DIREITO & TI
ISSN 2447-1097
WWW.DIREITOETI.COM.BR

WebScraping e acesso à informação: superando barreiras na extração de dados judiciais

WebScraping and access to information: overcoming obstacles in judicial data extraction

Mariela Campos Rocha[1]
Mariana Elis Campos Gomes[2]
Marcella Queiroz de Castro[3]

Recebido em: 17.08.2024
Aprovado em: 16.12.2024

RESUMO

Este artigo busca demonstrar como técnicas de tecnologia podem auxiliar cientistas sociais a realizarem estudos empíricos e quantitativos, superando obstáculos como a falta de acesso a dados estruturados pelos Tribunais de Justiça, ainda que requeridas essas informações pela Lei de Acesso à Informação. O estudo almeja contribuir no incentivo à pesquisa, especialmente por meio da disponibilização de scripts automatizados que utilizam da técnica de WebScraping, promovendo a possibilidade de realização de pesquisas acadêmicas com quantitativo significativo de dados judiciais. São compartilhados, enquanto anexos a este artigo, os códigos e tutoriais de utilização visando auxiliar outros pesquisadores a replicar o processo, com exemplos práticos dos tribunais de Minas Gerais, Rondônia e Mato Grosso. Embora o trabalho explore a extração e organização de dados públicos disponíveis nos sites das Cortes, destaca-se pela inovação na forma de obtenção e agrupamento dessas informações, viabilizando uma variedade de análises qualitativas e quantitativas. Este estudo defende a necessidade de maior transparência e acessibilidade dos dados judiciais e incentiva a colaboração entre academia e instituições públicas para aprimorar o acesso a dados e, assim, contribuir para a formulação e avaliação de políticas públicas no sistema de Justiça.

[1] Doutora em Ciência Política pela Universidade Federal de Minas Gerais (UFMG), com formação sanduíche na University of Texas em Austin. Mestrado em Ciência Política pela UFMG. Especialização em informática em Educação pela Universidade Federal de Lavras. Pesquisadora de Pós-doutorado no Instituto da Democracia e da Democratização da Comunicação (INCT/UFMG). Pesquisadora do Centro de Estudos sobre Comportamento Político (CECOMP do DCP/UFMG). E-mail: mariela.rocha@gmail.com. ID Lattes: 9177350771497696

[2] Mestranda em Ciência Política na Universidade Federal de Minas Gerais (UFMG); Graduada em Direito pela Pontifícia Universidade Católica de Minas Gerais (PUC Minas); Pesquisadora no Observatório da Justiça Brasileira (OJB-UFMG). E-mail: mariana.ecg@hotmail.com. ID Lattes: 8502972378621189

[3] Graduada em Direito pela Universidade de Brasília (UnB) e em Sistemas para Internet pelo Instituto Federal de Brasília (IFB). Analista de Tecnologia no Banco do Brasil. Consultora em Bruno Bioni Consultoria - Direito Digital. Pesquisadora bolsista pelo CNPq no Observatório da Justiça Brasileira (OJB-UFMG). E-mail: mqcastro@gmail.com. ID Lattes: 7277540757618435

Palavras-chave: Acesso à informação, Acesso a dados públicos, Ciência de dados, Tribunais de Justiça, WebScraping.

ABSTRACT

This article aims to demonstrate how technology can assist social scientists in conducting empirical and quantitative studies, overcoming obstacles such as the lack of access to structured data from Courts of Justice, even when such information is requested under the Brazilian Freedom of Information Act. The study seeks to encourage research, particularly through the provision of automated scripts that use WebScraping techniques, promoting the possibility of academic research with a significant amount of judicial data. Attached to this article are the codes and usage tutorials to assist other researchers in replicating the process, with practical examples from the courts of Minas Gerais, Rondônia, and Mato Grosso. Although the work explores the extraction and organization of publicly available data from court websites, its innovative method of obtaining and organizing this information is what enables a variety of qualitative and quantitative analyses. This study advocates for greater transparency and accessibility of judicial data and encourages collaboration between academia and public institutions to improve data access and thus contribute to the formulation and evaluation of public policies in the justice system.

Keywords: Access to Information, Public Data Access, Data Science, Courts of Justice, WebScraping.

1 INTRODUÇÃO

O presente estudo advém de frustrantes respostas à Lei de Acesso à Informação (LAI) por parte de Tribunais de Justiça (TJs) estaduais, algo recorrente para pesquisadores do Poder Judiciário (Cruz; Zuccolotto, 2021). Realizando pesquisas em bases de jurisprudência, as quais não permitem o download de todo o conteúdo pesquisado, apenas acórdão por acórdão, buscou-se os canais de Serviço de Informação ao Cidadão (SIC) para a disponibilização do banco de dados, já público, em um período específico, com fins acadêmicos. Não era o objetivo acessar processos em segredo de justiça, mas sim julgados já disponíveis de forma mais simples, sem necessidade de acesso pela interface dos sites, otimizando o tempo dos pesquisadores.

Em respostas, recebeu-se negativas atrás de negativas, muitas até sob a justificativa de ameaça a segurança do tribunal, embora os dados fossem públicos. À medida que os e-mails chegavam, nos perguntávamos como seguiríamos a pesquisa

adiante, sem precisarmos copiar e colar documento por documento; como podia o Poder Judiciário ignorar um instituto normativo tão importante para o país quanto a LAI; o que de errado havia como o pedido de acesso à dados públicos.

Em vários trabalhos pesquisadores tratam como dados as sentenças e decisões proferidas. Estas são objeto de estudo fundamental para entender o comportamento e o entendimento dos tribunais em determinado assunto (Castro Júnior; Calixto, 2022). Dados estatísticos, como quantidade de processos julgados, também são analisadas pelos estudiosos do Poder Judiciário, mas esses têm o acesso mais facilitado, uma vez que os TJs os fornecem ao Conselho Nacional de Justiça (CNJ), em conformidade com a Resolução nº 215/2016. Estes, anualmente, são publicados no Justiça em Números, relatório produzido pelo CNJ (Brasil, 2016; Oliveira; Cunha, 2020).

