UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Raylison Nunes dos Santos

Uma Solução para Verificação do Uso de Dados Pessoais em Formulários Web

Recife

2022

Raylison Nunes dos Santos

Uma Solução para Verificação do Uso de Dados Pessoais em Formulários Web

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Fernando Antonio Aires Lins

Recife

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

N972s Santos, Raylison Nunes dos
Uma solução para verificação do uso de dados pessoais em formulários Web / Raylison Nunes dos Santos. - 2022.
47 f. : il.

Orientador: Fernando Antonio Aires Lins.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, , Recife, 2022.

1. Sistemas Web. 2. Privacidade. 3. Crawler. 4. LGPD. 5. Proteção de Dados. I. Lins, Fernando Antonio Aires, orient. II. Título

CDD

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Raylison Nunes dos Santos às 15 horas do dia 26 de maio de 2022, no link https://meet.google.com/gte-ibtf-rja, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado "Uma Solução para Verificação do Uso de Dados Pessoais em Formulários Web", orientado por Fernando Antonio Aires Lins e aprovado pela seguinte banca examinadora:

Fernando Antonio Aires Lins
DC/UFRPE

George Augusto Valença Santos
DC/UFRPE

Agradecimentos

Agradeço primeiramente aos meus pais, em especial a minha mãe Risolania, que se dedicou tanto pela minha educação. Agradeço também aos meus professores que ao logo de meu curso contribuíram na minha formação acadêmica, em especial ao meu orientador, Fernando Aires, pela paciência, conhecimento e empenho dedicado à elaboração deste trabalho. Assim como, aos meus colegas de curso, pelo companheirismo durante toda a graduação. Por fim minha esposa Mariana, que esteve sempre ao meu lado me apoiando e incentivando a nunca desistir.

“As oportunidades multiplicam-se à medida que são agarradas.”
(Sun Tzu)

Resumo

Com o crescimento e disseminação da Web e das redes sociais, surgiu a necessidade de se conhecer melhor os usuários, para assim, conseguir os atrair com as melhores estratégias possíveis. Existe um esforço considerável por parte das empresas para conseguir dados que possam contribuir para essa base de conhecimento. Pensando em proteger os dados pessoais das pessoas, o governo propôs e aprovou a Lei Geral de Proteção de Dados (LGPD), que visa proteger os dados pessoais dos cidadãos. Contudo, mesmo com a citada lei, empresas vêm pedindo continuamente dados pessoais protegidos pela LGPD, sem atentar para os princípios descritos na citada lei. Neste contexto, este trabalho tem como objetivo a proposição de uma solução de verificação de formulários Web, que analisa os dados pessoais requisitados. Esta solução apresenta o desenvolvimento de um Crawler específico que extrai e categoriza os dados, além de disponibilizar os dados pessoais encontrados para análise. Esta solução avaliou 50 sites de diferentes segmentos, e os resultados apresentados indicam não apenas a preocupante quantidade de dados pessoais comumente requisitados em formulários simples, mas também quais os dados mais pedidos.

Palavras-chave: Sistemas Web, Privacidade, Crawler, LGPD, Proteção de Dados.

Abstract

With the growth and dissemination of the Web and social networks, the need arose to get to know users better in order to attract them with the best possible strategies. There is considerable effort on the part of companies to obtain data that can contribute to this knowledge base. Thinking about protecting people's personal data, the government proposed and passed the General Data Protection Law (LGPD), which aims to protect citizens' personal data. However, even with the aforementioned law, companies have been continuously asking for personal data protected by the LGPD, without paying attention to the principles described in the aforementioned law. In this context, this work aims to propose a solution for verification of Web forms, which analyzes the requested personal data. This solution features the development of a specific Crawler that extracts and categorizes the data, in addition to making the personal data found for analysis available. This solution evaluated 50 websites from different segments, and the results presented indicate not only the worrying amount of personal data commonly requested in simple forms, but also which data are most requested.

Keywords: Web Systems, Privacy, Crawler, LGPD, Data Protection.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CCBB	Centro Cultural Banco do Brasil
LGPD	Lei Geral de Proteção de Dados
CSS	Cascading Style Sheets
HTML	HyperText Markup Language
BD	Base de dados
SGBD	Sistemas de gestão de bases de dados
RGDP	Regulamento Geral a respeito da Proteção de Dados
CCPA	California Consumer Privacy Act
PIPEDA	Personal Information Pro-tection and Electronic Documents Act
DLT	Distributed Ledger Technologies
GDPR	General Data Protection Regulation
XSS	Cross-site Scripting
RFI	Remote File Inclusion
API	Interface de Programação de Aplicações
GCP	Google Cloud Platform
ORM	Object Relational Mapping

Sumário

1 Introdução

Ao longo dos anos a demanda por sistemas Web tem crescido cada vez mais (NASCIMENTO, 2020). Empresas tentam atrair seus clientes para consumirem seus produtos ou serviços através de conteúdo personalizado e análise de perfis (SILVEIRA et al., 2016). Com isso é iniciada uma busca por dados que possam ser mapeados para se conseguir criar uma publicidade direcionada e exclusiva, assim como o desenvolvimento de produtos e serviços. Tais atividades fazem com que a coleta de dados pessoais seja realizada de maneira muitas vezes exagerada, gerando um desconforto nas pessoas em compartilhar suas informações. Exemplo disso pode ser visto em (GLOBO, 2021) quando uma professora realizou uma denúncia informando que antes da visita a exposições no CCBB (Centro Cultural Banco do Brasil), as pessoas precisam fornecer a uma empresa terceirizada os dados pessoais. Esta situação inclusive levou o Ministério Público a instaurar um inquérito para apurar se o CCBB condiciona a entrada de visitantes a um cadastramento.

Este exemplo de situação acontece comumente ao se realizar uma compra e ser imediatamente abordada pelo atendente solicitando o CPF para se obter um desconto, ou até fazer o cadastro no aplicativo da empresa com promessas de promoções (SILVEIRA et al., 2016). Até recursos de geolocalização são usados pelas empresas para produzir marketing digital como descrito em (MEDIA, 2020). Isso mostra que não apenas os dados, mas nossos hábitos, estão sendo mapeados pelas corporações.

Em 2015, (SILVA; LUCIANO; MAGNAGNAGNO, 2015) apresentou uma pesquisa exploratória voltada a analisar o grau de preocupação das pessoas com seus dados pessoais espalhados na rede, feita com dados de 5 regiões brasileiras e um total de 1104 questionários respondidos. Nos quais mostraram muita preocupação com parte da população brasileira com como estão usando suas informações. Reforçando nosso interesse em conseguir levantar estatísticas sobre o uso dos dados solicitados por essas empresas.

Desta forma, fica evidente a necessidade de proteção dos dados compartilhados pelas pessoas. Com essa preocupação, em agosto de 2018, a lei n° 13.709 foi aprovada, entrando em vigência em agosto de 2020. A Lei Geral de Proteção de Dados (LGPD), como é chamada, dispõe sobre o tratamento de dados, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público, ou privado. Ela protege os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural (BRASIL, 2018). Esta lei visa criar um cenário de segurança jurídica, com a padronização de normas e práticas, para promover a proteção,

de forma igualitária e dentro do país e no mundo, aos dados pessoais de todo cidadão que esteja no Brasil (RIBEIRO, 2014).

Para contribuir com essa iniciativa, é importante existir uma ferramenta de acompanhamento para formulários digitais que possa validá-los e monitorá-los, observando assim, o que estão solicitando e se de fato aquela informação é necessária.

1.1 Justificativa e Motivação

Como dito, muitas empresas registram nas suas bases de dados variadas categorias de dados preenchidos em seus sistemas. Informações solicitadas em seus formulários sem possuírem justificativas para sua solicitação. Se faz necessário acompanhar quais informações as empresas andam requisitando no dia a dia das pessoas, e entender quais categorias de dados elas procuram.

Tendo em vista tecnologias usadas para mineração de dados e rastreamento Web, como Crawlers (SINGH; VARNICA, 2014), à criação de uma solução que consiga extrair de maneira automática as entradas dos formulários, agrupá-las e categorizá-las pode contribuir para disponibilização de uma plataforma de acompanhamento e monitoramento das informações presentes em formulários Web. Além de conseguir montar uma base de dados com informações sobres as categorias de dados mais usados pelas empresas para conhecer seus usuários.

O presente trabalho tem por motivação realizar as validações de dados requisitadas pelos formulários eletrônicos para cadastros em sites, considerando a aprovação da lei geral de proteção de dados e seus critérios quanto às categorias de dados. Desta maneira, esperamos contribuir com o tema desenvolvendo uma ferramenta que possa extrair e analisar esses formulários e assim levantar estatísticas, apontando que dados estão sendo requisitados, validando assim as informações que estão sendo pedidas.

1.2 Objetivos

Nesta seção estão expostos os objetivos propostos para este trabalho.

1.2.1 Objetivo Geral:

O presente trabalho tem por finalidade a construção de uma ferramenta de análise que possa realizar extração automatizada de dados presentes em formulários eletrônicos de cadastros e agrupá-los por tipos. Desta forma, será possível analisar que dados estão sendo pedidos e verificar se existem dados protegidos pela LGPD.

1.2.2 Objetivos Específicos:

1. Analisar os principais dados solicitados em formulários WEB
2. Criação de uma base de palavras categorizadas a partir das definições da LGPD no que diz respeito as categorias de dados.
3. Analisar a eficácia do crawler em identificar as entradas contidas nos formulários.
4. Especificar os detalhes de construção de uma plataforma para extração e análise de dados presentes em fomulários WEB.

1.3 Estruturação do trabalho

O restante do documento está estruturado da forma que se segue. O Capítulo 2 aborda e explica os conceitos da LGPD, sua origem e o seu papel na utilização de dados compartilhados. No Capítulo 3 são encontrados os trabalhos relacionados à proteção de dados. O Capítulo 4 explica a ferramenta proposta para extração de formulários. No Capítulo 5 são encontrados os resultados obtidos por este trabalho. Finalmente o Capítulo 6 contém as conclusões, discussões e trabalhos futuros a partir dos resultados obtidos neste trabalho.

