DESCOBERTA E IDENTIFICAÇÃO AUTOMATIZADAS DE PORTAIS DE DADOS PARA APOIAR O BENCHMARKING DE DADOS ABERTOS GOVERNAMENTAIS

MAZZOTTI, Victor Lucas[1], CORRÊA, Andreiwid Sheffer[2]

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

RESUMO: A existência de catálogos globais que indicam a existência e funcionamento de portais de dados é um desafio a ser vencido. Apesar de existirem alguns, não são atualizados e, na maioria das vezes, não são construídos de maneira automatizada. Este projeto busca, identifica e classifica de maneira automatizada portais de dados para que seja possível reunir todos os portais do mundo em um repositório único e confiável, promovendo maior entendimento e facilitando o acesso a estes.

PALAVRAS-CHAVE: Benchmarking; Dados abertos; Internet; Portais.

AUTOMATED DISCOVERY AND IDENTIFICATION OF DATA PORTALS TO SUPPORT GOVERNMENTAL OPEN DATA BENCHMARKING

ABSTRACT: The existence of global catalogues is a challenge to be overcome. Although there are some, they are not updated and, most of the time, are not built in an automated way. This project searches, identifies and classifies data portals, therefore it is possible to gather all the portals of the world in a single and reliable repository, promoting greater understanding and facilitating access to them.

KEYWORDS: Benchmarking; Open Data; Internet; Portals.

INTRODUÇÃO

O Benchmarking de dados abertos se resume em avaliar e classificar países, organizações e projetos baseado no modo em que são disponibilizados publicamente. Ele auxilia no entendimento e na comunicação para meios mais eficientes de utilizar dados abertos para solução de problemas. Os portais de dados abertos costumam utilizar plataformas de dados abertos. Dentre as plataformas mais utilizadas, observa-se: Arcgis; Ckan; Dkan; Junar; Opendatasoft; Pmydata; Socrata; Udata. Essas são justamente as plataformas que esse projeto visa identificar.

Existem diversos portais de dados abertos que não seguem os padrões exigidos, descumprindo os princípios estabelecidos pelos dados abertos. Segundo os princípios dos dados abertos, os dados devem ser disponibilizados de modo que possam ser facilmente lidos e manipulados (FOUNDATION; Open. 2019). Para a busca dos portais de dados abertos, utiliza-se um *crawler* web aberto denominado *Common Crawl* — Organização sem fins lucrativos que visa disponibilizar uma cópia da internet para qualquer usuário de forma gratuita — para descobrir URLs globalmente e, em seguida, trata os mesmos nos algoritmos desenvolvidos utilizando a linguagem de programação Python. Depois de tratados pelos algoritmos, os resultados são expostos mostrando quais das páginas web identificadas são portais de dados abertos e, também, qual plataforma de dados abertos cada um utiliza. Deste modo, já que estão reunidos e expostos em um só lugar, obtém-se uma visão mais ampla dos portais de dados abertos no mundo todo, facilitando a comparação entre eles. Este projeto contribui para o

desenvolvimento de uma maneira automatizada de construir um único repositório atualizado e confiável (CORREA e SILVA, 2019).

MATERIAL E MÉTODOS

Cada uma das plataformas trabalhadas neste projeto possui uma assinatura, que se trata de uma API (*Application Programming Interface*) da plataforma que é utilizada para identificá-la unicamente, considerando os dados retornados por ela. Quando uma API é utilizada em uma página web e obtém-se uma resposta, é obtido um retorno em JSON, que trata-se de um formato de arquivo fácil de ser interpretado por qualquer linguagem de programação (W3C, 2011). Este retorno indica, sobretudo, que esta página é provavelmente um portal de dados abertos. Além disso, na maioria dos casos, a própria assinatura indica qual plataforma este portal utiliza, já que cada plataforma tem sua assinatura própria.

TABELA 1. Cada portal e sua respectiva assinatura utilizada para acessar a API.

