DOI: 10.5433/1981-8920.2021v26n4p319

WEB SCRAPING EM DADOS PÚBLICOS: MÉTODO PARA EXTRAÇÃO DE DADOS DOS GASTOS PÚBLICOS DOS VEREADORES DA CÂMARA MUNICIPAL DE BELO HORIZONTE

PUBLIC DATA WEB SCRAPING: METHOD FOR EXTRACTION OF DATA FROM THE PUBLIC EXPENDITURE OF THE COUNCILORS OF THE CITY HALL OF BELO HORIZONTE

Wendel Vilaça de Assis[a]
João Victor Boechat Gomide[b]

RESUMO
Objetivo: Demonstração que o método de *web scraping* na linguagem de programação *python* é capaz de extrair e transformar os dados desestruturados de custeio parlamentar do portal de transparência da Câmara Municipal de Belo Horizonte, em dados abertos estruturados. **Metodologia:** Está apoiada em pesquisa bibliográfica de dados públicos da Câmara Municipal de Belo Horizonte (CMBH), sob o ponto de vista de dados abertos no contexto da LAI, e análise qualitativa na extração de dados via *web scraping*. **Resultados:** Demonstra a eficácia do método de *web scraping* na extração de dados e na transformação em dados abertos estruturados. Isso permite o compartilhamento dos dados, possibilitando a produção de novas soluções para o protótipo de *Chat Bot* Sumé, apresentado neste trabalho. **Conclusão:** Eficácia do novo método de *web scraping* para extração de dados, seguida de manipulação para transformá-los em dados abertos, bem como apresentação do protótipo *Chat Bot* Sumé.

Descritores: *Chat Bot*. Dados Abertos. Dados Públicos. Inteligência Artificial. *Web Scraping*.

1 INTRODUÇÃO

O acesso dos cidadãos às prestações de contas dos gastos públicos pode

[a] Cientista de dados. Mestre pelo Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC. E-mail: wendelredes@gmail.com

[b] Doutor em Artes pela Universidade Federal de Minas Gerais (UFMG). Doutor em Física pela Universidade Estadual de Campinas (UNICAMP). Professor do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC. E-mail: jvictor@fumec.br

ser incentivado por meio da internet. As operações financeiras dos gastos públicos são capazes de serem monitoradas em qualquer momento desde que estes estejam acessíveis na internet. De acordo com Eaves (2009), a disponibilização de dados públicos geralmente ocorre mediante alguns parâmetros existentes na legislação vigente de cada país. Para garantir ao cidadão o acesso às informações no Brasil, foi instituída a Lei de Acesso à Informação (LAI). Essa lei ratifica a universalização do acesso aos dados da gestão pública. Paralelamente à LAI, há o Decreto nº 7.724, de 16 de maio de 2012, que a regulamentou na esfera federal e foi importante para estabelecer o acesso à informação pública no Brasil (SÁ; MALIN, 2012).

A Lei de Acesso à Informação (LAI) afirma que as instituições públicas, órgãos públicos ou órgãos independentes, devem possuir portal próprio de transparência, com ressalva aos municípios com menos de 10 mil habitantes (BRASIL, 2011). A LAI possui duas classificações de transparência, a Ativa e a Passiva. De acordo com (YAZIGI, 1999), “a transparência ativa é uma maneira proativa da administração pública divulgar informações sem a necessidade de algum pedido ou solicitação da sociedade”. Em concordância com (LOPES; ASSUMPÇÃO, 2013) “a transparência passiva possibilita que o cidadão solicite de maneira simples informações ao governo, sendo um pilar para fomento do controle social”.

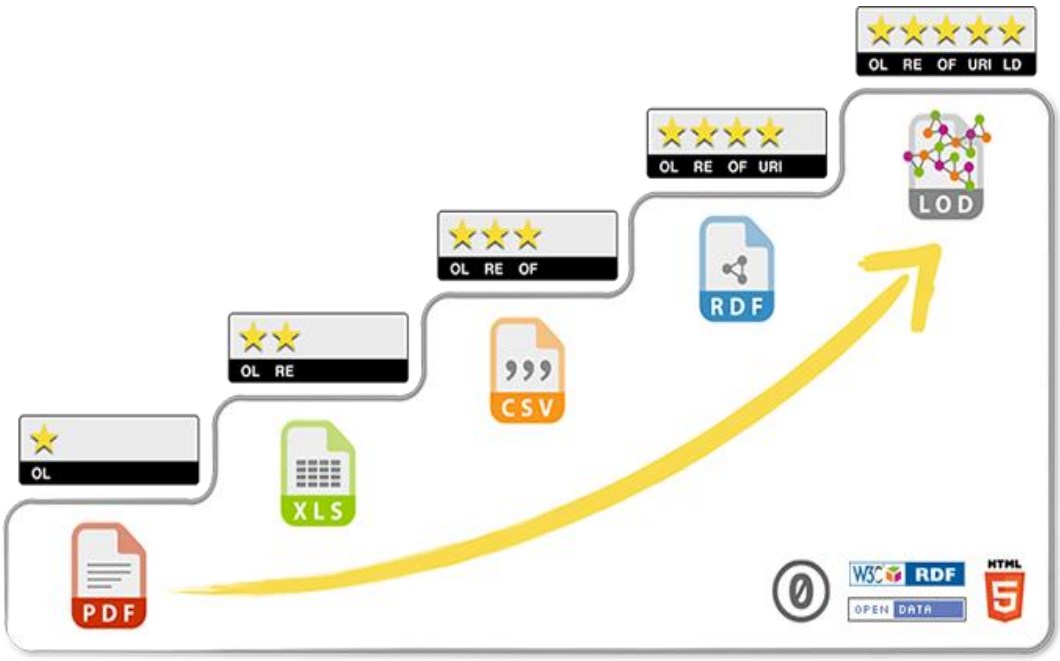
Os dados de custeio parlamentar do portal da Câmara Municipal de Belo Horizonte (CMBH) são estudados nesse artigo e estão relacionados às premissas da LAI referente a dados abertos nos municípios em que a lei se aplica. O presente trabalho tem o objetivo de demonstrar que a técnica de *web scraping* é eficaz na transformação de dados desestruturados em dados abertos, que serão definidos a seguir, e apresentar o agente de conversação *Chat Bot* Sumé desenvolvido nesta pesquisa. De acordo com Assis (2021), esse agente de conversação permitirá aos cidadãos consultarem os gastos mensais de custeio parlamentar dos vereadores da CMBH.