2 Fundamentação Teórica

Nesta seção serão apresentados os conceitos básicos necessários para o entendimento deste trabalho.

2.1 Web Crawler

Um, Web Crawler é um robô que possui funções como: procurar, coletar, classificar, organizar e disponibilizar dados. Esta ferramenta pode ser usada para busca em sites, documentos ou até mesmo banco de dados. São semelhantes aos motores de busca encontrados na Web como Google, o Yahoo, o Bing onde ele identifica links, tags ou palavras chaves para com isso coletar as informações. Web Crawlers também são personalizáveis, construídos com propósitos únicos e específicos; isto facilita a automatização na mineração de dados e busca de informações. Sendo assim, uma ferramenta prática para conseguir extrair informações de sites e estruturá-las em uma base de dados.

2.1.1 Categorias de Crawler

Existem muitos rastreadores (Crawlers) desenvolvidos que possuem arquiteturas e propósitos distintos, como descrito em (SINGH; VARNICA, 2014). Por exemplo, o DeepBot (ÁLVAREZ et al., 2007) tem como propósito a extração de conteúdo oculto, informações geradas dinamicamente por tecnologias como Javascript ao interagir com os sites. Ele se utiliza de heurísticas para identificar automaticamente os formulários de consulta relevantes e aprender como executar consultas neles.

Semelhantemente existe o H1We (RAGHAVAN; GARCIA-MOLINA, 2000), mas que se classifica como um Crawler focado, pois, visita sites na web através da busca de palavras chaves pré-fornecidas, e com isso interagir com seus formulários e opções disponíveis, navegando em suas páginas e acessando suas informações ocultas. A cada descoberta ele armazena as páginas e as indexa para aprendizado de novas interações e para conseguir disponibilizar um repositório de cache dessas páginas em aplicações militares ou de inteligência que não terão acesso à internet diretamente. Já outros como detalhado em (WANG, 2010) rastreiam mensagens na plataforma Twitter e aplicam métodos de aprendizagem de máquina para distinguir automaticamente contas de spam de contas normais.

Trabalhos como (WU et al., 2020) mostram um Crawler de propósito geral onde seu foco é a extração de conteúdos estatísticos, acessado diretamente a página e

extraído sua informação apenas identificando seus textos sem realizar interações. O processo foi usado para conseguir informações da loja de jogos online Steam, utilizando a biblioteca Selenium para identificação dos componentes das páginas e do framework Python Scrapy. Ao final do processo foi apresentado um relatório completo sobre vendas de jogos, opiniões de jogadores, médias de elogios, etc.

Em (FISCHER, 2006) observamos uma solução em que foi desenvolvido um Crawler do tipo incremental e distribuído. O Crawler distribuído é caracterizado pela quebra de múltiplos processos para realizar o rastreamento e é geralmente usado em conjunto com alguma outra categoria de Crawler. Já o incremental ele é um rastreador tradicional que substitui sempre os seus resultados a cada interação mantendo a informação extraída constantemente atualizada. A solução de (FISCHER, 2006), rastreou páginas Webs brasileiras (.br), as indexou e armazenou em um banco de dados relacional e disponibilizou uma plataforma de acesso rápido a informação se assemelhando aos motores de busca já mencionados.

2.1.1.1 Desenvolvimento de Crawlers

Como visto, o uso de Crawlers atendem diversas categorias de propósitos definidos pelos seus criadores, e se adéquam aos critérios especificados de cada situação. Será mostrado, a seguir, boas práticas e ferramentas para auxiliar no desenvolvimento de um Web Crawler sendo usadas para o desenvolvimento desta solução.

Para conseguir realizar a extração da informação desejada, deve-se ter uma noção de como o dado é exibido, e com isso conhecer as palavras chaves, tags, links e até mesmo o CSS (Cascading Style Sheets) da página alvo. Com isto, é possível navegar e consultar as páginas e colher a informação. Alguns destes itens são simples de conseguir e outros são mais complexos de identificar, pois, dependem muito de como o sistema Web foi implementado e principalmente como o seu HTML (HyperText Markup Language) é renderizado nos navegadores.

Após a pré análise das páginas alvo, é escolhida a linguagem de programação em que se adapta melhor para implementação e tratamento do dado colhido. Geralmente é implementado em Python por possuir muitas bibliotecas que vão auxiliar no tratamento do dado, mas podem ser usados muitas outras linguagens como Java e Javascript.

Outro ponto importante é possuir um extrator HTML para realizar o download das páginas Web ou um Webdriver, motor que consegue se comportar como um Browser. O extrator HTML é limitado e não possibilita interagir tornado a extração estática. Já o Webdriver conseguirá ser mais dinâmico e reproduzir interação com as páginas, navegar entres links e tags, assim como realizar ações de clique e de escrita. Um dos Webdriver mais usados é o Selênio, uma biblioteca desenvolvida para automatização de testes em páginas Web. Ele é bastante usado para extração de dados e possui

muitas funções como pesquisar URL (Uniform Resource Locator), reproduzir cliques, inserir informações em campos, etc.

2.2 Privacidade na era digital

Com o desenvolvimento das tecnologias, se tornou possível o armazenamento de dados em grande escala. Chamamos Base de dados (BD) essas coleções eletrônicas que armazenam abundantemente informação, organizadas estruturadamente, possibilitando a consulta rápida e facilitada a diversos dados, informações e documentos (BCIJO, 2020).

Este processo de armazenamento é muito usado nos softwares e ele vem desde a década de 60, quando surgiram os primeiros SGBD (sistemas de gestão de bases de dados), tendo como principal função a abstração da responsabilidade de gerir o acesso, persistência e a manipulação dos dados.

Ao olharmos para o conceito de “privacidade” definido como: qualidade do que é privado, do que diz respeito a alguém em particular (DICIO, 2017), e refletirmos sobre como as empresas estão usando essas bases de dados de seus clientes e o que elas sabem sobre nós, esta situação nos faz questionar sobre se temos de fato privacidade nessa era digital.

Continuando, a privacidade digital pode ser mais bem definida como a proteção das informações de cidadãos particulares que usam sistemas digitais (NEITINBAG, 2022). No entanto, quando as pessoas falam sobre privacidade digital, geralmente se referem a ela como, sua relação com o uso da Internet (NEITINBAG, 2022). Contudo, é fundamental, mecanismos e meios para serem garantidos às pessoas, privacidade com sua informação compartilhada.

2.3 Lei geral de proteção de dados.

No objetivo de proteger os dados pessoais, garantir transparência no seu uso e definir regras para utilização dos mesmos, foi aprovada pela Presidência da República em agosto de 2018, a Lei Geral de Proteção de Dados Pessoais, que afeta diferentes setores, serviços e pessoas seja no papel de indivíduo, empresa ou governo.

Entrando em vigência em 2020, a LGPD traz um cenário de segurança jurídica, padronização de normas para o uso de dados de todos os brasileiros estando ou não no país, assim como seus dados estando ou não sendo processados por um servidor no país (SERPRO, 2021). A lei possui fundamentos como privacidade, liberdade de expressão, inviolabilidade e livre iniciativa, defesa do consumidor, direitos humanos, dignidade e exercício da cidadania. Com isso vieram marcos importantes para que as

pessoas e empresas se adaptarem a essa nova realidade como descritos em (RODOTÀ, 2015):

- As entidades devem coletar dados apenas considerados necessários para a realização da tarefa que oferecem ao indivíduo. Dados como orientação sexual, saúde e religião não podem ser usadas para abuso ou discriminação;
- A anonimidade de dados é garantida pela lei, sempre que possível, quando coletados por organizações de pesquisa;
- O usuário que cede seus dados a algum serviço deve ter acesso facilitado ao tratamento deles e à finalidade, bem como saber quem irá manipulá-los.

O não atendimento a LGPD por acarretar punições como:

- Reparação de danos ao indivíduo que se sentir violado e multa de até 2% do lucro da instituição infratora;
- Advertência e bloqueio da informação coletada, podendo gerar a suspensão do acesso ao banco de dados por um período de até seis meses.

Vale salientar a maneira com a qual a LGPD trata as informações compartilhadas pelas pessoas, definindo as categorias de dados em 4 tipos: pessoas, sensíveis, públicos e anonimizados. Abaixo detalharemos cada um deles como explicados em (SERPRO, 2021):

- **Dados Pessoas** contemplam todas as informações que de maneira direta ou indireta conseguem identificar uma pessoa que esteja viva. Sendo assim dados como: nome, RG, CPF, gênero, data e local de nascimento, telefone, endereço residencial, localização via GPS, retrato em fotografia, prontuário de saúde, cartão bancário, renda, histórico de pagamentos, hábitos de consumo, preferências de lazer, endereço de IP (Protocolo da Internet) e cookies, entre outros.
- **Dados Sensíveis** assim como os demais ainda são dados pessoas, mas que demanda uma atenção maior: eles dispõem de informações com origem racial ou étnica, convicções religiosas ou filosóficas, opiniões políticas, filiação sindical, questões genéticas, biométricas e sobre a saúde ou a vida sexual de uma pessoa. Além de informações sobre crianças e adolescentes que são obrigatórias possuírem a autorização dos pais para uso dos mesmos.
- **Dados Públicos** são definidos pela lei (BRASIL, 2018) como “dados pessoais cujo acesso é público”. Eles devem ser tratados conforme sua finalidade que

justificaram sua disponibilização. Empresas podem utilizar esses dados sem pedir consentimento, mas em caso de querer compartilhar esses dados deverão solicitar permissão para os devidos donos.

- **Dados Anonimizados** são dados que antes eram vinculados a uma pessoa, mas que por etapas de desvinculação deixou de ser. Se um dado for anônimo a LGPD não se aplica ao mesmo. Salientando que, um dado só é realmente considerando anônimo se de fato ele não consegue reconstruir o caminho para descobrir sua origem.