Cada portal eletrônico de cada tribunal mantém suas decisões armazenadas em páginas de buscas de jurisprudências, não havendo uma centralização dos dados, o que dificulta o trabalho do pesquisador. Além disso, por não possuírem *application programing interface* (API), o acesso é individual (Oliveira; Cunha, 2020), ou seja, não há como compilar todos os acórdãos pesquisados em um único arquivo para análise, igualmente retardando a pesquisa a ser desenvolvida.

Com isso, o acesso aos dados públicos dos tribunais é penoso, sendo uma solução, para tanto, a utilização de técnicas de *WebScraping*. A partir destas, o pesquisador consegue coletar dados, estruturados ou não, em sites de forma otimizada, ou seja, reduzindo seu tempo de trabalho, uma vez que a máquina consegue minerar os dados de maneira mais rápida e eficiente (Rodrigues et al, 2021). Dessa forma, estudos quantitativos e qualitativos envolvendo muitos casos são facilitados, pois as decisões judiciais passam a ser coletadas de forma automática, reduzindo o viés de seleção de casos, bem como os erros durante a coleta dos dados.

Pesquisas voltadas para a análise da transparência e eficiência do Poder Judiciário, por exemplo, são promovidos a partir dessas técnicas, como em metodologias que usam a jurimetria. Assim, o acesso integral aos dados dos tribunais, os quais já são públicos em razão do dever de transparência judicial, por meio de *WebScraping*, possibilita a

realização de estudos empíricos no Direito, área ainda em ascensão no país (Maia; Bezerra, 2023).

Diante disso, busca-se responder à seguinte questão: como realizar pesquisas com análise de dados de Tribunais de Justiça diante das dificuldades impostas pelas negativas de acesso à informação? A partir de técnicas de *WebScraping*, objetivamos apresentar uma forma alternativa à Lei de Acesso à Informação para a coleta de dados jurisprudenciais em Tribunais de Justiça do Brasil. Considerando que cada TJ possui competência para formular e manter seu site, apresentamos no estudo de caso deste artigo, códigos em Python para três tribunais (Minas Gerais, Mato Grosso e Rondônia), os quais podem ser utilizados como base para a construção de códigos para outros tribunais.

Para além da introdução e conclusão, este trabalho conta com mais três seções, fazendo o caminho do Judiciário à raspagem. Primeiramente, discute-se a importância do acesso à informação para fins acadêmicos. Na sequência, apresenta-se a técnica de raspagem de dados. Por fim, aplica-se a referida técnica aos TJs supramencionados, a partir de códigos de *WebScraping* em linguagem Python.

2 A IMPORTÂNCIA DOS DADOS E A DIFICULDADE DE ACESSO

A Lei de Acesso à Informação (nº 12.527/2011) garante o direito fundamental de acesso à informação previsto na Constituição Federal aos cidadãos brasileiros (Brasil, 2021). É a partir dela que os interessados podem requerer a órgãos e entidades, inclusive o Poder Judiciário, informações que lhes sejam importantes, não podendo o solicitado exigir justificativa para tanto (Brasil, 2021) e, em caso dessas já estarem publicadas, é dever do requerido informar o local e forma de acesso.

A LAI é a representação da efetividade aos princípios de publicidade e moralidade da administração pública, sendo um mecanismo de garantia de transparência (Veloso, 2023) e fundamental para democracia (Bobbio, 2015). A partir de sua implementação, "[...] o Estado passou da posição de detentor do monopólio de 'documentos oficiais' para guardião de 'informações públicas'" (Michener; Contreras; Niskier, 2018, p. 611). Assim,

> [...] o acesso à informação pública é um requisito indispensável para o próprio funcionamento da democracia, maior transparência e boa gestão pública, e [...] em um sistema democrático representativo e participativo, a população exerce seus direitos constitucionais através da ampla liberdade de expressão e do livre acesso à informação. (Santana; Pamplona, 2019).

Em uma breve síntese, a Lei de Acesso à Informação é aplicada aos órgãos e entidades da administração pública direta e indireta dos três poderes em âmbito municipal, estadual e federal. A partir de sua promulgação, o sigilo de dados públicos passou a ser a exceção no cenário brasileiro (Santana; Pamplona, 2019).

Ao longo de seus artigos, a LAI estabelece duas formas de transparência: ativa e passiva (Brasil, 2021; Cruz; Zuccolotto, 2021; Santana; Pamplona, 2019; Michener; Contreras; Niskier, 2018). A primeira diz respeito aos dados escolhidos pela administração pública e seus representantes para serem publicados em seus sites, ou seja, é feita uma triagem previamente à publicização, sendo constantemente objeto de estudos em pesquisas sobre tribunais (Michener; Contreras; Niskier, 2018).

A segunda, por sua vez, refere-se aos pedidos de acesso à informação feitos pelos cidadãos à administração pública, marcada pelo formalismo excessivo no âmbito do Judiciário (Cruz; Zuccolotto, 2021). Por não ser previamente estabelecida, em razão de os órgãos e entidades não conseguirem prever os questionamentos a serem recebidos, "[...] representa um 'teste mais exigente' dos compromissos com o acesso à informação pública" (Michener; Contreras; Niskier, 2018, p. 611).

Considerando o objetivo apresentado na introdução, o qual consiste em apresentar uma forma alternativa para a coleta de dados jurisprudenciais em Tribunais de Justiça do Brasil, uma vez que as cortes não disponibilizam os acórdãos de forma compilada para facilitar o processo de pesquisa, o presente artigo perpassa pelas duas formas de transparência, focando na passiva. Em razão disso, cabe descrever pormenorizadamente o estabelecido no capítulo III da LAI, intitulado "do procedimento de acesso à informação" (Brasil, 2021).

Neste ponto, é importante destacar que apesar de ser mencionado diretamente como subordinado da LAI no artigo 1º, inciso I, é apenas em 2015 que o Judiciário passa a adotá-la (Santana; Pamplona, 2019), a partir de Resoluções do Conselho Nacional de

Justiça (CNJ), como a nº 215/2015 (Brasil, 2018). Assim, valemos da Lei nº 12.527/2011 e da referida resolução para explicitar o percurso dos pedidos de acesso à informação no âmbito dos Tribunais.

