Scraper: Mapping FOIA requests and FOIA denials in Brazil

Luiz Fernando Toledo

Context for scraper 1

Brazil enacted access to information law, the same as FOIA in the US, in 2012. Since then, more than 1 million requests were made by citizens.

But what kind of questions do people ask?

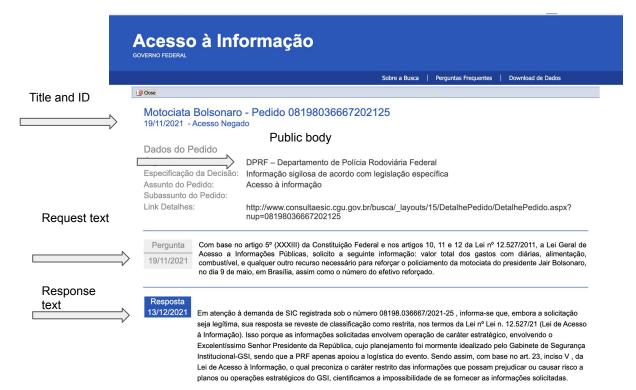
Repository

The Brazilian government has a website where all requests are kept. You can search by keyword, by public body's name and by date

Busca de pedidos e respostas

Busque aqui as respostas dadas aos pedidos de informação feitos a órgãos e entidades do Poder Executivo Federal. A pesquisa pode ser feita por palavra-chave ou por filtros específicos Órgão: Decisao: Busca Avançada

How requests are kept



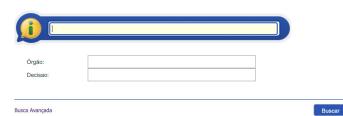
Scraper 1

My goal was to create a scraper that:

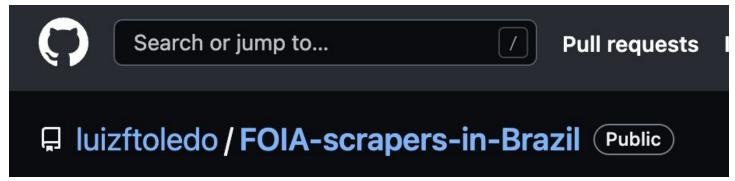
- -> Looks for a specific keyword in this website in a given period (let's say last 30 days)
 -> Scrapes all FOIA requests in that period containing that specific keyword (ex: Bolsonaro)
- -> Saves the results in a CSV file

Busca de pedidos e respostas

Busque aqui as respostas dadas aos pedidos de informação feitos a órgãos e entidades do Poder Executivo Federal. A pesquisa pode ser feita por palavra-chave ou por filtros específicos



Scraper steps



```
# import everything that I will use
import re
import pandas as pd
import time
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.support.ui import Select
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.support.ui import WebDriverWait
from datetime import date
import requests
from tqdm.auto import tqdm
```

Time range using the hyperlink + date.today()

```
#find a link that gives me the last 30 days FOIA requests with the keyword 'Bolsonaro'
# if you want to change the keyword, just change the variable search words
# creating a link
keyword = 'bolsonaro'
today = date.today()
days = datetime.timedelta(30)
thirty days ago = today - days
thirty days ago = str(thirty days ago)
year regex = r''(\frac{d}{d})-''
year only = re.findall(year regex, thirty days ago)[0]
month regex = r"....(..)"
month only = re.findall(month regex, thirty days ago)[0]
day regex = r".....(..)"
day only = re.findall(day regex, thirty days ago)[0]
### gambiarra monstra pra pegar a URL pg ainda to aprendendo, rs
useful link1 = 'http://www.consultaesic.cqu.qov.br/busca/SitePages/resultadopesquisa.aspx?k=ALL('
useful link2 = keyword
useful link3 = ')%20DataPedido%3E%3D'
useful link4 = day only
useful link5 = '%2F'
useful link6 = month only
useful link7 = '%2F'
useful link8 = year only
useful link = useful link1 + useful link2 + useful link3 + useful link4 + useful link5 + useful link6 + useful link7 + useful link8
```

