15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

CLASSIFICADOR PARA IDENTIFICAÇÃO AUTOMATIZADA DE DADOS PESSOAIS NO CONTEXTO DA LGPD

NANCY M. YUZAWA[1], ANDREIWID S. CORREA [2]

[1] Graduanda em Tecnologia de Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Campus Campinas, nancy.m@aluno.ifsp.edu.br.
[2] Docente, IFSP, Campus Campinas, andreiwid@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: Com a vigência da Lei Geral de Proteção de Dados (LGPD) desde 2020, tornou-se necessário implementar mecanismos para o adequado tratamento de dados pessoais e sensíveis. No entanto, muitos desses dados tratados anteriormente à lei permanecem disponíveis e de forma inadequada na internet. Diante disso, este projeto tem como objetivo identificar a divulgação de dados pessoais de forma automatizada e de acordo com a LGPD, tendo como o estudo de caso as páginas do site de uma prefeitura. Por meio do uso de *web crawling* (navegação automatizada em sites) e *scraping* (extração de informações de páginas *web*), foi possível identificar e verificar páginas da web com a divulgação de dados pessoais e sensíveis. Na primeira rodada obteve-se 87% de "falsos positivos", ou seja, apesar de remeterem a potenciais dados pessoais e sensíveis, essas páginas não divulgavam ou solicitavam dados de abrangência da LGPD, o que foi importante para o refinamento do algoritmo. Assim, na segunda rodada, com foco nas páginas HTML e PDF, apesar de constatado um aumento na porcentagem de "falsos positivos" em HTML, as páginas com PDFs obtiveram resultado melhor, perfazendo cerca de 48% de "falsos positivos".

PALAVRAS-CHAVE: algoritmo; crawler; scraping; html; pdf

CLASSIFIER FOR AUTOMATED IDENTIFICATION OF PERSONAL DATA IN THE CONTEXT OF THE LGPD

ABSTRACT: With the General Data Protection Law (LGPD) in effect since 2020, it has become necessary to implement mechanisms for the proper handling of personal and sensitive data. However, many of these data processed prior to the law remain available and inadequately displayed on the internet. In light of this, the objective of this project is to identify the disclosure of personal data in an automated manner and in accordance with the LGPD, with the case study focusing on the pages of a municipal website. Through the use of web crawling (automated browsing of websites) and scraping (extraction of information from web pages), it was possible to identify and verify web pages that disclosed personal and sensitive data. In the first round, 87% of the results were "false positives," meaning that although they pointed to potential personal and sensitive data, these pages did not disclose or request data covered by the LGPD, which was important for refining the algorithm. Thus, in the second round, focusing on HTML and PDF pages, despite an increase in the percentage of "false positives" in HTML, the pages with PDFs showed better results, with around 48% of "false positives".

KEYWORDS: algorithm; crawler; scraping; html; pdf

INTRODUÇÃO

A introdução da LGPD foi um grande marco histórico na regulamentação que visa garantir o direito à proteção de dados pessoais, tanto em meio físico quanto nos meios digitais (Brasília, 2024). É importante observar que o conceito de dado pessoal sensível já estava presente na Lei 12.527/2011 (Lei de Acesso à Informação - LAI) e foi mantido pela LGPD (Minas Gerais, 2024). Sua implementação, Lei Federal nº 13.709/2018, entrou em vigor em setembro/2020 demandando esforços para essa nova adaptação dos sistemas de informação e dos processos institucionais (Santa Catarina, 2024).

Apesar da adoção da lei vigente e tendo em vista o legado de páginas da web, sobretudo de prefeituras, ainda há grande potencial da existência de páginas contendo dados pessoais divulgados ou solicitados por meio de formulários e/ou outros meios.

Considerando o problema, este projeto tem por objetivo desenvolver o processo de identificação dos dados pessoais para o adequado tratamento de acordo com a legislação, utilizando classificação automatizada a partir de conteúdo da web.

MATERIAL E MÉTODOS

Neste projeto foi utilizado a linguagem Python e suas bibliotecas para a realização da análise e manipulação das páginas web contendo as potenciais fontes de dados pessoais divulgados ou solicitados.