Portal	**Assinatura**
Udata	/api/1/datasets/?page_size1
Socrata	/api/catalog/v1
Arcgis/ OpenDataSoft	/api/v2
Junar	/manageDeveloper/create/’
Dkan/Ckan	/api/3/action/site_read
Pmydata	/sparql.json?querySELECT+%2A+WHERE+%7B%3Fs+%3Fp+%3Fo%7D+LIMIT+1

Como se observa na TABELA 1, algumas plataformas utilizam a mesma assinatura. Neste caso, a diferenciação de uma plataforma para outra se dá através do retorno que a API fornece. Para utilizar as assinaturas, basta unir a assinatura que se deseja ao endpoint (porta de entrada das requisições) de uma determinada página web (ex: www.linkdosite.com/assinatura).

Para obter uma cópia da internet, faz-se a utilização do *Common Crawl,* baixando os arquivos que a organização disponibiliza. Com estes arquivos, é possível listar todas as páginas web que a organização identificou.

Assumindo-se que se deseja identificar apenas os portais de dados abertos que utilizam das oito plataformas apresentadas na TABELA 1 a partir das páginas web identificadas pelo *Common Crawl*, as páginas web são submetidas aos três algoritmos deste projeto.

O primeiro algoritmo deve reunir todos as páginas web que o *Common Crawl* disponibilizou em URLs, separando-as e organizando-as em arquivo de texto. Durante a etapa de separação, são escolhidas apenas as páginas web que possuem potencial para serem portais de dados abertos. Assume-se que, uma vez que uma determinada página web possua um determinado domínio ou uma determinada palavra-chave(opendata, dados, ..) , possui potencial para ser um portal de dados abertos.

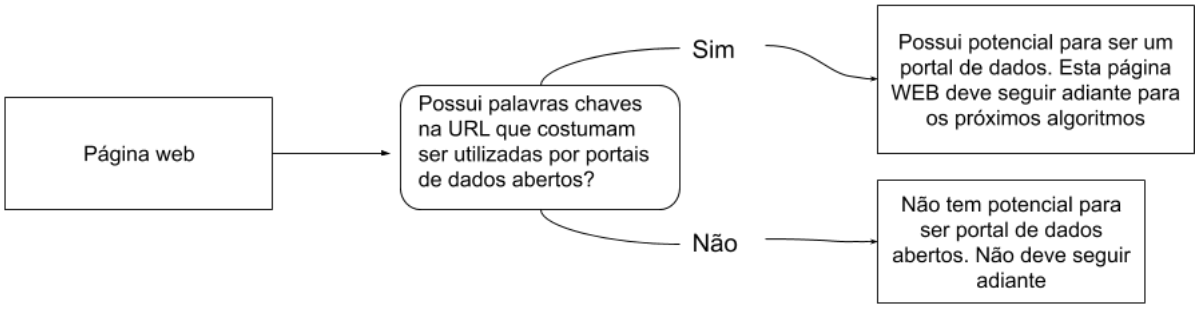

FIGURA 1. Representação de como as páginas web disponibilizadas pelo Common Crawl são tratadas na primeira etapa deste projeto.

Conforme se identifica na FIGURA 1, as páginas web passam por uma espécie de filtro, onde se identifica quais destas são prováveis portais de dados abertos ou não. Neste momento, não foi feita nenhuma identificação de plataforma de dados abertos, apenas a separação e organização das páginas web disponibilizados do *Common Crawl* , para só então poder passar para a próxima etapa onde se utilizam as assinaturas, que é responsabilidade do segundo algoritmo.

No segundo algoritmo é feita a criação de dois arquivos de texto. Um deles, é designado para armazenar aquelas URLs que foram identificados como portal de dados abertos e, juntamente com a URL, armazena-se qual plataforma esse portal de dados abertos utiliza. Já o outro é responsável por armazenar aquelas páginas web que não foram identificados como portal de dados abertos e, portanto, assume-se que não são portais de dados abertos. Uma vez que estes arquivos tenham sido criados, é feita a leitura do arquivo gerado pelo algoritmo anterior que possui as URLs dos possíveis portais de dados identificados. Cada linha é lida e, para cada linha, faz-se a tentativa de acessar cada página web com cada uma das assinaturas apresentadas na TABELA 1. Esta tentativa é feita em forma de requisição *HTTP*. Para realizar esta requisição via algoritmo Python, faz-se o uso de uma biblioteca chamada *URLLIB*, biblioteca disponível com diversas funções para manipulação de URL (FOUNDATION; PYTHON, 2020). Uma vez que uma página web tenha retorno para uma das assinaturas da TABELA 1, assume-se que este indica um portal de dados abertos pela presença da respectiva plataforma e, então, faz-se a inserção desta página web no arquivo contendo portais de dados abertos. Do contrário, faz-se a inserção desta página no arquivo que é designado para páginas que não foram identificados como portal de dados abertos.