> O Chat Bot Sumé é uma solução fundamentada em dados abertos, tendo em vista que sua base de dados será o resultado do método de Web Scraping aplicado aos dados de custeio parlamentar no portal de transparência da CMBH (ASSIS, 2021, p. 52).

1.1 Dados Abertos

Os dados abertos são classificados como uma “Informação que tenha se tornado pública, mesmo que seja uma foto de uma digitalização de um fax de uma mesa - se a fotografia tiver uma licença aberta” (BERNERS-LEE, 2009). Berners-Lee (2009) desenvolveu a classificação por estrelas com o propósito de identificar a maturidade dos dados abertos. A classificação das estrelas proposta por Berners-Lee (2009) relaciona as estrelas como é retratado abaixo:

Figura 1 – Escala de desenvolvimento e enriquecimento dos dados abertos



Fonte: 5Stars (2012)

A escala de enriquecimento dos dados abertos é conceituada como “quanto mais estrelas você obtém, à medida que o torna progressivamente mais poderoso e mais fácil para as pessoas usarem” (BERNERS-LEE, 2009).

> 1 Estrela: Disponível na web (em qualquer formato), mas com uma licença aberta, para ser Open Data; 2 Estrelas: Disponível como dados estruturados legíveis por máquina (por exemplo, Excel em vez de digitalização de imagem de uma tabela); 3 Estrelas: Como (2) mais formato não proprietário (por exemplo, CSV em vez de Excel); 4 Estrelas: Além disso, usa os padrões abertos do W3C (RDF e SPARQL) para identificar os objetos,

> para que as pessoas possam apontar para seus objetos; 5 Estrelas: Todos os itens acima, mais: Vincule seus dados aos dados de outras pessoas para fornecer contexto (BERNERS-LEE, 2009, online, tradução nossa).

A LAI (BRASIL, 2011) aponta, em seu art. 8º, § 3º, que os órgãos e entidades públicas deverão divulgar informações na internet com o requisito "possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina; (BRASIL, 2011)". Segundo o apontamento da LAI (BRASIL, 2011), para ser considerado Dado Aberto, ele deve estar em formato aberto, estruturado e legível por máquina, ou seja, essa definição se encaixa exatamente no esquema de três estrelas de classificação do Berners-Lee (2009). Esse estudo adotou a definição de dados abertos da LAI (BRASIL, 2011) com a premissa de que os dados disponibilizados por instituições públicas só podem ser considerados dados abertos se estiverem classificados em no mínimo três estrelas do esquema proposto por Berners-Lee (2009).

1.2 DADOS DE CUSTEIO PARLAMENTAR DA CMBH

A Câmara Municipal de Belo Horizonte (CMBH), de acordo com as informações disponíveis em seu portal de Transparência, afirma que:

> O Portal da Transparência promove a ampliação das ações de divulgação dos gastos empreendidos pela Câmara, com a atualização constante de sua base de dados, de modo a favorecer a compreensão da sociedade sobre o funcionamento da Casa Legislativa (CMBH, 2020a, online).

No portal da CMBH existem três tipos de gastos de gabinete dos Vereadores. Eles são referentes aos gastos de custeio parlamentar, de serviços postais e de gastos com telefonia. Segue abaixo um exemplo de como acessar os gastos de custeio parlamentar dos vereadores:

Figura 2 – Custeio parlamentar

Fonte: CMBH (2020)

O portal da CMBH tem como propósito disponibilizar a prestação de contas dos gastos públicos dos vereadores:

> O Portal da Transparência da Câmara Municipal de Belo Horizonte – CMBH -constitui-se como um importante instrumento por meio do qual esta Casa realiza o processo de prestação de contas ao cidadão belo-horizontino, promovendo o acesso a dados e informações sobre a gestão administrativa e a execução orçamentária e financeira da CMBH (CMBH, 2020a, online).

Apesar da disponibilização dos dados, esses, por sua vez, não estão acessíveis (RODRIGUES; FONTES, 2018), tendo em vista que os dados estão desestruturados e não são classificados como dados abertos de acordo com a LAI. Até o presente momento, a CMBH não está de acordo com a legislação no que que diz respeito à acessibilidade dos dados e não está cumprindo os deveres inerentes à disponibilização de dados abertos conforme a LAI (BRASIL, 2011).

A ausência de dados abertos disponibilizados pelos órgãos públicos prejudica o acesso universal a esses dados de gastos públicos. Um exemplo de

inovação tecnológica amparada em dados abertos é o projeto *Serenata do Amor*, que é um sistema que possibilita controle social através da exposição de gastos públicos. Segundo Rodrigues e Fontes (2018), criadores da solução, “o objetivo desse projeto era expor para as pessoas informações sobre gastos dos deputados que já são públicas, porém não tão acessíveis” (RODRIGUES; FONTES, 2018).

A LAI explicita que o acesso aos dados abertos é um direito dos cidadãos, não sendo facultativo ao município, desde que este seja elegível para aplicação da lei. De acordo com o referencial teórico utilizado neste artigo e a LAI, com seus princípios de dados abertos, é possível afirmar que os dados disponibilizados no portal da CMBH, relativos aos gastos de custeio parlamentar, não são dados abertos, e, dessa maneira, não atende à LAI. Essa inconformidade se dá pelo fato de a CMBH não disponibilizar os dados de custeio parlamentar como dados abertos. Além dessa contradição, a CMBH também não atende aos padrões de transparência do Estado, que são um dos pilares do governo aberto (OGP, 2011).