2.4 Cenário internacional de legislações sobre proteção de dados pessoais

A LGPD foi inspirada no Regulamento Geral a respeito da Proteção de Dados Europeia (RGDP), legislação da União Europeia que dispõe sobre o tema e demonstra a preocupação quanto ao tratamento dos dados na Europa. Este regulamento foi aprovado no dia 27 de abril de 2016 e estabelece as regras alusivas ao tratamento, por uma pessoa, uma empresa ou uma organização, de dados pessoais relativos a pessoas na União Europeia. Em (CONSUMIDORMODERNO, 2021) encontramos alguns exemplos deste movimento de proteção aos dados em outros países pelo mundo.

O primeiro deles é nos Estados Unidos, que mesmo não possuindo uma regulamentação para todo país, estados como a Califórnia, aprovaram uma legislação própria sobre tratamento de dados. A lei aprovada na Califórnia foi a Califórnia Consumer Privacy Act em 2018, e funciona como um método de defesa ao consumidor. Outro exemplo de legislação é oriundo do Canadá, que possui desde 2000 a PIPEDA (Personal Information Protection and Electronic Documents Act), ou Lei de Proteção de Informações Pessoais e Documentos Eletrônicos em sua legislação nacional. No Japão, a Lei de Proteção de Informações Pessoais, de 2003, teve devido à emenda APPI (Act on the Protection of Personal Information), em 2017 uma ampliação a ponto de a União Europeia considerar o país totalmente adequado quanto à proteção de dados. Por fim, pode-se citar também a Nova Zelândia e a Argentina como países com legislações específicas na área. Na Nova Zelândia, uma Lei de Privacidade de 1993 regula a segurança de dados, e possui uma atualização sendo proposta. Na Argentina a Lei de Proteção de Dados Pessoais, teve aprovação em 2000, e limita o uso dos dados apenas às atividades consentidas pelos cidadãos.

A Figura a seguir exibe o mapa do cenário internacional e o grau de adequação para políticas de tratamento de dados pessoais.

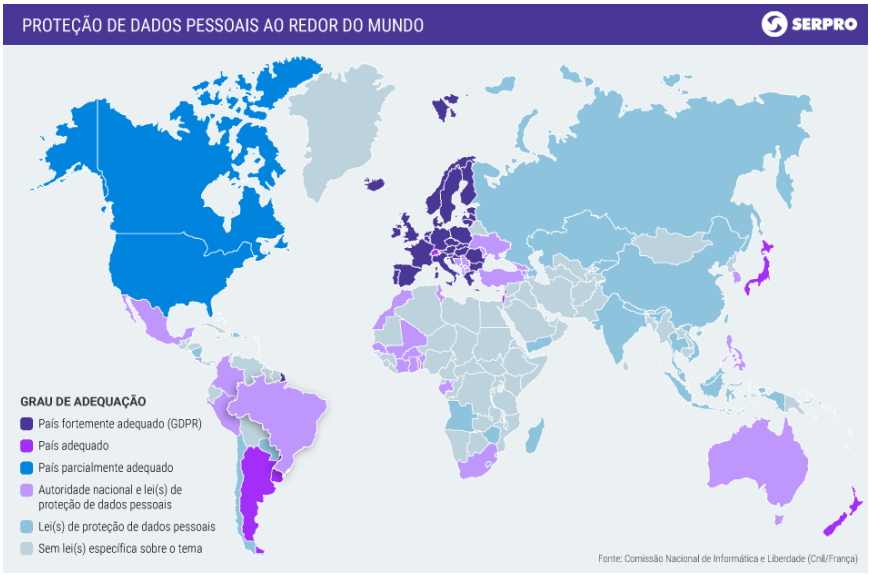


Figura 1 – Proteção de dados pessoais ao redor do mundo. (SERPRO, 2021)

De maneira geral a implementação da LGPD é muito similar a GDPR, tendo diferenças sutis, mas apresentando consequência diferente. Sendo assim a GDPR pode ser considerada mais restrita e detalhada do que a LGPD, apresentando mais restrições no tratamento dos dados e especificando melhor como deve ser o tratamento de dados pelas empresas. Contudo, as lacunas atuais da LGPD quanto ao tratamento dos dados deve ir sendo preenchidas a medida que a lei for evoluído, assim como a GDPR vem evoluído da Europa.

3 Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados com a temática deste projeto. Tendo em vista que a LGPD entrou em vigor recentemente, ao realizar buscas em sites como Google Scholar, IEEExplore e SciELO não foram encontrados trabalhos voltados diretamente para avaliação de dados pessoas em formulário Web. Contudo, foram encontrados trabalhos voltados à proteção de dados pessoais e aplicação da LGPD.

O trabalho apresentado por (CARVALHO, 2021) introduziu uma proposta de implementação de um framework de compliance à LGPD voltado a prevenir fraudes no contexto de big data. A pesquisa foi desenvolvida com o método de design science research usando estudos de caso e técnicas de pesquisa empíricas. Teve seu foco no tratamento de prevenção de fraude no uso dos dados pessoais com o uso de boas práticas de governança de dados e segurança da informação. Contudo, esse manual compilado de boas práticas para a gestão de dados ficou restrito apenas ao uso de empresas públicas e ao setor financeiro, mas se mostrou uma ferramenta interessante para auxiliar as empresas que estão coletando dados a evitarem possíveis fraudes.

(MORTE et al., 2020) propõem a realização de uma análise de DLTs (Distributed Ledger Technologies) para tratamento de dados pessoais. As DLTs são contabilidades distribuídas, e uma muita usada hoje é o blockchain para validação de transações digitais com moedas como o Bitcoin. O trabalho focou em analisar documentos regulatórios e relacionados com a GDPR (General Data Protection Regulation) entre outras fontes e teve como objeto de análise principal a aplicação Datavalid. A Datavalid é uma aplicação desenvolvida pelo Serviço Federal de Processamento de Dados. A análise teve sua conclusão de que o uso de blockchain privadas conseguem atender o tratamento de dados pessoas conciliando com as DLTs. No entanto, em blockchain públicas se obtêm dificuldade em responsabilizar o uso dos dados.

(MACHADO et al., 2016) propõe o desenvolvimento de um Crawler para análise de vulnerabilidades de segurança em páginas Web. Este Crawler conseguia realizar varredura e com isso disponibilizar as informações coletadas para que os administradores dos sistemas analisados pudessem tomar medidas de melhoria. A inspeção ocorreu em 30 sites que fazem parte dos sites mais acessados do Brasil, sendo coletados 591 páginas com problemas como vulnerabilidades de SQL Injection, Cross-site Scripting e Remote File Inclusion.

(SILVA et al., 2021) Desenvolve um Framework para identificação de conformidade na adequação da LGDP, em empresas do setor químico brasileiro. A trabalho apresenta um estudo dos principais desafios encontrados na indústria química para se

adequar a lei além de realizar um estudo em outros Frameworks que poderiam ajudar neste processo. A validação do Framework aconteceu em uma empresa de teste onde ao final do processo efetuaram um levantamento através de formulários de pesquisa, que ao analisados, conseguiram evidencia um deficit na adequação da LGPD na gestão e segurança de dados presentes na corporação.

Estes trabalhos mostram a necessidade de aplicação da LGPD, assim como sua validação nas empresas que solicitam dados pessoais. Também é visto como o uso de ferramentas como Web Crawlers consegue contribuir no auxílio de identificação de dados pessoas presentes formulários eletrônicos.

4 Metodologia

Este capítulo apresenta a metodologia utilizada para desenvolvimento e experimento da solução, de modo a verificar a qualidade da abordagem utilizada no problema proposto: extração e categorização das entradas presentes nos fomulários Web. Existindo assim um ponto a ser avaliado, a eficiência em obter as entradas presentes nos fomulários avaliados.

Para a execução do experimento, foram selecionadas e analisados 50 sites de diversos domínios como: saúde, tecnologia, construção, esportes, etc. Para assim garantir uma maior diversidade de categorias de sites que o Crawler consegue atuar. Após a execução do crawler, os resultados foram comparados com a extração manual destas entradas presentes nos formulários e obtendo assim sua precisão de extração.

4.1 Variáveis

Na avaliação dos dados obtidos, foi utilizado o cálculo de Precisão como métrica de avaliação de desempenho em extrair dados. Além de, definir as variáveis que serviram de entrada para a precisão.

- Presentes — **Total de inputs presentes no formulário.**
- Obtidos — **Total de inputs extraídos nos formulários.**

4.2 Métricas

$$\left(\frac{obtidos}{presentes}\right) * 100 \tag{4.1}$$

Equações 1 – Calculo de Precisão

4.3 Fases

A construção da solução pode ser particionadas nas seguintes etapas:

- **Fase 1** — Estudo e levantamento da arquitetura a ser utilizada.

  Esta etapa dispõem de analisar a melhor maneira de desenvolver a solução, de modo a garantir flexibilidade e escalabilidade para construção dos módulos que vão compor a aplicação.

- **Fase 2** — Desenvolvimento.

  Esta etapa corresponde a implementação das APIs e bancos de dados presentes no Backend, assim como, a implementação do Dashboard no Frontend.

- **Fase 3** — Execução e Analise.

  Esta etapa realiza a execução da solução. Ela contou com o levantamento dos sites a serem avaliados, criação da base de palavras conhecidas e Avaliação de resultado.

