CONGRESSO NACIONAL de PESQUISA e ENSINO em CIÊNCIAS V CONAPESC
ISSN 2525-3999
www.conapesc.com.br

Web Scraping e Análise de dados

Ana Beatriz Gomes Barbosa [1]
Alexsandro Bezerra Cavalcanti [2]

INTRODUÇÃO

A técnica W*eb Scraping* pode ser definida como "raspagem" de dados diretamente da *web*, onde extraímos informações relevantes de sites através de *bots*, submetendo os dados à análise posterior. Essa técnica é importante pois a análise auxilia a encontrar padrões e a tomar decisões com maior probabilidade de acerto.

Sabemos que, em muitos casos é possível fazer essa extração manualmente, mas quando trabalhamos com um grande volume de dados espalhados em vários *sites* ou em diversas abas de um mesmo *site*, a extração manual acaba sendo inviável. Nesses casos, precisamos automatizar a coleta. Para isso, fazemos uso de *bots*, os quais conseguem extrair imagens, arquivos, tabelas e rastrear todo o site. Todavia, é necessário que insiramos no *script* expressões regulares para que o *bot* consiga identificar no código fonte o padrão, conseguindo então extrair do *site* os dados relevantes.

Após a extração, é possível armazenarmos os dados em tabelas ou *dataframes* para facilitar nossa análise. Feito isso, podemos escolher como prosseguir com o estudo: o próprio Python dispõe de módulos para análise descritiva, mas nesse trabalho usamos o *software* R.

METODOLOGIA

O presente trabalho teve início a partir de estudos realizados através do projeto de Iniciação Científica vinculado ao PET- Matemática e Estatística da Universidade Federal de Campina Grande. Primeiramente, fizemos um estudo bibliográfico do assunto. Utilizamos o *software* Python para fazer as extrações dos dados referentes às rodadas da temporada 2018 da *Premier League.*

A coleta dos dados foi feita através do *Web Scraping,* utlizando os seguintes módulos Python: *Beautiful Soup* e *Selenium. A*dicionamos os dados coletados em arquivos *.csv*. Posteriormente fizemos uso do *software* R para fazer uma análise descritiva dos dados

[1] Graduando do Curso de Estatística da Universidade Federal de Campina Grande -UFCG, e bolsista do PET-Matemática e Estatística anabeatbarbosa.machado@gmail.com;
[2] Professor orientador: Doutor, Universidade Federal de Campina Grande - UFCG, alexdme@gmail.com.

ISSN 2525-3999
www.conapesc.com.br

extraídos. Então, fizemos uma comparação entre os resultados dos times da casa e os visitantes, para assim podermos analisar se o time que está jogando em casa tem alguma vantagem sobre o visitante para conseguir a vitória.

REFERENCIAL TEÓRICO

Apesar dessa técnica ser pouco conhecida no Brasil e no meio acadêmico brasileiro, diversos autores ressaltam seu potencial e também suas aplicações no nosso dia a dia. Entre eles, podemos ressaltar Calò (2014) e Mitchell (2019).

Mitchell (2019) destaca em seu livro como o *Web Scraping* funciona e também quais são suas potencialidades. Com isso, diz que

> [...] os *web scrapers* podem acessar lugares que as ferramentas de pesquisa tradicionais não conseguem. Um pesquisa no *Google* por "voos mais baratos para Boston" resultará em uma grande quantidade de anúncios publicitários e *sites* populares para busca de voos. O *Google* sabe apenas o que esses sites dizem em suas páginas de conteúdo, mas não os resultados exatos de várias consultas fornecidas a uma aplicação de busca de voos. No entanto, um web scraper bem desenvolvido pode colocar em um gráfico o custo de um voo para Boston ao longo do tempo para uma variedade de sites e informar qual é o melhor momento para comprar uma passagem. (Mitchell, 2019, p. 12-13)

Calò (2014) em sua monografia exorta sobre a importância de ferramentas que consigam extrair o grande volume de dados presentes nas redes para obter uma análise mais precisa, segundo ele

> A evolução da computação, nos últimos anos, incentivou a criação de grandes volumes de dados em todas as áreas de conhecimento. Observou-se um grande esforço no sentido de digitalizar, organizar e disponibilizar o histórico de organizações, pesquisas, e processos produtivos de inúmeros campos. Atualmente, grande parte destes dados encontra-se disponível para consulta, porém cada elemento tem pouca relevância quando analisado individualmente. Desde então, foram concebidos novos métodos de análise cuja finalidade principal é gerar novos significados a partir de grandes quantidades de dados brutos. (Calò, 2014, p. 2)

ISSN 2525-3999
www.conapesc.com.br

RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados da análise descritiva com base nos dados extraídos. Conseguimos obter uma tabela contendo os dados de todas as 38 rodadas da temporada de 2018 da *Premier League,* um total de 380 jogos.

Os dados não estavam juntos em uma mesma página no site, o *bot* navegou por algumas páginas para conseguir obter todas as informações, e finalmente organizou todo o material em um único arquivo *.csv*.

Após terminada a extração, nos dirigimos ao *software* R para fazermos as análises descritivas. Nosso intuito com a análise era encontrar um padrão para podermos generalizar para situações futuras. Queríamos descobrir se os times que jogam em casa recebem algum tipo de vantagem sob os visitantes, utilizando os dados da temporada de 2018 da *Premier League.*

Identificamos que o número médio de gols por partida foi de 3.08, o número médio de gols dos times quando jogam em casa foi de 1.56 e o número médio de gols dos times quando jogam fora de casa foi de 1.25. Podemos notar que o número de gols está bem baixo, no caso dos times que jogam fora de casa está ligeiramente inferior ao que jogam em seu próprio terreno, porém nada de muito gritante.

Obtemos também o percentual de vitórias dos times quando jogam em casa, que foi de 0.48 e o percentual de jogos com mais de 2.5 gols, que foi de 0.54. Considerando a hipótese que os times que jogam em seu território teriam mais chances de vitória, notamos que o percentual obtido através dos resultados da temporada de 2018 foi bem abaixo do esperado.

CONSIDERAÇÕES FINAIS

Os resultados obtidos com a análise descritiva nos levam a crer que os times que jogam em casa não recebem vantagem sobre os visitantes. Porém, não podemos fazer uma generalização desse resultado para as demais temporadas ou para outros campeonatos de mesma categoria.

Para tomarmos conclusões mais precisas, é necessário que façamos uma análise mais profunda utilizando os dados de outras temporadas, para verificarmos se o resultado se repete. Também é pertinente fazermos uma comparação com os resultados de outros campeonatos.

ISSN 2525-3999
www.conapesc.com.br

Palavras-chave: *Web Scraping*, Python, *software* R, Análise descritiva, *Premier League.*

REFERÊNCIAS

CALÒ, Alessandro. Extração e análise de informações jurídicas públicas. **USP**, 2014. Disponível em: <https://bcc.ime.usp.br/tccs/2014/sandro/Monografia.pdf/>. Acesso em 20 de jun. de 2020.

MITCHELL, Ryan. **Web Scraping com Python.** 2ed. São Paulo: Novatec, 2019.