Nos termos do artigo 10 da Lei nº 12.527/2011 qualquer pessoa, seja física ou jurídica, pode requisitar informações aos órgãos do Poder Judiciário, normalmente via Serviço de Informações ao Cidadão (SIC), bastando identificar-se e especificar o pedido, sendo vedado exigir justificativa para tanto por parte do solicitante. O órgão solicitado tem o prazo de 20 (vinte) dias para responder a requisição ou negar o acesso, neste caso apresentando as razões. Ressalta-se que, conforme disposto no artigo 12, os Tribunais podem negar pedidos insuficientemente claros, desproporcionais; exigentes de trabalhos adicionais de análise ou tratamento dos dados; de informações já descartadas, legalmente sigilosas, como processos em segredo de justiça, e que ameacem a segurança institucional do Tribunal; e, por fim, dados pessoais, nos termos da Lei Geral de Proteção de Dados.

Valendo-se da transparência passiva, como descrito na introdução, este artigo advém de inúmeras solicitações de acesso a parte das bases de dados dos acórdãos publicados pelos Tribunais de Justiça dos estados de Minas Gerais, Rondônia e Mato Grosso para fins acadêmicos. Como mencionado, os sites dos TJs disponibilizam para consulta e download estes documentos, mas de forma individual, um por um, respeitando a transparência passiva. Encaminhamos pedidos solicitando acesso aos resultados da base a partir de uma requisição específica, de modo a facilitar o tratamento e uso dos dados para as pesquisas a serem desenvolvidas, ou seja, apenas requisitamos o compartilhamento de algo público e estruturado, o compilado dos acórdãos encontrados nas buscas de jurisprudência. Contudo, como também relatado anteriormente, recebemos negativas, sob a justificativa de os dados já estarem disponíveis ou pelo pedido trazer supostos riscos à segurança do tribunal, sem maiores esclarecimentos.

Diante da barreira do sistema e das negativas por parte dos Tribunais, restam aos pesquisadores opções alternativas para a extração dos dados que lhes servirão em seus trabalhos. Castro (2022); Carvalho (2021), Oliveira (2017) e Calò (2014) utilizam técnicas de *WebScrapping* para coletar os acórdãos de diferentes Tribunais. Nos próximos

tópicos demonstraremos como aplicar esta técnica em Tribunais estaduais, pouco explorados na literatura de *judicial politics*.

3 TÉCNICA DE RASPAGEM DE DADOS

A forma mais simples de extrair dados de um sistema, enquanto usuário externo, é com requisições estruturadas a um servidor, recebendo de volta informações que foram geradas pela deflagração de rotinas internas do servidor após o chamado realizado (Oliveira e Cunha, 2020). A descrição acima se refere a um estilo de arquitetura de serviços de tecnologia chamada API (Interface de Programação de Aplicações), a qual usa dos protocolos web HTTP e permite a comunicação entre sistemas distintos (AWS, 2024). Um exemplo notável da disponibilização de dados por uma API pública é o programa do Conselho Nacional de Justiça (CNJ) chamado de *DataJud*.

Neste programa os metadados das decisões judiciais, como tribunal que a proferiu, número do processo, nível de sigilo, formato, classe, entre outros, são disponibilizados para que desenvolvedores e pesquisadores tenham acesso facilitado a informações processuais públicas, originárias da Base Nacional de Dados do Poder Judiciário (Portaria do CNJ nº 160 de 9 de setembro de 2020). Contudo, como apontam Oliveira e Cunha (2020), "o Poder Judiciário não dispõe de *application programing interface* (API) para facilitar o acesso do público às informações processuais, mantendo e privilegiando o acesso individual de cada processo por advogados, partes, magistrados ou pesquisadores".

Apesar das diversas informações disponibilizadas pela plataforma *DataJud*, as quais, sem dúvida, auxiliam no desenvolvimento da pesquisa acadêmica a partir do acesso à informação e análise dos padrões da Justiça, não são ainda disponibilizadas as decisões em si, contendo o inteiro teor de seu conteúdo, pelo menos não de uma forma facilitada como a API pública do CNJ. Para contornar a falta de acessibilidade de dados estruturados, pesquisadores e desenvolvedores necessitam recorrer a técnicas mais antiquadas e antigas de extração de dados para chegar a essas informações, sendo a mais

comum a extração manual, técnica que exige muitos pesquisadores e muito tempo dos profissionais (Oliveira e Cunha, 2020).

Uma técnica alternativa, a qual será apresentada neste estudo, é o *WebScraping*. Com ela, os dados são extraídos dos sites de forma automatizada, copiando a forma que um humano poderia extraí-los de uma página na *web* e tornando dados não-estruturados em informações que podem ser salvas e analisadas em uma planilha (Sirisuriya, 2018). A tarefa realizada por um programa de *WebScraping* poderia ser realizada por uma pessoa que copiasse e colasse as informações obtidas do site, à maneira da extração manual, porém, além de ser uma tarefa tediosa e cansativa, seria também extremamente propensa a erros em razão da repetitividade e do tamanho da tarefa (Sirisuriya, 2018).

Essa tecnologia, a qual permite automatizar a organização e extração de informações em materiais analisáveis, pode ser dividida em três fases (Kheder, 2021). Primeiro, a fase de busca (ou *fetching*), que consiste em realizar a requisição no protocolo HTTP[4] para o site onde se encontra a informação relevante. Para essa fase são usadas bibliotecas e ferramentas que recriam requisições do tipo GET[5] no protocolo *web*, ou seja, somente simulam uma pessoa acessando o site que será a fonte dos dados. A segunda fase, chamada de extração de dados (ou *extraction*) consiste no uso de ferramentas de busca por padrões, como expressões regulares ou *queries* para encontrar a informação importante dentro de todo o conteúdo da página. Por fim, chega-se à fase da transformação dos dados (*transformation stage*) em que os dados anteriormente dispersos em cantos de um site são filtrados por relevância e convertidos para um formato estruturado de apresentação e armazenamento (Kheder, 2021).

Muitos são os usos possíveis para a técnica de raspagem de dados aqui discutida, incluindo o monitoramento de tendências de preços de produtos em sites de comércio

[4] Requisição no protocolo HTTP é uma mensagem enviada pelo cliente ao servidor, pedindo a realização de uma ação específica, como a obtenção de um recurso ou a submissão de dados e que segue o protocolo padrão da internet chamado HTTP (AWS, 2024).

