Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Uma biblioteca para anonimização de dados pessoais brasileiros em textos

Raylan da Silva Sales
Stefano Luppi Spósito

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientadora
Prof.a Dr.a Edna Dias Canedo

Coorientadora
Prof.a MSc Geovana Ramos Sousa Silva

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Uma biblioteca para anonimização de dados pessoais brasileiros em textos

Raylan da Silva Sales
Stefano Luppi Spósito

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof. Dr. Rodrigo Bonifácio de Almeida
CIC/UnB

Prof. Dr. Dibio Leandro Borges
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 11 de dezembro de 2023

Dedicatória

Dedicamos este trabalho à Deus, aos amigos e às nossas famílias, que nos apoiaram em todo momento e nos deram forças para continuar.

Agradecimentos

Agradecemos primeiramente a Deus, que nos permitiu e nos deu forças para alcançar nossos objetivos durante todos nossos anos de estudos. Agradecemos imensamente as nossas famílias, pelo apoio e incentivo nos momentos difíceis e que sempre lutaram por nossa educação e sempre acreditaram em nós. Agradecemos aos nossos amigos e a todos que nos apoiaram. Agradecemos as professoras Edna Dias Canedo e Geovana Ramos Sousa Silva, por terem nos orientado e exercido este papel com dedicação, paciência e amizade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Atualmente, uma grande quantidade de dados pessoais de indivíduos está em posse de grandes empresas. A Lei Geral de Proteção de Dados (LGPD) foi criada para regularizar o uso destes dados e reduzir esta posse ao máximo, para que sejam usados apenas os dados necessários e, se necessário, anonimizá-los de acordo com os regulamentos estabelecidos, entretanto, não há ainda uma ferramenta específica para anonimização de dados pessoais brasileiros. O objetivo deste trabalho é criar uma biblioteca para a anonimização de dados pessoais, de forma a atender às especificidades dos dados pessoais brasileiros. A ferramenta tem como meta proporcionar um meio eficiente e seguro para a remoção de informações sensíveis e identificáveis presentes nos documentos, preservando a privacidade e a confidencialidade dos dados, de acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD). Para este trabalho, foi utilizado uma abordagem mista, combinando elementos de pesquisa experimental e aplicada. A pesquisa experimental avaliou as bibliotecas existentes por meio de testes em cenários de anonimização de dados brasileiros, enquanto a pesquisa aplicada desenvolveu uma biblioteca específica para dados brasileiros em Python, considerando técnicas existentes e requisitos locais. A metodologia incluiu uma revisão bibliográfica, com pesquisa em bases científicas e a avaliação prática da biblioteca, destacando critérios como eficácia, preservação da utilidade dos dados e conformidade com regulamentações de proteção de dados. Após a compreensão das técnicas de anonimização de dados, foi desenvolvida uma biblioteca na linguagem de programação Python para a anonimização de dados pessoais brasileiros, de acordo com a LGPD. Foram realizados testes com as diferentes funções da biblioteca, utilizando os dados pessoais encontrados através de pesquisas na internet e dados gerados artificialmente em programas criados com a biblioteca para anonimizar dados pessoais. Os resultados demonstram a eficácia e funcionalidade da biblioteca. Assim, a biblioteca pode ser utilizada em qualquer código criado na linguagem Python. Este trabalho apresenta a criação de uma biblioteca com o intuito de contribuir e auxiliar no cumprimento das regulações estabelecidas pela LGPD, bem como promover a segurança e privacidade das informações pessoais. Como trabalhos futuros, o suporte a novos formatos será incluído na biblioteca, juntamente com uma possível integração a ferramentas de Inteligência Artificial que servirão para uma melhor

detecção de dados pessoais e que permitirá a anonimização desses dados.

Palavras-chave: Anonimização, Dados, Lei Geral de Proteção de Dados, Quasi-Identificadores, Supressão

Abstract

Currently, a large amount of individuals' personal data is in the possession of large companies. The General Data Protection Law (LGPD) was created to regularize the use of this data and reduce this ownership as much as possible, so that only the necessary data is used and, if necessary, anonymize it in accordance with established regulations, however, There is still no specific tool for anonymizing Brazilian personal data. The objective of this work is to create a library for the anonymization of personal data, in order to meet the specificities of Brazilian personal data. The tool aims to provide an efficient and safe means for removing sensitive and identifiable information present in documents, preserving the privacy and confidentiality of data, in accordance with the General Personal Data Protection Law (LGPD). For this work, a mixed approach was used, combining elements of experimental and applied research. The experimental research evaluated existing libraries through tests in Brazilian data anonymization scenarios, while the applied research developed a specific library for Brazilian data in Python, considering existing techniques and local requirements. The methodology included a bibliographic review, with scientific research and practical evaluation of the library, highlighting criteria such as effectiveness, preservation of data usefulness and compliance with data protection regulations. After understanding data anonymization techniques, a library was developed in the Python programming language for the anonymization of Brazilian personal data, in accordance with the LGPD. Tests were carried out with the different functions of the library, using personal data found through internet searches and data artificially generated in programs created with the library to anonymize personal data. The results demonstrate the effectiveness and functionality of the library. Thus, the library can be used in any code created in the Python language. This work presents the creation of a library with the aim of contributing and assisting in compliance with the regulations established by the LGPD, as well as promoting the security and privacy of personal information. As future work, support for new formats will be included in the library, along with a possible integration with Artificial Intelligence tools that will serve to better detect personal data and allow for the anonymization of this data.

Keywords: Anonymization, Data, Quasi-Identifiers, Supression

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este capítulo apresenta os principais problemas que este trabalho visa resolver, bem como breves explicações sobre o tema principal, assim como soluções propostas e métodos a serem utilizados para a realização das soluções.

1.1 LGPD e a Anonimização de Dados

A Lei Geral de Proteção de Dados Pessoais(LGPD) [1], sancionada em 2018 e em vigor desde setembro de 2020, regulamenta como as empresas que atuam no Brasil devem agir em relação à coleta, tratamento e compartilhamento de dados pessoas e sensíveis. Em suma, a LGPD recomenda que seja feita a anonimização sobre os dados pessoais por parte da empresa, em relação a seu cliente. Segundo a lei, dados pessoais são todas aquelas informações capazes de identificar o seu titular, tais quais nome, RG, telefone, data, local de nascimento, e-mail, dados a respeito das contas nas redes sociais, entre outras informações que sejam capazes de levar à identificação do titular de dados. Vale ressaltar também, que para a LGPD, um dado só é considerado anonimizado quando ele perdeu definitivamente a possibilidade de identificar uma pessoa.

Anonimização de Dados

A anonimização de dados é o processo de remover informações pessoais identificáveis de conjuntos de dados, a fim de proteger a privacidade dos indivíduos. O objetivo é permitir que os dados sejam usados para fins legítimos, como pesquisas, análises estatísticas e outras atividades que possam beneficiar a sociedade ou uma organização, sem violar a privacidade dos indivíduos. [2]

A anonimização de dados pode ser aplicada em qualquer contexto que envolva dados pessoais de um ou mais indivíduos, sendo não apenas necessária, mas indispensável, principalmente para questões nos setores da saúde [3], redes sociais e empresariais. No

setor da saúde, técnicas de anonimização podem ser utilizadas para proteger informações sensíveis dos pacientes, ao mesmo tempo em que permitem análises e pesquisas [4]. Por exemplo, uma revisão sistemática das técnicas de anonimização de dados para a área da saúde constatou que diferentes abordagens podem ser usadas para diferentes tipos de dados, como dados demográficos, diagnósticos e procedimentos. No setor financeiro, técnicas de anonimização podem ser usadas para proteger os dados dos clientes, protegendo empresas de ataques que possam levar ao vazamento de dados sensíveis, que sirvam como identificadores para pessoas físicas (CPF, RG) [5]. No âmbito das redes sociais, a anonimização de dados pode ser utilizada para manter a privacidade de seus usuários enquanto a plataforma (Rede Social) compartilha datasets a respeito dos mesmos, para terceiros [6]. Técnicas de anonimização de dados pessoais, como mascaramento, generalização e distorção de dados, podem ser utilizadas para proteger informações sensíveis, como detalhes bancários e endereços residenciais [2]. As técnicas específicas utilizadas podem depender do tipo de dado que está sendo anonimizado e das necessidades específicas da organização, portanto, é necessário encontrar um equilíbrio entre a proteção da privacidade e a utilidade dos dados para análises e pesquisas.

1.2 Anonimização de Dados Brasileiros

De acordo com a Lei Geral de Proteção de Dados [1]:

> Art. 46. Os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de tratamento inadequado ou ilícito.

Onde, estas medidas de seguranças, dentre várias, consistem em ferramentas de anonimizaçao, sejam elas softwares individuais ou bibliotecas de linguagens de programação já existentes. Para este artigo, foram verificadas duas bibliotecas para a linguagem Python, Anonimization [7] e Anonimpy [8], juntamente de um software de anonimização, ARX [9]. Atualmente, diversas ferramentas de anonimização de dados estão disponíveis em várias fontes da internet, onde cada uma realiza as respectivas operações utilizando técnicas únicas e eficazes. Entretanto, as ferramentas mais eficientes não atendem às necessidades de casos brasileiros, não oferecendo suporte nem para a linguagem portuguesa, nem para o formato de dados pessoais brasileiros, como CPF e RG. Portanto, é necessário que exista uma biblioteca open source, segura e eficaz, que possa ser aproveitada por qualquer organização ou indivíduo, garantindo assim, uma anonimização padronizada e segura, de acordo com a LGPD.

1.3 Objetivos

O objetivo principal deste trabalho é criar uma biblioteca para a anonimização de dados pessoais, de forma a atender às especificidades dos dados pessoais brasileiros. A ferramenta tem como meta proporcionar um meio eficiente e seguro para a remoção de informações sensíveis e identificáveis presentes nos documentos, preservando a privacidade e a confidencialidade dos dados, de acordo com as normas e diretrizes aplicáveis à proteção de informações pessoais e à privacidade dos indivíduos. Assim, a solução tecnológica desenvolvida irá contribuir para o cumprimento de regulamentações e políticas relacionadas à proteção de dados, bem como para a promoção da segurança e privacidade das informações.

1.3.1 Objetivo Específico

Para atingir o objetivo geral os seguintes objetivos específicos foram definidos:

- Realizar um levantamento bibliográfico das principais técnicas e algoritmos de anonimização de dados pessoais existentes na literatura.
- Criar uma biblioteca para incorporar as necessidades específicas da anonimização de dados pessoais brasileiros.
- Avaliar a eficácia da biblioteca criada por meio de testes e métricas.

Os testes incluíram documentos com variados tipos de dados pessoais, abrangendo diferentes contextos e cenários. O processo de anonimização foi aplicado a esses documentos, e os resultados foram analisados quanto à preservação da privacidade e à efetividade na remoção de informações sensíveis.

Durante a avaliação da eficácia da biblioteca, foram consideradas as seguintes métricas:

- Preservação da Integridade Estrutural: verificação da manutenção da estrutura e formato original dos documentos após o processo de anonimização.
- Efetividade na Remoção de Dados Sensíveis: avaliação da capacidade da biblioteca em remover informações pessoais sensíveis sem comprometer a utilidade dos dados.