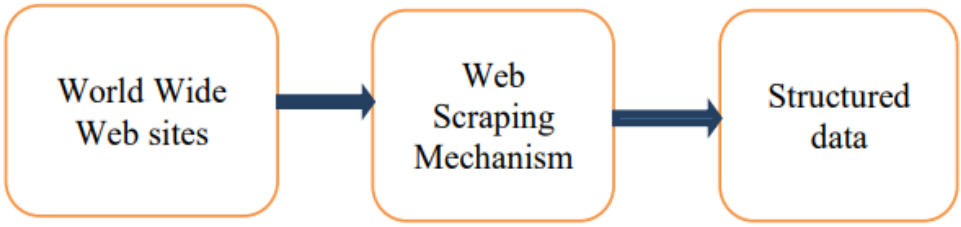
Considerando essas situações supracitadas, foi solicitado, junto à CMBH, a disponibilização de dados abertos relativos aos gastos dos vereadores. A solicitação formal ocorreu em 30 de setembro de 2020, por meio da Solicitação 64668. Apesar dos requerimentos, as respostas apresentadas pela CMBH no tocante aos dados abertos não foram atendidas. Vale ressaltar que, até o presente momento, nem previsão de projeto relacionado ao tema foi apresentado pela CMBH (ASSIS, 2021).

1.3 WEB SCRAPING

De acordo com o autor (HERNÁNDEZ *et al.,* 2015) “*web scraping,* ou extração de dados da *web*, é o processo de rastreamento e download de sites de informações e extração de dados não estruturados para um formato estruturado”. *Web scraping* pode ser considerado como uma técnica de mineração de dados que pode ser realizada na internet. Essa técnica permite a estruturação de dados após a realização da busca, extração e manipulação dos dados desestruturados.

Os dados provenientes dos métodos de *web scraping* podem ser estruturados em planilhas, bancos de dados ou arquivos no formato de *comma-separated values* (CSV), como em Mattosinho (2010). Ainda, de acordo com o autor, "*web scraping* é uma técnica usada para cortar informações de páginas da web com base em rotinas de script". O diagrama da figura 3 representa a arquitetura básica do *web scraping*:

Figura 1 – Arquitetura Básica de *web scraping*



Fonte: Mattosinho (2010)

Diouf *et al.* (2019) reveem as diferentes formas de se fazer *web scraping*, com suas categorias e ferramentas, assim como as áreas de aplicação. O *web scraping* é uma tarefa complexa e consome muito tempo de trabalho e recursos, especialmente se ela é feita manualmente. Tem muitas soluções automatizas propostas, como a apresentada no presente trabalho.

2 OBJETIVOS

Esse estudo tem como objetivo principal demonstrar a aplicação de técnicas de *web scraping* para solucionar inconformidades no portal da Câmara Municipal de Belo Horizonte (CMBH) de acordo com as diretivas da LAI. A aplicação da técnica tem, como finalidade, o aumento da transparência na publicação de gastos públicos no portal da CMBH. A linguagem de programação *Python* foi utilizada para desenvolver esse método de *web scraping*. Essa técnica foi empregada nos gastos de custeio parlamentar dos vereadores no portal de transparência da CMBH. Como alternativa para a inconformidade anteriormente abordada, foram adotadas as técnicas de *web scraping* para transformar dados desestruturados em dados abertos e foi realizada uma análise da aplicação do conceito de dados abertos de acordo com a LAI.

A intenção é criar alternativas para aumentar a transparência na divulgação dos dados públicos que estão disponíveis e, entretanto, não estão legíveis por máquina e estão desestruturados, o que dificulta o acesso e a análise desses dados pelos cidadãos. Ressalta-se a importância da adoção dessas técnicas para se utilizar em outros municípios e fomentar esse tipo de participação para aumentar a transparência pública e possibilitar acesso à informação aos cidadãos.

3 METODOLOGIA

A metodologia adotada nesse artigo contempla dois pontos principais. O primeiro está apoiado na pesquisa bibliográfica dos dados públicos sob a perspectiva de dados abertos relacionada aos requisitos da LAI aplicada aos dados de custeio parlamentar da CMBH. De acordo com (GIL, 2007, p. 44) “a pesquisa bibliográfica é desenvolvida com base em material já elaborado, constituído de livros e artigos científicos”.

O segundo ponto apresenta o método de *web scraping*, tema central nesse estudo e está pautado na demonstração detalhada da sua elaboração, desenvolvimento, extração e transformação dos dados desestruturados fornecidos pelo portal da CMBH. A Análise Qualitativa do método de *web scraping* desenvolvido tem como objetivo avaliar sua eficácia. De acordo com Bardin (2011), a pesquisa de análise qualitativa pode ser classificada como “... aquela que se fundamenta principalmente em análises qualitativas, sendo assim caracterizada de maneira geral pela não utilização de instrumental estatístico na análise dos dados”.

Para esse estudo foi utilizado uma nova concepção de transparência que complementa as existentes, nomeada no artigo como Transparência Ativa Reversa. Esse conceito traduz-se na transformação, em dados abertos legíveis por máquina, daqueles dados que antes não estavam em formato aberto e não eram legíveis por máquina (ASSIS, 2021). Podemos afirmar que:

> Transparência Ativa Reversa ocorre em situações que os próprios cidadãos desenvolvem métodos, técnicas, algoritmos, ferramentas ou soluções para transformar dados desestruturados e as informações públicas desordenadas em

> dados abertos. O nome Transparência Ativa Reversa faz analogia à Engenharia Reversa pelo fato de criar algo baseado em alguma coisa já existente (ASSIS, 2021, p. 30).

3.1 MÉTODO WEB SCRAPING

Na construção do método de *web scraping* é considerada primeiramente a análise do conteúdo do site que se deseja extrair dados. A partir dessa etapa, é possível mapear os dados que precisam ser extraídos. Assim, conseguimos obter indicadores para desenvolver o *script.* Nesse projeto foi utilizado a linguagem *python* na versão 3.6 (PSF, 2001). As bibliotecas utilizadas foram a *Python Selenium WebDriver, Beautiful Soup* e *Pandas*. O método de *web scraping* foi dividido em seis etapas, em que estão incluídas a extração, a manipulação dos dados desestruturados do custeio parlamentar e a transformação em dados abertos estruturados.