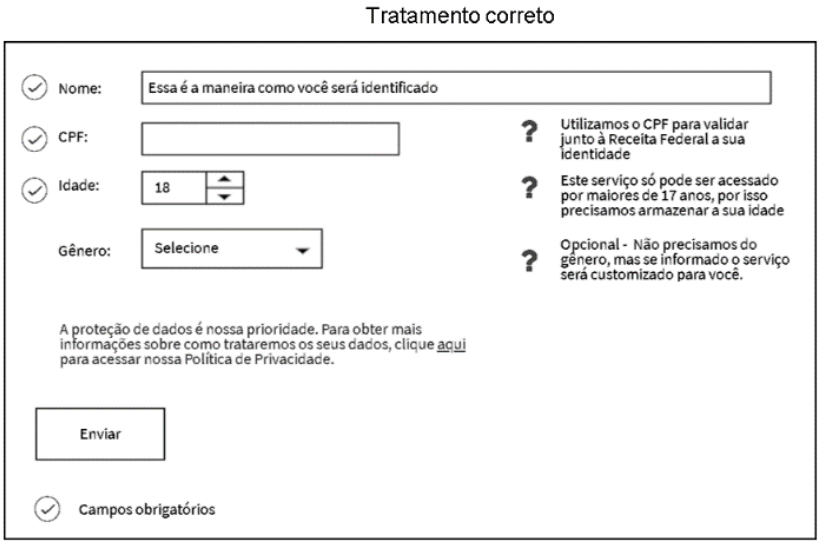
5 LGPD Form Checker: Uma Solução para Verificação do Uso de Dados Pessoais em Formulários Web

Este capítulo apresenta a contribuição principal deste trabalho, o LGPD Form Checker, uma aplicação desenvolvida para verificar o uso de dados pessoais protegidos pela LGPD em Formulários Web. Este capítulo apresenta as diferentes camadas da aplicação e suas respectivas capacidades para o reconhecimento de dados pessoais em formulários Web.

5.1 Visão Geral

Os formulários Web são a porta de entrada para a informação das pessoas em diversos sistemas. Como visto, através deles as empresas conseguem efetuar diversas ações como: validação de informações do usuário, segurança, análise de crédito, pagamento e campanhas de marketing (RESOLVARAPIDO, 2022).

Contudo, cada dado solicitado deve estar justificado à luz dos princípios da Lei Geral de Proteção de Dados. Entretanto, muitas vezes os cadastros são realizados sem uma justificativa clara para cada informação pedida. A Figura 2 mostra um exemplo de um formulário Web nos padrões ideais previstos pela LGPD. Para cada campo presente no formulário, é apresentado de maneira clara o seu propósito.



Fonte: Produção do autor

Figura 2 – Exemplo de Formulário Web conforme LGPD (SERPRO, 2021)

Infelizmente, sabemos que este não é o padrão de apresentação dos formulários Web, em geral. Pensando nisso, foi implementado uma solução que extrai estas entradas, as agrupa em dados pessoais e sensíveis, e assim é montado uma base de dados para análise da informação solicitada aos usuários no momento do cadastro nas plataformas digitais.

5.2 Arquitetura

A Figura 3 mostra a arquitetura da solução desenvolvida neste projeto. A aplicação conta com um Dashboard como Frontend onde estão sendo exibidos todos os gráficos e relatórios desenvolvidos a partir dos resultados tratados. Este dashboard também possibilita cadastrar novos sites para novas análises. Já no Backend da aplicação, se apresenta o desenvolvimento de uma arquitetura baseada em microsserviço com três APIs (Interface de Programação de Aplicações) e dois bancos de dados.

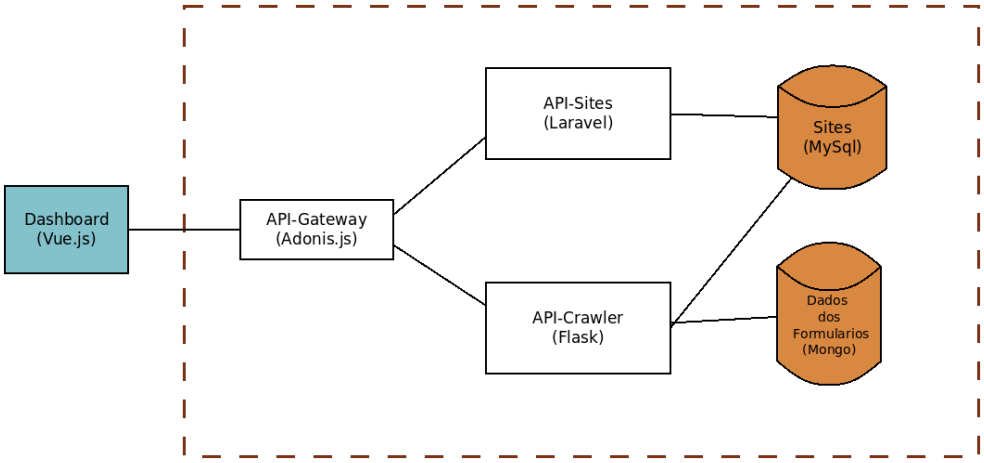


Figura 3 – Arquitetura LGPD Form Checker

A motivação do uso dessa arquitetura é conseguir disponibilizar de forma desacoplada os serviços que compõem a aplicação. Isto acaba tornando possível escalar e adicionar novos recursos para novas funcionalidades. A API de entrada será a API Gateway, onde a mesma efetuará todo o processo de gerenciamento de requisições assim como o controle de autenticação e gerenciamento de usuário. Seguindo temos a API Sites, onde o gerenciamento de todos os dados nas tabelas estruturadas é realizado. Este é o serviço usado para gerar os relatórios e inserir os novos sites que a aplicação irá consumir. Por fim, a API Crawler possui toda a lógica para extração dos formulários e tratamento dos dados.

A persistência dos dados é realizada através de dois bancos de dados, sendo um relacional (Mysql) e o outro não relacional (Mongodb). O Mysql é responsável pelas informações tratadas, sites para análise e base de palavras. A base de palavras é o conjunto de entradas categorizadas conhecidas, ao qual o sistema consulta no momento de comparar com as entradas dos sites em análises. Já o Mongodb, é para armazenamento dos elementos não estruturados obtidos pela extração.

Para o gerenciamento da arquitetura proposta, foi decidido o uso de tecnologia de containers Docker, escolhido por apresentar facilidades para gerenciamento de ambiente de desenvolvimento e, porque o mesmo facilita a implantação da aplicação em ambiente de produção. Cada contêiner isola um módulo da aplicação e limita o acesso às APIs e bancos, garantindo segurança aos dados extraídos. Por fim, a implantação da aplicação foi feita na GCP (Google Cloud Platform), nuvem atualmente consolidada no mercado e recomendada para implantação de aplicações (MULTIEDRO, 2022). Nesta nuvem, foram implantados tanto o Backend como o Frontend de nossa solução.

5.3 Micro serviços

Arquitetura micro serviços vem com a proposta de tornar sistemas mais flexíveis, escaláveis e com manutenção mais simples. Foi criada em 2011 em uma conferência de arquitetos de software e representa um estilo de arquitetura de sistema e não exatamente o tamanho dos serviços que lhe contemplam com dito em (SOFTWARE, 2021). Fugindo do padrão monolítico dos sistemas, seu desenvolvimento constitui da quebra da aplicação em módulos que irão se comunicar entre si e terão recursos compartilhados.

5.3.1 API Gateway

Como API de entrada na aplicação foi desenvolvido um Gateway para controlar e gerir as requisições externas e interna, esta API utiliza o framework Javascript Adonis JS, um framework robusto e confiável que consegue ser escalável com o nível de desenvolvimento da aplicação (BEZERRA, 2021). A principal motivação da escolha desta tecnologia nesta API se dá na utilização de chamadas assíncronas e o start package (estado inicial da aplicação ao iniciar o projeto) que o framework possui, além de todos os recursos que o Node JS consegue entregar como flexibilidade, leveza e produtividade.

Além disso, a API realiza o controle de autenticação da aplicação através da disponibilização de login com JWT (JSON Web Token). A criação do usuário é realizada por e-mail e password e com isso já se consegue gerar um token de acesso para acessar os recursos da aplicação. Uma observação neste ponto é que não expandimos completamente a camada de autenticação e segurança, por não ser o foco do trabalho, mas deixamos implementado por e-mail e senha um ponto de partida que pode ser desenvolvido e escalado em necessidades futuras.

5.3.2 API Sites

Através desta API, são gerados os relatórios para o dashboard, o gerenciamento dos dados analisados tratados pela API Crawler e o cadastro dos sites para análise. Sua implementação utiliza a linguagem PHP com o framework Laravel.

O Laravel é um framework completo e robusto que consegue facilitar o desenvolvimento de APIs REST e oferece também um ORM (Object Relational Mapping), que é o Eloquent. Com o uso do Eloquent, é possível acessar e manipular o banco de dados de forma simplificada e, com isso, realizar a análise das informações. Através dele construímos Migrações (procedimento que altera o estado de um banco de dados criando tabelas e relacionamentos) e Seeds (classes responsáveis por inserir os dados).

Gerando assim, as tabelas e dados iniciais que se fazem necessários, como as bases de palavras para dados pessoais e sensíveis que serão utilizados na API Crawler.

Outro fato importante para o uso do framework se refere ao seu processo de implantação ser também simplificado e seu custo em recursos computacionais (processamento, espaço de disco, consumo de memória RAM) ser considerado baixo. Além disto, o suporte desta ferramenta é um diferencial, pois a mesma possui uma comunidade ativa e é relativamente bem documentada.

O cadastro dos sites para análise recebe os seguintes itens:

- **Name** — Informa o nome do site.
- **Link** — Informa o link do site com o formulário para o acesso do Crawler.
- **Selector** — Informa a referência CSS para achar o formulário.
- **Xpath** — Informa a referência Xpath para achar o formulário.

Cada site cadastrado possui duas flags: “run” e “error”. A “run” é usada para validar quem a rotina processou. A “error” é utilizada para informar que existiram erros na extração e processamento.

5.3.3 API Crawler

Esta API é responsável pelo processo de extração dos dados pessoais solicitados nos formulários, assim como categorizar os dados em sensíveis e pessoais, conforme definido na LGPD. Seu desenvolvimento foi em Python na versão 3.7 em conjunto com o micro framework flash, e este desenvolvimento se constitui como o último microsserviço implementado na solução.