1.4 Resultados Esperados

Espera-se que este estudo sirva como parte da documentação para a versão final da biblioteca de anonimização que suporta dados brasileiros. Assim como referência para futuras bibliotecas ou estudos na mesma área. Além disso, espera-se que a biblioteca criada seja

capaz de anonimizar dados pessoais em diferentes tipos de arquivos, de acordo com a preferência do usuário.

1.5 Metodologia de Pesquisa

Será adotada uma metodologia de pesquisa que engloba elementos de pesquisa experimental e pesquisa aplicada. A pesquisa experimental será utilizada para realizar testes e avaliações das bibliotecas de anonimização existentes, a fim de comparar sua eficácia e desempenho. Esses testes envolverão a aplicação das bibliotecas em cenários de anonimização de dados brasileiros, com a finalidade de verificar sua capacidade de proteger informações sensíveis e identificáveis. Para isso, serão definidos critérios de avaliação e métricas apropriadas para a análise comparativa das bibliotecas.

Além disso, a pesquisa também envolverá a criação de uma biblioteca de anonimização específica para dados brasileiros em Python. Nesse aspecto, a pesquisa aplicada desempenhará um papel importante, pois foram realizados estudos e análises sobre as técnicas de anonimização existentes, bem como as particularidades e requisitos específicos relacionados aos dados brasileiros. Essa etapa envolveu a revisão de literatura, coleta de informações relevantes sobre os métodos de anonimização utilizados no contexto brasileiro e o desenvolvimento de uma solução que atenda às necessidades específicas da anonimização de dados brasileiros.

A metodologia adotada também compreende a revisão bibliográfica , na qual serão realizadas pesquisas em bases de dados científicos, artigos, livros e outras fontes relevantes para identificar as bibliotecas de anonimização existentes e os aspectos relacionados à anonimização de dados brasileiros. Essa revisão fornece uma base sólida para a compreensão das metodologias, técnicas e melhores práticas utilizadas nesse campo.

No que diz respeito à avaliação da biblioteca de anonimização adaptada, a pesquisa seguirá uma abordagem prática, envolvendo a aplicação da biblioteca em cenários reais de anonimização de dados brasileiros. Essa etapa pode incluir a definição de estudos de caso, a coleta de dados relevantes e a análise dos resultados obtidos. Os critérios de avaliação serão definidos com base em métricas relevantes, como a eficácia da anonimização, a preservação da utilidade dos dados e a conformidade com as regulamentações e políticas de proteção de dados.

1.6 Estrutura do TCC

No Capítulo 2, foi realizado um referencial teórico abordando a Lei Geral de Proteção de Dados (LGPD) e as técnicas de anonimização. Foram explorados conceitos fundamentais,

como os princípios da LGPD e as diferentes abordagens para a proteção de dados pessoais. Além disso, o capítulo inclui uma revisão dos trabalhos relacionados na academia e na indústria, proporcionando uma visão abrangente do estado atual das práticas de anonimização.

No Capítulo 3, apresentamos a proposta desenvolvida neste trabalho. Inicialmente, são discutidos os detalhes da anonimização de dados pessoais brasileiros, seguido por uma visão geral das tecnologias e ferramentas utilizadas no desenvolvimento da biblioteca. O capítulo aborda aspectos como a arquitetura da aplicação, requisitos de software, e detalhes relacionados ao deploy da biblioteca. Também são apresentados os códigos-fonte e a estrutura da biblioteca, proporcionando uma compreensão abrangente de sua implementação.

O Capítulo 4 está dedicado aos testes realizados para validar a eficácia da biblioteca de anonimização. Os testes incluem cenários variados, como anonimização de dados em formatos PDF, DOCX e XLSX, bem como a anonimização de strings comuns de dados. A discussão dos resultados aborda o desempenho geral da biblioteca, sua robustez e confiabilidade. O capítulo também destaca ameaças e limitações identificadas durante a validação, proporcionando uma visão crítica do alcance e das capacidades da biblioteca.

No Capítulo 5, apresentamos as conclusões derivadas do trabalho realizado. São discutidos os resultados obtidos nos testes, as contribuições da biblioteca desenvolvida, e possíveis direções para trabalhos futuros na área de anonimização de dados pessoais brasileiros.

Capítulo 2

Referencial Teórico

2.1 Lei Geral de Proteção de Dados

No contexto da sociedade digital, o crescente fluxo de informações pessoais e a necessidade de proteger a privacidade das pessoas têm se tornado questões de extrema relevância [10]. Nesse sentido, a Lei Geral de Proteção de Dados (LGPD) surge como uma importante regulamentação no Brasil, estabelecendo diretrizes para o tratamento de dados pessoais por organizações públicas e privadas [11].

A LGPD foi promulgada em 2018 e tem como objetivo principal garantir aos indivíduos o controle sobre seus dados pessoais, estabelecendo regras claras sobre como esses dados devem ser coletados, armazenados, tratados e compartilhados [1]. A lei visa proteger a privacidade, a liberdade e a dignidade dos cidadãos, além de promover a transparência nas atividades de tratamento de dados.

Principais características da LGPD [11]:

- Proteção dos direitos fundamentais: A LGPD tem como objetivo proteger os direitos fundamentais de liberdade, privacidade e livre desenvolvimento da personalidade das pessoas naturais.
- Segurança jurídica: A lei busca criar um cenário de segurança jurídica, estabelecendo regulamentos e práticas para promover a proteção dos dados pessoais de todos os cidadãos que estejam no Brasil, de acordo com os parâmetros internacionais existentes.
- Definição de dados pessoais: A LGPD define o que são dados pessoais e estabelece que alguns deles estão sujeitos a cuidados específicos, como os dados pessoais sensíveis e os dados pessoais de crianças e adolescentes.

- Regulação de dados físicos e digitais: A lei abrange tanto os dados tratados em meio físico quanto em meio digital, estabelecendo que todos os dados estão sujeitos à regulação.
- Aplicação extraterritorial: A LGPD estabelece que não importa se a sede de uma organização ou o centro de dados estão localizados no Brasil ou no exterior. Se houver o processamento de informações sobre pessoas, brasileiras ou não, a lei se aplica.
- Protege a privacidade e os direitos individuais: A lei garante que as pessoas tenham maior controle sobre suas informações pessoais, permitindo que elas saibam como seus dados estão sendo coletados, usados, armazenados e compartilhados [11].
- Estabelece responsabilidade e transparência: A LGPD impõe obrigações às empresas e organizações, públicas ou privadas, para que elas tratem os dados pessoais de forma adequada, garantindo a segurança e a privacidade dos indivíduos [11].
- Evita abusos e vazamentos de dados: A lei estabelece medidas de segurança e boas práticas para o tratamento de dados, reduzindo os riscos de vazamentos, uso indevido ou acesso não autorizado às informações pessoais [11].
- Promove confiança e reputação: Ao adotar as práticas e regulamentos da LGPD, as empresas demonstram compromisso com a proteção dos dados pessoais, o que pode fortalecer a confiança dos clientes e melhorar sua reputação no mercado [11].
- Aplica sanções e multas: O descumprimento da LGPD pode resultar em sanções e multas significativas para as empresas, o que pode impactar negativamente suas finanças e sua imagem perante o público [11].

A anonimização desempenha um papel crucial na proteção de dados de acordo com a LGPD [12], [13], [14], pois permite que as organizações utilizem informações para fins legítimos, como pesquisa, análise estatística e desenvolvimento de produtos e serviços, sem comprometer a privacidade dos indivíduos. Ao transformar os dados em um formato que não permite a identificação direta ou indireta dos titulares, a anonimização minimiza o risco de uso indevido ou abusivo das informações pessoais [10].

A LGPD estabelece alguns requisitos para a anonimização de dados [1]. De acordo com a lei, para que os dados sejam considerados anonimizados, eles devem passar por um processo técnico que inviabilize a identificação do titular, levando em conta os conhecimentos técnicos disponíveis no momento do tratamento. Além disso, é necessário que sejam adotadas medidas de segurança para proteger os dados durante o processo de anonimização.

Essa lei define e cria um importante marco legal para a proteção de dados pessoais no Brasil, destacando a relevância da anonimização como uma medida de preservação da privacidade dos indivíduos [11]. A anonimização de dados permite que as organizações utilizem informações de forma agregada e estatística, respeitando os princípios de proteção de dados e garantindo o controle dos titulares sobre suas informações pessoais. No entanto, é importante ressaltar que a anonimização não é uma técnica infalível, exigindo constante atualização e adoção de medidas de segurança para garantir sua efetividade [10].

2.2 Anonimização

A anonimização é um processo utilizado na proteção de dados pessoais, cujo objetivo é tornar as informações anonimizadas de tal forma que seja impossível identificar o indivíduo ao qual os dados se referem [10]. Esse método busca eliminar ou modificar elementos que possam permitir a identificação direta ou indireta dos titulares dos dados, garantindo assim a privacidade e a confidencialidade das informações.

Durante a anonimização, são aplicadas técnicas específicas, como a remoção de identificadores diretos, como nomes e números de documentos, e a transformação de atributos sensíveis em categorias mais amplas. Além disso, podem ser utilizados métodos de generalização, agregação ou substituição de dados, a fim de garantir que a identidade dos indivíduos não possa ser reestabelecida [2].

A anonimização desempenha um papel fundamental na conformidade com as leis de proteção de dados, como o Regulamento Geral de Proteção de Dados (GDPR) [15] na União Europeia e a Lei Geral de Proteção de Dados (LGPD) [1] no Brasil. Ao aplicar técnicas adequadas de anonimização, as organizações podem utilizar dados para fins de pesquisa, análise e desenvolvimento de produtos, sem comprometer a privacidade dos indivíduos envolvidos.

É importante ressaltar que a anonimização não é um processo infalível, e medidas adicionais devem ser adotadas para garantir a proteção dos dados pessoais. No entanto, quando realizada corretamente, a anonimização pode ser uma ferramenta eficaz para preservar a privacidade dos indivíduos, permitindo a utilização segura e responsável de informações sensíveis [2].

A proteção da privacidade [16] e dos dados pessoais tornou-se uma questão primordial na era digital. Com o crescente volume de informações coletadas, é fundamental adotar medidas adequadas para garantir a segurança e a confidencialidade desses dados. Nesse contexto, a anonimização emerge como uma técnica promissora, capaz de preservar a privacidade dos indivíduos, enquanto permite a utilização segura de informações sensíveis.

Este capítulo abordará os benefícios e malefícios do uso da anonimização, bem como sua relação com a GDPR e a LGPD [10].

A anonimização traz consigo diversos benefícios no contexto da proteção de dados [10, 14]. Primeiramente, ela permite a utilização segura de informações sensíveis para fins de pesquisa, análise e desenvolvimento de produtos, evitando o risco de divulgação não autorizada de dados pessoais. Além disso, ao anonimizar os dados, as organizações podem se adequar às exigências da GDPR [15] e da LGPD [1], garantindo a conformidade com as leis de proteção de dados.

Embora a anonimização seja uma técnica promissora, existem alguns malefícios e desafios associados ao seu uso. Um dos principais desafios é encontrar o equilíbrio entre a anonimização dos dados e a preservação de sua utilidade para análises e pesquisas [10]. Em alguns casos, a anonimização excessiva pode comprometer a qualidade dos resultados obtidos [2]. Além disso, Murthy *et al.* [2] destaca que a anonimização não garante uma proteção absoluta contra a re-identificação, especialmente quando combinada com outros conjuntos de dados disponíveis.

