“Web scraping” e Programas de Estudo

Artur Hosoi, Débora van Putten, Pedro Thomazelli, Pedro Zanineli
Ilum School of Science, CNPEM, Campinas - Brasil

RESUMO: Diante da proposta da disciplina de Prática em Ciência de Dados de se elaborar um programa, a partir da escrita do código, na linguagem de programação Python, de acordo com a finalidade escolhida pelo grupo, o presente artigo apresenta a escolha de se realizar um “web scraping” em busca de dados acerca de oportunidades, (inter)nacionais, de crescimento acadêmico. Essas oportunidades envolvem de estágios a intercâmbios e as informações foram retiradas de diferentes domínios da internet. Após a coleta e o tratamento dos dados, eles foram dispostos em um site de livre acesso.
Palavras-chave: Atividades extracurriculares e extraclasse; Web scraping; Python; Dados;

ABSTRACT: The purpose of the current article is to explain and demonstrate the creation process of the project originally proposed during our Data Science course and further improved and adapted by our group. The project aims to utilize the "web-scraping" technique, coded in Python (programming language), in order to search for and extract international internship; scholarship, and job opportunity related data for further processing, organization and disponibilization in a free-access website developed by us.
Keywords: *Web-scraping; Data extraction; Python, Extracurricular activities; Professional opportunities*

Sumário

1. Introdução

Atividades de caráter extracurricular e extraclasse são relevantes para a vivência acadêmica de qualquer indivíduo[1, 2]. Durante uma graduação, a depender da instituição a qual está vinculado, um estudante se depara com diversas oportunidades: empresas juniores, grupos de extensão, estágios, intercâmbios, movimentos estudantis e populares.

Cada uma dessas experiências, seja ela nacional ou internacional, contribui para a sua formação pessoal, profissional e estudantil, na medida em que coloca a pessoa em contato com novas culturas; novas dinâmicas de trabalho; permite a construção de relações interpessoais inéditas e permite a aplicação de conhecimentos obtidos na prática, tornando uma aprendizagem ainda mais significativa. No que tange à formação de futuros cientistas, essas vivências assumem caráter extremamente positivo, pois, carregam o potencial de transformar, de ampliar o ponto de vista do indivíduo, possibilitando que ele consiga olhar para problemas antigos de maneiras novas e que consiga propor novos métodos, teorias e ações de enfrentamento. Em outras palavras, o terreno das atividades de caráter extracurricular e extraclasse é profícuo para a inovação.

Segundo Filho e Jacinto (2021)[2], a participação da instituição de ensino é fundamental para a construção de experiências acadêmicas livres que contribuam para a vivência e a qualificação acadêmica. Com isso em mente, e pensando também que os discentes em si são elementos essenciais da composição de um espaço de ensino e devem se comportar como tal, consideramos ser adequado tomar uma iniciativa no sentido de criar um espaço de possibilidades no contexto da Ilum. Por meio dessa pequena e inicial ação, esperamos colocar e apresentar tais atividades como possibilidades no contexto da faculdade. Ou, ao menos, desejamos suscitar certo querer nos demais graduandos da Escola de Ciência.

Para tanto, optou-se pela realização de um "web scraping", seguido pelo desenvolvimento de um site. Tal plataforma reúne dados de suma relevância no momento da escolha de uma tarefa extraclasse, como o título do anúncio da vaga, a data de divulgação e o link para o outro domínio, do qual retiramos a informação, que contém a postagem completa.

O foco do trabalho são oportunidades de estágios, programas de verão, intercâmbio, bolsas de estudo, 'work and study' de cunho nacional e internacional. Visando alcançar discentes da Ilum e demais interessados.

2. Metodologia

De modo a recolher informações acerca de atividades extracurriculares e extra sala de aula, optou-se pela técnica de "web scraping". Posteriormente, desenvolveu-se um site que abriga todos os dados coletados.