[5] Uma requisição que segue o protocolo HTTP pode ser de diferentes tipos, uma delas é o tipo GET. Este é utilizado para solicitar dados de um servidor, sem alterar o estado do recurso solicitado, servindo majoritariamente para recuperar informações (AWS, 2024).

digital, controle da variação do preço de ações de empresas, elaboração de relatórios de monitoramento e levantamentos de metadados para análise de tendências ou análises bibliométricas (Kheder, 2021), dentre outros. O *WebScraping*, muito comum comercial e academicamente, é uma forma de lidar com a falta de dados estruturados e com a necessidade de varredura de grandes quantidades de informação típica da atualidade (Sirisuriya, 2018). Por exemplo, as empresas conhecidas como *lawtechs* e *legaltechs*[6] usam de técnicas de *WebScraping*, em escala de magnitude maior que a deste artigo, para criar seu próprio banco de dados de precedentes e processos em trâmite, agregando informações em um único local para facilitar o acesso e, no fim, transformando os dados em produtos para venda.

Apesar da habitualidade da prática, não deve deixar de ser debatida a questão ética da aplicação dessa técnica, ainda mais se tratando de raspagem de dados pessoais e sensíveis cujo armazenamento e tratamento está incluso nos regramentos da Lei Geral de Proteção de Dados ou equivalente do país de uso dos dados.

Ao extrair grandes quantidades de dados, os desenvolvedores podem encontrar problemas éticos e legais. O fato de os dados estarem disponíveis na internet não significa que a sua captura automatizada e estruturação para análise não vá ferir a ética e a regulação de dados do país. Por exemplo, o uso dos dados extraídos por raspagem automática pode gerar quebras contratuais, violação de direitos de propriedade intelectual ou divulgação, ainda que não intencional, do segredo de negócio de uma empresa (Krotov; Johnson; Silva, 2020).

Considerar esses dilemas éticos é importante para a pesquisa e para o uso comercial de técnicas de *WebScraping* de forma a se evitar o vazamento de informações e a violação de direitos relacionados a proteção de dados dos indivíduos ligados aos dados extraídos. Embora não seja esse o escopo e objetivo principal deste artigo, é importante ponderar a legalidade e a validade ética da extração automatizada de dados do inteiro teor

[6] De acordo com a Associação Brasileira desse ramo de empresa, *legaltechs* e *lawtechs* "são entidades que desenvolvem e fornecem produtos e serviços na seara jurídica com implementação de tecnologia agregada ao produto" (AB2L, 2024).

de decisões judiciais como aqui se propõe de forma alternativa para manutenção da pesquisa frente às negativas de acesso institucional via Lei de Acesso à Informação.

Certo é que a ética e a legalidade da raspagem de dados podem ser primeiramente analisadas pela possibilidade de acesso e tratamento dos próprios dados em si. Como já ponderado no tópico anterior, o inteiro teor das decisões judiciais é informação pública, disponibilizada diariamente via publicação nos Diários de Justiça de cada Tribunal e, posterior à publicação, nas abas de consulta jurisprudencial de cada Corte, como prevê e estimula a legislação pertinente (Lei de Acesso à Informação, Decreto Regulamentar nº 7.724 de 2012 e o Marco Civil da Internet).

Ainda que públicos os dados, caso extraídos por *WebScraping* o nível de litigância de pessoas, especialmente na esfera trabalhista, seria considerado essa ação uma violação frontal à Proteção de Dados Pessoais. Não sendo este o caso, é seguro assumir uma postura de que se está agindo dentro dos ditames da Lei, posto ainda quando não são expostos, como no caso desta proposta, dados pessoais das partes envolvidas no processo ou reveladas tendências de litigância de pessoas físicas de dados não anonimizados, mas somente são extraídos os dizeres judiciais de cada decreto judicial.

Ademais, é importante ressaltar que os códigos de *WebScraping* não burlam a segurança dos sites das Cortes, não invadem dispositivos e tampouco acessam dados sigilosos, mas somente acessam informação pública, fornecida pelo titular dessa informação, de maneira eficaz, rápida e, frisa-se, ética, tornando possíveis estudos sobre o Sistema de Justiça para sua compreensão e melhoramento.

Assim, superando o ditame ético levantado, desde que sejam respeitados os limites da privacidade e proteção de dados pessoais, e que as técnicas de WebScraping sejam empregadas conscientemente, é possível afirmar que tais práticas estão alinhadas com as normas vigentes. Além disso, o uso ético e eficiente dessas técnicas pode contribuir significativamente para a pesquisa e aprimoramento do Sistema de Justiça, proporcionando um melhor entendimento e desenvolvimento institucional, como abaixo será explorado em prática.

4 ACESSO ALTERNATIVO AOS DADOS: A PESQUISA RESISTE

O intuito principal deste artigo é demonstrar que mesmo diante das negativas de acesso aos dados de forma estruturada por parte dos requeridos, é possível, com pouca incursão nos ramos da tecnologia, que cientistas sociais se aventurem na seara dos estudos empíricos e quantitativos, lidando com dados de maior volume sem necessariamente depender de ampla mão de obra para obtenção das informações.

Como apresentado, este artigo surge da necessidade destas pesquisadoras em se construir uma base de dados de decisões judiciais dos Tribunais Estaduais. Diante das negativas via LAI, recorreu-se a técnicas de *WebScraping* para varrer o site dos TJs estaduais em busca dos dados de classe processual, número do processo, órgão julgador, câmara julgadora, comarca de origem, nome do Desembargador relator, data de julgamento, ementa e inteiro teor do acórdão. Ao fim era necessário chegar a uma forma estruturada dos dados, dispostas em uma planilha para fácil manipulação e análise de diversos pesquisadores sociais.

Para levantar as informações necessárias para o trabalho, ante a negativa dos Tribunais ao pedido feito por Lei de Acesso à Informação, foi adotada a metodologia de extração de dados com *WebScraping* que, contudo, não é feita sem desafios, os quais podem incluir a alteração de layout e código dos sites sendo raspados, o bloqueio de IP de um usuário que sobrecarrega os servidores e CAPTCHAS para verificação de robôs.