Primeiramente, definiu-se a base de palavras-chave que auxiliaram o algoritmo na busca dos dados pessoais e sensíveis. Foi utilizado o Censo Demográfico de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE) que listou as principais palavras-chave por segmento, por exemplo a convicção religiosa, as etnias, entre outros. Além disso, para expandir a base de palavras-chave e aumentar as chances de sucesso na busca, foram utilizados outros meios, tais como os sites da prefeitura, jornalísticos, jurídicos, Portal do Governo Federal, site do TSE e a própria LGPD.

Visando organizar a análise das páginas, foi definida a separação do código-fonte de acordo com o tipo de página: HTML, PDF, Google Forms, CSV, RTF, DOC, DOCX, XLS e XLSX. Portanto, foi montado um filtro que tinha como entrada URLs previamente salvas em um arquivo de texto, verificando-se as últimas quatro letras, como mostra a Figura 1. Posteriormente, produziu-se um novo arquivo texto de acordo com o tipo de página.

```
while line:
   line = fp.readline()
   cnt += 1
   a=len(line)
   print("Line {}: {}".format(cnt, line.strip()))
   if(a!=0):
       e=line[a-4]
       b=line[a-3]
       c=line[a-2]
       d=line[a-1]
       print("{}\t{}\t{}\t{}".format(e,b, c,d))
```

FIGURA 1. Código em Python para mostrar as últimas letras das URLs de entrada.

Em cada tipo de página, foi realizado o *scraping* para identificar de forma automatizada os dados pessoais com base nas palavras-chave definidas.

Para as páginas HTML utilizou-se as bibliotecas e funcionalidades da linguagem de programação Python, com destaque para as bibliotecas BeautifulSoup e RegularExpression, para a extração de elementos textuais e a realização de buscas das palavras-chave (Beautiful-Soup, 2024).

Nas páginas com formulários Google Forms, primeiramente, foi analisado e encontrado os padrões comumente usados, que foram:

- http://goo.gl/forms/;
- https://docs.google.com/forms;
- forms.gle.

Posteriormente, foi implementado um algoritmo para a extração das *tags* <a>, que são usadas em páginas web para criar links, buscando os endereços nos padrões mencionados anteriormente.

Referente aos demais tipos de páginas, com destaque aos documentos em PDF, foi usada a biblioteca Tika para a extração de todo o texto do arquivo (Dias, 2022).

Com a mesma biblioteca Tika foi possível verificar os arquivos em DOC e RTF, apenas com a alteração do nome e do tipo de arquivo. A respeito do DOCX, a verificação foi através da biblioteca Docx.

Para as páginas contendo XLS, XLSX e CSV, realizou-se as buscas com auxílio da biblioteca Pandas (IOtools 2024). Em virtude de serem trabalhadas em formato tabular, possuindo linhas e colunas, foi necessário percorrer cada célula para buscar as informações e alocar os dados textuais em uma variável do mesmo tipo. Para tal, usou-se um método similar às demais, como mostra a Figura 2 a seguir.

```
raw_excel = pd.read_excel("output_xlsx.xlsx", index_col=None,
header=None)

raw_excel = pd.read_excel("output_xls.xls", index_col=None, header=None)

df = pd.read_csv('output_csv.csv', encoding='UTF-8', engine='python',
error_bad_lines=False, sep=',', chunksize=100000)
```

FIGURA 2. Trechos do algoritmo que busca nos arquivos xlsx, xls e csv.

Sendo que para todos os tipos, após a extração do texto, foi criado um mecanismo para registrar a pontuação, salvando-as em uma planilha. Isso permitiu dar mais atenção às palavras que apareciam com maior frequência, anotando tanto a quantidade de vezes que cada palavra(s)-chave aparecia(m) quanto a sua frequência, para posteriormente verificar o algoritmo.

Após a geração da planilha com os registros das pontuações e a identificação das primeiras páginas pelo código em HTML, foi realizada uma amostragem para verificação manual devido ao grande número de URLs obtidas, a fim de averiguar se havia alguma divulgação ou solicitação de dados pessoais ou sensíveis.

RESULTADOS E DISCUSSÃO

A primeira rodada contava com 1.315 páginas HTML, porém, após a análise, constatou-se que 917 delas eram da Biblioteca Jurídica (repositório de leis do município), que não contém informações de caráter pessoal por se tratar de legislação. Portanto, as 398 páginas restantes foram analisadas manualmente e classificadas em:

- positivo: páginas que possuíam dados pessoais ou sensíveis sendo divulgados ou solicitados;
- falso positivo: páginas que, apesar de gerarem resultados pelo algoritmo (ocorrência de palavras-chave, por exemplo), em sua estrutura não divulgam ou solicitam dados de cunho pessoal;
- negativo: páginas que não possuem nenhum dado pessoal ou sensível;

- falso negativo: páginas em que o algoritmo não apontou a ocorrência de nenhum dado pessoal, porém possuem dados pessoais;
- erros: páginas que apresentaram erros em sua utilização (por exemplo, sites com redirecionamento, para outra página, mal-sucedido).