A construção desta API em Python foi motivada pelo fato de a linguagem possuir uma ampla variedade de bibliotecas que conseguem facilitar o tratamento de dados e realizar o processo de extração (MATOS, 2019). Uma destas bibliotecas é o Selenium, que tem como finalidade principal a criação de testes automatizados em aplicações WEB; contudo, consegue ser também útil para criação de Crawlers, pois consegue interagir com os componentes HTML presentes nas páginas dos sites.

Para o uso do Selenium se faz necessário possuir conhecimento sobre as tags HTML. Tags são componentes usados para informar ao navegador a estrutura do site (HOMEHOST, 2022). A principal característica das tags é estarem sempre dentro dos sinais de Chevron (sinal de “maior que” e “menor que”), ou seja: < > (HOMEHOST, 2022).

Para facilitar o processo de encontrar as tags <form>, local ao qual se encontra o formulário, seus inputs e labels, usamos o Xpath e o CSS Seletor dos sites. O Xpath é o conjunto de nós em documento XML e funciona como um caminho de etapas para o componente escolhido como uma árvore do pai até o filho. O CSS Seletor é a referência de estilização que aquele componente possui.

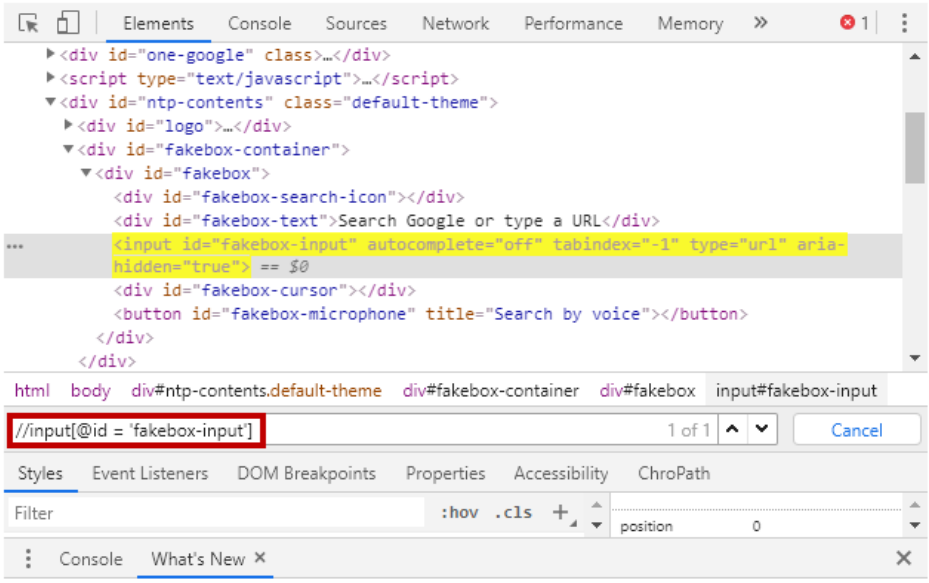


Figura 4 – Referência Xpath do formulário (SINGODIYA, 2022)

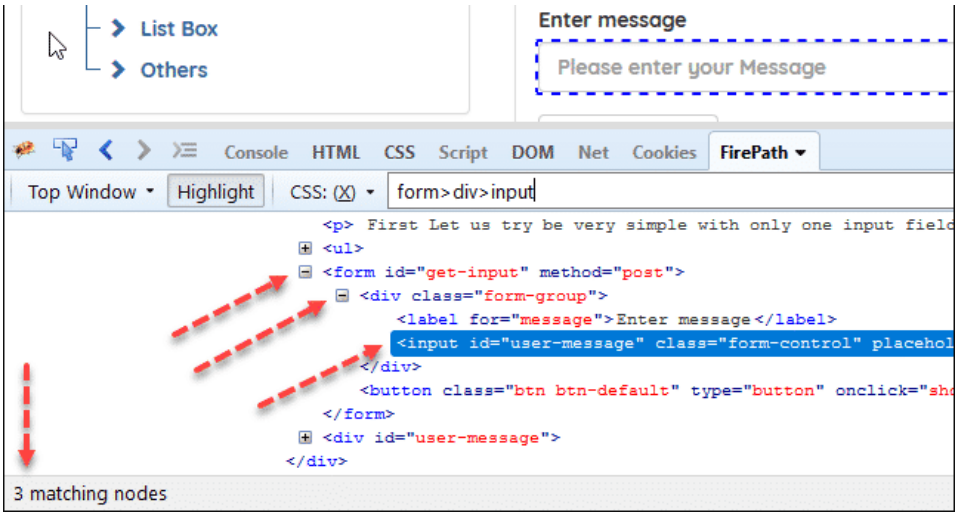


Figura 5 – Referência CSS Seletor do formulário (SWTESTACADEMY, 2019)

Cada um destes elementos (Xpath e CSS Seletor) pode ser obtido através das ferramentas de desenvolvedores disponíveis nos navegadores.

5.3.3.1 Algoritmo

Tendo os sites a serem analisados previamente cadastrados através do dashboard, junto ao seu Xpath e o CSS seletor, é possível iniciar o processo na API Crawler por um botão de início de rotina. Esta solicitação começa com o carregamento de todos os sites cadastrados que possuem sua flag “run” como falso. A API Sites fornece uma lista de todos os sites a serem processados, ao qual será analisado site a site.

Com o Xpath e o CSS seletor do formulário carregado pelo Selenium, o algoritmo executa duas análises do formulário, uma para cada um destes recursos. Primeiramente é carregado duas listas, com as tags <label> e <input> vazias. Com isso, começa a navegação entre os componentes internos partindo da tag pai <form> até seu componente mais interno, sendo ele um <label> ou um <input>. Ao ser encontrado o componente, é extraído seu texto e seu atributo nome, e assim armazenamos esta informação em suas respectivas listas. O processo de navegação se repete até o fim do fechamento da tag </form>.

Em caso de falha na identificação dos componentes, é escrito um LOG de erro e a flag de erro para o site fica com o valor de “true”. Desta forma se registra o problema e se realiza ações que possam resolver a falha.

Cada execução com sucesso tem seus dados armazenados em uma coleção no banco do Mongo, pois os dados ainda não foram tratados. Posteriormente, ao final do processamento, uma nova rotina é iniciada para realizar o processo de categorização e separação dos elementos em pessoais e sensíveis.

5.3.3.2 Tratamento dos elementos

Antes de iniciar esta seção, é importante esclarecer que não foi o foco deste projeto desenvolver uma forma de categorização das entradas extraídas (especialmente utilizando técnicas de machine learning ou similares). Para isto, seria necessário um vasto conjunto de bases de treinamento para criar um modelo de análise de dados, fugindo do escopo deste projeto. Desta forma, para auxiliar o processo de categorização, foram criadas tabelas que possuem o “tipo do input” e deixamos a categoria de dados divididas em três formas: pessoais, sensíveis e não definidos.

A categorização dos dados se inicia com o carregamento das palavras chaves presentes na base de palavras. Após o carregamento da base de palavras, é carregada a lista das entradas de cada site, armazenadas anteriormente através do processo de extração. Estas entradas passam por 3 processos: limpeza das palavras, comparação das palavras e contenção. Todos esses passos fazem parte do cálculo de similaridade realizado ao comparar as palavras encontradas com a base de palavras. Esse processo é essencial para classificar as entradas e evitar repetição de dados. A seguir será

detalhado cada passo.

(i) Limpeza das palavras — Processo que realiza o tratamento das entradas as tornando-as compatíveis com base de palavras. Este passo conta com a aplicação de três filtros nas palavras, definidos abaixo:

- O primeiro filtro remove caracteres especiais, através do uso da seguinte expressão regular (“[â-zA-Z0-9]”).
- O segundo filtro remove os espaços contidos nas palavras.
- Por fim, o terceiro filtro altera as palavras em minúsculas aplicando um “Lower case” (caixa baixa) de modo a garantir igualdade quando comprar elementos.

(ii) Comparação das palavras — Nesta etapa e na seguinte, se utiliza a biblioteca do Python Sklearn, uma biblioteca que aplica técnicas de aprendizado de máquina. Ela dispõe de ferramentas simples e eficientes para análise preditiva de dados, é reutilizável em diferentes situações, possui código aberto, acessível a todos e construída sobre os pacotes “NumPy”, “SciPy” e “Matplotlib” (DIDATICA, 2022). Com a sua utilização, foi possível calcular a similaridade das palavras presentes em nossa base com as entradas a serem categorizadas. Desta forma, foi possível reconhecer palavras presentes nos formulários que sejam similares às contidas na base de palavras. Um exemplo é a entrada “nome”, sendo usualmente escrita de múltiplas formas, como: nome completo, name, nome do usuário, user name, etc. A seguir, são apresentados dois conceitos: n-gramas e vocabulário. Estes conceitos são fundamentais, pois a partir deles conseguiremos compreender como adquirir o valor de similaridade usado na comparação das palavras.

Os n-gramas são a sequência de elementos dentro de uma frase, podendo estes elementos serem palavras, símbolos, letras ou até mesmo classificação gramatical. Para este processo, foi definido que um n-grama equivale a uma palavra (SILVA; SOUZA, 2014). Para a criação dos n-gramas é necessário obter um contador de n-gramas. Este contador será usado para vetorizar as palavras analisadas e gerar com isso os vocabulários. A instância do contador é realizada através da função “CountVectorizer” passando dois parâmetros: O “analyzer” que recebe o valor “word”, informando assim que será analisado cada palavra e o parâmetro “ngram_range=(n,n)” onde “n” é igual a 1, tornando assim cada palavra um n-grama. A Figura 6 a seguir ilustra a maneira de instância o contador de n-gramas a ser usado na montagem dos vocabulários.

```
counts = CountVectorizer(analyzer="word", ngram_range=(n, n))
```

Figura 6 – Contador de n-gramas.

Os vocabulários são os dicionários responsáveis por armazenar os valores correspondentes de cada n-grama e sua respectiva palavra. Eles são formados a partir do contador de n-gramas utilizando a função abaixo descrita na Figura 7.

```
counts.fit([inputCompare, input]).vocabulary_
```

Figura 7 – Criação do dicionário de n-gramas.