A aplicação do WebScraping, frisa-se, realizada somente em razão das negativas de acesso via LAI e em razão da inexistência de uma API pública que permita o acesso aos dados de forma direta. Embora a requisição constante de acessos ao site com *scripts* possa até mesmo gerar um bloqueio do endereço IP do usuário ou fazer com que o *script* esbarre em CAPTCHAS voltados a diminuir a carga de acessos com que servidor tem que lidar, esses desafios técnicos são contornáveis com a inclusão de esperas programadas dentro do código e com a implementação de leitores de imagens como o *pytesseract* para perpassar os CAPTCHAS incluídos nos sites. No caso deste estudo, identificou-se que, em respeito à LAI, os sites dos tribunais em análise são raspáveis sem necessidade de

grandes contornos técnicos, tendo sido as técnicas supramencionadas suficientes para a lograr a raspagem dos tribunais escolhidos.

Nestas páginas, visando a reprodutibilidade científica desse aporte metodológico que possibilita a construção de uma base de dados para estudos acadêmicos, serão explicadas a linguagem, bibliotecas e lógica por trás do desenvolvimento do código, o qual possibilita a construção de uma base de dados de decisões judiciais de qualquer tema. Visando a transparência e a reprodução em outros estudos, são disponibilizados os códigos (*scripts*) para extração das decisões judiciais nos TJs de Minas Gerais[7], Rondônia[8] e Mato Grosso[9], bem como, para exemplificação, a planilha advinda da execução do *script* voltado à extração de todas as decisões judiciais resultantes da busca por "'passagem aérea' e 'consumidor'" no TJMG nos anos de 2022 e 2023[10]. Além disso, com o intuito de facilitar os trabalhos dos iniciantes em programação, apresenta-se um tutorial mais detalhado e instrutivo dos requisitos de instalação e demonstração da execução dos *scripts*[11].

Os *scripts* de *WebScraping* para extração de decisões judiciais foram desenvolvidos em linguagem Python, simples e poderosa para a realização da tarefa, com auxílio das bibliotecas Selenium[12], NLTK[13], OpenPyxl (Felix, 2023) e RE (Python

[7] Disponível em: www.drive.google.com/drive/folders/1YrChcgMnDkGFUh5RuN5oPrnhoo3ZqtbP?usp=share_li. Acesso em: 7 jul. 2024.

[8] Disponível em: www.drive.google.com/drive/folders/1_m5rGSLB4qnujSYH5GhNb1qVpzZH53kW?usp=share_link . Acesso em: 7 jul. 2024.

[9] Disponível em: www.drive.google.com/drive/folders/1rI9rLquoRJFk2udJ1vOSDmwiwHIy_EOr?usp=share_link Acesso em: 7 jul. 2024.

[10] Disponível em: www.drive.google.com/drive/folders/1nr2h_5V2YWUdzkFAfh7d1WBuQfvbU5q7?usp=share_link

[11] Tutorial WebScraping disponível em: www.drive.google.com/drive/folders/1nr2h_5V2YWUdzkFAfh7d1WBuQfvbU5q7?usp=share_link Acesso em: 2 ago. 2024.

[12] Selenium é uma biblioteca para testes automatizados e de automação de navegação a web. Ela permite que desenvolvedores criem códigos que simulam interações humanas com o navegador. Em termos simplificados, a biblioteca é como um "robô" que pode abrir um navegador, clicar botões, digitar, fazer buscas e extrair informações. (Selenium, 2023)

[13] NLTK (Natural Language Toolkit) é uma biblioteca (um conjunto de programas) que facilita funções e recursos para processamento de linguagem natural. Ela permite que desenvolvedores reaproveitem

Software Foundation, 2023). Selenium é uma biblioteca para testes que permite automatizar interações com um navegador *web*, neste caso, o *Google Chrome*. Com o auxílio de um *WebDriver* (arquivo compilado de programas em diversas linguagens) é possível controlar o navegador e realizar tarefas típicas de navegação e interação com um site, como rolagem de tela, cliques e escrita. Já a NLTK (*Natural Language ToolKit*), assim como a RE (*Regular Expressions*) são bibliotecas *Python* facilitadoras da manipulação de textos, permitindo realizar a separação de trechos e identificar padrões específicos no conteúdo das decisões. Ainda, foi também usada a biblioteca *openpyxl* para permitir que os resultados obtidos da extração de dados fossem expostos de maneira fácil no software Excel.

Através do *WebScraping*, foi desenvolvida uma rotina automatizada que se utiliza da página de resultados da busca jurisprudencial, após inseridas as palavras-chave do assunto, para extrair informações jurisprudenciais, simulando uma interação humana com o site de pesquisa de precedentes. Por se tratar de uma busca automatizada, sem requerer o labor ostensivo e constante de um indivíduo, se garante uma alta precisão na extração dos dados, bem como a possibilidade de execução em larga escala e para fins de análise acadêmica (Kheder, 2021).

A primeira parte da raspagem inclui, como supramencionado, a fase de "*fetching*". O *WebDriver*, indicado o caminho dele no computador do usuário, aciona o método *.get()* para abrir o link desejado pelo desenvolvedor.

```
# Abre navegador já na página de resultado de UM acórdão
navegador = webdriver.Chrome(service=caminho)
navegador.get(link_da_pesquisa)
```

Após entrar no site, como simulação de uma interação humana, o código aguarda alguns segundos o carregamento do site antes de iniciar a varredura por informações. Essa pausa prévia ao ingresso no estágio de extração é importante para evitar a sobrecarga dos

estudos de outros pesquisadores em processamento de linguagem natural e apliquem as já consolidadas técnicas de forma fácil e prática. (Bird; Klein; Loper, 2021)

servidores do tribunal e gerar algum tipo de penalização ao IP do usuário em razão da quantidade de acessos e, também, para impedir que se inicie a busca por informações sem que o site esteja já carregado com todos seus dados.