3.1.1 Scraping Vereador

Logo após o mapeamento dos dados que precisam ser extraídos do portal, a opção de “ferramentas do desenvolvedor” no navegador Google Chrome nos permite visualizar todos os elementos e marcadores HTML da página presentes no portal. Na extração do nome do vereador e do seu partido político é preciso identificar o “id” dentro do HTML. O “id” com as informações supracitadas estão na “*view-content*”, dentro de uma “*div class”* no HTML. Segue abaixo o exemplo:

Figura 4 – Div “view-content” nome do partido e vereador

Fonte: CMBH (2020)

Foi desenvolvido o *web scraping* por meio da biblioteca *Selenium Web Driver.* Com essa biblioteca foi implementado um robô para capturar os dados do id “view-content”, como no código a seguir:

Figura 5 – Código *Scraping* Vereador

```
import time
from bs4 import BeautifulSoup
from selenium import webdriver

driver = webdriver.Firefox(executable_path="C:/Users/Wendel/PycharmProjects/pythonProject/Driver/geckodriver.exe")
driver.get("https://www.cmbh.mg.gov.br/vereadores")
time.sleep(3)
dados_0 = driver.find_element_by_class_name("view-content")
html_0 = dados_0.get_attribute("innerHTML")
soup = BeautifulSoup(html_0, 'html.parser')
resultado = (soup.get_text())
print(soup.get_text())
with open("Vereador.csv", "w", encoding="utf-8") as f:
    s = "".join(resultado)
    f.write(s + "\n")
time.sleep(2)
driver.close()
```

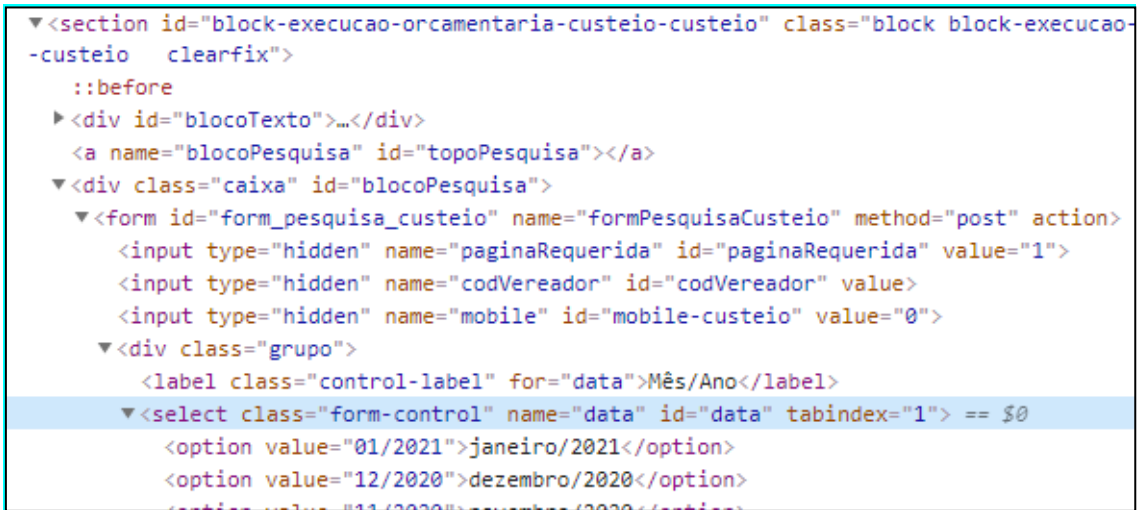
Fonte: ASSIS (2021)

3.1.2 Scraping custeio parlamentar

Nessa etapa foi realizado o mapeamento dos dados no portal referente ao

gasto mensal de custeio parlamentar. Como no exemplo anterior, do *scraping* Vereador, é preciso identificar qual “id” se refere ao mês com os respectivos gastos. O “id” referente à opção mensal está nomeado como “data”, id=“data”., como abordado a seguir:

Figura 6 – Id “data” custeio parlamentar



Fonte: CMBH (2020)

O próximo passo é identificar a *tag* HTML “id”. Esta, por sua vez, possui o resultado com os dados pesquisados anteriormente.

Após a posse dos dados supracitados, é possibilitada a criação do método de Scraping que utilizará a linguagem Python, munida da biblioteca *Selenium WebDrive.* Essa biblioteca permite a criação de um robô em que se é possível simular uma interação humana com os cliques do mouse nas páginas dos sites, por intermédio do navegador.

O robô criado acessará, por cliques, as *tags* que representam os meses, em “Mês/Ano”, e a opção de “Pesquisar”, e, assim, capturará os dados resultantes da pesquisa. O código a seguir demonstra as etapas para conseguir extrair os dados referente aos gastos do custeio parlamentar:

Figura 7 – Código Scraping custeio parlamentar

```
import time
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.support.ui import Select

driver = webdriver.Firefox(executable_path="C:/Users/Wendel/PycharmProjects/pythonProject/Driver/geckodriver.exe")
driver.get("https://www.cmbh.mg.gov.br/transparencia/vereadores/custeio-parlamentar")
time.sleep(3)
clicar0 = driver.find_element_by_id("data").click()
time.sleep(1)
mes = ["05/2017","06/2017","07/2017","08/2017","09/2017","10/2017","11/2017","12/2017",
       "04/2018","05/2018","06/2018","07/2018","08/2018","09/2018","10/2018","11/2018","12/2018",
       "02/2019","03/2019","04/2019","05/2019","06/2019","07/2019","08/2019","09/2019","10/2019","11/2019","12/2019",
       "01/2020","02/2020","03/2020","04/2020","05/2020","06/2020","07/2020","08/2020","09/2020","10/2020"]
for a in mes:
      select = Select(driver.find_element_by_id("data"))
    select.select_by_value(a)
    filtrar = driver.find_element_by_id("pesquisar-custeio")
    filtrar.click()
    time.sleep(2)
    dados = driver.find_element_by_id("resultadoPesquisa_custeio")
    html = dados.get_attribute("innerHTML")
    soup = BeautifulSoup(html, "html.parser")
    table = soup.select_one("table")
    headers = [header.text+";" for header in table.select("tr.success td")]
    print(headers)
    with open("Custeio_Parlamentar.csv", "a") as f:
        s = "".join(headers)
        f.write(s + "\n")
time.sleep(2)
driver.close()
```