De maneira breve, o "web scraping"[3] consiste em extrair dados e informações

provenientes de sites da “web”, a fim de armazená-las e convertê-las em um sistema estruturado para posterior análise e disponibilização. Em português, podemos chamá-lo “mineração” ou “raspagem” de dados. Alguns exemplos da aplicação dessa coleta de dados no cotidiano são: sites como 123 Milhas, Google Voos e demais endereços de “web” que agreguem dados.

Para a elaboração do projeto fazendo uso dessa técnica, estruturou-se um sistema na linguagem de programação Python, sendo o código escrito na plataforma Jupyter Notebook.

O objetivo era de que o programa, ao receber um link de algum “web” site, conseguisse extrair os dados solicitados e definidos a partir de alguns critérios delimitados pelo grupo. No conjunto das informações recolhidas deveria constar: área, instituição, orçamento, país e continente, prazo, descrição e link.

Para desenvolvimento do código, foi necessária a utilização de três bibliotecas . A primeira delas foi a “*feedparser*”, cuja função é coletar informações de sites alimentados por *RSS* - do inglês, *Rich Site Summary* ou *Really Simple Syndication* - que exibe, através de HTML, um grande número de informações de maneira reduzida. Além disso, há o *PyGitHub*, responsável por fazer o “*commit”* para o GitHub, após feito o “*web scrap*” do site raspado. Por último, a biblioteca “*googletrans*”, responsável pela tradução das informações para o português, já que nem todos os sites minerados eram nacionais.

Usando inicialmente o “*feedparser”*, criou-se uma lista vazia para armazenar links que satisfizessem aquilo que era proposto, ou seja, que fossem ao formato RSS. A partir disso, os dados foram coletados e, usando a própria biblioteca, nomeados, de modo a organizar as informações. Posteriormente, percorreu-se a lista e criou-se um *“feed”* para cada link, exibindo somente suas saídas, a princípio.

Feito isso, o código indicava que deveria ser gerado um arquivo de texto (“.csv”) com o nome de “data”, onde são armazenadas as informações coletadas dos sites. O formato do arquivo utilizado foi o CSV (Valores separados por vírgulas, ou *Comma-separated Values*, em inglês), porque ele permitia o armazenamento e organização de dados de maneira semelhante a uma tabela.

A partir desse arquivo, as informações foram transportadas para o GitHub pela biblioteca “*PyGitHub*”. Para tanto, o primeiro passo consistiu na sincronização da conta ao Jupyter, o que exigia uma chave de acesso. Com isso, tornou-se possível acessar o repositório desejado, que foi o criado exclusivamente para a realização do projeto na conta de um dos integrantes da equipe. No repositório foram armazenadas as informações até então coletadas.

Em outra instância, para a criação do site pertencente o grupo, foi usado o “GitHub Pages”, feito a partir do Jekyll - um gerador de sites estáticos, ou seja, sites que exibem as mesmas coisas para todos os usuários. A criação da nossa página “web” também exigia a criação do repositório. Com esses dois elementos, o site já estava praticamente pronto para armazenar as informações coletadas e também para ser editado. O “index.md” foi responsável pela criação da página inicial do site. Por outro lado, no arquivo “programas.md” criou-se outra página, na qual foram armazenados os programas encontrados. Para esse armazenamento, as instruções também foram dadas no próprio código.

Finalmente, obteve-se o site do projeto finalizado com todas as informações de atividades extracurriculares coletadas.

3. Resultados E Discussão

Ao rodar o código pela primeira vez e fazendo isso a partir da mineração de três sites, obtivemos 487 resultados. Desses, a maioria tratava de oportunidades de emprego e vagas de estágio, variando entre oportunidades (inter)nacionais.

Todavia, muitos dos resultados se relacionavam com atividades extracurriculares e extraclasse de maneira destoante ao objetivo do trabalho, o que é consequência direta dos endereços de ‘web’ escolhidos para se fazer a raspagem. Um exemplo claro disso é: havia sido selecionado um blog, que contava sobre as experiência de intercâmbio, estágio e emprego da autora. Todavia, embora se encaixassem no assunto geral, os resultados não contemplavam o projeto de maneira satisfatória porque as postagens deste ‘blog’ não tratavam do anúncio de vagas e oportunidades, mas sim de relatos pessoais.