Selenium: click + go to next page until the end

```
#step one: create a list with links from all the pages
list of links = []
#try to open up to 100 pages in order to find those links
for i in tqdm(range(100)):
    for pedido in driver.find elements(By.CLASS NAME, 'ms-srch-ellipsis'):
        links = pedido.find elements(By.TAG NAME, 'a')
        for link in links:
            caso = link.get attribute('href')
            list of links.append(caso)
    #click on the next page
    try:
        driver.find element(By.XPATH, '//*[@id="PageLinkNext"]').click()
        time.sleep(2)
    except:
        break
```

Create a list with the results and then a dataframe. Save it into a csv

```
# create a list and fill it with request title, request link and request text

lista_de_pedidos = []
for link in tqdm(list_of_links):
    lista_do_pedido = []
    driver.get(link)
    resumao = driver.find_element(By.XPATH, '/html/body/form/div[4]/div[10]/div[1]/div[2]/div[2]/div[2]/div[1]/table/tbody/tr/td/div/div_lista_do_pedido.append(resumao.text)
    link_do_pedido = driver.find_element(By.XPATH, '/html/body/form/div[4]/div[10]/div[1]/div[2]/div[2]/div[2]/div[1]/table/tbody/tr/td_lista_do_pedido.append(link_do_pedido.text)
    texto_do_pedido = driver.find_element(By.XPATH, '/html/body/form/div[4]/div[10]/div[1]/div[2]/div[2]/div[2]/div[1]/table/tbody/tr/td_lista_do_pedido.append(texto_do_pedido.text)
    lista_de_pedidos.append(lista_do_pedido)

lista_de_pedidos
```

```
import pandas

df = pd.DataFrame(lista_de_pedidos, columns = ['resumo', 'link', 'texto_do_pedido'])

df.to_csv('lista_de_pedidos.csv')
```

Magic happens

② ② lista_de_pedidos.csv						
0	Motociata Bolsonaro - Pedido 08198036667202125	http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=08198036667202125	Com base no artigo 5º (XXXIII			
1	Transmissão eventos EBC - Pedido 53125001689202182	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=53125001689202182001689200169920016992001699200169920016992001699200169920016992001699000169900016990001699000169900016990000169900000000$	Bom dia, visto que nos últimos			
2	Motociata Paraná - Pedido 00137020260202199	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=00137020260202199$	Com base no artigo 5º (XXXIII			
3	- Pedido 48003007432202181	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=48003007432202181$	Solicito acesso as providencias			
4	Voos da FAB - Pedido 03005233291202100	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=030052332912021000000000000000000000000000000$	Boa noite. Solicito a descrição			
5	Autoria de relatório - Pedido 08198036590202193	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=08198036590202193$	Prezados, venho requerer, com			
6	Produção de biodiesel Pronaf - Pedido 21210015332202126	http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=21210015332202126	Gostaria de obter a participaçã			
7	acesso - Pedido 25072032776202117	$http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=25072032776202117$	Peço acesso à íntegra do Proce			
8	Regulamentação da Lei 14.166, de 10 de junho de 2021 - Pedido 59011000053202187	http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=59011000053202187	Prezados, gostaria de saber se			
9	uso de recursos federais para apae - Pedido 71003100014202183	http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=71003100014202183	Gostaria de saber se há um ca			
10	Acesso aos dados - Pedido 03005227873202149	http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=03005227873202149	Prezados, Gostaria de ter aces			

Why is it important?

- It helps journalists to monitor a specific subject based on how the government is responding
- It makes SO MUCH easier to read FOIA requests and have ideas for new requests or even using existing databases for potential stories and projects
- The most important is that...

Why is it important?

It makes FOIA PEOPLE'S LIFE EASIER!!!



Scraper 2: FOIA denials and appeals

Context for scraper 2

I run a NGO in Brazil called Fiquem Sabendo (it could be translated as "Get in the know"), a data agency specialized in the Brazilian FOIA and public data in general. One of our a projects is Don't LAI to Me, a newsletter focused on disclosing new data and documents through FOIA (LAI means Lei de Acesso à Informação, the brazilian FOIA)

a Saúde habilitou mais 540 UTI exclusivos para o tratar no Sistema e (SUS). Das unidades hat m na Paraíba, 10 em Perná, 78 no Rio Grande Catarina, 88 em SF so,18 em A' Federal, 8 ações fo overno Fe s para refc objetivo foi c vidas. Para as has mento foi de R\$ 77,7 i. ela única pelo Ministério c a previsão de 140 es da vacina ara Covid-19