Fonte: ASSIS (2021)

No código acima, o robô captura o resultado da pesquisa de cada mês, armazena temporariamente na memória, com estrutura de repetição *for* até a próxima repetição da estrutura elaborada. Todo o conteúdo HTML referente aos dados de gastos do custeio parlamentar é extraído e armazenado na variável “html”. A partir dessa etapa, passa a se utilizar a biblioteca *Beautiful Soup* para tratamento dos dados extraídos. Na próxima variável que for criada no código, nomeada como “soup”, os dados da variável “html” são recebidos para tratamento por meio da *Beautiful Soup*. O método *select_one* da variável “table” seleciona o texto que deve ser extraído. Na etapa de extração de dados desestruturados, é uma boa prática definir um delimitador para a separação dos dados, tendo em vista que, posteriormente, sua manipulação será mais fácil para

transformá-los em dados estruturados no formado CSV. O delimitador escolhido foi o caractere ponto e vírgula (;).

Na variável “headers” são armazenados os valores dos resultados obtidos pela estrutura *for*. Cada resultado dos gastos advém dos elementos “tr” da classe “success”, e da *tag* “td”, em formato de tabela, com colunas e linhas. Por fim, agrupando o resultado das etapas anteriores de extração, é realizada a gravação dos dados em formato CSV no arquivo “Custeio_Parlamentar.csv”, por intervenção do “with open”, concluindo assim a intervenção do robô *Selenium* e encerrando a seção do navegador. Vale ressaltar que todos esses procedimentos para extrair os dados do partido do vereador ocorrem da mesma forma, armazenando-os no arquivo “Vereador.csv”.

3.1.3 Manipulação de dados no *Pandas*

A biblioteca Pandas é utilizada para tratamento e análise de dados, tendo como principal biblioteca a *DataFrame* (PANDAS, 2020). Com o *Pandas* os dados que foram extraídos são trabalhados e armazenados em formato CSV, haja vista que esta biblioteca é extremamente eficiente e rápida para a manipulação de dados. Foram criados três *DataFrames* distintos para tratar os dados extraídos, cada um realizando uma tarefa específica para agrupamento de dados dos vereadores e seus respectivos gastos. Os *DataFrames* são “Partido_Vereador.csv”, “Vereador_Custeio.csv” e o “dataset-bot.csv”, este último usado como *DataSet* do *Chat Bot* Sumé.

3.1.4 DataFrame partido e vereador

A partir da fonte do arquivo “Vereador.csv “, foi necessário a manipulação dos dados no arquivo para agrupar de maneira estruturada, em colunas e linhas, o nome do vereador e do seu partido político. Para alcançar esse objetivo foi utilizada a estrutura “while” com a propriedade “df.shape”, para concatenar os dados, tendo em vista que eles se encontravam em uma única coluna. Através do código de concatenação, produziu-se a geração do arquivo “Partido_Vereador.csv”.

3.1.5 DataFrame vereador e custeio parlamentar

O *DataFrame,* que gerará como resultado o arquivo “Vereador_Custeio.csv”, teve tratamento do *Pandas* oriundo do arquivo “Custeio_Parlamentar.csv”, levando em consideração que seria necessário agrupar os dados em colunas distintas para unificar, posteriormente, o vereador com o gasto do custeio parlamentar. Essa concatenação gerou o arquivo “Vereador_Custeio.csv”, com os dados de gastos agrupados de todos os vereadores.

3.1.6 DataFrame Chat Bot Sumé

O último *DataFrame* e etapa que concluem o *web scraping* trataram, por meio do *Pandas,* os resultados dos *DataFrames* "Partido_Vereador.csv" e "Vereador_Custeio.csv". Essa etapa unificou os dados dos *DataFrames* através do id *Vereador*, para a realização da soma dos valores referente aos gastos de cada vereador e partido, como demonstrado no código a seguir:

Figura 8 – Código DataFrame *Chat Bot* Sumé

```
import pandas as pd

df_vereador = pd.read_csv("Partido_Vereador.csv", sep=';', encoding='cp1252')
df_gasto = pd.read_csv("Vereador_Custeio.csv", sep=';', encoding='cp1252')

df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('R','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('$','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('.','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace(',','.').astype(float)

df_gasto_agregado=df_gasto.groupby(by="Vereador").Gasto.sum().reset_index()

df_gasto_agregado = round(df_gasto_agregado, 2)
df_vereador_partido_gasto=df_gasto_agregado.merge(df_vereador,on='Vereador')

df_vereador_partido_gasto=df_vereador_partido_gasto[['Vereador','Partido','Gasto']]

print(df_vereador_partido_gasto)

df_vereador_partido_gasto.to_csv('dataset-bot.csv',index=False,sep=';',encoding='cp1252')
```

Fonte: ASSIS (2021)

A seguir está o exemplo do resultado gerado após a execução do código acima:

Figura 9 – Resultado DataFrame *Chat Bot* Sumé

Fonte: ASSIS (2021)

Concluindo todas as etapas do *web scraping* e da manipulação dos dados, procedeu-se à criação do arquivo “dataset-bot.csv”, que contém, de forma estruturada, todos os dados de gastos de custeio parlamentar de todos os vereadores e partidos da CMBH, no mandato de 2017 até 2020.