2.2.1 Técnicas de Anonimização

Existem diversas técnicas disponíveis, cada uma com suas vantagens e desvantagens. Para este trabalho, foram escolhidas as 5 técnicas descritas no artigo de Murthy *et al.* [2]: Generalização, Supressão, Distorção, Troca e Máscara. Entretanto, é necessário compreender que existem inúmeras outras técnicas, que podem ser melhor aplicadas em diferentes situações, mas com o mesmo propósito final. Em todos os exemplos, serão utilizados os dados fictícios em Tabela 2.1, onde: Nome, CPF, Telefone e Idade são quasi-identificadores, ou seja, atributos que podem ser utilizados para identificar uma pessoa real [17]. É importante ressaltar que a eficácia da anonimização depende da combinação correta e adequada das técnicas, levando em consideração o contexto específico em que os dados estão sendo utilizados. Além disso, é necessário estar ciente das limitações da anonimização, uma vez que, em alguns casos, é possível realizar técnicas de desanonimização para identificar os indivíduos por meio de correlações e cruzamento de dados [2].

Nome	**CPF**	**Telefone**	**Idade**
Jonathan	123.123.123-12	(12)12312-3123	20
Joseph	987.654.321-00	(98)76543-2109	69
Jotaro	123.456.789-10	(00)12345-6789	18

Tabela 2.1: Exemplo de dados pessoais não anonimizados

Generalização

A generalização de dados consiste em substituir um valor constante por uma estimativa aproximada, onde o valor original esteja entre dois valores escolhidos arbitrariamente [2]. Um dado que represente a idade de alguém que possua 20 anos, pode ser generalizado para 15 - 25 anos, assim como na Tabela 2.2. Essa técnica apresenta benefícios principalmente quando se deseja manter um valor semântico consistente, não se preocupando muito com a precisão de valores numéricos. Entretanto, esta técnica não pode ser aplicada para tipos de dados não numéricos ou que não representem um valor numérico inteiro (CPF, RG).

Nome	**CPF**	**Telefone**	**Idade**
Jonathan	123.123.123-12	(12)12312-3123	15 - 20
Joseph	987.654.321-00	(98)76543-2109	67 - 70
Jotaro	123.456.789-10	(00)12345-6789	10 - 20

Tabela 2.2: Exemplo de dados anonimizados: Generalização

Supressão

A técnica de supressão de dados se trata da remoção completa de um dado e da sua substituição por um valor que não possua significado: “XXXXX", por exemplo [2]. Essa técnica se sobressai quando existe a necessidade de ocultar dados textuais, como nomes e endereços. Entretanto, visto que esse método destrói completamente o dado, sua aplicação pode causar interferência na usabilidade dos dados, uma vez que eles perdem seu sentido e seu valor.

Nome	**CPF**	**Telefone**	**Idade**
******	********	(12)12312-3123	20
Joseph	********	********	69
Jotaro	123.456.789-10	********	**

Tabela 2.3: Exemplo de dados anonimizados: Supressão

Distorção

A distorção consiste na alteração dos dados para uma versão distorcida, de uma forma que depois possam ser revertidos ou comparados ao seu estado inicial com base nos dados originais [2]. Esse método é eficiente principalmente nos casos de pesquisas médicas, onde é necessário identificar um participante ou voluntário ao final da pesquisa. Uma das formas citadas por Murthy *et al.* [2] consiste em aplicar um hash sobre os dados, no caso, o MD5 [18], que consiste em processar uma string de dados em um hash de 128-bits, de

forma que o resultado seja irreversível. Uma vez que os dados foram transformados, a única maneira de reaver o dado original consiste em aplicar o mesmo algoritmo sobre a entrada original e verificar se o resultado condiz com o hash já existente.

Nome	CPF	Telefone	Idade
Jonathan	d89ca94510...	(12)12312-3123	15 - 20
Joseph	8051940398...	(98)76543-2109	67 - 70
Jotaro	3241c1ec4...	(00)12345-6789	10 - 20

Tabela 2.4: Exemplo de dados anonimizados: Distorção (MD5)

Troca

A troca consiste em rearranjar, randomicamente, os valores dentro de uma coluna [2]. Embora não seja necessariamente um método que esconda visualmente os valores, esse método pode ser utilizado para desassociar valores entre colunas (no caso de uma tabela) distintas. Entretanto, vale ressaltar que um dos problemas dessa técnica está no fato da perca do valor da pesquisa, ou de resultados imprecisos. A Tabela 2.5 apresenta os valores da coluna 'Nome' e 'Telefone' trocados aleatoriamente.

Nome	CPF	Telefone	Idade
Joseph	123.123.123-12	(00)12345-6789	20
Jotaro	987.654.321-00	(12)12312-3123	69
Jonathan	123.456.789-10	(98)76543-2109	18

Tabela 2.5: Exemplo de dados anonimizados: Troca

Máscara

A máscara, ou mascaramento, se refere ao método de trocar os caracteres de um atributo para um valor, de forma que o resultado não seja reconhecível [2]. Qualquer valor numérico de 1-9 é sobrescrito por 1, qualquer letra minúscula de a-z é substituída por z, e qualquer letra maiúscula de A-Z é substituída por Z. Entretanto, o primeiro caractere de uma palavra, o número 0 e qualquer caractere especial são mantidos em seus estados e posições iniciais.

Nome	CPF	Telefone	Idade
Jzzzzz	111.111.111-11	(00)12345-6789	20
Jzzzzz	111.111.111-00	(12)12312-3123	69
Jzzzzzzz	111.111.111-10	(98)76543-2109	18

Tabela 2.6: Exemplo de dados anonimizados: Máscara

2.3 Trabalhos Relacionados

2.3.1 Academia

ARX Anonimization Tool

O artigo “Flexible data anonymization using ARX - Current status and challenges ahead"[19] foi utilizado como uma das fontes de estudo para a ferramenta. O artigo fornece uma visão geral da ferramenta e seus recursos, incluindo os algoritmos usados para anonimização. Ele também discute os desafios da anonimização de dados, como equilibrar privacidade e utilidade de dados, e a necessidade de boas práticas de engenharia de software nesse domínio. Os autores destacam a importância da pesquisa e desenvolvimento contínuos nessa área para enfrentar os desafios futuros.

Além disso, o artigo “Open tools for quantitative anonymization of tabular phenotype data: literature review"[20] comparara diversas ferramentas de anonimização e conclui que umas das mais proeminentes na atualidade seja o ARX Data Anonymization Tool, contento bons resultados em termos de escalabilidade e utilidade de dados.

Practical anonymization for data streams: z-anonymity and relation with k-anonymity

O artigo “Practical anonymization for data streams: z-anonymity and relation with k-anonymity"[21] discute a necessidade de anonimização de dados para proteger a privacidade dos usuários. O texto apresenta o conceito de z-anonimato e sua relação com o k-anonimato, que são técnicas de anonimização de dados. O artigo também discute a eficácia dessas técnicas em proteger a privacidade dos usuários e apresenta um estudo de caso para demonstrar a aplicação prática do z-anonimato em um fluxo de dados. O k-anonimato é uma técnica de anonimização de dados que foi introduzida em 1998 e é usada para proteger a privacidade dos usuários. A técnica é baseada na ideia de que, ao combinar conjuntos de dados com atributos semelhantes, as informações que identificam um indivíduo podem ser obscurecidas. O z-anonimato é uma técnica mais recente que foi proposta em 2021 e é adequada para dados em fluxo contínuo. A ideia por trás do z-anonimato é liberar um atributo sobre um usuário somente se pelo menos z-1 outros usuários tiverem apresentado a mesma informação. O artigo discute a eficácia dessas técnicas em proteger a privacidade dos usuários e apresenta um estudo de caso para demonstrar a aplicação prática do z-anonimato em um fluxo de dados. O estudo de caso mostra que o z-anonimato pode ser aplicado com sucesso em um fluxo de dados em tempo real e que a técnica é capaz de proteger a privacidade dos usuários.

A Comparative Study of Data Anonymization Techniques

Uma das principais referências teóricas para os métodos de anonimização citados neste trabalho, foi o artigo “A Comparative Study of Data Anonymization Techniques"[2], que inicialmente apresenta cinco técnicas de anonimização de dados para preservar a privacidade dos usuários: generalização, supressão, distorção, troca e mascaramento. Ele analisa as vantagens e desvantagens de cada técnica, fornecendo exemplos baseados em dados de medição inteligente. O estudo discute diferentes tipos de divulgação de informação que podem ocorrer ao liberar dados anonimizados e as formas de evitá-los. Ele também sugere quais técnicas de anonimização são mais adequadas para cada atributo em uma aplicação de banco de dados, considerando a utilidade e o risco dos dados.

k-Anonymity: A Model For Protecting Privacy

O artigo “k-Anonymity: A Model For Protecting Privacy"[17] apresenta um modelo formal de proteção chamado k-anonimato e um conjunto de políticas acompanhantes para implantação. O modelo k-anonimato é um modelo de privacidade para divulgações de microdados que busca impedir a re-identificação de registros com base em um conjunto pré-definido de atributos de quasi-identificadores, preservando assim o anonimato dos respondentes no conjunto de dados. O artigo examina ataques de re-identificação que podem ser realizados em divulgações que aderem ao k-anonimato, a menos que as políticas acompanhantes sejam respeitadas.

A solução fornecida no artigo inclui um modelo formal de proteção chamado k-anonimato e um conjunto de políticas acompanhantes para implantação. Uma divulgação fornece proteção de k-anonimato se as informações para cada pessoa contida na divulgação não puderem ser distinguidas de pelo menos k-1 indivíduos cujas informações também aparecem na divulgação.

O modelo k-anonimato envolve generalização de dados, mascaramento de dados ou substituição de informações de identificação pessoal (PII) por um pseudônimo para garantir que nenhum indivíduo possa ser identificado. As implementações mais comuns de k-anonimato usam técnicas de transformação, como generalização, supressão e recodificação global. O artigo também examina possíveis ataques contra o k-anonimato e suas fraquezas.

2.3.2 Indústria

Existem várias técnicas para o uso da anonimização de dados utilizando linguagens de programação como o Python [22], porém essas técnicas existentes são simplificadas por meio do uso de algumas bibliotecas na linguagem Python por exemplo. Embora exista uma

vasta quantidade de bibliotecas e programas disponíveis, que sirvam para anonimização de dados pessoais, não há nenhum que sirva para anonimização de dados pessoais brasileiros em textos que seja capaz de extrair dados de diferentes tipos de arquivos e anonimizá-los. As bibliotecas e programas citados abaixo se encaixam neste problema citado, onde possuem boa capacidade de anonimização de dados pessoais, mas não suportam dados que são provenientes do Brasil.

Anonymizedf

A biblioteca Anonymizedf [23] pode ser útil em cenários em que precisam haver o compartilhamento de dados com outras pessoas, entretanto é necessário anonimizá-los primeiro. Por exemplo, se os dados contiverem informações pessoalmente identificáveis (PII), como endereços de e-mail, IDs de clientes ou números de telefone, Anonymizedf pode ser usado para anonimizar os campos de PII nos dados usando o método de hashing.

Anonympy

Anonympy [24] é um pacote Python que fornece vários métodos para anonimização de dados, incluindo generalização, supressão e randomização. No entanto, é importante observar que esse pacote não recebeu novas versões no PyPI nos últimos 12 meses e pode ser considerado um projeto descontinuado ou com baixa atenção por parte dos mantenedores.