No terceiro e último algoritmo, reúne-se apenas os portais que foram identificados na etapa anterior e, faz-se requisições HTTP para eles, porém dessa vez utilizando técnicas para identificar a região de cada portal de dados identificado e a quantidade de datasets. Ao fim desta etapa, obtém-se dois arquivos de texto. Um deles representa aqueles portais que obtiveram um retorno JSON conhecido. Já o outro, serve para aqueles que não tiveram um retorno JSON ou que tiveram, mas na verdade se tratam de falsos positivos.

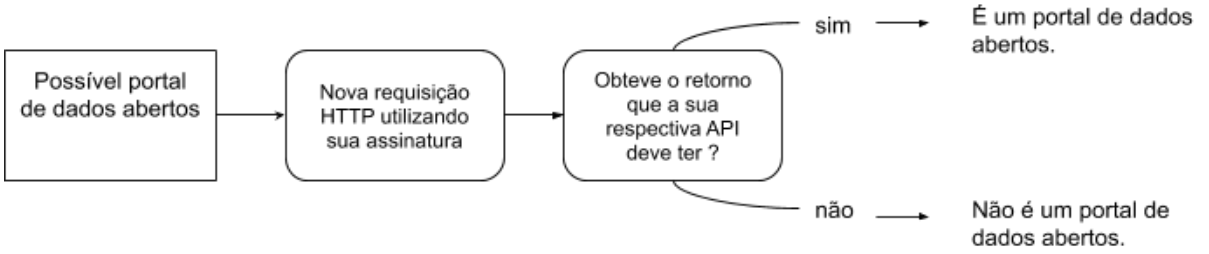

FIGURA 2. Representação da forma como o algoritmo trata as páginas que foram identificadas como possível portal de dados abertos na etapa anterior.

Conforme analisa-se na FIGURA 2, faz-se uma nova requisição HTTP utilizando a assinatura específica de cada página web que tem potencial para ser um portal de dados abertos. Ao fim desta requisição, se verifica se esta teve o retorno que deveria ter ao utilizar sua assinatura. Se sim,

classifica-se como portal de dados abertos. Do contrário, classifica-se como uma página web que não é um portal de dados abertos.

RESULTADOS E DISCUSSÃO

Para fins deste trabalho, será apresentado os resultados de agosto de 2019 até abril de 2020. Para cada mês trabalhado no decorrer do projeto, foi feita uma análise apontando a frequência de cada portal e seu respectivo idioma.

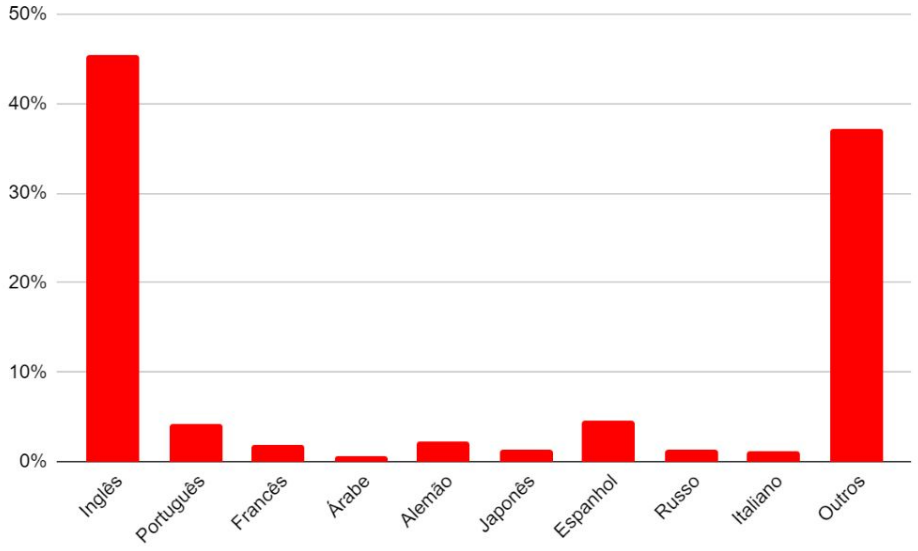

FIGURA 3. Gráfico representando frequência de cada idioma nos portais de dados identificados de ago/19 até abr/20.