Dessa maneira, foi possível demonstrar a eficácia do método de *web scraping* criado para realizar a extração e manipulação dos dados de gastos do custeio parlamentar e estruturá-los num formato que atende as exigências da LAI referente a dados abertos. De acordo com os dados extraídos do portal de transparência da CMBH, o arquivo criado possibilitou identificar a soma de todos os gastos de custeio parlamentar dos vereadores, chegando ao valor de R$ 1.222.690,08 (Um milhão e duzentos e vinte e dois mil e seiscentos e noventa reais e oito centavos).

O resultado desse método também permitiu elaborar um *ranking* com os vereadores e partidos que mais gastaram verbas de custeio parlamentar em seus respectivos mandatos. Segue abaixo exemplo:

Tabela 1 – *Ranking* dos 5 vereadores que mais gastaram

Posição	Vereador	Partido	Gasto
1 °	Catatau do Povo	PSD	R$ 61.071,08
2 °	Juninho Los Hermanos	AVANTE	R$ 58.540,10
3 °	Gilson Reis	PCdoB	R$ 51.409,24
4 °	Dr. Nilton	PSD	R$ 45.804,42
5 °	Irlan Melo	PSD	R$ 45.641,15

Fonte: ASSIS (2021)

Tabela 2 – *Ranking* dos 5 partidos que mais gastaram

Posição	Partido	Gasto
1 °	PSD	R$ 345.802,13
2 °	AVANTE	R$ 99.219,33
3 °	CIDADANIA	R$ 95.465,69
4 °	PSB	R$ 81.799,73
5 °	PSC	R$ 80.262,43

Fonte: ASSIS (2021)

Os dados originais referentes aos gastos dos vereadores estão disponíveis no portal de transparência da CMBH. Entretanto, como explicado anteriormente, os dados não estão acessíveis em formato aberto conforme a LAI. É possível consultar todos os scripts do método *web scraping* apresentados neste artigo no GitHub, plataforma onde esses dados estão disponíveis e acessíveis a qualquer pessoa:

Link do Git Hub - https://github.com/wendelvilaca/chat_bot_sume

Ressalta-se que com o resultado desses dados extraídos do método de *web scraping* foi possível realizar a criação do *Chat Bot* Sumé, um agente de conversação que permitirá aos cidadãos consultar os gastos mensais de custeio parlamentar dos vereadores de acordo com (ASSIS, 2021). Segue abaixo o link do *Chat Bot* Sumé:

Chat Bot Sumé - https://bot.dialogflow.com/dc4da3ec-e99e-4cc3-ae4c-61fff833579a

4 DADOS EXCLUÍDOS DO PORTAL DE TRANSPARÊNCIA

Depois da conclusão do método de *web scraping* e da criação dos *Data*

Set, foi possível identificar algumas inconsistências dos dados disponibilizados no portal. Após várias tentativas de extração os dados, eles estavam inconsistentes e os valores da soma dos gastos dos vereadores não compactuavam com os resultados extraídos anteriormente. Após várias tentativas e revisão dos códigos para identificar algum erro, foi observado que o problema era que alguns dados de gastos do custeio parlamentar de vereadores não estavam mais disponíveis no portal, ou seja, tinham sido excluídos do portal de transparência. Após alguns dias, os dados faltantes, que antes estavam inacessíveis, ficaram disponíveis novamente. Essa ausência de dados no portal pode ter ocorrido por causa de alguma manutenção do sistema, ou alteração no banco de dados ou retirada proposital. Infelizmente, não foi possível saber o que ocorreu naqueles dias específicos, mas isso comprovou que o método construído funcionou efetivamente.

Vale ressaltar que só foi possível identificar essas inconsistências, tendo em vista que os dados eram extraídos mensalmente através do web *scraping* e eram armazenados localmente nos computadores e na nuvem, como uma boa prática de backup. Foi realizado, naquele período, um levantamento de cunho acadêmico dos dados faltantes, com o objetivo de contabilizar e comparar os resultados antigos aos valores que estavam ausentes. O resultado desse levantamento chegou ao valor de R$ 194.489,24 (Cento e noventa e quatro mil e quatrocentos e oitenta e nove reais e vinte e quatro centavos). Serviu como opção ranquear os valores e os vereadores que tiveram seus gastos retirados, conforme a tabela a seguir:

Tabela 3 – Dados excluídos do portal CMBH

Vereador	Partido	Gasto
Professor Wendel Mesquita	PSB	R$ 41.586,95
Jair Di Gregório	PP	R$ 35.676,16
Cláudio Duarte	PSL	R$ 31.501,47
Rafael Martins	PSD	R$ 22.726,21
Osvaldo Lopes	PHS	R$ 16.383,87
Cesar Gordin	PROS	R$ 16.278,51
Áurea Carolina	PSOL	R$ 12.913,67
Doorgal Andrada	PSD	R$ 8.706,73
Wellington Magalhães	PTN	R$ 6.146,92

Ronaldo Batista	PCC	R$ 2.128,97
Mateus Simões	NOVO	R$ 384,88
Ricardo da Farmácia	PMN	R$ 54,90
	TOTAL	**R$ 194.489,24**

Fonte: ASSIS (2021)

Se não existisse esse método de *web scraping*, em que foi possível extrair e armazenar os dados coletados do passado referente aos gastos, não seria possível que estas informações pudessem ser questionadas pelos cidadãos algum dia. Isso é uma prova de que esse método de *web scraping* é eficaz.

5 ANÁLISE DOS RESULTADOS

O portal de transparência da CMBH disponibiliza dados referente aos gastos dos vereadores. Após as análises realizadas neste artigo, é possível afirmar que nenhum dado disponibilizado pela CMBH está em formato aberto, ao menos até a coleta de dados do presente trabalho. Diante disso, a CMBH não atende aos requisitos da LAI (BRASIL, 2011) no que se refere a dados abertos. Segundo a classificação de cinco estrelas proposta por Berners-Lee (2009), a CMBH possui maturidade mínima referente a dados abertos, ou seja, possui somente uma estrela. Os dados de custeio parlamentar da CMBH são expostos na página do portal e, para acessar esses dados, é necessário selecionar o mês e o ano que se deseja. Em outros termos, os dados apresentados são desestruturados no que diz a respeito do conceito de dados abertos da LAI (BRASIL, 2011).