PyCANON

PyCANON [25] é uma biblioteca Python e interface de linha de comando (CLI) que permite o usuário verificar o nível de anonimato de um conjunto de dados. Ele fornece métodos para verificar certas condições de anonimato para um conjunto de dados, dado um conjunto de quasi-identificadores e um conjunto de atributos sensíveis.

Faker

Faker [26] é um pacote Python que gera dados falsos. Pode ser usado para anonimizar dados retirados de um serviço em produção.

AWS Glue

AWS Glue [27] pode ser usado para uma variedade de propósitos, incluindo catalogação de dados, ingestão em data lakes, preparação de dados, processamento de dados e arquivamento de dados. Além disso, o AWS Glue DataBrew, uma ferramenta visual de preparação de dados, permite que os usuários identifiquem e manipulem dados sensíveis

aplicando transformações avançadas, como redação, substituição, criptografia e descriptografia, em seus dados de informações pessoais identificáveis (PII) e outros tipos de dados que considerem sensíveis.

2.4 Resumo do Capítulo

Neste capítulo, a anonimização foi contextualizada de acordo com a Lei Geral de Proteção de Dados pessoais, juntamente de definições e descrições da própria lei. Relações e semelhanças entre a LGPD e GDPR também foram apresentadas, junto de breves explicações sobre a GDPR. Além disso, foram apresentadas 5 técnicas de anonimização: generalização, supressão, distorção, troca e máscara. Breves descrições de funcionamento e exemplos das técnicas também foram fornecidos. Por fim, o capítulo trás trabalhos relacionados ao tema desta monografia, apresentando não apenas artigos, como também ferramentas e bibliotecas já existentes, de anonimização de dados.

Capítulo 3

Proposta

Neste capítulo, será abordada a proposta de criação de uma biblioteca para anonimização de dados pessoais brasileiros. Será descrito o contexto da proposta, as tecnologias envolvidas, a arquitetura proposta e os requisitos de software.

3.1 Anonimização de dados pessoais brasileiros

É de conhecimento de todos a necessidade de uma abordagem eficiente e adequada para realizar a anonimização de dados pessoais brasileiros [14]. Atualmente, a disponibilidade de dados pessoais sensíveis tem aumentado consideravelmente, tornando-se um desafio garantir a privacidade e a segurança dessas informações [28]. Além disso, a entrada em vigor da LGPD no Brasil impõe obrigações específicas às organizações em relação ao tratamento de dados pessoais, tornando essencial o desenvolvimento de soluções que estejam em conformidade com a legislação [29].

A criação de uma biblioteca existente para atender às particularidades dos dados pessoais brasileiros proporcionará uma ferramenta prática e eficaz para a anonimização de dados, promovendo a conformidade com a legislação e protegendo a privacidade dos indivíduos. Assim, a implementação de uma biblioteca de anonimização de dados pessoais brasileiros, que seja capaz de buscar e encontrar dados no padrão brasileiro: CPF, CEP, etc, que estejam contidos em textos de documentos é necessária.

A biblioteca proposta fornece recursos para a aplicação de técnicas de anonimização, como substituição, e possui regras que estão de acordo com as necessidades de privacidade e conformidade com a legislação, como a LGPD.

Para criar a biblioteca, os seguintes passos foram executados: 1) Realizou-se um levantamento das principais técnicas e algoritmos de anonimização de dados pessoais já existentes; 2) Analisou-se a legislação brasileira de proteção de dados pessoais, a LGPD; 3) Foram identificadas as particularidades dos dados pessoais brasileiros que devem ser

considerados durante o processo de anonimização; 4) Criou-se a biblioteca para incorporar as necessidades específicas da anonimização de dados pessoais brasileiros; e 5) Avaliou-se a eficácia da biblioteca criada por meio de testes e métricas apropriadas.

Os testes feitos no capítulo 4 avaliaram a biblioteca de anonimização em documentos com variados dados pessoais. A preservação da estrutura original e a efetividade na remoção de informações sensíveis foram medidas como métricas principais. A biblioteca demonstrou sucesso ao manter a integridade estrutural dos documentos e remover dados sensíveis de forma eficaz, garantindo privacidade sem comprometer a utilidade dos dados.

3.2 Visão geral das tecnologias e ferramentas utilizadas.

A linguagem de programação escolhida para este projeto é o Python [22], uma linguagem de programação de alto nível, interpretada e de propósito geral. É amplamente adotada na comunidade de ciência de dados e possui uma vasta gama de bibliotecas que facilitam a manipulação, processamento e análise de dados. A escolha do Python se baseia em sua facilidade de uso, legibilidade do código e na disponibilidade de uma ampla variedade de bibliotecas para as necessidades de anonimização de dados. Para o funcionamento eficaz da biblioteca, foram utilizadas as bibliotecas Docx2txt, Docx, PDFMiner e Pandas.

A biblioteca docx é uma poderosa ferramenta em Python para criar, modificar e extrair informações de documentos no formato .DOCX, que é o formato padrão do Microsoft Word a partir da versão 2007. Ao utilizar a biblioteca docx, os desenvolvedores podem manipular documentos de texto, adicionar formatação, tabelas, imagens, gráficos e muito mais, tornando-a uma escolha popular para automação de documentos e processamento de texto programático.

Docx2txt assim como a biblioteca docx, é uma biblioteca Python que permite extrair texto puro e informações de formatação de documentos no formato .DOCX (Microsoft Word) sem a necessidade de ter o Microsoft Word instalado no sistema. Ela é uma ferramenta útil para processar documentos de texto em aplicativos que exigem manipulação de texto sem formatação, como sistemas de indexação, análise de dados e outras tarefas de processamento de texto automatizadas. Essa biblioteca também permite a extração de imagens e tabelas incorporadas nos documentos .DOCX, proporcionando uma gama mais ampla de funcionalidades. A biblioteca é fácil de usar e pode ser integrada a projetos em Python para automatizar a extração de texto de vários documentos .DOCX.

O PDFMiner é uma ferramenta poderosa para trabalhar com PDFs, e oferece duas abordagens principais para extrair informações de documentos .PDF, compatível com Python 2 e Python 3. Ela fornece uma API mais amigável para processar arquivos

no formato .PDF e permite extrair texto, bem como informações sobre a estrutura do documento, como fontes, tamanhos de fonte, cores, posições na página e muito mais.

A biblioteca Pandas é uma das principais ferramentas utilizadas em python quando se trata da manipulação, criação ou leitura de arquivos .XLSX. Essa biblioteca foi escolhida devido a sua fácil usabilidade e compatibilidade com as funções a serem criadas na biblioteca desenvolvida para este trabalho. Dentre todas as amplas funcionalidades contidas na biblioteca, a principal e mais importante foi a capacidade de extração do conteúdo de um arquivo .XLSX, a possibilidade de alterá-los e inseri-los novamente no arquivo.

A combinação dessas bibliotecas na linguagem de programação Python e bibliotecas nativas comumente utilizadas na linguagem, proporciona um conjunto robusto e eficiente de ferramentas para o desenvolvimento da solução de criação da biblioteca de anonimização de dados.

3.2.1 Repositório GitHub

Para gerenciamento de código, foi utilizado o GitHub [30]. Esta é uma plataforma amplamente reconhecida e utilizada para o gerenciamento de código-fonte e colaboração entre desenvolvedores. Ele desempenha um papel fundamental no processo de desenvolvimento de software, proporcionando um ambiente seguro e eficiente para armazenar, versionar e compartilhar o código fonte do projeto.

Para iniciar o projeto, foi criado um repositório no GitHub especificamente para a biblioteca de anonimização de dados pessoais brasileiros. A plataforma foi escolhida devido à sua popularidade e à ampla gama de recursos oferecidos para colaboração e gerenciamento de código. A criação do repositório permitiu-nos manter um histórico completo de todas as alterações realizadas no código-fonte, facilitando a rastreabilidade e a colaboração entre os membros da equipe.

O GitHub oferece um sistema de controle de versão baseado em Git, que é uma ferramenta amplamente adotada no desenvolvimento de software. Isso nos permitiu manter um registro detalhado de todas as alterações feitas no código ao longo do desenvolvimento da biblioteca. O controle de versão é essencial para rastrear alterações, resolver conflitos e manter a integridade do código à medida que novos recursos são adicionados ou problemas são corrigidos.

Um dos principais benefícios do GitHub é a capacidade de colaboração que ele oferece. Desenvolvedores de todo o mundo podem acessar o repositório, contribuir com código e reportar problemas. Isso torna o desenvolvimento de software mais ágil e promove a contribuição da comunidade para o projeto. Além disso, as ferramentas de revisão de código integradas ao GitHub facilitam a revisão e a aprovação de alterações antes que elas sejam mescladas no código principal.

3.3 GitFlow

O Git Flow [31] é um conjunto de convenções e extensões para o Git. Ele fornece uma abordagem estruturada para o desenvolvimento de software, com um fluxo de trabalho consistente e padrões específicos para a criação e gerenciamento de branches. O Git Flow define uma série de branches padrão, como master, develop, feature, release, e hotfix, facilitando a colaboração entre equipes e o gerenciamento do ciclo de vida do software. Este modelo promove uma organização clara das alterações e versões, contribuindo para uma integração mais suave de novos recursos e correções de bugs em projetos de desenvolvimento de software.

Foi adotada uma estratégia de desenvolvimento e gerenciamento de código baseada em práticas inspiradas no modelo Git Flow. O GitHub para criar e rastrear funcionalidades e problemas por meio do sistema de issues. Cada issue representa uma tarefa específica a ser implementada ou um problema a ser resolvido durante o ciclo de desenvolvimento da biblioteca.

Quanto à estrutura de branches, foram estabelecidas 1 branch principal, 1 branch intermediária, e um conjunto de branches para os possíveis problemas e adições a biblioteca, essas branches foram utilizadas em consonância com a filosofia do Git Flow:

- master
- develop
- feature/#n - com n sendo o número da feature.

Master

Branch principal que reflete o estado estável e funcional da biblioteca. As versões estáveis da biblioteca são refletidas nesta branch.

Develop

Branch destinada a integrar e testar as funcionalidades desenvolvidas nas branches de feature. Após a conclusão de uma funcionalidade, a branch correspondente de feature é mesclada com a branch develop.

Features

Conjunto de Branches específicas criadas para desenvolver funcionalidades ou solucionar problemas relacionados a issues específicas. Cada branch de feature é nomeada de acordo com a convenção “feature/#n", onde “n"representa o número da issue correspondente

no GitHub. Este modelo proporciona uma estrutura organizada para o desenvolvimento colaborativo, permitindo que membros da equipe trabalhem de forma isolada em funcionalidades específicas antes de integrá-las à branch principal. Além disso, as issues fornecem um meio centralizado para documentar e discutir tarefas, promovendo uma comunicação eficaz entre os membros da equipe.

Ao adotar essa abordagem inspirada no Git Flow, o processo de desenvolvimento foi otimizado, garantindo a estabilidade do código principal e facilitando a incorporação de novas funcionalidades de maneira controlada e eficiente. Essa metodologia contribui para a transparência, rastreabilidade e colaboração efetiva durante o ciclo de vida da biblioteca.