Esta etapa é finalizada com a montagem do dicionário e sua inclusão no vetor de comparação, onde a primeira linha é o texto a ser comparado e a segunda linha o texto fonte. Cada coluna é um número representante da palavra descrita no dicionário. Este vetor será passado como parâmetro para próxima etapa.

(ii) Contenção — A contenção é a etapa final e ela é composta pela normalização do vocabulário. A normalização será o valor calculado a partir da interseção da contagem de n-gramas nas duas palavras sendo comparadas. Este cálculo é dividido em 3 partes:

1. Calcular a interseção de n-gramas entre os textos.
2. Adicionar o número de termos comuns.
3. Normalizar o valor na etapa 2 pelo número de n-gramas no texto base.

Estes passos podem ser representados pela seguinte equação:

$$\frac{\sum count\,(ngramA) \bigcap count\,(ngramB)}{\sum count\,(ngramA)} \tag{5.1}$$

O cálculo da interseção de n-gramas gera uma lista onde cada posição corresponde a quantidade de interseções contadas e com isso somamos os valores desta lista e o dividimos pelo somatório dos valores contidos no n-grama da palavra a ser comparada. Este resultado nos fornece uma normalização de dados gerando assim o grau de similaridade dos textos.

Através do grau de similaridade gerado, é possível determinar a categoria da palavra. Desta forma, se grava em uma tabela a entrada tratado e o site do qual ele veio, com sua categoria. Com essas informações, se produz relatórios para exibição no Dashboard.

5.4 Dashboard

O desenvolvimento do Dashboard teve sua implementação usando o Vue.js, um framework Javascript usado em Frontend que permite flexibilidade na montagem

e desenvolvimento de páginas Web (PICOLLO, 2020). O Dashboard conta com três telas:

- Dashboard;
- Cadastro de Sites;
- Tabela de sites cadastrados.

A Figura 8 mostra a tela inicial, o Dashboard. Nela se encontra uma visão geral dos sites analisados e dos dados encontrados, representados a partir de 3 gráficos:

1. **Dados Pessoais** — Quantitativo dos dados pessoais encontrados nos sites analisados.
2. **Dados Sensíveis** — Quantitativo dos dados sensíveis encontrados nos sites analisados.
3. **Sites** — Quantitativo geral de dados encontrados em cada site analisado, separados por site. Este gráfico possui 3 filtros para ajustar sua visualização. São eles: Dados Pessoais, Dados Sensíveis e Indefinidos. Valores indefinidos são as entradas não categorizadas.

Figura 8 – Tela Dashboard

A Figura 9 mostra o formulário de cadastro de sites. Usado para gravar que sites serão analisados. Seus campos foram apresentados na seção 4.2.2.

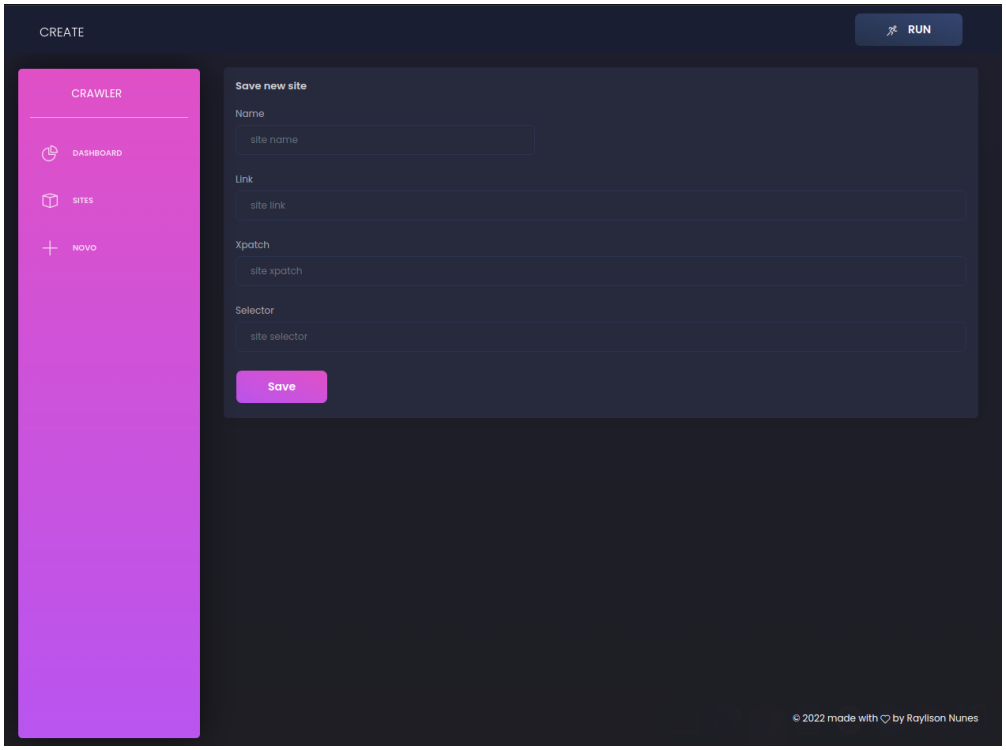


Figura 9 – Formulário para cadastro dos sites

Por fim temos a tela de sites cadastrados exibidos na Figura 10. Esta tela exibe uma tabela paginada com todos os sites cadastrados na aplicação.

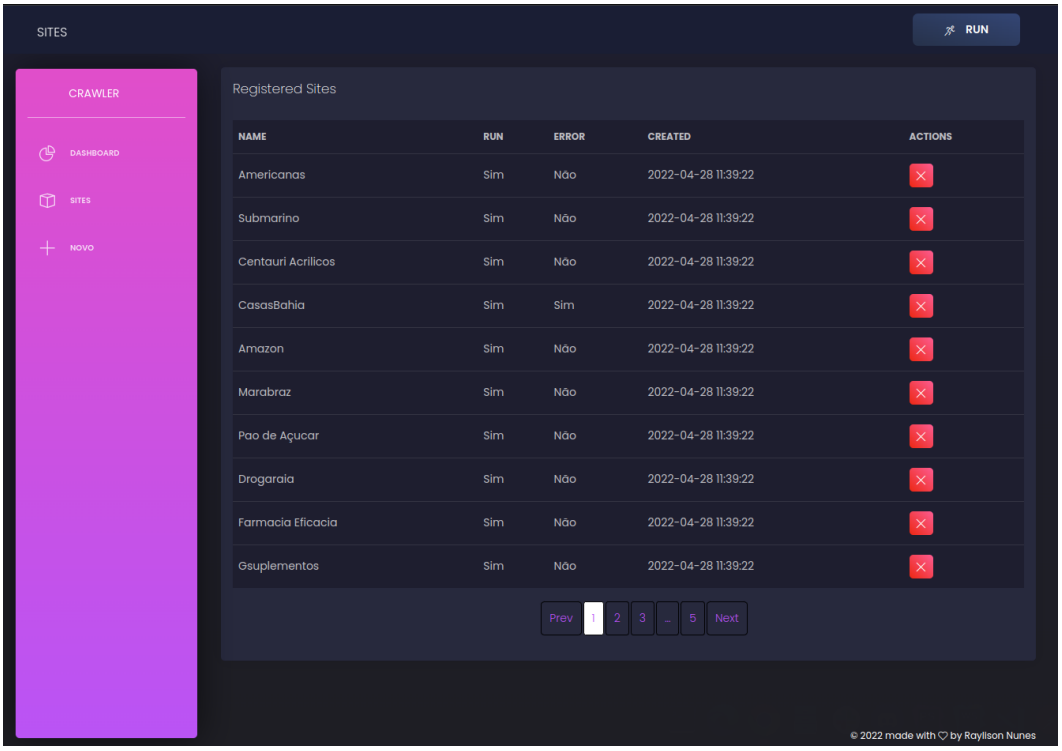


Figura 10 – Tabela dos Sites Cadastrados para Análise.

A interface conta no topo direito com um botão de “RUN”, usado para disparar a requisição que dará início a rotina de extração e ao final iniciar a rotina de tratamento de dados. Este botão também pode ser visualizado nas figuras anteriormente exibidas.

6 Avaliação

Nesta seção, são apresentados os resultados obtidos pela execução da solução proposta. Também são apresentados resultados referentes ao desempenho da ferramenta.

6.1 Visão geral da avaliação

Os sites avaliados foram cadastrados através do formulário de cadastro presente no Dashboard da aplicação. Os resultados apresentados foram obtidos através da exportação dos dados catalogados presentes em nosso banco de dados “Sites”, importados em uma planilha para montagem dos gráficos.

O Crawler foi executado uma única vez, após o cadastro dos sites e com isso podemos avalizar seu desempenho em extrair, categorizar e reconhecer palavras novas. O fluxo completo para geração dos relatórios pode ser constatado na Figura 11.

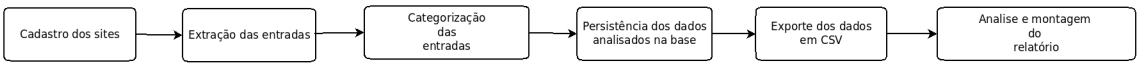


Figura 11 – Fluxograma para extração dos resultados

6.2 Resultados

Ao todo foram identificados 145 dados pessoais solicitados as pessoas. A quantidade mínima de dados pessoais solicitados foi de 1 (um), e a máxima foi de 15 (quinze). De maneira geral a média foi de 4 (quatro) solicitações por site.

A Figura 12 apresenta uma visão geral dos sites e a quantidade de informações solicitadas.