Pode-se observar na FIGURA 3 que o idioma inglês lidera em número de ocorrências e, então, chega-se à conclusão de que o idioma inglês é o mais frequente em todos os portais de dados abertos identificados neste projeto.

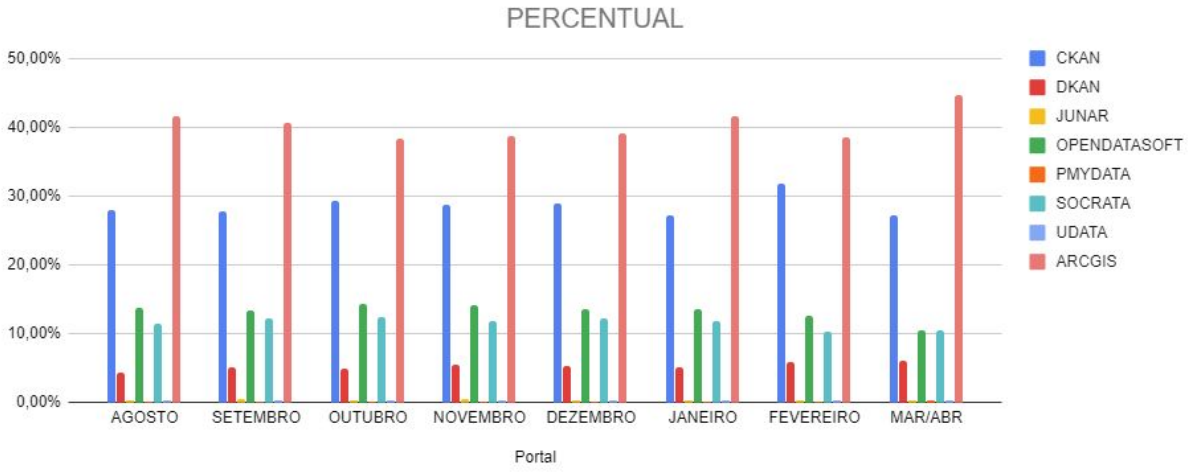

FIGURA 4. Representação quantitativa de frequência de plataformas dos meses de AGO/19 até ABR/20.

Conforme conclui-se com a análise da FIGURA 4, observa-se também que a plataforma ARCGIS é a que mais se destaca em número de instalações.

CONCLUSÕES

Com as informações obtidas ao longo deste trabalho, pode-se reunir todo o resultado obtido e catalogar cada um deste, organizando e classificando cada portal identificado. Deste modo, há dados suficientes para a elaboração de um único repositório. Este, contendo, além de confiabilidade e organização, fácil entendimento para quaisquer estudos futuros que sejam realizados acerca do assunto. Os resultados obtidos possibilitam também o ranqueamento dos países identificados, podendo-se identificar aqueles países que mais seguem os princípios dos dados abertos.

AGRADECIMENTOS

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pela bolsa de iniciação científica e, sobretudo, ao meu orientador que, além de prestar apoio e incentivo para a realização das atividades propostas, confiou a mim a responsabilidade de seguir este projeto.

REFERÊNCIAS

CORREA, A. S.; SILVA, F. S. C. Laying the foundations for benchmarking open data automatically: a method for surveying data portals from the whole web. Proceedings of the 20th Annual International Conference on Digital Government Research: Governance in the Age of Artificial Intelligence. Anais...Dubai, United Arab Emirates: ACM, 2019. Disponível em: <http://doi.acm.org/10.1145/3325112.3325257>

FOUNDATION, Open Knowledge. What is open? 2019. Disponível em: <https://okfn.org/opendata/>. Acesso em: 24 maio. 2020.

FOUNDATION, Python Software. urllib - URL handling modules. 2020. Disponível em: <https://docs.python.org/3/library/urllib.html>. Acesso em: 23 set. 2020.

W3C .Manual dos Dados Abertos: Governo. 2011. Disponível em: <https://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 24 set. 2020.