MÉTODO PARA PROCESSAMENTO AUTOMATIZADO DE FORMULÁRIOS PARA ALCANÇAR A DEEP GOVERNAMENTAL WEB

TORRES, Ana Hely Carvalho[1], CORRÊA, Andreiwid Sheffer[2]

[1] Graduando em Tecnologia de Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Câmpus Campinas, hely.ctorres@gmail.com.
[2] Docente, IFSP, Câmpus Campinas, andrewid@ifsp.edu.br.
Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

Apresentado no
10° Congresso de Inovação, Ciência e Tecnologia do IFSP
27 e 28 de novembro de 2019- Sorocaba-SP, Brasil

RESUMO: Segundo os princípios de dados abertos governamentais, é dever das instituições públicas fornecer o acesso irrestrito e reuso dos seus dados para a população, disponibilizando-os em formato legível por máquina. Porém, algumas instituições públicas locais não têm implementado de forma correta esses princípios, uma vez que seus portais de dados disponibilizam dados em formatos considerados não abertos e, principalmente, requerem o preenchimento de parâmetros em formulários web antes de apresentar os dados que residem na parte oculta da web (*deep web)*, limitando o processamento por máquina. Assim, visando garantir a acessibilidade dos dados governamentais, este trabalho objetiva comparar e aplicar métodos para processamento automatizado de formulários web, a fim de alcançar os dados governamentais que estão na *deep web*. Foram analisados os portais de transparência de 20 municípios do Brasil, que continham formulários para consulta, a fim de identificar e preencher seus respectivos campos a partir da interação automatizada com os elementos HTML da página, utilizando a linguagem de programação Python. O processamento automatizado de formulários web se mostra possível, porém apresenta alguns impasses devido a grande diversidade presente nas interfaces de busca de cada portal de transparência.

PALAVRAS-CHAVE: dados abertos; web governamental profunda; web scraping.

EVALUATION OF METHODS FOR AUTOMATED FORM PROCESSING TO ACHIEVE THE DEEP GOVERNAMENTAL WEB

ABSTRACT: According to the principles of open government data, it is the duty of public institutions to provide unrestricted access and reuse of their data to the population, making it available in machine readable format. However, some local public institutions have not implemented these principles correctly, as their data portals typically require filling in multiple field web forms before presenting the data. In order to ensure the accessibility of government data, this paper aims to compare and apply automated methods for web forms processing. We analyzed the transparency portals of twenty municipalities in Brazil, that apply HTML forms as tools for query public expenses or revenues, in order to identify and fill in their respective fields from the automated interaction with the HTML elements of the pages. Automated processing of web forms is possible but has some impasses due to the great diversity present in the query interfaces of each transparency web portal.

KEYWORDS: open data; deep governmental web; web scraping.

INTRODUÇÃO

Dados Abertos Governamentais ou simplesmente dados abertos, são termos que dizem respeito à publicação e divulgação de informações e dados do setor público, por meio da Internet, para livre utilização pela sociedade (AGUNE; GREGORIO FILHO; BOLLIGER, 2010). A área de Dados Abertos estabelece alguns requisitos conceituais e técnicos a fim de guiar a abertura dos dados governamentais com o uso das Tecnologias da Informação e Comunicação. Desde que a Lei do Acesso à Informação foi promulgada (BRASIL, 2011), as equipes políticas têm desenvolvido soluções técnicas referentes à disponibilização de seus dados, porém de forma independente e heterogênea, sem atender aos princípios de dados abertos governamentais. Não apenas, surgiu uma espécie de portal de dados cujo modelo baseia-se em um website clássico, no qual os dados residem na *deep web,* uma parte da internet não indexada, que não pode ser encontrada, pois seu conteúdo é gerado dinamicamente após uma consulta em um formulário web, porém diferente de *dark internet,* que está relacionada a parte da web que não pode ser acessada por meios convencionais . Em contraste, portais de dados abertos tornam os dados facilmente encontráveis na parte superficial da web, através dos tradicionais motores de busca, como o Google ou Bing. Ademais, estes portais requerem o processamento de formulários antes de apresentarem os dados nos formatos, por exemplo, PDF (*Portable Document Format*), HTML (*Hypertext Markup Language*) ou CSV (*Comma-separated values*), e por isso comprometem dois princípios de Dados Abertos, que são o acesso irrestrito e a possibilidade de processamento por máquina.