Superado o carregamento, o navegador passa a procurar, na estrutura de todo o site, por identificadores únicos da estrutura do site que demonstrem onde cada informação está. Por exemplo, no trecho exemplificativo de código abaixo, o código realiza a busca pela classe processual, número do processo e nome do desembargador relator a partir do "seletor CSS"[14] de cada item de informação no site.

```
        # ACHANDO AS INFORMAÇÕES
        infoProcesso = navegador.find_element(By.CSS_SELECTOR,
'#tabelaEspelho > div:nth-child(4)')

        infoProcessoTokenizado =
RegexpTokenizer(r'\w+').tokenize(infoProcesso.text.lower())

        classeProcessual = ' '.join(infoProcessoTokenizado[0:-13])

        numeroProcessoCNJ = '.'.join(infoProcessoTokenizado[-7:-1])

        numeroProcessoTJMG = transformar_numero(numeroProcessoTJMG)

        infoRelator = navegador.find_element(By.CSS_SELECTOR,
'#tabelaEspelho > div:nth-child(7)')
```

Por fim, alcança-se a fase de transformação dos dados. Nessa etapa, as informações brutas obtidas pela performance automatizada do código são tratadas e padronizadas antes de serem incluídas numa planilha que esquematize e possibilite a visão macro dos dados. A forma de tratamento dos dados é feita por manipulação de *strings*, busca de padrões com expressões regulares e verificações dos dados em cláusulas

[14] Simplificadamente falando, seletor CSS é um identificador de uma parte do site dentro de toda a estrutura dele. Os desenvolvedores de site, para conseguir indicar para o navegador qual parte será estilizada ou modificada, colocam uma 'etiqueta' no elemento para identificar que algo deve ser feito nele, como, por exemplo, deixar as bordas arredondadas, trocar a cor ou mudar o local. Esse identificador pode também ser usado no WebScraping para achar os elementos dentro da página.

lógicas de *if-else* (como testes para ver se a informação que chegou do código confere com o desejado).

No código abaixo, por exemplo, usa-se do costume do Tribunal de Justiça de Minas Gerais de separar as informações processuais da decisão do inteiro teor pela expressão "A C Ó R D Ã O| ACÓRDÃO", escrita dessa forma, como um marco para procurar informações. Em razão desse padrão da Corte, permite-se usar técnicas de expressões regulares para procurar exatamente essa maneira de escrever e também técnicas de manipulação de *strings*, como os métodos "*split()*" ou o corte inverso de listas, para destacar informações relevantes de ruídos do texto.

```
        inteiroTeoreEmenta = navegador.find_element(By.CSS_SELECTOR,
'#panel1 > div:nth-child(3)').text
        inteiroTeorSplit = inteiroTeoreEmenta.split(r'A C Ó R D Ã O|
ACÓRDÃO')
        acordaoInteiroTeor = inteiroTeorSplit[-1]
```

Além disso, a variável "inteiroTeoreEmenta" é manipulada até se alcançar uma variável nova, "limpa", só com o inteiro teor do acórdão chamada "acordaoInteiroTeor". Importante destacar que essa técnica não usa inteligência artificial no reconhecimento de padrões, ou seja, ela não "aprende" com o erro em que informações não são capturadas. No caso de um erro, como caso o Tribunal tenha decidido seguir outro padrão para uma das decisões, o *script* somente segue adiante e deixa a informação vazia no lugar de tentar preenchê-la e acabar por cometer erros que seriam de difícil identificação no momento da análise dos dados.

Em última etapa, após a padronização e transformação dos dados, para que seja permitida a análise, os dados são colocados em planilhas de Excel com o auxílio da biblioteca *Python OpenPyXl*. Para possibilitar que os resultados da extração de dados sejam exibidos de maneira simples, inclusive para aqueles sem domínio em tecnologias de bancos de dados tradicionais, optou-se por essa ferramenta de trabalho que é conhecida e dominada corporativamente. Nos *scripts* disponibilizados pelas autoras deste artigo para

extração de dados para três distintos tribunais, foi usada a mesma função que insere informações no Excel:

```
def coloca_excel(folha, linha, coluna, valor):
celula = folha.cell(row=linha, column=coluna)
celula.value = ILLEGAL_CHARACTERS_RE.sub(r'', valor)
```

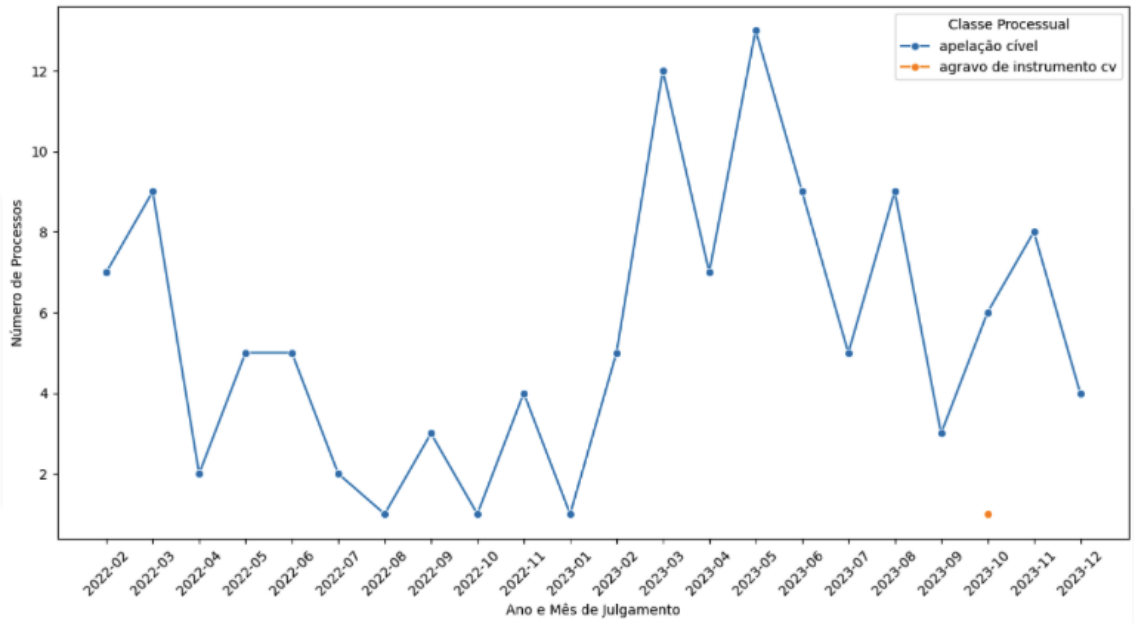
Como resultado do código, obtém-se uma tabela em que cada coluna traz informações de metadados do processo (como órgão julgador, classe processual, e números identificadores), bem como dados de mérito do processo em análise (súmula do julgamento, ementa e inteiro teor do acórdão).