Nenhum falso negativo foi constatado, sugerindo que a maioria ou possivelmente todas as palavras-chave foram apontadas. No entanto, 347 resultaram em falsos positivos, que representam 87% da amostra, como é possível ver na Figura 3. Isso ocorreu devido às informações contidas no cabeçalho e no rodapé das páginas web, apresentando os próprios dados da prefeitura como endereço e telefone, mas que não eram nenhum dado pessoal dos cidadãos.

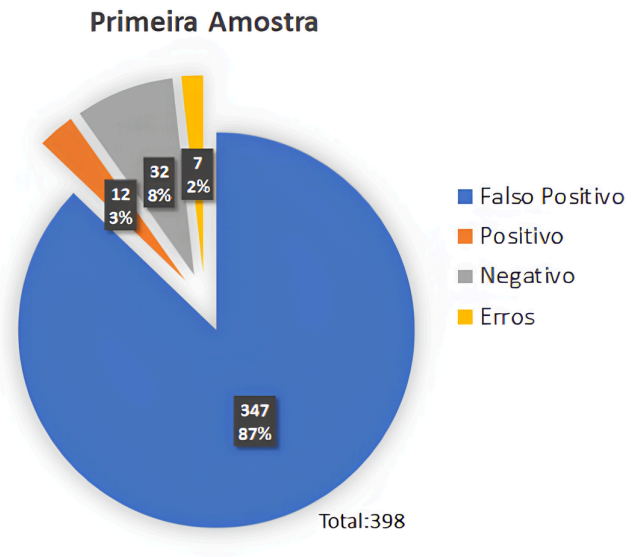

FIGURA 3. Gráfico do resultado da primeira amostra.

Com o término da primeira rodada submetida, realizou-se melhorias no código. Dentre elas, a desconsideração da Biblioteca Jurídica por conta que não apresentavam dados pessoais. Além disso, foram descartados a análise de sites com conteúdo dedicado a contratos, à exposição de imagens e/ou vídeos e os cabeçalhos e rodapés presentes na estrutura do site.

Após o refinamento, foi utilizado o mecanismo de registros de pontuações em 66.000 URLs de conteúdo HTML e 22.000 de PDFs aproximadamente. Posteriormente, adotou-se a subdivisão das páginas de acordo com o tema abordado (saúde, turismo, educação, governo, entre outros tópicos), para que possibilitasse ser checado manualmente uma quantia de amostragem e identificado possíveis melhorias de acordo com a temática, além de mostrar à prefeitura as áreas que deveriam concentrar seus esforços.

Com a checagem manual e a categorização, as classificações foram realizadas de acordo com os critérios estabelecidos anteriormente (positivo, falso positivo, negativo, falso negativo e erro). No entanto, algumas modificações foram feitas: apenas as classificações de falso positivo e erro foram mantidas, e a categoria positiva foi dividida em três novas possibilidades:

- divulgação: página/arquivo que apresenta em sua estrutura divulgação de dados pessoais e/ou dados pessoais sensíveis;
- solicitação: página/arquivo que apresenta em sua estrutura formulários ou algum outro meio para coleta de dados pessoais e/ou dados pessoais sensíveis;
- ambos: página/arquivo que apresenta em sua estrutura tanto divulgação quanto solicitação de dados pessoais e/ou dados pessoais sensíveis

E os resultados obtidos da checagem manual são apresentados nas Tabela 1 e Tabela 2.

TABELA 1. Resultado da amostra das páginas HTML.

Classificação	Total (páginas)	Porcentagem (%)
Divulgação	17	0,68
Solicitação	18	0,72
Ambos	0	0,00
Falso Positivo	2.396	96,42
Erro	54	2,17
Amostragem	2.485	100,00

TABELA 2. Resultado da amostra das páginas PDF.