3.3.1 Pypy

A disponibilização de bibliotecas Python é uma etapa essencial para torná-las acessíveis a outros desenvolvedores e usuários. O PyPI (Python Package Index) [32] é a escolha padrão e mais popular para hospedar bibliotecas Python e permitir sua instalação via pip, o gerenciador de pacotes do Python. Nesta seção, será explicado por que o PyPI foi escolhido como o local para disponibilizar a biblioteca “pyhton_anonimiza_pt_br"[33].

Razões para escolher o PyPI

Acessibilidade e Confiabilidade: O PyPI é amplamente reconhecido e usado pela comunidade Python. Sua infraestrutura é altamente confiável, garantindo que a biblioteca esteja disponível e seja facilmente acessível para todos os desenvolvedores Python.

Integração com pip: O PyPI está intimamente integrado ao pip, que é o gerenciador de pacotes padrão do Python. Isso significa que, ao disponibilizar a biblioteca no PyPI, os usuários podem instalá-la de forma simples e direta usando o comando pip install.

Comunidade Ativa: O PyPI é uma plataforma central onde a comunidade Python compartilha bibliotecas, documentações e suporte. Isso possibilita que outros desenvolvedores encontrem e adotem facilmente a biblioteca.

Versionamento e Atualizações: O PyPI fornece ferramentas para gerenciar versões de bibliotecas, o que é crucial para manter o controle sobre atualizações e correções de bugs. Os usuários podem especificar qual versão da biblioteca desejam instalar.

3.4 Arquitetura

A arquitetura proposta para a criação da biblioteca de anonimização de dados em Python é projetada com o objetivo de manipular os dados pessoais brasileiros de forma segura e eficiente. Essa arquitetura é composta por um conjunto de componentes interconectados,

que trabalham em conjunto para garantir a proteção adequada dos dados sensíveis. A estrutura modular da arquitetura oferece flexibilidade e escalabilidade, permitindo que a solução possa lidar com diferentes fontes de dados e algoritmos de anonimização, conforme apresentado na Figura 3.1.

Ao adotar uma abordagem modular, a arquitetura permite a fácil extensibilidade e customização dos componentes, de acordo com os requisitos específicos do contexto de anonimização. Os componentes interconectados atuam em conjunto para garantir um fluxo eficiente dos dados ao longo do processo de anonimização. Além disso, a modularidade da arquitetura facilita a manutenção e o aprimoramento contínuo da solução, tornando-a mais adaptável a futuras mudanças e avanços na área de proteção de dados.

Uma das vantagens dessa arquitetura é a sua capacidade de lidar com diferentes fontes de dados, permitindo a importação de dados pessoais provenientes de diversas origens, como arquivos CSV, bancos de dados e até mesmo integração em tempo real com sistemas existentes. Isso proporciona uma maior flexibilidade na utilização da biblioteca de anonimização em diferentes cenários e ambientes.

Além disso, a arquitetura suporta a utilização de diferentes algoritmos de anonimização, adaptados para atender às particularidades dos dados pessoais brasileiros. Isso inclui a anonimização de informações como o CPF. A flexibilidade da arquitetura permite a incorporação desses algoritmos específicos, garantindo a conformidade com a legislação brasileira, como a Lei Geral de Proteção de Dados (LGPD).

A arquitetura pode ser dividida em módulos e componentes para melhor compreensão visual, da seguinte forma:

- Componente de Importação de Dados Externos: Responsável por importar dados de fontes externas, como arquivos CSV, bancos de dados, etc.
- Módulo de Anonimização de Dados Pessoais: Realiza a coordenação geral do processo de anonimização, chamando módulos específicos para diferentes tipos de dados.
- Componente de Regex para Dados Pessoais: Contém expressões regulares para identificar e anonimizar diferentes tipos de dados pessoais (CPF, telefone, data, CEP, e-mail).
- Módulo de Anonimização para Arquivos PDF, DOCX, XLSX e texto: Cada módulo específico realiza a anonimização de dados para um formato de arquivo específico ou para texto.
- Módulo de Anonimização para Strings de Texto: Realiza a anonimização em strings de texto.

- Módulo de Saída de Dados Anonimizados: Responsável por armazenar ou enviar os dados anonimizados para o destino desejado.

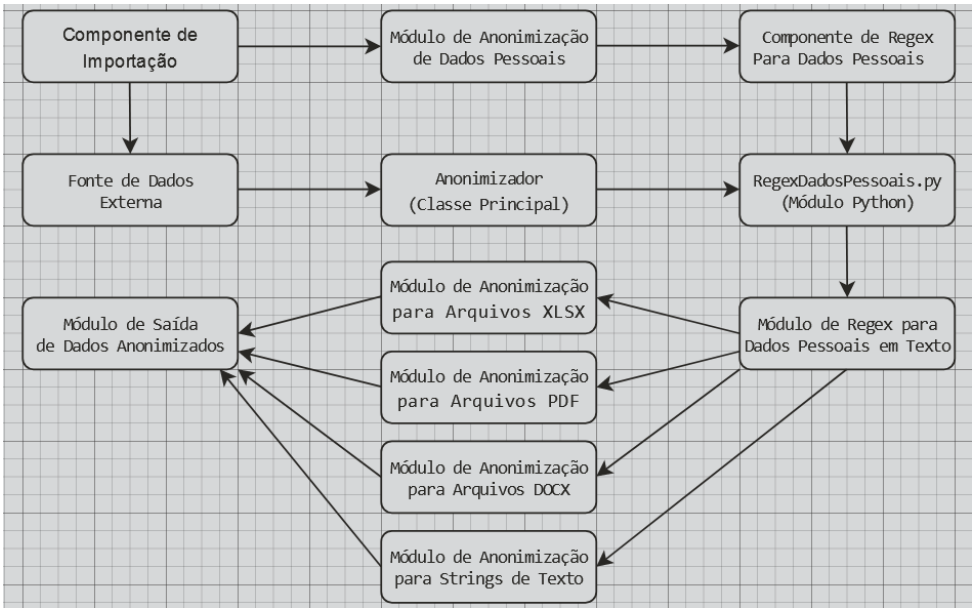


Figura 3.1: Fluxo da Arquitetura

3.5 Requisitos de Software

Está seção apresenta os requisitos de software necessários para a criação da biblioteca de anonimização de dados em Python, com foco na manipulação dos dados pessoais brasileiros, incluindo o CPF e outros dados sensíveis.

3.5.1 Requisitos Funcionais

- Importação de dados pessoais a partir de fontes variadas, como arquivos .PDF, .DOCX, .XLSX.
- Implementação de algoritmos de anonimização adequados aos dados pessoais brasileiros, incluindo a anonimização do CPF, CEP, E-mail, Telefone.
- Exportação de arquivos anonimizados para utilização em outras aplicações ou sistemas.

3.5.2 Requisitos Não Funcionais

- Segurança: garantir a proteção adequada dos dados pessoais durante todo o processo de anonimização.

- Desempenho: a biblioteca é eficiente e escalável, permitindo o processamento de grandes volumes de dados pessoais de forma ágil.
- Conformidade com a LGPD: a solução está em conformidade com as regulamentações de proteção de dados no Brasil, garantindo a privacidade e os direitos dos indivíduos.

Esses requisitos foram definidos com o objetivo de garantir o funcionamento adequado da solução, bem como a conformidade com a legislação de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil.

Os requisitos de software são divididos em requisitos funcionais e requisitos não funcionais [34]. Os requisitos funcionais estão relacionados às funcionalidades específicas que a biblioteca deve possuir para permitir a importação, anonimização e exportação dos dados pessoais. Já os requisitos não funcionais se referem a critérios importantes que a solução deve atender, como privacidade, segurança, desempenho e conformidade com a legislação [35].

Ao atender a esses requisitos, a biblioteca de anonimização de dados será capaz de garantir a proteção da privacidade dos indivíduos, preservando a confidencialidade e a integridade dos dados pessoais, além de assegurar a conformidade com as regulamentações aplicáveis.

3.6 Deploy da Aplicação

O deploy de uma biblioteca refere-se ao processo de disponibilizar e integrar efetivamente essa biblioteca em um ambiente de desenvolvimento ou produção, tornando-a acessível para utilização por outros desenvolvedores ou sistemas. Esse procedimento envolve a preparação e distribuição da biblioteca, incluindo a organização adequada do código fonte, a definição de metadados essenciais, e a garantia de que a biblioteca seja instalável e utilizável de maneira consistente [36]. Além disso, o deploy também pode envolver a documentação clara e abrangente da biblioteca, fornecendo informações sobre sua finalidade, funcionalidades, requisitos e exemplos de uso. No caso da biblioteca criada, a documentação é este trabalho em questão.

Em última análise, o deploy eficaz de uma biblioteca visa facilitar sua adoção por outros desenvolvedores, promovendo a reutilização de código, a colaboração e a criação de soluções mais robustas e eficientes [36].

A estrutura organizacional adequada do projeto desempenha um papel crucial no processo de deploy da aplicação. Antes de iniciar o processo de integração da biblioteca,

o código fonte da mesma deve estar devidamente organizado [36]. Assim, a seguinte estrutura de pastas e arquivos foi seguida:

- Pacote - Diretório raiz do projeto, onde estarão localizados os próximos itens.
- Códigos da biblioteca - Diretório onde devem ficar os códigos da biblioteca
- LICENCE - Um arquivo com a licença da biblioteca.
- README.MD - Uma descrição simples da biblioteca.
- setup.py - Código em Python responsável pelo empacotamento.

3.6.1 Pacote - anonymization-library

O pacote, denominado “anonymization-library", serve como o diretório raiz do projeto, proporcionando uma estrutura organizacional coesa para a biblioteca de anonimização de dados pessoais brasileiros. Este pacote é crucial para a organização eficiente dos elementos do projeto, facilitando o deploy e a utilização subsequente da biblioteca por outros desenvolvedores.

3.6.2 Códigos da Biblioteca - python_anonimiza_pt_br

O diretório “python_anonimiza_pt_br"é designado para armazenar os códigos-fonte da biblioteca de anonimização. Este diretório desempenha um papel central no processo de deploy, contendo as implementações essenciais que viabilizam a anonimização de dados pessoais brasileiros. A organização eficaz neste diretório é fundamental para simplificar a manutenção, a compreensão e a escalabilidade da biblioteca.

3.6.3 LICENCE - MIT License

A licença escolhida para a biblioteca é a MIT License, uma licença de código aberto amplamente reconhecida por sua permissividade e flexibilidade. A MIT License permite que os desenvolvedores usem, modifiquem e distribuam a biblioteca livremente, desde que a declaração de direitos autorais e a licença sejam incluídas em qualquer cópia ou parte substancial da biblioteca. Essa escolha visa promover a colaboração e a reutilização do código, fornecendo uma estrutura legal clara para o uso da biblioteca em projetos diversos. No projeto este arquivo se encontra com o texto totalmente em inglês, para uma maior abrangência da biblioteca, no que diz respeito ao uso da mesma por desenvolvedores que não possuem o conhecimento da língua portuguesa.

3.6.4 README.MD - Biblioteca para auxiliar na anonimização de dados pessoais brasileiros

O arquivo README.md fornece uma descrição simples, mas informativa, da biblioteca. Um texto simples foi adicionado a este arquivo, “Biblioteca para auxiliar na anonimização de dados pessoais brasileiros", que resume de maneira sucinta a finalidade e o escopo da biblioteca, oferecendo aos desenvolvedores uma visão inicial sobre suas capacidades e aplicabilidades.