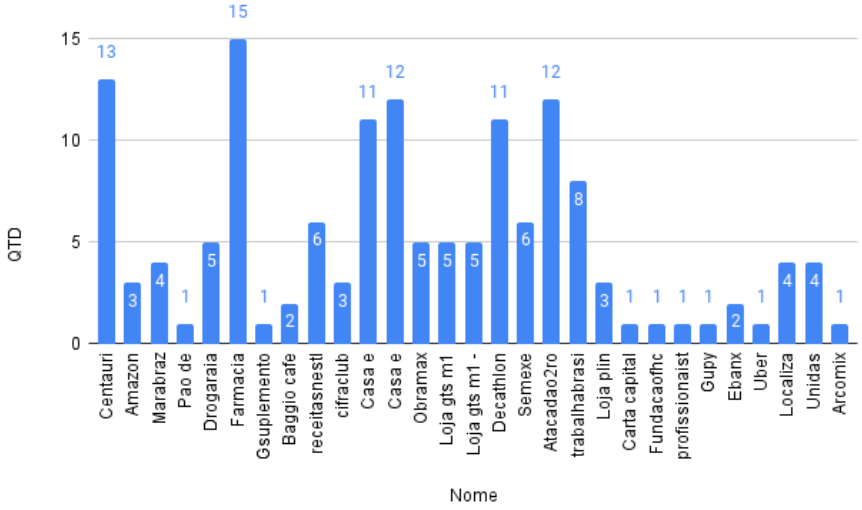


Figura 12 – Visão Geral de Dados Solicitados em Sites

Por sua vez, a Figura 13 apresenta o quantitativo separado por cada elemento solicitado. Dados como CPF, data de nascimento e nome estão no topo da lista dos mais solicitados, representando 41,6% dos dados pessoais.

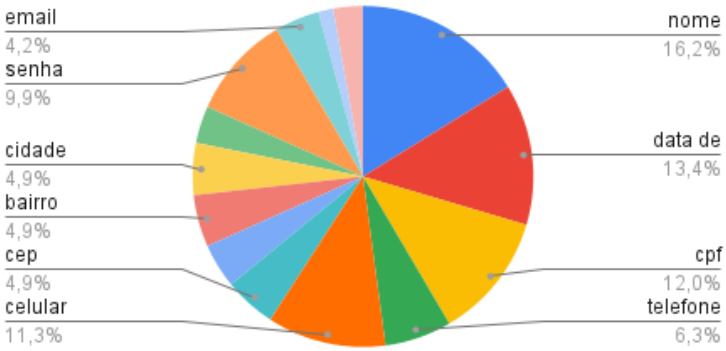


Figura 13 – Gráfico Geral dos Inputs Solicitados.

É importante ressaltar que nenhum destes sites analisados possuía informações sobre para que seria usado cada um destes dados. Sobre as categorias de dados (pessoais e sensíveis), foram obtidos um total de 94,5% de solicitações em dados pessoais e 5,5% para sensíveis como demonstrado na Figura 14.

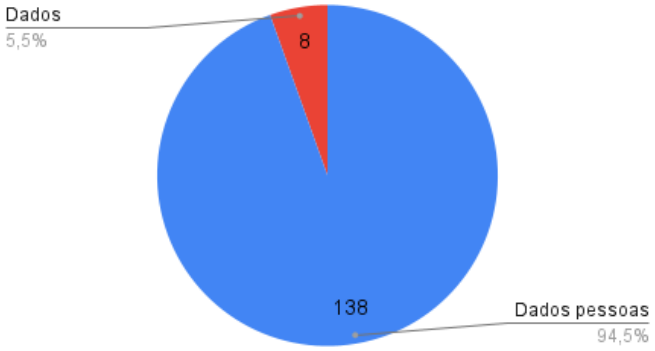


Figura 14 – Dados Pessoais x Dados Sensíveis.

É válido ressaltar que estas informações foram retiradas a partir de uma execução única da aplicação, e nesta execução se observou como a mesma conseguiu obter e processar os dados. Neste processo algumas palavras foram aprendidas e, se tratadas, conseguiriam melhorar ainda mais a capacidade de reconhecer entradas e capturar dados. A Tabela 1 apresenta a lista das palavras aprendidas nesta execução.

ID	Name
1	rg
2	e mail
3	n
4	referencia
5	genero
6	sobrenome
7	tipo de endereço
8	telefone comercial
9	natureza cliente
10	tipo cliente
11	possui deficiencia
12	nivel formacao
13	funcao
14	nacionalidade
15	nome mae
16	passaporte

Tabela 1 – Palavras aprendidas

A seguir serão evidenciadas informações relacionadas ao desempenho do sistema em extrair a informação desejada. Iniciaremos com o percentual de acerto em obter os valores com um comparativo de sucesso e erro. O gráfico exibido na Figura refvisao-geral-sucesso-exec mostra que em 80% dos sites foram processados sem erro. Já os 20% de erros aconteceram ao tentar acessar o elemento HTML que não existia, e assim era levantando uma exceção no sistema.

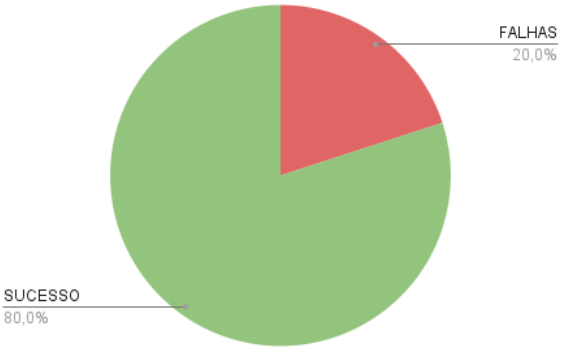


Figura 15 – Visão geral do sucesso das execuções.

Considerando as execuções bem-sucedidas (que não apresentaram erro no processamento), foi obtido uma taxa de 72,5% de extrair ao menos 1 dado pessoal e 27,5% em não conseguir nada.

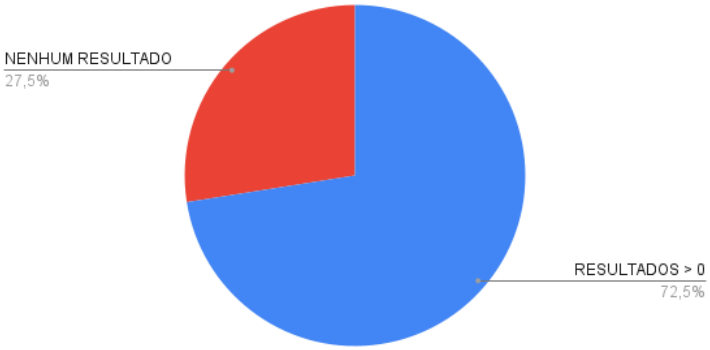


Figura 16 – Visão geral da obtenção ou não de dados nas execuções.

De forma geral, se conseguiu extrair um total de 147 inputs em todos os sites, onde o total de dados presentes era 234. A Figura 17 ilustra este resultado.

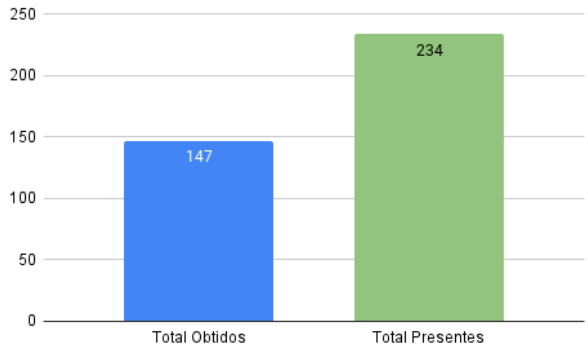


Figura 17 – Precisão na captura de entradas.

Também se avaliou a precisão de captura das entradas. A Tabela 2 mostra a precisão de maneira individual em cada site.

Nome	Achado	Esperado	Precisão
Centauri Acrilicos	13	15	86,67%
Amazon	3	3	100,00%
Marabraz	4	14	28,57%
Pao de Açucar	1	5	20,00%
Drogaraia	5	7	71,43%
Farmacia Eficacia	15	15	100,00%
Gsuplementos	1	7	14,29%
Baggio cafe	2	3	66,67%
receitasnestle	6	6	100,00%
cifraclub	3	4	75,00%
Casa e construção	11	18	61,11%
Casa e contrução - PJ	12	18	66,67%
Obramax	5	7	71,43%
Loja gts m1	5	7	71,43%
Loja gts m1 - PJ	5	7	71,43%
Decathlon	11	15	73,33%
Semexe	6	6	100,00%
Atacadao2rodas	12	16	75,00%
trabalhabrasil	8	13	61,54%
Loja plin cosmeticos	3	4	75,00%
Carta capital	1	3	33,33%
Fundacaofhc	1	4	25,00%
profissionaisti	1	1	100,00%
Gupy	1	3	33,33%
Ebanx	2	7	28,57%
Uber	1	5	20,00%
Localiza	4	8	50,00%
Unidas	4	6	66,67%
Arcomix	1	7	14,29%

Tabela 2 – Precisão individual.

Em termos gerais, a precisão foi de 61,97%. Este número é considerado interessante para um estudo inicial, pois, ele tende a ser melhorado em futuras execuções/estudos e permitiu se ter uma boa noção sobre como estão sendo pedidos os dados pessoais em formulários Web atualmente (considerando os sites selecionados).

7 Considerações finais e trabalhos futuros

Este trabalho apresentou uma solução para análise de dados pessoais pedidos em formulários Web através da construção de um Web Crawler que, por duas rotinas distintas, extrai e categoriza os dados. Os resultados obtidos mostraram não apenas que a solução proposta é viável, mas também ilustraram como está sendo pedido, em termos quantitativos, dados pessoais em formulários Web atualmente. Além desta contribuição, também se destaca as contribuições parciais a seguir:

1. Implementação de um Web Crawler integrado a uma arquitetura de microsserviços para extração de entradas em formulários Web;
2. Levantamento de dados pessoais e dados sensíveis comumente utilizados em formulário eletrônicos;
3. Implementação de uma base de dados com as entradas dos formulários categorizados pelas definições da LGPD.

7.1 Impacto para pesquisa e prática

De maneira geral, o uso da ferramenta consegue transparecer para sociedade, quais informações as empresas conhecem e requisitam a ela. Desta forma, fica mais aparente a necessidade de realizar ações que venham cobrar a justificativa dos dados, assim como, exigir segurança no tratamento dessas informações.