Portanto, medidas se tornam necessárias para resolver a problemática. Para isso, o presente trabalho objetiva comparar e aplicar métodos de processamento automatizado de formulários web para descoberta de dados governamentais que estão na *deep web,* realizar testes nos portais de transparência presentes nos websites oficiais dos órgãos públicos e analisar os resultados obtidos.

MATERIAL E MÉTODOS

A técnica desenvolvida utiliza métodos de web scraping (MITCHELL, 2015) para extrair, de cada website escolhido, o conteúdo da página HTML, identificar os atributos e os campos do formulário para então preencher os campos encontrados, submeter o formulário preenchido e, por fim, coletar os dados retornados do banco de dados web. Até o momento, o método permite a submissão do formulário com os valores extraídos do HTML, ou seja, campos de seleção são preenchidos com valores padrão e genéricos e campos de texto vazios. Os campos de texto ainda não estão sendo tratados, causando uma queda do desempenho caso o preenchimento correto desse tipo de campo seja obrigatório para submissão. Os algoritmos utilizados neste projeto foram desenvolvidos na linguagem de programação Python com o auxílio do IDE (*Integrated Development Environment*) PyCharm, em sua edição Community; gratuita. A figura 1 representa o ciclo de vida do método proposto e as ferramentas utilizadas:

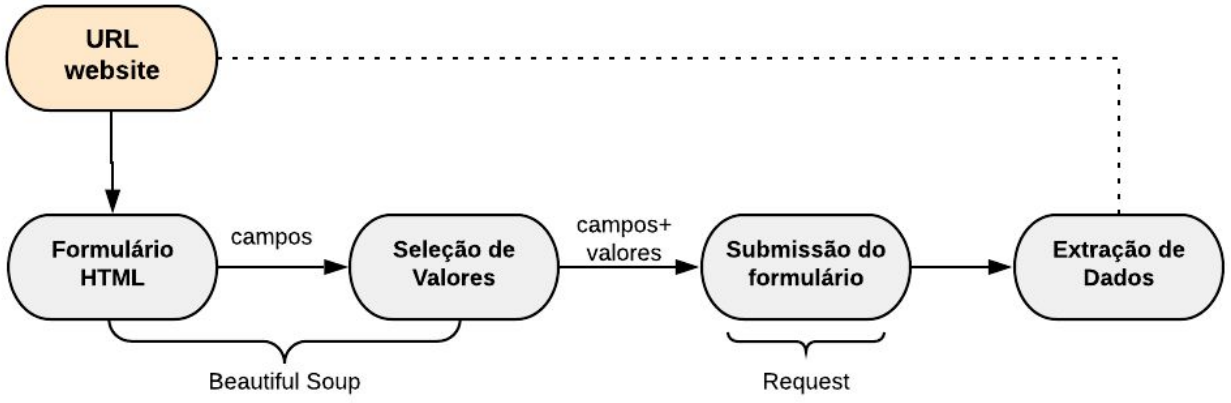

FIGURA 1. Ciclo de vida do método.

Para extrair o conteúdo da página, foi feita uma representação da árvore DOM *(Document Object Model*) do documento HTML utilizando a biblioteca *Selenium Webdriver,* e a partir dela foi criado um objeto de uma biblioteca utilizada em Python para *web scraping*, denominada *Beautiful Soup* (MITCHELL, 2015; RICHARDSON, 2015), que busca por todos elementos *form* da página, e extrai seus atributos(*method* e *action)* e campos(de texto e de seleção). Para cada tipo de campo é retirado seu atributo *name,* que servirá para identificação dos campos e posterior submissão. Após a extração dos campos, é feita a seleção dos seus respectivos valores. Para os campos de seleção(select, radio e checkbox) os possíveis valores podem ser extraídos facilmente do código HTML da página, utilizando a biblioteca anteriormente mencionada. Já para os campos de texto é necessário uma arbitragem de valores aplicando técnicas de Machine learning, que por enquanto não está sendo utilizada. Sendo assim, esses campos são preenchidos com valores padrão, que no caso é uma string vazia. Uma vez extraídos os campos do formulário e selecionados seus valores, é criado um conjunto de pares (campo/valor). Esses conjuntos de pares são necessários para formar um dicionário (*dict)* que será utilizado pela biblioteca *Requests* para submeter o formulário ao servidor através de uma requisição HTTP (REITZ, 2013).