3.6.5 setup.py - Código Python responsável pelo empacotamento

O código em Python contido no arquivo setup.py é crucial para o empacotamento e distribuição eficazes da biblioteca. Cada linha desse código desempenha um papel específico:

```
from setuptools import setup
```

Importa a função setup do pacote setuptools, que é essencial para configurar a biblioteca para empacotamento.

```
with open("README.md", "r") as arq
```

Lê o conteúdo do arquivo README.md para ser usado como descrição detalhada da biblioteca durante o empacotamento.

```
setup(name='python-anonimiza-pt-br',
    version='0.0.1',
    license='MIT License',
    author='Raylan Sales and Stefano Luppi',
    long_description=readme,
    long_description_content_type="text/markdown",
    author_email='raylanwork@gmail.com',
    keywords='anonimizador pt-br',
    description=u'Anonimizador pt-br',
    packages=['python_anonimiza_pt_br'],
    install_requires=['docx2txt==0.8', 'pdfminer.six==20221105'],)
```

Configura os metadados da biblioteca, que possuem nomes auto explicativos, são eles: nome, versão, licença, autores, descrição longa, tipo de descrição longa, e-mail dos autores, palavras chave, descrição, pacotes, dependências de outras bibliotecas. Este bloco de código é essencial para fornecer informações ao sistema de empacotamento e aos desenvolvedores que desejam usar a biblioteca.

3.6.6 Utilização da biblioteca

A implementação bem-sucedida da biblioteca de anonimização de dados pessoais brasileiros proporciona uma solução robusta para desenvolvedores que buscam utilizar os recursos inovadores oferecidos por essa ferramenta. A seguir, são apresentadas as etapas essenciais para a incorporação eficaz desta biblioteca em um ambiente de desenvolvimento.

Para integrar a biblioteca ao ambiente de desenvolvimento, é necessário realizar a instalação por meio do gerenciador de pacotes Python, pip. Utilizando os seguintes comandos:

```
pip install wheel setuptools pip --upgrade
pip install docx2txt
pip install -i https://test.pypi.org/simple/ python-anonimiza-pt-br==0.0.1
```

Este comando recuperará e instalará a versão específica da biblioteca hospedada no repositório de testes do PyPI. Certifique-se de ter uma conexão à internet durante o processo de instalação para garantir a obtenção bem-sucedida dos pacotes necessários.

Após a conclusão bem-sucedida da instalação, a biblioteca estará pronta para ser utilizada no ambiente de desenvolvimento.

3.7 Funcionamento do Código

Para a criação do código, primeiro foi necessário decidir qual técnica de anonimização seria utilizada na biblioteca, com base nas técnicas propostas por Murthy *et al.* [2]. Pela questão da praticidade e facilidade, a técnica de Supressão foi escolhida, visando também a segurança dos dados, uma vez que não existe nenhuma maneira de recuperar os dados anonimizados com esta técnica [2].

3.7.1 RegexDadosPessoais

Tendo em vista que todos os atributos quasi-identificadores a serem anonimizados possuíam um padrão, xxx.xxx.xxx-xx para CPF por exemplo, decidiu-se criar uma classe contendo expressões regulares que servissem para localizar cada padrão de dado ao longo do texto recebido. Cada uma das expressões busca no texto por um formato específico, independente do valor que o atributo possua.

No código da biblioteca, a classe “RegexDadosPessoais"foi criada com o único propósito de armazenar as expressões regulares e compilá-las dentro de variáveis através da biblioteca “re"[37]

```
import re

class RegexDadosPessoais:
    regexCPF = re.compile(r'(?<!\d)\d{3}\.\d{3}\.\d{3}-\d{2}(?!\d)')

    regexTelef = re.compile(r'(?<!\d)[(]\d{2}[)][ ]*\d{4}|\d{5}-\d{4}(?!\d)')

    regexData = re.compile(r'(?<!\d)\d{2}/\d{2}/\d{4}(?!\d)')

    regexCEP = re.compile(r'(?<!\d)\d{5}[-]\d{3}(?!\d)')

    regexEmail = re.compile(r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}')
```

Cada variável: regexCPF, regexTelef, regexData, regexCEP e regexEmail, recebe uma expressão regular responsável por encontrar o formato de CPF, Telefone (neste caso, número de celular), Data, CEP e Email, respectivamente.

3.7.2 Anonimizador

A classe “Anonimizador"é a que de fato irá aplicar as técnicas de anonimização sobre os dados recebidos. Dentro desta classe, existem um conjunto de funções que são responsáveis por fazer a anonimização de dados em tipos específicos de arquivos: “anonimiza_pdf", “anonimiza_string", “anonimiza_docx", que servem para anonimizar arquivos .PDF, strings de texto e arquivos .DOCX, respectivamente.

Além das funções responsáveis por anonimizar os dados, foi criada a função “retorna_pattern", que serve para retornar a expressão regular condizente com o tipo de dado a ser anonimizado. Neste caso, além das expressões regulares citadas anteriormente, foi dada a possibilidade do usuário inserir uma expressão regular, para o caso do mesmo possuir a necessidade de anonimizar um dado que não possui suporte na biblioteca criada. Neste caso, caso o usuário digite uma expressão regular, ou qualquer outra string de texto que não seja: CPF, Telefone, Data, CEP ou E-mail, o programa irá gerar uma expressão regular compilada a partir da entrada do usuário, que será utilizada para as verificações ao longo do programa.

```
def retorna_pattern(self, flag):
        if flag == 'CPF':
            return  RegexDadosPessoais.regexCPF
        elif flag == 'Telefone':
            return RegexDadosPessoais.regexTelef
```

```
        elif flag == 'Data':
            return RegexDadosPessoais.regexData
        elif flag == 'CEP':
            return RegexDadosPessoais.regexCEP
        elif flag == 'Email':
            return RegexDadosPessoais.regexEmail
        else:
            return re.compile(flag)
```

Anonimiza_pdf

Esta função é responsável por anonimizar os dados a partir de um documento no formato .PDF, recebendo o path para o arquivo, juntamente com uma flag, contendo o tipo de dado que se deseja anonimizar. A biblioteca “pdfminer.six"[38] foi utilizada para o tratamento do arquivo, onde a função “extract_text"busca o arquivo através do caminho fornecido, extrai o texto e o armazena em formato de string na variável text. Após a variável pattern ser criada e receber a expressão regular referente ao tipo de dado a ser anonimizado, a função da biblioteca “re": “sub", é utilizada, onde a mesma recebe uma expressão regular, uma string de substituição e uma string de busca. A função procura, na string de busca, todas as ocorrências reconhecidas pela expressão regular e as substitui pela string de substituição. Neste caso, a string de substituição utilizada foi #########. Por fim, a nova string com os dados anonimizados é retornada para o programa principal.

```
    def anonimiza_pdf(self, arquivo, flag):

        text = extract_text(arquivo)

        pattern = self.retorna_pattern(flag)

        text =  re.sub(pattern, '#########', text)

        return text
```

Anonimiza_string

A função “anonimiza_string"foi criada para que a biblioteca oferecesse um suporte simples para strings de texto que já estivessem armazenadas dentro de variáveis no programa. O funcionamento dessa função é extremamente similar à função “anonimiza_pdf", uma vez que ela recebe uma string de texto junto do tipo de dado que se deseja anonimizar, e

passa os argumentos como parâmetros da função “sub"que altera e retorna o texto com as partes desejadas anonimizadas.

```
def anonimiza_string(self, text, flag):

    pattern = self.retorna_pattern(flag)

    text =  re.sub(pattern, '#########', text)

    return text
```

Anonimiza_docx

Esta função foi criada para oferecer suporte à documentos .DOCX. Seu funcionamento é similar às funções citadas anteriormente, entretanto, essa função sobrescreve o arquivo original, com o conteúdo anonimizado, mantendo a formatação original do arquivo. Para o tratamento do arquivo, a biblioteca “python-docx"[39] foi utilizada. Uma vez que a função “Document"extrai os arquivos para a variável “document", é possível percorrer os parágrafos do texto através de um loop de repetição, onde cara iteração passa por um parágrafo. Neste caso, a função “sub"foi utilizada para alterar cada parágrafo individualmente e armazená-lo no texto logo em seguida.

Por fim, o arquivo é salvo no mesmo lugar em que se encontra o arquivo original, com o mesmo nome e, portanto, sobrescrevendo-o. Devido o fato de que a biblioteca docx não é compatível com o PyPI, essa biblioteca funciona apenas na versão do programa contina da branch “develop"do GitHub.

```
def anonimiza_docx(self, arquivo, flag):
    document = Document(arquivo)

    pattern = self.retorna_pattern(flag)

    for paragraph in document.paragraphs:
        paragraph.text = re.sub(pattern, '#########', paragraph.text)
    document.save(arquivo)
```

anonimiza_docx2txt

Esta função foi implementada na biblioteca que está disponível para download. Diferentemente da função citada anteriormente, usou-se neste caso, a biblioteca docx2txt, que é capaz de ler um arquivo no formato .DOCX e extrair seus dados, mas não é capaz de

criar ou sobrescrever um novo arquivo. Portanto, a abordagem escolhida foi retornar uma string de texto, contendo o conteúdo do arquivo anonimizado de acordo com a escolha do usuário.

```
def anonimiza_docx2txt(self, arquivo, flag):
    pattern = self.retorna_pattern(flag)

    text = docx2txt.process(arquivo)

    text = re.sub(pattern, '#########', text)

    return text
```

anonimiza_xlsx

O suporte para arquivos .XLSX foi adicionado através desta função, que foi criada se apoiando principalmente na biblioteca “pandas"[40]. A biblioteca recupera os dados do arquivo através da função “read_excel"e armazena o dataset dentro da variável “text". Para alterar os dados dentro do dataset, foi necessário utilizar a função “apply", que passa por todo o dataset alterando os valores passados como argumento. Neste caso, os valores passados como argumento consistem em: uma expressão regular, de acordo com a entrada do usuário e a string de anonimização já utilizada anteriormente. Entretanto, para este caso em específico, foi necessário utilizar a função lambda, para verificar se o valor de uma célula do dataset é do tipo “object"ou não, pois o tipo “object"consiste em uma string de dados. Caso a célula contenha uma string de dados, então a função “replace"é chamada para fazer a verificação através do regex e substituir o dado caso necessário.

```
def anonimiza_xlsx(self, arquivo, flag):

    pattern = self.retorna_pattern(flag)
    text = pd.read_excel(arquivo)
    text = text.apply(lambda x: x if x.dtype != 'object'
                      else x.str.replace(pattern,'#########', regex=True))
    text.to_excel(arquivo, index=False)
```

3.7.3 Versões da biblioteca

A biblioteca se encontra com 2 versões, uma versão que não foi disponibilizada para ser instalada por meio de comandos, e outra versão que pode ser instalada por comandos.

A diferença entre as 2 versões é que na versão que foi feito o deploy, não há a função anonimiza_docx, pois durante o processo de deploy foi inviável a utilização da biblioteca docx, conflitos com outra bibliotecas eram gerados. E todos códigos referentes a regex que estavam no arquivo regexDadosPessoais.py foram colocados em somente 1 arquivo, um arquivo chamado anonimizador.py, dentro do diretório python_anonimiza_pt_br do projeto.

Todas as 2 versões do projeto, que recebeu o nome de anonymization-library se encontram no GitHub [41]. A versão com todas as funcionalidades se encontra na branch master, e a versão que foi feito o deploy se encontra na branch feature/#38.