A percepção desse erro motivou o grupo a revisar todos os dados encontrados pelo programa e dispostos na nossa página. Outrossim, também nos incentivou a buscar por novos sites, ainda no formato RSS, para serem minerados, de modo a conseguir reunir informações que fossem mais condizentes com o propósito do trabalho. Por último, também reforçou a relevância do tratamento de dados.

Ao refazer a coleta, partindo então de dois novos endereços da “internet” dessa vez, obtivemos 20 resultados. Desses, a quase totalidade se relacionava com bolsas de estudo internacionais. Embora a redução tenha sido brusca se compararmos a primeira coleta e a segunda, refazer essa parte do trabalho foi a melhor decisão, visto que nos forneceu dados mais válidos e em adequação com a proposta.

É interessante comentar aqui também algumas das dificuldades e limitações deste trabalho. Contudo, antes disso, é pertinente pontuar que se trata de um projeto de primeiro período da graduação e, por isso, já esperávamos enfrentar alguns obstáculos, afinal, programar é um exercício desafiador [4].

Mais do que isso, já esperávamos que nem todos os resultados obtidos estivessem absolutamente consoante às metas iniciais propostas durante a idealização da atividade. Lidar com a possibilidade de não conseguir suprir todas as expectativas também faz parte da aprendizagem.

Em primeiro lugar, as categorias definidas inicialmente (área, instituição, orçamento, país e continente, prazo, descrição e link) não foram todas obtidas. Por sua vez, a apresentação final das oportunidades contou somente com o título dos anúncios, a data de publicação deles e o link que dava acesso ao site raspado.

Implementar a busca por todas essas questões intensificaria a complexidade do código. Não teria sido impossível fazê-lo, mas optamos por priorizar a qualidade das informações que estavam sendo disponibilizadas. Por conseguinte, focamos nossas atividades em buscar bons sites para mineração, assegurando que informações essenciais (título do anúncio, sua atualidade e o link para acessá-lo) estivessem disponíveis.

4. Conclusão

O "web scraping" é uma técnica indubitavelmente útil para diversas finalidades. Com a enorme quantidade de informações disponibilizadas na "internet" atualmente, utilizar um recurso que facilite a navegação por ela é fundamental.

Dessa maneira, desenvolver um projeto que envolvesse não somente uma metodologia ainda desconhecida pelos membros do grupo, mas também um assunto cativante foi uma experiência proveitosa. Ademais, tal trabalho foi muito fiel à própria proposta de aprendizagem e "ensinagem" da Ilum, o que colabora inclusive para a consolidação da proposta da instituição de ensino.

5. Referências

[1] SIMÃO, P. A., BUSNELLO, M. B. **Estágio Curricular e Extracurricular na Formação Profissional: Relato de Experiência.** Disponível em: <https://www.publicacoeseventos.unijui.edu.br/index.php/salaoconhecimento/article/view/7653/6390>. Acesso em: 17 mai. 2022.

[2] FILHO, Adelmo dos Santos, JACINTO, P. O impacto das atividades extracurriculares no desenvolvimento estudantil. **Rev. Educação para a Diferença**, v.2, n.3, 2021. Disponível em: <https://www.revistas.uneb.br/index.php/abatira/article/view/10226. Acesso em: 17 mai. 2022.

[3] ZHAO, Bo. **Web Scraping**. 2017. Disponível em: <https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf> . Acesso em: 21 e jun. 2022.

[4] SILVA, Walquiria dos Santos *et al*. **Levantamento sobre as dificuldades dos discentes nas disciplinas de programação no curso técnico de Informática.** Disponível em: <https://www.diversitasjournal.com.br/diversitas_journal/article/view/616/659>. Acesso em: 22 jun. 2022.