos, SP.
3. BORGES, Thiago da Cunha; GANIMI, Zeus Olenchuk. **Extração de dados com web scraping para análise da variação de preço de veículos automotores**, 2018. Dissertação (Tecnólogo em Sistemas de Computação) - Curso de Tecnologia em Sistemas de Computação, UFF, Niterói, RJ.
4. **Yahoo! Finanças** <<https://br.financas.yahoo.com/>> Acesso em 09 de Maio de 2020
5. ABREU, Karen Kraemer. **História e usos da Internet**, 2009. Biblioteca On-line de Ciências da Comunicação, p. 2-4
6. KABIR MUHAMMAD SAJJAD, Syed. **Basic Guidelines for Research: An Introductory Approach for All Disciplines**, Edition: First, Chapter: 9, Publisher: Book Zone Publication, Chittagong-4203, Bangladesh, p. 201-275.
7. ZOZUS NAHM, Meredith. **The data book collection and management of research data**-Chapman and Hall CRC Press, 2017, p. 36.
8. BERNERS-LEE, Tim. **WWW: past, present, and future**, 1996. Publicação acadêmica – Computer, vol. 29, no. 10, pp. 69-77.
9. LONGMAN WESLEY, Addison. **A History of HTML**, 1998. <https://www.w3.org/People/Raggett/book4/ch02.html> Acesso em 24 mai. 2020.
10. BERNERS-LEE, Tim. **Information Management: A Proposal**, 1989. <https://www.w3.org/History/1989/proposal.html> Acesso em 24 mai. 2020.
11. LAWSON, Bruce; SHARP, Remy. **Introducing HTML5, 2nd Edition**, 2011. Livro. New Riders Press, Berkeley, Canadá.
12. **What is the difference between webpage, website, web server and search engine?** <<https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Pages_sites_servers_and_search_engines>> Acesso em 24 mai. 2020.
13. **CSS: Cascading Style Sheets** <<https://developer.mozilla.org/en-US/docs/Web/CSS> > Acesso em 24 mai. 2020.
14. **Javascript** <<https://developer.mozilla.org/en-US/docs/Web/JavaScript>> Acesso em 24 mai. 2020.
15. **Google Chrome** <https://www.google.com/chrome> Acesso em 24 mai. 2020.
16. **Mozilla Firefox** <https://www.mozilla.org/en-US/firefox/new> Acesso em 24 mai. 2020.
17. **Microsoft Internet Explorer** <https://www.microsoft.com/pt-br/download/internet-explorer.aspx> Acesso em 24 mai. 2020.
18. OLSEN, Wendy; **Coleta de Dados**: Debates e Métodos Fundamentais em Pesquisa Social. 1. ed. Porto Alegre: Penso, 2015. p. 15.