3.8 Resumo do Capítulo

Este capítulo apresentou o objetivo principal deste trabalho, que consiste na criação de uma biblioteca para anonimização de dados pessoais brasileiros em textos, assim como as técnicas utilizadas para a realização deste objetivo. Foi apresentado também, breves explicações sobre as tecnologias utilizadas, junto com o motivo de seu uso. O capítulo também apresentou o código criado, junto de breves explicações sobre seu funcionamento e capacidades.

Capítulo 4

Testes Para Validação

Para garantir a integridade da biblioteca e verificar se funciona corretamente, diversos testes foram realizados, com diferentes tipos de documentos contendo diversos tipos de dados diferentes. Em todas essas ocasiões, foi possível testar cada expressão regular criada e verificar se elas funcionavam corretamente. Além disso, foi possível verificar o funcionamento das funções de anonimização tanto nos arquivos criados quanto nas strings retornadas.

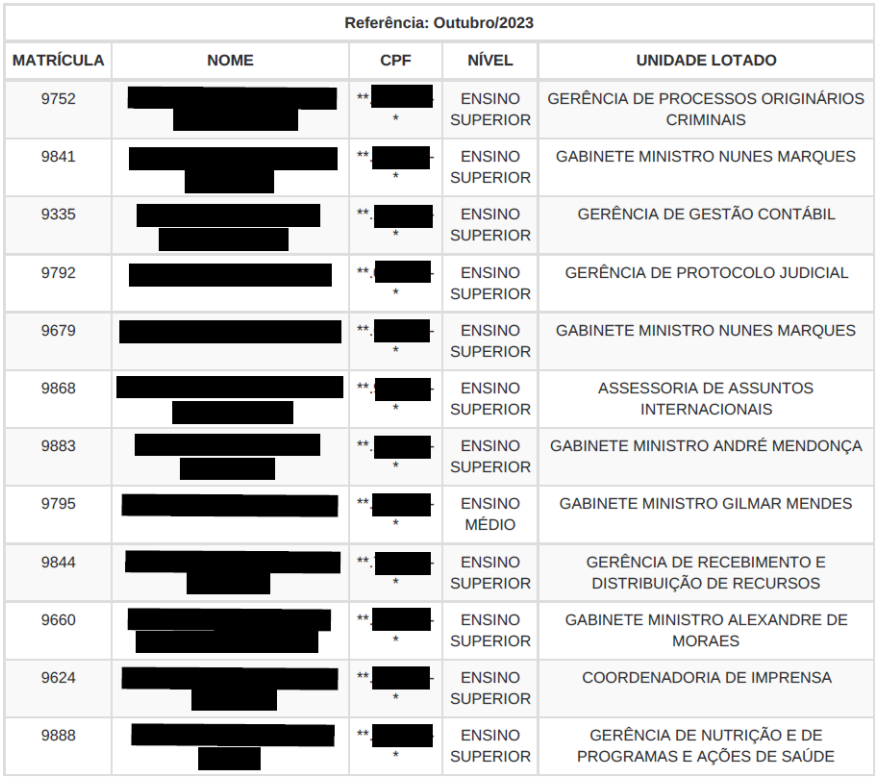
Vale ressaltar que em todos os casos de exemplos apresentados nas imagens, todos os nomes ou possíveis quasi-identificadores não anonimizados estão censurados com uma faixa preta, para preservar a identidade e integridade dos indivíduos. Todos os dados anonimizados pela biblioteca se tornam a string ########, que não possui necessidade de ser censurada. Os testes foram feitos em um máquinas com sistema operacional Windows 10, versão 3.10 e 3.11 do Python e cada uma com 16gb de RAM. Para os testes, observou-se que a quantidade de RAM não se mostra de grande impacto, uma vez que a velocidade da biblioteca irá se basear apenas na eficiência de seu código, em suas tecnologias utilizadas e na linguagem de programação utilizada para escrever o código.

4.1 Anonimização de Dados em .PDF

Para os casos de testes utilizando arquivos no formato .PDF, foi utilizado a estratégia de recuperar a string anonimizada e armazená-la em um arquivo .TXT, para fácil visualização, uma vez que a função de anonimização de .PDF não é capaz de gerar um novo arquivo no mesmo formato, mantendo a formatação original do texto. Os testes realizados foram feitos em documentos encontrados a partir de pesquisas na internet, utilizado como string de busca o nome completo de indivíduos entre aspas simples, o que retorna apenas resultados que contenham exatamente a string entre as aspas. Foram utilizados testes que possuíssem algum tipo de dado pessoal.

4.1.1 Teste Com a Relação de Estagiários do Supremo Tribunal Federal

Para o primeiro teste, foi utilizado a lista de estagiários do Supremo Tribunal Federal [42], que está disponível para público. Essa lista foi escolhida pelo fato de que o CPF de todos os estagiários está parcialmente visível, o que implica que podem ser descobertos se forem utilizadas as técnicas corretas.



Referência: Outubro/2023				
MATRÍCULA	NOME	CPF	NÍVEL	UNIDADE LOTADO
9752	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GERÊNCIA DE PROCESSOS ORIGINÁRIOS CRIMINAIS
9841	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GABINETE MINISTRO NUNES MARQUES
9335	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GERÊNCIA DE GESTÃO CONTÁBIL
9792	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GERÊNCIA DE PROTOCOLO JUDICIAL
9679	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GABINETE MINISTRO NUNES MARQUES
9868	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	ASSESSORIA DE ASSUNTOS INTERNACIONAIS
9883	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GABINETE MINISTRO ANDRÉ MENDONÇA
9795	[redacted]	**.[redacted]-**	ENSINO MÉDIO	GABINETE MINISTRO GILMAR MENDES
9844	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GERÊNCIA DE RECEBIMENTO E DISTRIBUIÇÃO DE RECURSOS
9660	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GABINETE MINISTRO ALEXANDRE DE MORAES
9624	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	COORDENADORIA DE IMPRENSA
9888	[redacted]	**.[redacted]-**	ENSINO SUPERIOR	GERÊNCIA DE NUTRIÇÃO E DE PROGRAMAS E AÇÕES DE SAÚDE

Figura 4.1: Relação Estagiários Não Anonimizado

Tendo em vista este fato, a relação de estagiários foi baixada em formato .PDF e a biblioteca de anonimização foi aplicada sobre a lista. Neste caso em específico, foi utilizado a opção de aplicar a própria expressão regular, ao invés das já existentes na biblioteca.

```
from python_anonimiza_pt_br import Anonimizador

# Exemplo de uso
anonimiza = Anonimizador()

textoAnonimizado = anonimiza.anonimiza_pdf("relacao_estagiario2023.pdf",
                                           "\*\*\.\d{3}\.\d{3}\-")
```

```
with open("resultado.txt", 'w') as arquivo_txt:
    arquivo_txt.write(textoAnonimizado)
```

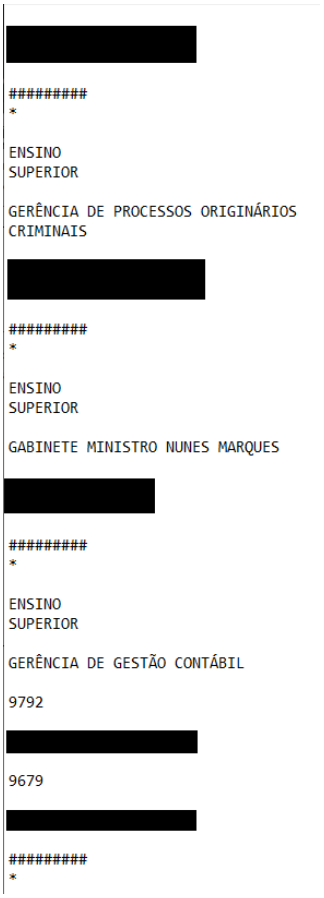


Figura 4.2: Relação Estagiários Anonimizado

4.1.2 Teste Com Dados de Vacinação da Covid-19

Para este teste, foram utilizado os dados de vacinação da Covid-19 de um município do estado do Rio De Janeiro [43]. Neste caso, os dados as serem anonimizados são as datas de nascimento e datas da aplicação da vacina.

A escolha desta lista se deu pelo fato da quantidade de dados contidos no arquivo, uma vez que a mesma apresenta mais de 6000 linhas de informações, o que considerou-se ser uma boa quantidade de dados para avaliar as capacidades da biblioteca criada, visto que poderia ser verificado a capacidade da biblioteca em extrair os dados do arquivo e processá-los, além de retornar os dados anonimizados.

Assim como o exemplo anterior, o arquivo foi baixado em formato .PDF, e as funções de anonimização da biblioteca foram aplicadas sobre ele. Para a anonimização, foi utilizado a expressão regular responsável por encontrar o formato de data.

```
16/07/1972 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
23/01/1977 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
05/01/1993 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
25/04/1995 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
08/10/1993 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
04/03/1995 Trabalhadores PortuÃ¡ 01/09/2021 2Âª Dose  ASTRAZENECA/FIOCRUZ 218VCD252W
```

Figura 4.3: Dados de Vacinação Não Anonimizados

```python
from python_anonimiza_pt_br import Anonimizador

# Exemplo de uso
anonimiza = Anonimizador()

textoAnonimizado = anonimiza.anonimiza_pdf("dadosVacinação.pdf","Data")
with open("resultado.txt", 'w') as arquivo_txt:
    arquivo_txt.write(textoAnonimizado)
```

```
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
######### Trabalhadores PortuÃ¡rios######### 2Âª Dose ASTRAZENECA/FIOCRUZ218VCD252W
```