Este artigo demonstrou como o método de *web scraping* é eficaz na concepção de dados abertos, estruturados em formato CSV, proveniente de dados desestruturados. Isto significa que é possível adequar qualquer tipo de dado desestruturado para o patamar de 3 estrelas (BERNERS-LEE, 2009), podendo ser manipulável por máquina Eaves (2009) e atender aos requisitos da LAI (BRASIL, 2011). Vale destacar que, neste trabalho, uma base de dados foi criada através do método de *web scrapingí,* em que foi possível utilizá-lo como fonte de dados para o protótipo do *Chat Bot* Sumé. Este protótipo está disponível gratuitamente para a população como um agente de conversação, podendo ser

realizada interação e visualização dos gastos durante o mandato dos vereadores de Belo Horizonte e permite, com isso, o exercício do controle social (ASSIS, 2021).

Depois de inúmeras pesquisas e análises dos dados de gastos, um consumo específico chamou a atenção, o de Serviço Postal, que é ofertado pela CMBH como exposto abaixo:

> Além do gabinete, a Câmara oferece serviços e materiais complementares a cada um dos vereadores, mediante processos de aquisição definidos nos termos da legislação federal de licitações:
>
> VII - Serviços Postais - Serviço fornecido pela Empresa Brasileira de Correios e Telégrafos, disponibilizado aos gabinetes dos vereadores, por meio do Contrato nº 9912468166/2019. O contrato está em vigor desde setembro de 2019 (CMBH, 2020b, online).

Por ser um serviço que pode ser totalmente dispensável aos cofres públicos, tendo em vista que pode ser substituído por diversos recursos tecnológicos ou meios de comunicação digitais mais baratos ou de graça, o Serviço Postal foi utilizado inúmeras vezes. Salienta-se que somente no mandato de 2017 a 2020 foram gastos, no mínimo, R$ 701.845,29 (setecentos e um mil oitocentos e quarenta e cinco reais e vinte e nove centavos) com este item. Esse é um exemplo que demonstra que a população deve exercer fiscalização dos gastos públicos e, consequentemente, o controle social.

Uma comparação pertinente relativa aos gastos de Serviço Postal, que pode ser realizado no cenário em que vivemos atualmente, é referente aos meses de fevereiro e maio de 2020. Somente em três meses, esse serviço consumiu R$ 99.875,77 (noventa e nove mil oitocentos e setenta e cinco reais e setenta e sete centavos).

Até o presente momento da publicação deste artigo, não foi obtido nenhum resultado relativo a projetos ou adequação da CMBH relacionados à disponibilização de dados abertos dos gastos dos vereadores.

6 CONCLUSÃO

Mediante análise dos dados disponíveis no portal da CMBH relacionados aos gastos do custeio parlamentar, foi possível identificar irregularidades na disponibilização desses dados que estavam desestruturados e em formatos que não atendem aos requisitos da LAI no tocante a dados abertos. Para tratar essa inconsistência, o método de *web scraping* foi produzido, de maneira independente, para realizar a extração dos dados desestruturados, manipulação e transformação em dados abertos estruturados. O resultado da aplicação desse método possibilitou que os dados de custeio parlamentar entrassem em conformidade com a LAI e pudessem estar disponíveis para o acesso da população através do *Chat Bot* Sumé e de pesquisadores, pelo Git Hub. Este artigo demonstrou a eficácia do método de *web scraping* em tornar os dados desestruturados do portal da CMBH em dados abertos. Esse método foi testado e viabilizou o acesso aos gastos de custeio parlamentar dos Vereadores em formato de dados abertos estruturados.

Esse artigo demonstra o espaço existente para inovações empregadas em dados desestruturados na internet, utilizando tecnologias como o *web scraping* e o *Chat Bot* Sumé, com foco nos dados públicos, tendo por finalidade o aumento da transparência pública. A contribuição, demonstrada nesse trabalho, é de que é possível fomentar o aumento na transparência pública através da Transparência Ativa Reversa aqui proposta. Com relação ao aspecto social, o *Chat Bot Sumé* pode ser um percursor, no município de Belo Horizonte, de um agente para consulta dos gastos de custeio parlamentar e uma interface de conversação entre a CMBH e os cidadãos do município. Com isso, a população pode exercer o controle social com mais facilidade e recursos. Também, no campo social, espera-se que esse trabalho gere debates e discussões na população acerca dos gastos dos vereadores, incentivando, assim, a verificação da necessidade de utilização e prioridade dos serviços utilizados pelos vereadores.

AGRADECIMENTOS

Os autores gostariam de agradecer ao Conselho de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa de Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro recebido.

REFERÊNCIAS

5 STARS OPEN DATA. **5 Stars Open Data**. 2012. Disponível em: https://5stardata.info/en/. Acesso em: 15 set. 2020.

ASSIS, W. V. **Chat Bot Sumé:** web scraping em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte. Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) – Faculdade de Ciências Empresariais, Universidade Fumec, Belo Horizonte, p. 90. 2021.

BARDIN, L. **Análise de conteúdo**. São Paulo: Edições 70, 2011.

BERNERS-LEE, T. **Linked Data**. 2009. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 18 ago. 2020.

BRASIL. **Lei nº 12.527**, de 18 de novembro de 2011. Presidência da República. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 05 set. 2020.

CÂMARA MUNICIPAL DE BELO HORIZONTE – CMBH, 2020a. **Transparência.** Belo Horizonte, 07 de julho de 2020. Disponível em: https://www.cmbh.mg.gov.br/transparencia-principal. Acesso em: 20 de junho de 2020.