Por meio do trabalho feito a partir do *WebScraping*, é possível a produção de gráficos a partir dos dados coletados e organizados. Como exemplo das possibilidades que a tabela resultado do *WebScraping* permite, exibem-se aqui, possíveis produções a partir dos dados. Neste artigo não se possui o objetivo de analisar com profundidade as razões dos resultados, mas apenas explorar, expor e exemplificar a técnica metodológica que aqui se defende.

Para a pesquisa exemplificativa de Direito do Consumidor que foi feita[15], foram encontrados 122 acórdãos, cuja raspagem completa demorou 48 minutos. A distribuição desses julgamentos por ano pode ser graficamente demonstrada pelo gráfico abaixo:

[15] A pesquisa foi feita no site de busca jurisprudencial do TJMG pelas expressão de busca "'passagem aérea' e 'consumidor'" com recorte de tempo para os anos de 2022 e 2023. A pesquisa foi realizada em 11 de julho de 2024 e retornou 122 acórdãos compatíveis com a busca.

Gráfico 1 - Análise de quantidade de processos por mês, ano e classe processual



Fonte: elaborado pelas autoras (2024).

No gráfico 1, dos 122 (cento e vinte e dois) acórdãos extraídos somente um era da classe processual de Agravo de Instrumento, enquanto todos os demais eram Apelações Cíveis. Ainda, percebe-se um significativo aumento de casos na virada do ano de 2022 para 2023, o que permitiria levantamento de hipóteses sobre este fato estar associado somente à aproximação do recesso forense e uma diminuição natural da produtividade dos servidores ou pode estar ligado a um início dos julgamentos associados a situações que impactaram diversos consumidores, possivelmente oriundas da Pandemia de Covid 19.

Caso se quisesse fazer uma incursão entre as palavras usadas nos julgamentos, seria possível usar a técnica de “nuvem de palavra”. No caso abaixo, para construção da figura 1 usou-se da biblioteca Python *WordCloud* para verificar quais as palavras com

maior ocorrência dentre os julgados colacionados, permitindo gerar inferências sobre os tópicos mais abordados nas decisões daquele período analisado. Nota-se uma ocorrência significativa da expressão “cancelamento de voo” junto com “falha na prestação de serviço”. Ressalta-se o fato de somente ser possível construir tal imagem em razão da coleta da ementa e inteiro teor dos julgamentos, os quais também foram organizados na planilha, como anteriormente descrito.

Figura 1 - Nuvem de palavras das decisões extraídas com *WebScraping*

Fonte: elaborado pelas autoras (2024).

A figura e gráfico apresentados possuem somente o objetivo de exemplificar formas de tratamento dos dados que podem permitir análises sociais e visualização de tendências no julgamento do Judiciário por meio dos dados obtidos através do *WebScraping*. A partir dos dados coletados e organizados, é possível que pesquisadores desenvolvam outras abordagens a partir dos resultados encontrados. O aqui apresentado trata-se apenas de uma forma para enfrentar as negativas respostas dadas pelo Poder Judiciário aos estudiosos e acresce-se ao amplo conjunto de ferramentas de resistência do meio acadêmico.

5 CONSIDERAÇÕES FINAIS

O intuito principal deste artigo foi demonstrar que mesmo diante de percalços no acesso aos dados de forma estruturada por parte dos Tribunais de Justiça, existem técnicas e formas de os ramos da tecnologia auxiliarem os cientistas a se aventurem na seara dos estudos empíricos e quantitativos, lidando com raspagem, extração e tratamento de dados sem depender de ampla mão de obra para obtenção das informações. Este estudo é, em seu fim, uma contribuição ferramental visando o incentivo à pesquisa e buscando demonstrar, explicar e tornar reprodutível a forma de se fazer pesquisa que se encontrou viável após se deparar e superar as "pedras do caminho" da falta de acesso.

Almejou-se neste artigo construir explicações sobre a transparência devida e, contudo, inacessível do poder judiciário, compartilhando a prática da busca de dados pela L.A.I e trazer exemplos que permitam a reprodutibilidade científica da técnica de *WebScraping* por cientistas sociais interessados em análises de dados judiciais. Contudo, sabe-se que o espaço limitado das páginas pode ser insuficiente para compreensão completa da técnica e dos códigos aqui em parte trazidos para explicação. Com isso em mente, as autoras disponibilizam também, como anexos a este artigo, os códigos para os tribunais de Minas Gerais, Rondônia e Mato Grosso, a tabela resultado da execução do código para o Tribunal de Minas Gerais e um tutorial mais detalhado e instrutivo dos requisitos de instalação e demonstração da execução dos *scripts*. Assim, advoga-se pela reprodutibilidade científica e pela resistência da pesquisa quantitativa apesar dos desafios e da falta de transparência do Judiciário.

O resultado de um esforço metodológico de *WebScraping*, embora só traga informações públicas e que - de fato - estão disponíveis no site de cada Corte, é inovadora em sua forma de extração e de agrupamento dos dados e por possibilitar análises diversas pelos pesquisadores sociais com os dados coletados. Gráficos e distintas visualizações dos dados como, por exemplo, de nuvens de palavras, recorrência de termos, quantitativo de julgamentos daquele tema por ano e por classe processual e diversas outras análises qualitativas e quantitativas se tornam possíveis em razão da extração e da forma

organizada que os dados são dispostos, demonstrando a relevância do trabalho exposto e do ferramental apresentado.

Entretanto, é necessário avançar. Ainda é preciso explicar a razão da negativa ao fornecimento de dados pelas Cortes Judiciais, bem como o incentivo a formas mais fáceis de obtenção de dados públicos, sem imposição de barreiras tecnológicas. É essencial que informações precisas, atualizadas e confiáveis sobre o judiciário estejam acessíveis a todos os atores sociais, de maneira cada vez mais coerente, exata e livre de obstáculos técnicos.