Don't LAI to me

a newsletter da Fiquem Sabendo para quem quer informação direto da fonte

leitos de UTI exclusivos para o pacientes com Covid-19 no Si 'SUS). Das u rraíba, 10 em Pernambuco, Grande do Sul, 45 em Sar 11 em Mato Gross Distrito Feder ilitações forc no Fed ra reforçar o etivo foi cuid 'ara as habilita oi de R\$ 77.7 m ão de 140 milhõe Além desse mo - ----- da 7

LAI/FOIA is under threat due to privacy and secrecy laws



I want to monitor when FOIA requests are being denied due to a privacy law. It is called LGPD - Brazilian General Data Protection Law

There is another website with all FOIA appeals

Busca de decisões da CGU e da CMRI

Busque aqui as decisões da Controladoria-Geral da União e da Comissão Mista de Reavaliação de Info aos recursos apresentados a essas instâncias no âmbito da Lei de Acesso à Informação. A busca pode somente o campo textual ou combinando o texto com os demais filtros. Exemplo: para consultar decisõe das instâncias, digite o texto desejado no campo "Pesquisa" e selecione a opção desejada no filtro "Inst campo de pesquisa. A busca pode ainda ser refinada utilizando a opcão "Busca Avancada".



4 08198029455202191 CGU

o qual limita-se saber quais informações foram removidas em virtude da vidência da LGPD ... foram retiradas do site em decorrência da LGPD, afastando a compreensão de que se trata de ...

buscaprecedentes.cgu.gov.br/busca/.../08198029455202191 CGU.pdf

Scraper 2 steps:

Almost the same, but with pdfs like this



LEI DE ACESSO À INFORMAÇÃO - RECURSO SUBMETIDO À CGU

PARECER Nº 1459/2021/CGRAI/OGU/CGU

Número do processo:	08198.029455/2021-91		
Órgão:	Ministério da Justiça e Segurança Pública - MJSP		
Assunto:	Recurso contra negativa a pedido de acesso à informação.		
Data do Recurso à CGU:	19/10/2021		
Restrição de acesso no recurso à CGU (Fala.BR):	Não.		
Requerente:	Não identificado.		
Opinião técnica:	Opina-se pelo não conhecimento do recurso devido à inexistência da informação, nos termos da Súmula CMRI nº 6/2015.		

How? **Tika** to read and parse + **Regex** to get only the parts that I need

```
1: # step2: loop through those links, open the pdfs, read them, get the part of the text where the final decision...
   # ...is written and then save it into a list.
   # count the decision number
   list of contents = []
   request n = 0
   # loop through every link, get the text from PDF using Tika, use regex to extract only the parts that I need
   # and then only print those parts
   for a in list of links[:20]:
       request n = request n + 1
       url = a
       pdfFile = parser.from file(url)
       final = (pdfFile["content"])
       final = final.replace('\n', '')
       somente decisao = r"\w\s\w\s\w\s\w\s\w\s\w\s\w\s((.|\n)*)"
       decisao = re.findall(somente decisao, final)
       try:
           list of contents.append(decisao[0])
       except:
           print(f'It did not work for request number {request n}' )
         somente inicial = r"Inicial\:(.*\n.*\n.*\n.*\n)"
         pedido inicial = re.findall(somente inicial,textooo)
         pedido inicial[0]
         print(f'-----Este foi o pedido inicial: ----- {pedido inicial[0]}-----')
```