CÂMARA MUNICIPAL DE BELO HORIZONTE – CMBH, 2020b. **Custeio Parlamentar**, 2020. Belo Horizonte, 07 de julho de 2020. Disponível em: https://www.cmbh.mg.gov.br/perguntas-frequentes/vereadores-sal%C3%A1rio-presen%C3%A7a-custeio-do-mandato-gabinetes/como-s%C3%A3o-custeados-os. Acesso em: 20 de junho de 2020.

DIOUF, R.; SARR, E. N.; SALL, O.; BIRREGAH, B.; BOUSSO, M.; MBAYE, S. N. Web Scraping: State-of-the-Art and Areas of Application. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 7., Los Angeles, CA, USA, 2019. **Proceedings** […] Los Angeles: IEEE, 2019. p. 6040-6042, doi: 10.1109/BigData47090.2019.9005594.

EAVES, D. **The Three Laws of Open Government Data**, 2009. Disponível em: http://eaves.ca/2009/09/30/three-law-of-open-government-data/. Acesso em: 3 dez. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

HERNÁNDEZ, A.; GÓMEZ VÁZQUEZ, E.; RINCÓN, C. A. B.; GARCÍA, J. M.; MALDONADO, A. C.; IBARRA-OROZCO, R. Metodologías para análisis político utilizando Web Scraping. **Research in Computing Science**, [*S. l.*], v. 95, p. 113-121, 2015. DOI: 10.13053/rcs-95-1-9.

LOPES, K. M. G.; ASSUMPÇÃO, R. C. Processos e solução tecnológica para implementação da lei de acesso à informação (LAI). *In*: CONGRESSO CONSAD DE GESTÃO PÚBLICA, 6., Brasília, 2013. **Anais** [...] Brasília: 2013.

MATTOSINHO, F. J. A. P. **Thesis on Mining Product Opinions and Reviews on the Web.** Technische Universitat Dresden ,2010.

OPEN GOVERNMENT PARTNERSHIP (OGP). **Declaração de governo aberto**, 2011. Disponível em: www.opengovpartnership.org/open-government-declaration. Acesso em: 10 set. 2020.

PANDAS. **About Pandas**. Disponível em: https://pandas.pydata.org/about/. Acesso em: 8 de jun. 2020.

PYTHON SOFTWARE FOUNDATION (PSF). **What is python?** 2001. Disponível em: https://docs.python.org/3/faq/general.html#what-is-python. Acesso em: 18 ago. 2020.

RODRIGUES, J. C.; FONTES, C. Estudo de Caso “Operação Serenata de Amor”: a análise de Big Data no combate à festa dos gastos públicos. *In*: CONGRESO DE LA ASOCIACIÓN LATINOAMERICANA DE INVESTIGADORES DE LA COMUNICACIÓN, 14., San Pedro, 2018. **Anais** [...] San Pedro: Universidade da Costa Rica, 2018.Disponível em: https://www.researchgate.net/publication/323585318_Estudo_de_Caso_Operacao_Serenata_de_Amor_a_analise_de_Big_Data_no_combate_a_festa_dos_gastos_publicos. Acesso em: 18 out. 2020.

SÁ, M. I. F.; MALIN, A. M. B. Lei de Acesso à Informação: Um Estudo Comparativo com Outros Países. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 13., Rio de Janeiro, 2012. **Anais** [...] Rio de Janeiro: Fio Cruz, 2012.

YAZIGI, A. F. Dinero, política y transparencia: el imperativo democrático de combatir la corrupción. INTERNATIONAL ANTI-CORRUPTION CONFERENCE, 9., África do Sul, 1999. **Anais** […]. África do Sul, 1999. p. 10-15.

PUBLIC DATA WEB SCRAPING: METHOD FOR EXTRACTION OF DATA FROM THE PUBLIC EXPENDITURE OF THE ALTERNATORS OF THE CITY HALL OF BELO HORIZONTE

ABSTRACT
Objective: Demonstration of the Web Scraping method in Python capable of extracting and transforming the unstructured Parliamentary Costing data from the transparency portal, of the Belo Horizonte City Hall, into structured open data. **Methodology**: It is governed by a bibliographic search of Public Data from the Municipality of Belo Horizonte (CMBH), from the point of view of Open Data in the context of LAI and qualitative analysis in the extraction of data via Web Scraping. **Results**: Efficacy of the Web Scraping method in data extraction and transformation into structured open data, which allows data sharing, enabling the production of new solutions, the Chat Bot Sumé prototype, presented in this work. **Conclusion**: Efficacy of the new method of Web Scraping for data extraction, followed by manipulation to transform them into Open Data as well as presentation of the prototype Chat Bot Sumé.

Descriptors: Artificial Intelligence. Chat Bot. Government Data. Open Data. Web Scraping.

SCRAPING WEB DE DATOS PÚBLICOS: MÉTODO DE EXTRACCIÓN DE DATOS SOBRE GASTO PÚBLICO DEL CONSEJO MUNICIPAL DE BELO HORIZONTE

RESUMEN
Objetivo: Demostración del método Web Scraping en Python capaz de extraer y transformar los datos no estructurados de Costo Parlamentario del portal de transparencia, del Ayuntamiento de Belo Horizonte, en datos abiertos estructurados. **Metodología**: Se rige por una búsqueda bibliográfica de Datos Públicos del Municipio de Belo Horizonte (CMBH), desde el punto de vista de Datos Abiertos en el contexto de LAI y análisis cualitativo en la extracción de datos vía Web Scraping. **Resultados**: Eficacia del método Web Scraping en la extracción y transformación de datos en datos abiertos estructurados, que permite compartir datos, posibilitando la producción de nuevas soluciones, el prototipo Chat Bot Sumé, presentado en este trabajo. **Conclusión**: Eficacia del nuevo método de Web Scraping para la extracción de datos, seguido de manipulación para transformarlos en Open Data así como presentación del prototipo Chat Bot Sumé.

Descriptores: Chat Bot. Datos públicos. Información abierta. Inteligencia artificial. Web Scraping.

Recebido em: 13.07.2021
Aceito em: 09.12.2021