14° Seminário de Iniciação Científica e Tecnológica

Extração de dados via web scraping como suporte na análise do crescimento de geração fotovoltaica

Renan Moreira Soares (IC)*, Guilherme Rezende Pereira (IC), Leonardo Garcia Marquess (PQ), Marcelo Escobar de Oliveira (PQ)

PIBIC
Campus Itumbiara
** renanmoreirasoares@gmail.com*

***Palavras Chave**: Estatística; Extração de dados; Fatores socioeconômicos; Geração Distribuída; Web scraping.*

Introdução

A quantidade de dados disponíveis na internet vem aumentando com o passar dos anos. O processo de coletar estes dados pode ser feito de forma manual, no entanto, despende tempo, esforço e está susceptível a erros. Com isto, a prática do *web scraping* se mostra como uma opção para coletar estes dados de forma automatizada. Dados coletados utilizando esta ferramenta possuem maior confiabilidade. Diversas análises no setor elétrico vêm sendo realizadas com dados disponíveis publicamente, como a realizada em [1], em que o autor verifica se há uma correlação entre indicadores socioeconômicos e geração distribuída no estado de Goiás. Com base nisso, visando dar suporte para futuras análises, este trabalho apresenta a construção, em linguagem *Python*, de ferramentas *web scraper*s para coletar dados de geração distribuída e indicadores socioeconômicos. Além disso, como forma de apresentar a robustez das ferramentas, uma análise similar é realizada, no entanto, considerando todo o território brasileiro.

Metodologia

Existem diferentes técnicas utilizadas para realizar o *web scraping*, comparadas em diversos estudos. Em [2] e outros estudos, algumas técnicas foram comparadas, levando a conclusão de que a escolha da técnica depende do formato do site.

Para os dados socioeconômicos, coletados do site do Instituto Brasileiro de Geografia e Estatística (IBGE), teve que ser levado em consideração que os dados estão dispostos em diferentes páginas, construídas em HTML, uma linguagem de marcação de texto. De tal forma, a ferramenta *web scraper* deveria percorrer diversas páginas, localizando os dados, dispostos sempre nas mesmas posições, extraí-los e salvá-los em uma planilha. Devido a característica de percorrer diversas páginas, a ferramenta ainda deve ser capaz de evitar bloqueios, cumprindo diretrizes éticas e legais relacionadas ao *web scraping,* descritas em [3] e [4].

Já os dados de geração distribuída foram coletados do site da Agência Nacional de Energia Elétrica (ANEEL), dispostos em uma única página, porém, com conteúdo dinâmico. A localização dos dados no site não é fixa, e apenas 20 linhas de dados são apresentados por vez, exigindo que a página seja rolada para apresentar mais. Para isso, técnicas diferentes do *scraper* do IBGE foram necessárias.

Com os dados coletados, para realizar a correlação, procedimentos semelhantes aos encontrados em [1] foram realizados. No entanto, neste caso, a análise foi realizada para todos os estados brasileiros, com dados de mais de 5 mil munícipios, coletando os dados de potência instalada fotovoltaica, população, PIB per capita e Índice de Desenvolvimento Humano (IDH) para cada um. Além disso, apenas a correlação de Pearson foi utilizada.

Resultados e Discussão

O *web scraper* IBGE permite coletar dados de quase todas as pesquisas disponíveis no site, ampliando as análises possíveis. No entanto, devido as características do site, a sua extração despende maior tempo, podendo levar horas. Já o *web scraper* ANEEL coleta os dados de geração distribuída levando segundos ou minutos, a depender da quantidade de dados disponíveis. Ambas as ferramentas retornam planilhas.

Com base na análise realizada, foi possível verificar que a geração fotovoltaica, na maioria dos estados, possui correlação maior com o IDH, demonstrando que investir na geração fotovoltaica pode possuir maior impacto positivo neste indicador. No entanto, conforme ressaltado em [1], isto deve ser verificado por outros métodos estatísticos.

Conclusões

As ferramentas construídas permitiram que a coleta de dados para pesquisas sobre crescimento e outras análises sobre geração distribuída fosse aprimorada. Diminuindo esforços, aumentando confiabilidade e trazendo a possibilidade de que novos estudos sejam realizados, auxiliando na tomada de decisões.

Agradecimentos

Este projeto foi financiado pelo CNPq.

[1] BORGES, Lucas da Mata Santana et al. Análise de fatores socioeconômicos em relação ao crescimento da Geração Distribuída no estado de Goiás. 2020.

[2] SIRISURIYA, De S. et al. A comparative study on web scraping. Proceedings of 8th International Research Conference of KDU. General Sir John Kotelawala Defence University, p. 135–140, 2015.

[3] KROTOV, Vlad; SILVA, Leiser. Legality and ethics of web scraping. Twenty-fourth Americas Conference on Information Systems. New Orleans, 2018.

[4] MITCHELL, Ryan. Web scraping com Python: Coletando mais dados na web moderna. 2ª Edição. Novatec Editora, 2019.