Apesar das limitações dos dados, as técnicas de WebScraping abrem um novo caminho para estudos, e espera-se que as explicações fornecidas neste artigo despertem a curiosidade e encorajem mais cientistas sociais e profissionais do Direito a se aventurarem na análise de dados, incentivando o diálogo entre instituições públicas e academia para possibilitar a elaboração e a avaliação de políticas públicas de melhoria e acompanhamento da prestação dos serviços de Justiça.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE LAWTECHS & LEGALTECHS. Disponível em: https://ab2l.org.br/ecossistema/radar-de-lawtechs-e-legaltechs/. Acesso em: 7 jul. 2024.

AWS. What is an API? **Amazon Web Services**, 2024. Disponível em: https://aws.amazon.com/pt/what-is/api/#:~=API%20significa%20Application%20Programming%20Interface,de%20serviço%20entre%20duas%20aplicações. Acesso em: 2 ago. 2024.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python:** Analyzing Text with the Natural Language Toolkit. Versão 3.6, 2021. Disponível em: https://www.nltk.org/book/. Acesso em: 7 jul. 2024.

BRASIL. Conselho Nacional de Justiça. **Portaria CNJ nº 160, de 9 de setembro de 2020**. Diário da Justiça, Brasília, DF, 10 set. 2020. Seção 1, p. 12. Disponível em: www.atos.cnj.jus.br/atos/detalhar/3453. Acesso em: 7 jul. 2024.

BRASIL. Conselho Nacional de Justiça. **DATAJUD:** Base Nacional de Dados do Poder Judiciário [recurso eletrônico]. Brasília: CNJ, 2024. Disponível em: https://www.cnj.jus.br/sistemas/datajud/. Acesso em: 7 jul. 2024.

BRASIL. **Decreto nº 7.724, de 16 de maio de 2012**. Regulamenta a Lei nº 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do caput do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição. Diário Oficial da União, Brasília, DF, 17 maio 2012. Disponível em: www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/decreto/d7724.htm. Acesso em: 7 jul. 2024.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005; e dá outras providências. Diário Oficial da União, Brasília, DF, 18 nov. 2011. Disponível em: www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.Acesso em: 7 jul. 2024.

BRASIL. **Lei nº 12.965, de 23 de abril de 2014**. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Diário Oficial da União, Brasília, DF, 24 abr. 2014. Disponível em: www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 7 jul. 2024.

CALÒ, Alessandro. Extração e Análise de Informações Jurídicas Públicas. 2014. 75 f. **Trabalho de Conclusão de Curso** (Bacharelado em Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014. Disponível em: https://bccdev.ime.usp.br/tccs/2014/sandro/Monografia.pdf. Acesso em: 17 ago. 2024.

CARVALHO, Taynara de Jesus. Pesquisa e Desenvolvimento de um Sistema de Automação de Jurimetria. 2021. 47 f. **Trabalho de Conclusão de Curso** (Bacharelado em Engenharia) – Universidade de Brasília, Faculdade UnB Gama, Brasília, DF, 2021. Disponível em: https://bdm.unb.br/bitstream/10483/30748/1/2021_TaynaraDeJesusCarvalho_tcc.pdf. Acesso em: 17 ago. 2024.

CASTRO, Marcella Queiroz de. Processamento de Linguagem Natural, Segurança Jurídica e Uniformidade da Jurisprudência: Um estudo sobre a viabilidade da aplicação de técnicas de Processamento de Linguagem Natural na identificação de divergências jurisprudenciais. **Monografia Final de Curso**, 2022, Faculdade de Direito, Universidade de Brasília, Brasília, DF.

Conselho Nacional de Justiça. **DATAJUD**: Base Nacional de Dados do Poder Judiciário [recurso eletrônico]. Brasília: CNJ, 2024. Disponível em: https://www.cnj.jus.br/sistemas/datajud/. Acesso em: 7 jul. 2024.

FELIX, Eric. Openpyxl: A Python library to read/write Excel 2010 xlsx/xlsm files. **Versão 3.0.8, 2023**. Disponível em: https://openpyxl.readthedocs.io/en/stable. Acesso em: 7 jul. 2024.

KHEDER, Moaiad. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. **International Journal of Advances in Soft Computing and its Applications**, v. 13, p. 145-168, dez. 2021.

KROTOV, V.; JOHNSON, L.; SILVA, L. Tutorial: Legality and Ethics of Web Scraping. **Communications of the Association for Information Systems**, v. 47, 2020. Disponível em: https://doi.org/10.17705/1CAIS.04724. Acesso em: 7 jul. 2024.

MAIA, Marcos; BEZERRA, Cicero Aparecido. Padrões nos acórdãos do Tribunal Regional Federal da Quarta Região. **Revista DireitoGV**, São Paulo, v. 19, 2023.

OLIVEIRA, F. L. de; CUNHA, L. G. **Os indicadores sobre o Judiciário brasileiro:** limitações, desafios e o uso da tecnologia. Revista DireitoGV, [S.l.], v.16, n.1, 2020.

OLIVEIRA, R. B. Utilização de Ontologias para Busca em Base de Dados de Acórdãos do STF. 2017. 58 f. **Dissertação** (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017. Disponível em: https://pdfs.semanticscholar.org/8072/b7b990d24d14e692b3bc8f16cf83639d1ea1.pdf. Acesso em: 18 ago. 2024.

RODRIGUES, Quemuel Baruque de Freitas *et al.* **Webscraping em R:** uma abordagem para investigação em ciências sociais. Simbiótica, Vitória, v. 08, n. 04, 2021.

SELENIUM. **Selenium:** Browser Automation. Versão 4.0.0-alpha-7, 2023. Disponível em: https://www.selenium.dev/. Acesso em: 7 jul. 2024.

SIRISURIYA, S. C. M. A Comparative Study on Web Scraping. *In*: INTERNATIONAL RESEARCH CONFERENCE - KDU, 11., 2018, Ratmalana, Sri Lanka. **Anais** [...]. Ratmalana: General Sir John Kotelawala Defence University, 2018. p. 59-65. Disponível em: http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y. Acesso em: 7 jul. 2024.

PYTHON SOFTWARE FOUNDATION. **re:** Regular expression operations. Versão 3.10.0, 2023. Disponível em: https://docs.python.org/3/library/re.html. Acesso em: 7 jul. 2024.