Já no âmbito acadêmico, a disponibilização de um sistema como esse, oferece maneiras de se analisar as categorias de dados comumente solicitados as pessoas de maneira simples e prática. Assim como, o uso da base de palavras, poderá servir de fonte de treinamento para algoritmos de Machine Learning, e contribuir com o desenvolvimento de ferramentas mais robusta e específicas.

7.2 Limitações e ameças à validade

Devido ao prazo curto, não foi possível melhorar o desempenho da solução em identificar e extrair as entradas nos fomulários melhorando assim sua precisão geral aplicando técnicas mais assertivas na identificação dos Labels e Inputs e obter assim seus valores para análise. Assim como, realizar mais rodadas de execução apos a categorização das palavras não conhecidas, de modo a conseguir um resultado mais assertivo na representação dos dados presentes nos sites.

7.3 Trabalhos futuros

No contexto deste projeto, os seguintes trabalhos futuros se destacam.

- Possibilitar o acesso a campos que necessitam de interação humana, como clicar ou informar um dado prévio, para serem exibidos.
- Realizar, de forma automatizada, a classificação de palavras não reconhecidas, usando uma estratégia mais robusta, que consiga aprender utilizando a base de dados da solução.
- Buscar informações, de maneira automatizada, que possam justificar o pedido dos dados pessoais solicitados.
- Disponibilizar uma API pública para utilização externa.

Referências

ÁLVAREZ, M. et al. Crawling the content hidden behind web forms. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2007. p. 322–333. Citado na página 14.

BCIJO. *O Que São Bases De Dados*. 2020. Disponível em: <https://biblioteca.pucrs.br/ufaqs/o-que-sao-bases-de-dados/>. Citado na página 16.

BEZERRA, G. *Visao Geral do Adonisjs*. 2021. Disponível em: <https://giuliana-bezerra.medium.com/visao-geral-do-adonisjs-c2096a329685>. Citado na página 27.

BRASIL. Lei n° 13.709 de agosto de 2018. lei geral de proteção de dados pessoais. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Citado 2 vezes nas páginas 11 e 17.

CARVALHO, A. P. Proposta de um framework de compliance à lei geral de proteção a dados pessoais (lgpd): um estudo de caso para prevenção a fraude no contexto de big data. 2021. Citado na página 20.

CONSUMIDORMODERNO. *Em que pé estão as leis de proteção de dados no mundo, Consumidor Moderno*. 2021. Disponível em: <https://digitalks.com.br/artigos/privacidade-digital-quais-sao-os-limites/>. Citado na página 18.

DICIO. *Privacidade*. 2017. Disponível em: <https://www.dicio.com.br/privacidade/>. Citado na página 16.

DIDATICA. *A biblioteca scikit-learn – Pyhton para machine learning*. 2022. Disponível em: <https://didatica.tech/a-biblioteca-scikit-learn-pyhton-para-machine-learning/>. Citado na página 31.

FISCHER, T. R. *TÍTULO: IMPLEMENTAÇÃO DE UM CRAWLER INCREMENTAL DISTRIBUÍDO: UM SISTEMA DE BUSCA DE PÁGINAS NA WEB (WEB SEARCH)*. Tese (Doutorado) — UNIVERSIDADE REGIONAL DE BLUMENAU, 2006. Citado na página 15.

GLOBO, O. *Informar o CPF não é obrigatório. Saiba seus direitos. O Globo*. 2021. Disponível em: <https://oglobo.globo.com/economia/defesa-do-consumidor/informar-cpf-nao-obrigatorio-saiba-seus-direitos-22666875>. Citado na página 11.

HOMEHOST. *Tags HTML*. 2022. Disponível em: <https://www.homehost.com.br/blog/tutoriais/tags-html/>. Citado na página 28.

MACHADO, C. C. et al. Um web crawler para projeções e análise de vulnerabilidades de segurança e consistência estrutural de páginas web. *Revista de Empreendedorismo, Inovação e Tecnologia*, v. 2, n. 2, p. 3–12, 2016. Citado na página 20.

MATOS, D. *Por que Cientistas de Dados escolhem Python?* 2019. Disponível em: <https://www.cienciaedados.com/por-que-cientistas-de-dados-escolhem-python/>. Citado na página 28.

MEDIA, I. L. *Meio e Mensagem.* 2020. Disponível em: <http://tech.meioemensagem.com.br/in-loco-media/>. Citado na página 11.

MORTE, A. B. et al. Uma análise sobre o uso de dlts no tratamento de dados pessoais: Aderência aos princípios e direitos elencados na lgpd. In: SBC. *Anais do III Workshop em Blockchain: Teoria, Tecnologia e Aplicações*. [S.l.], 2020. p. 74–87. Citado na página 20.

MULTIEDRO. *Google Cloud Platform o que é e quais as suas vantagens*. 2022. Disponível em: <https://blog.multiedro.com.br/google-cloud-platform-o-que-e-e-quais-as-suas-vantagens/>. Citado na página 26.

NASCIMENTO, A. L. *Softwares, Sistemas e Aplicações: a evolução no mundo das soluções.* 2020. Disponível em: <https://administradores.com.br/artigos/softwares-sistemas-e-aplicações-a-evolução-no-mundo-das-soluções>. Citado na página 11.

NEITINBAG. *What is digital privacy*. 2022. Disponível em: <https://www.netinbag.com/pt/internet/what-is-digital-privacy.html>. Citado na página 16.

PICOLLO, L. *Vue JS: o que é, como funciona e vantagens*. 2020. Disponível em: <https://blog.geekhunter.com.br/vue-js-so-vejo-vantagens-e-voce/>. Citado na página 33.

RAGHAVAN, S.; GARCIA-MOLINA, H. *Crawling the hidden web*. [S.l.], 2000. Citado na página 14.

RESOLVARAPIDO. *As empresas que coletam meus dados podem vendê-los a terceiros?* 2022. Disponível em: <https://resolvarapido.com/as-empresas-que-coletam-meus-dados-podem-vende-los-a-terceiros/>. Citado na página 24.

RIBEIRO, C. Big data: os novos desafios para o profissional da informação. *Universidade Federal do Estado do Rio de Janeiro(UNIRIO).*, 2014. Citado na página 12.

RODOTÀ, S. A vida na sociedade da vigilância: a privacidade hoje. In: *A vida na sociedade da vigilância: a privacidade hoje*. [S.l.: s.n.], 2015. p. 381–381. Citado na página 17.

SERPRO. *O que muda com a LGPD*. 2021. Disponível em: <https://www.serpro.gov.br/lgpd/menu/a-lgpd/o-que-muda-com-a-lgpd>. Citado 4 vezes nas páginas 16, 17, 19 e 25.

SILVA, E. M. da; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, Universidade Federal de Santa Catarina, v. 19, n. 40, p. 1–31, 2014. Citado na página 31.

SILVA, R. H. d. et al. Framework para identificar o nível de conformidade das empresas brasileiras do setor químico no processo de adequação à lei geral de proteção de dados pessoais. 2021. Citado na página 20.

SILVA, V. R. Britto-da; LUCIANO, E. M.; MAGNAGNAGNO, O. A. Preocupação com a privacidade na internet: uma pesquisa exploratória no cenário brasileiro. *Anais do V Encontro de Administração da Informação, 2015, Brasil.*, 2015. Citado na página 11.

SILVEIRA, S. A. et al. A privacidade e o mercado de dados pessoais| privacy and the market of personal data. *Liinc em Revista*, Instituto Brasileiro de Informação em Ciência e Tecnologia, v. 12, n. 2, 2016. Citado na página 11.

SINGH, M.; VARNICA, B. Web crawler: Extracting the web data. *International Journal of Computer Trends and Technology*, v. 13, n. 3, p. 132–137, 2014. Citado 2 vezes nas páginas 12 e 14.

SINGODIYA, M. *Introduction to Xpath*. 2022. Disponível em: <https://www.geeksforgeeks.org/introduction-to-xpath/>. Citado na página 29.

SOFTWARE, O. *Micro Serviços: Qual a diferença para a Arquitetura Monolítica?* 2021. Disponível em: <https://www.opus-software.com.br/micro-servicos-arquietura-monolitica/>. Citado na página 27.

SWTESTACADEMY. *CSS Seletor*. 2019. Disponível em: <https://www.swtestacademy.com/wp-content/uploads/2017/09/img_59cee3e5d9728.png>. Citado na página 29.

WANG, A. H. Don't follow me: Spam detection in twitter. In: IEEE. *2010 international conference on security and cryptography (SECRYPT)*. [S.l.], 2010. p. 1–10. Citado na página 14.

WU, H. et al. Data analysis and crawler application implementation based on python. In: IEEE COMPUTER SOCIETY. *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*. [S.l.], 2020. p. 389–393. Citado na página 14.

A APÊNDICE A — SITES CADASTRADOS

ID	Name
1	Americanas
2	Submarino
3	Centauri Acrilicos
4	CasasBahia
5	Amazon
6	Marabraz
7	Pao de Açucar
9	Drogaraia
10	Farmacia Eficacia
11	Gsuplementos
12	Studio3t
13	Cafe Store
14	Coffee Mais
15	avell
16	Baggio cafe
17	Tudo Gostoso
18	receitasnestle
19	cifraclub
20	Leroy merlin
21	Leroymerlin - Empresa
22	Casa e construção
23	Casa e contrução - PJ
24	Obramax
25	Loja gts m1
26	Loja gts m1 - PJ
27	Decathlon
28	Semexe
29	Atacadao2rodas
30	trabalhabrasil
31	Eudora
32	Ziluu
33	Loja plin cosmeticos
34	Carta capital
35	Nfse

ID	Name
36	Unasus
37	Fundacaofhc
38	Fiocruz
39	profissionaisti
40	Gupy
41	Udemy
42	Ebanx
43	Uber
44	Localiza
45	Unidas
46	Autoeurope
47	Movida
48	girafa
49	Arcomix
50	Armazempb
51	Gsuplementos