RESULTADOS E DISCUSSÃO

Para o objetivo deste trabalho, foram escolhidos os portais de transparência de 20 prefeituras. Cada portal escolhido contém um formulário para consultas ao banco de dados web, como por exemplo consultas em relação a receitas e despesas públicas, com tamanhos e números de campos variados. Em cada portal foi feita primeiramente uma análise visual de sua interface e código fonte e logo após executados experimentos com o método. O experimento consistiu em utilizar o URL da página no algoritmo e analisar seu desempenho e resultados.

Após executados os experimentos, foi possível identificar alguns fatores limitantes que se repetiam nas páginas analisadas, os quais afetam a eficiência do método desenvolvido, como mostra a tabela 1. O percentual indica a proporção de sites analisados de acordo com a amostra (20 prefeituras).

TABELA 1. Fatores limitantes para o método.

Fator	Descrição	Prefeituras	Porcentagem
Tabelas Dinâmicas	Tabela de dados resultante da requisição escrita em Javascript.	Tocantins, Recife, Jundiaí, Campinas, São Roque	25%
Campos Despadronizados	Campos que não estão na tag *form* ou *input*, ou que estão em tabelas(*table)*.	Rio de Janeiro, Feliz Natal, Jundiaí ,Nazaré do Piauí, São Roque	25%
Função escondida para requisição	A requisição só é feita com algum elemento *hidden* ou uma função.	Bom Jesus, Recife, São Carlos, Cosmópolis	20%
Campos *text* Obrigatórios	Só retorna dados se o campo *text* estiver preenchido.	Pindamonhangaba, Cosmópolis, Campos do Jordão, Campinas	20%
Form nao acessível através do URL	O formulário só fica disponível após clicar algum botão na tela.	Taubaté, Salvador	10%
Código de verificação	Necessário digitar o código de verificação para poder ter acesso aos dados.	Belo Horizonte	5%

Dentre os 20 portais de transparência analisados, apenas 30% resultaram na execução bem-sucedida do algoritmo, ou seja, após a submissão do formulário foi possível ter acesso aos dados que estavam no banco de dados do site.

CONCLUSÕES

O processamento automatizado de formulários web para alcançar dados governamentais que estão da *deep web* mostra-se possível, porém apresenta alguns desafios devido à falta de padronização na criação dos portais para atender os princípios dos Dados Abertos. As interfaces de consulta apresentam grande diversidade, sendo criadas exclusivamente para manipulação pelo ser humano.

Os resultados preliminares alcançados com esta análise, já permitem verificar desvios entre o que é esperado para divulgação dos dados e o que é encontrado nos websites governamentais brasileiros. Cada portal apresenta uma singularidade em seu sistema, tornando o processamento automatizado um trabalho complexo e talvez não compensatório. Até o final do projeto, deseja-se implementar o tratamento de campos do tipo texto, arbitrando seus possíveis valores utilizando técnicas de Machine learning.

AGRADECIMENTOS

Agradeço ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (PIBIFSP) pela bolsa de iniciação científica, ao orientador por toda a confiança que me foi prestada, além do incentivo para realizar as atividades propostas, permitindo-me o crescimento e o amadurecimento técnico e pessoal.

REFERÊNCIAS

AGUNE, R. M.; GREGORIO FILHO, A. S.; BOLLIGER, S. P. Governo aberto SP: disponibilização de bases de dados e informações em formato aberto. In: CONGRESSO CONSAD DE GESTÃO PÚBLICA, III, Brasília, 2010.

MITCHELL, R. Web scraping with Python: collecting data from the modern web.1.ed.Sebastopol:O'Reilly Media, 2015.

REITZ, K. Requests: HTTP for Humans. 2013. Disponível em: <https://2.python-requests.org/en/master/>. Acesso em: 15 jun. 2019.

RICHARDSON, L. Beautiful Soup Documentation. 2015. Disponível em: < https://www.crummy.com/software/BeautifulSoup/bs4/doc/ > . Acesso em: 03 jun. 2019.