Without a scraper I would have to read 225 PDFs and find the parts that I want to read manually...

```
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/08198029455202191 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137009925202111 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137010897202177 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137022808202054 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137003633202167 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137008313202101 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.qov.br/busca/dados/Precedente/00137004027202169 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137006543202128 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137010097202156 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.qov.br/busca/dados/Precedente/50001004764202175 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/60143000917202144 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/08198012100202163 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137003711202123 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137003719202190 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/01217006788202128 CGU.pdf#search=ALL%281qpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/53005004814202172 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/72020002003202011 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/25072015288202056 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/25072007041202147 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137012290202121 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/25072006338202195 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/03005106317202130 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/18840001129202061_CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/23480013430202093 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/25072013295202102 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/03005100806202188 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/23480024724201915 CGU.pdf#search=ALL%281qpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/23546048245202199 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/03005081868202183 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/23480002978202016 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/60143000705202167 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/53005000940202158 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/99928000062202089 CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/00137012258202146_CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/00137011518202166 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/23480023751201962 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/60000001146202137_CGU.pdf#search=ALL%281gpd%29
http://buscaprecedentes.cqu.qov.br/busca/dados/Precedente/18870003450202022 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/03005086770202112 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/03005086771202167 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cqu.gov.br/busca/dados/Precedente/6000001304202078 CGU.pdf#search=ALL%28lqpd%29
http://buscaprecedentes.cgu.gov.br/busca/dados/Precedente/25072016836202146 CGU.pdf#search=ALL%28lgpd%29
http://buscaprecedentes.cqu.qov.br/busca/dados/Precedente/53005000988202166 CMRI.pdf#search=ALL%281qpd%29
```

Now it is just one CSV

1	А	В	С	D	Е	F
		num_pedido	orgao	opiniao_tecn	pedido_e_re:	decisao_cgu
	0	08198.02945	MinistV@rio	Opina-se pel	Solicita lista	No exercício das
	1	00137.00992	Gabinete de	Opina-se: a	o requerente	No exercício das
	2	00137.01089	Gabinete de	Opina-se: a	o requerente	No exercício das
	3	00137.02280	Gabinete de	Opina-se pe	O requerent	No exercício da
	4	00137.00363	GSI-PR ,Äì G	Opina-se pe	Solicita-se o	No exercício das
	5	00137.00831	Gabinete de	Opina-se pe	Solicita-se o	No exercício das
	6	00137.00402	Gabinete de	Opina-se pe	O cidadV£o	No exercício da
	7	00137.00654	Gabinete de	Opina-se pe	Solicitou ace	No exercício das
1	8	00137.01009	Gabinete de	Opina-se pe	Solicitou os	No exercício das
	9		proces	so-50001004	764202175	CONTROLADORIA-GI
	10	60143.00091	CEX ,Äì Coma	Opina-se pe	O requerent	No exercício das
	11	08198.01210	MinistV@rio	Opina-se pe	solicitou info	No exercício das
	12	00137.0037	Gabinete de	Opina-se pe	O cidadV£o	No exercício da
	13	00137.0037	Gabinete de	Opina-se pe	O cidadV£o	No exercício da
;	14	01217.00678	MinistV@rio	Opina-se pel	Solicita lista	No exercício das
	15	53005.00481	Empresa Bra	Opina-se pe	Solicita aces	No exercício das
	16	72020.00200	MinistV@rio	Opina-se pe	Solicitou ace	No exercício das
	17	25072.01528	MinistV@rio	Opina-se pe	Solicitou-se	No exercício das
1	18	25072.00704	ANVISA,Äì A	Opina-se pe	Solicita-se o	No exercício das
	19	00137.01229	Gabinete de	Opina-se pe	Solicitou os	No exercício das
	20	25072.00633	MinistV@rio	Opina-se pe	a requerente	No exercício das
	21	03005.10631	MinistV@rio	Opina-se pe	Solicita aces	No exercício das
	22	18840.00112	Caixa Econô	Opina-se pe	O requerent	Parecer - Recurso
	23		23	480.013430/2	2020-93 CGL	J Controladoria-Gera
;	24	25072.01329	MinistV@rio	Opina-se pe	Questionou	No exercício das
	25	03005.10080	MinistV@rio	Opina-se pe	Solicita rela	Parecer - Recurso
	26		23	480.024724/2	2019-15 CGU	J Controladoria-Gera
1	27	23546.04824	Universidade	Opina-se pe	O requerent	No exercício das
1	28	03005.08186	MinistV@rio	Opina-se pe	Solicita-se o	No exercício das
	20		າລ	180 003078/	2020-16 CGI	Controladoria-Gar

Everything is available on Github (and I have already received one nice feedback and improvement to the original code)

https://github.com/luizftoledo/FOIA-scrapers-in-Brazil