Classificação	Total (páginas)	Porcentagem (%)
Divulgação	14	1,93
Solicitação	359	49,65
Ambos	0	0,00
Falso Positivo	348	48,13
Erro	0	0,00
Amostragem	723	100,00

Depois da segunda rodada de testes, verificou-se que a ocorrência de falsos positivos foi maior que o teste anterior, de 87% para 96,42% nas páginas HTMLs. Ademais, as amostras das páginas PDFs também demonstraram uma grande ocorrência de falsos positivos, com 48,13%.

CONCLUSÕES

Analisando os dados das duas rodadas, verificou que, embora a taxa de falsos positivos tenha aumentado significativamente na comparação entre a primeira e a segunda rodada das páginas HTML. E em respeito aos arquivos em PDF, também constatou-se uma alta porcentagem de falsos positivos, com 48,13%. Essa análise das subdivisões, permitiu ter melhor entendimento do comportamento da quantidade de dados pessoais em relação à temática. Dessa forma, é possível desconsiderar algumas delas que não contêm informações de caráter pessoal, como ocorreu com as bibliotecas jurídicas.

Ao comparar PDFs e HTMLs, da segunda rodada, notou-se uma quantidade significativamente maior de dados divulgados e solicitados no formato PDF, com 51,58%, em comparação a 1,4% no formato HTML. Através desse resultado, essa análise evidencia que o formato PDF possui a maior concentração de dados pessoais.

Assim, conclui-se que os resultados obtidos indicam perspectivas para a melhoria do algoritmo, visando uma maior acurácia na identificação de dados pessoais e sensíveis a partir das páginas em HTML e PDF. Como mencionado anteriormente, um dos pontos de melhoria é a desconsideração das páginas (URLs) que embora contenham dados, não incluem dados pessoais, como o tema relacionado aos arquivos (por exemplo, arquivos municipais). Além disso, é importante concentrar-se nas páginas PDFs por conterem maior quantidade de dados pessoais.

CONTRIBUIÇÕES DOS AUTORES

Nancy M. Yuzawa: contribuiu com a análise de dados, desenvolvimento, implementação, teste de software e com a redação do trabalho.

Andreiwid S. Correa: contribuiu com a curadoria, planejamento do projeto e correção do trabalho.

Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Agradeço ao IFSP e ao orientador pela oportunidade, além do acompanhamento e da orientação ao longo do projeto realizado.

REFERÊNCIAS

BEAUTIFUL-SOUP: How do I use regular expressions with Beautiful Soup?. How do I use regular expressions with Beautiful Soup?. Disponível em: https://webscraping.ai/faq/beautiful-soup/how-do-i-use-regular-expressions-with-beautiful-soup. Acesso em: 15 ago. 2024.

BRASÍLIA. STJ. . **LGPD**: lgpd: um marco na regulamentação sobre dados pessoais no brasil. LGPD: Um marco na regulamentação sobre dados pessoais no Brasil. Disponível em: https://www.stj.jus.br/sites/portalp/Leis-e-normas/lei-geral-de-protecao-de-dados-pessoais-lgpd#:~:text=A%20Lei%20Geral%20de%20Proteção,vigor%20em%20setembro%20de%202020. Acesso em: 03 ago. 2024.

DIAS, Tiago. **Extraindo Texto de Arquivos PDF com Python**: biblioteca python tika. Biblioteca Python Tika. 2022. Disponível em: https://dadosaocubo.com/extraindo-texto-de-arquivos-pdf-com-python/. Acesso em: 15 ago. 2024.

IOTOOLS (text, CSV, HDF5, ...): Reading Excel files. Reading Excel files. Disponível em: https://pandas.pydata.org/docs/user_guide/io.html. Acesso em: 15 ago. 2024.

MINAS GERAIS. SEFA. . **LGPD LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS**: tipologia dos dados pessoais. Tipologia dos dados Pessoais. Disponível em: https://www.fazenda.mg.gov.br/transparencia/lgpd/LGPD-SEF-Tipologia-dos-Dados.pdf. Acesso em: 03 ago. 2024.

SANTA CATARINA. UFSC. . **Proteção de Dados Pessoais - UFSC**: o que é a lei geral de proteção de dados?. O que é a Lei Geral de Proteção de Dados?. Disponível em: https://lgpd.ufsc.br/duvidas-frequentes/#:~:text=A%20Lei%20Geral%20de%20Proteção%20de%20Dados%20Pessoais-%20LGPD%20(Lei,da%20personalidade%20de%20cada%20indivíduo.. Acesso em: 15 ago. 2024.