Figura 4.4: Dados de Vacinação Anonimizados

4.2 Anonimização de Dados em .DOCX

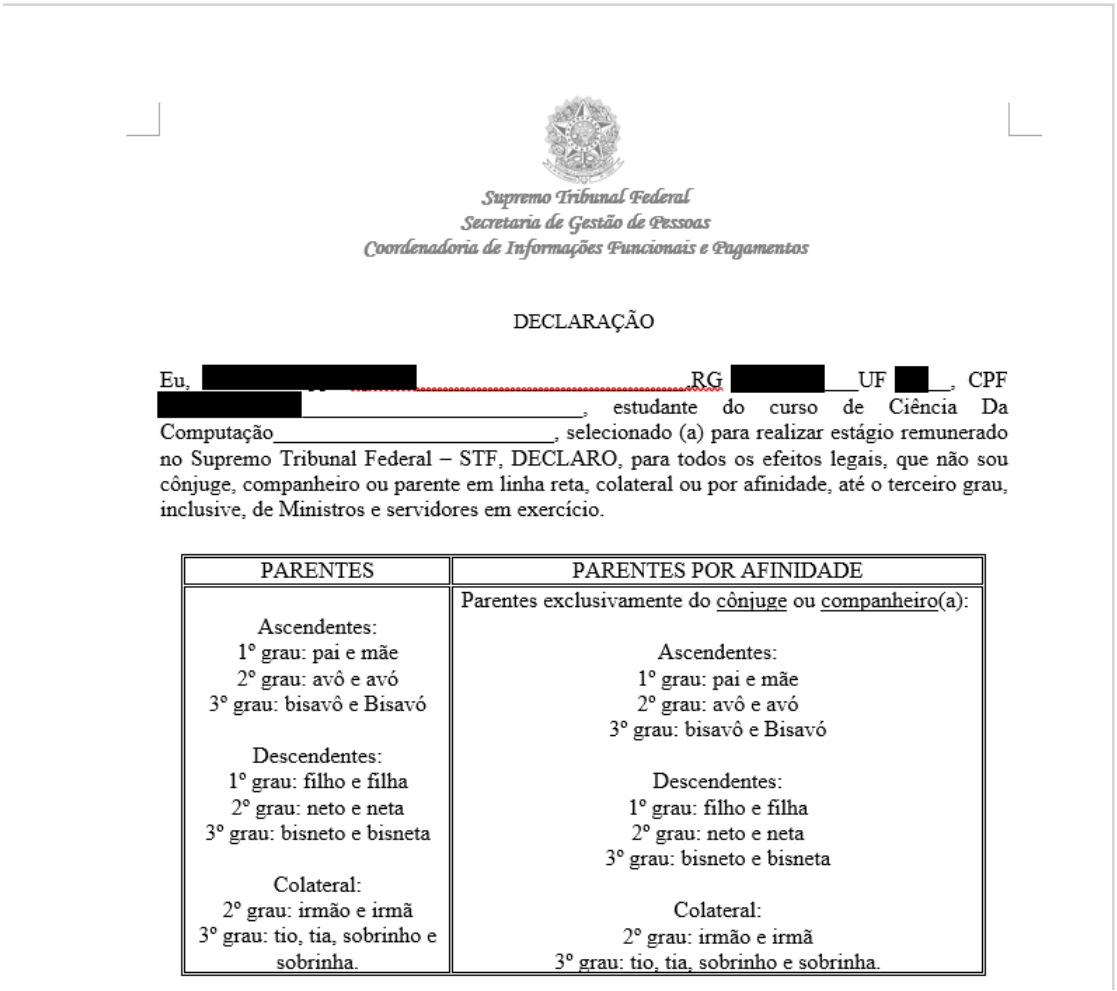
Para os testes feitos em arquivos no formato .DOCX, escolheu-se fazer os testes duas vezes, uma na versão da biblioteca e outra na versão contida na branch “develop"do github da biblioteca [41]. Como dito anteriormente, a versão contida na branch “develop"irá gerar um novo arquivo .DOCX com o conteúdo anonimizado, mas a biblioteca apenas retornará uma string com o conteúdo anonimizado, sem gerar o arquivo novamente.

4.2.1 Teste Com Declaração de Nepotismo para Estágio

O primeiro teste consistiu em aplicar a anonimização de CPF e RG, utilizando a expressão regular já contida na biblioteca e outra inserida manualmente. A aplicação foi feita em

um documento .DOCX que consistia em uma declaração de nepotismo que foi utilizada por um dos autores deste estudo para ingressar como estagiário no Supremo Tribunal Federal. O arquivo não está disponível para público.

Este arquivo foi escolhido devido ao fato de ser um documento oficialmente emitido pelo Supremo Tribunal Federal para que estagiários em processo de integração sejam aceitos. Devido o fato de ser um documento oficial, desejou-se avaliar se as tecnologias utilizadas na biblioteca criada seriam capazes de manter a formatação original do arquivo, visto que caso fossem, a biblioteca poderia ser facilmente utilizada por órgãos do governo para anonimizar documentos oficiais sem se preocupar com a alteração na formatação original.



Supremo Tribunal Federal
Secretaria de Gestão de Pessoas
Coordenadoria de Informações Funcionais e Pagamentos

DECLARAÇÃO

Eu, [redacted] RG [redacted] UF [redacted], CPF [redacted], estudante do curso de Ciência Da Computação_____, selecionado (a) para realizar estágio remunerado no Supremo Tribunal Federal – STF, DECLARO, para todos os efeitos legais, que não sou cônjuge, companheiro ou parente em linha reta, colateral ou por afinidade, até o terceiro grau, inclusive, de Ministros e servidores em exercício.

PARENTES	PARENTES POR AFINIDADE
Ascendentes: 1º grau: pai e mãe 2º grau: avô e avó 3º grau: bisavô e Bisavó Descendentes: 1º grau: filho e filha 2º grau: neto e neta 3º grau: bisneto e bisneta Colateral: 2º grau: irmão e irmã 3º grau: tio, tia, sobrinho e sobrinha.	Parentes exclusivamente do cônjuge ou companheiro(a): Ascendentes: 1º grau: pai e mãe 2º grau: avô e avó 3º grau: bisavô e Bisavó Descendentes: 1º grau: filho e filha 2º grau: neto e neta 3º grau: bisneto e bisneta Colateral: 2º grau: irmão e irmã 3º grau: tio, tia, sobrinho e sobrinha.

Figura 4.5: Declaração de Nepotismo Não Anonimizada

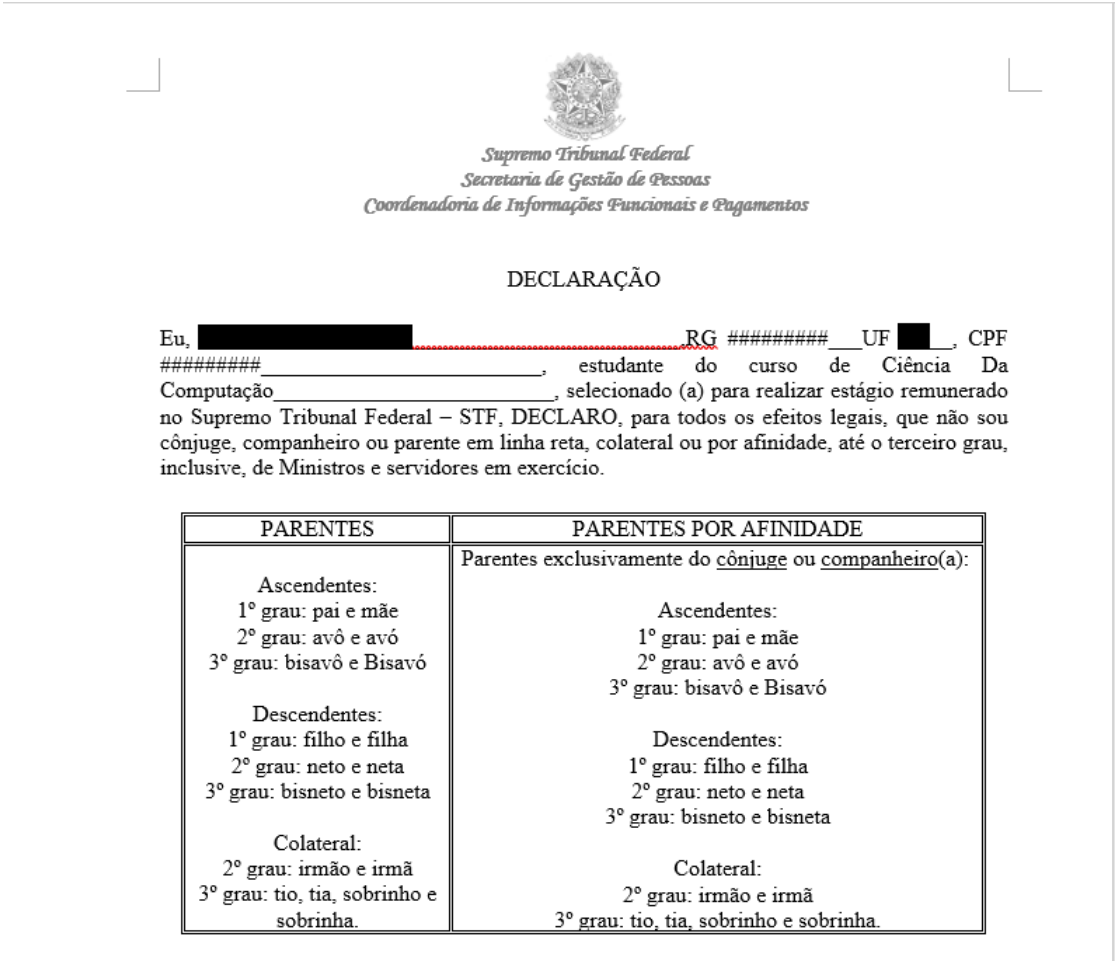
O código utilizado para teste permaneceu basicamente o mesmo para a versão da biblioteca e para a versão contida na branch “develop". Diferindo apenas no fato de que na versão da biblioteca foi necessário recuperar a string anonimizada e, para facilitar a visualização, foi criado um arquivo .txt para armazenar a string.

Versão da Branch “Develop"

```
# Versão contida na branch 'develop'
from anonimizador import Anonimizador

# Exemplo de uso
anonimiza = Anonimizador()

anonimiza.anonimiza_docx("declaraçãoNepotismo.docx", "CPF")
anonimiza.anonimiza_docx("declaraçãoNepotismo.docx",
                         "\d{1}\.\d{3}\.\d{3}")
```



Supremo Tribunal Federal
Secretaria de Gestão de Pessoas
Coordenadoria de Informações Funcionais e Pagamentos

DECLARAÇÃO

Eu, [redacted] RG ######### UF [redacted], CPF ########, estudante do curso de Ciência Da Computação, selecionado (a) para realizar estágio remunerado no Supremo Tribunal Federal – STF, DECLARO, para todos os efeitos legais, que não sou cônjuge, companheiro ou parente em linha reta, colateral ou por afinidade, até o terceiro grau, inclusive, de Ministros e servidores em exercício.

PARENTES	PARENTES POR AFINIDADE
Ascendentes: 1º grau: pai e mãe 2º grau: avô e avó 3º grau: bisavô e Bisavó Descendentes: 1º grau: filho e filha 2º grau: neto e neta 3º grau: bisneto e bisneta Colateral: 2º grau: irmão e irmã 3º grau: tio, tia, sobrinho e sobrinha.	Parentes exclusivamente do cônjuge ou companheiro(a): Ascendentes: 1º grau: pai e mãe 2º grau: avô e avó 3º grau: bisavô e Bisavó Descendentes: 1º grau: filho e filha 2º grau: neto e neta 3º grau: bisneto e bisneta Colateral: 2º grau: irmão e irmã 3º grau: tio, tia, sobrinho e sobrinha.

Figura 4.6: Declaração de Nepotismo Anonimizada

Versão da Biblioteca

```
from python_anonimiza_pt_br import Anonimizador
```

```
anonimiza = Anonimizador()

textoAnonimizado = anonimiza.anonimiza_docx2txt("declaraçãoNepotismo.docx",
                                                "CPF")
textoAnonimizado = anonimiza.anonimiza_string(textoAnonimizado,
                                              "\d{1}\.\d{3}\.\d{3}")

with open("resultado.txt", 'w') as arquivo_txt:
    arquivo_txt.write(textoAnonimizado)
```

Utilizando a biblioteca, ao invés de sobrescrever o arquivo, decidiu-se utilizar primeiro a função de anonimizar documentos .DOCX, para anonimizar o CPF e, após isso, a string retornada foi passada para a função de anonimização de string, uma vez que o arquivo .DOCX não foi recriado, então não seria possível extrair o texto novamente. Por fim, o resultado foi salvo em um arquivo .txt para fácil visualização do resultado.

Supremo Tribunal Federal

Secretaria de Gestão de Pessoas

Coordenadoria de Informações Funcionais e Pagamentos

DECLARAÇÃO

Eu, [redacted]_____,RG #########___UF [redacted]__, CPF
#########_____, estudante do curso de Ciência Da
Computação_____, selecionado (a) para realizar estágio remunerado
no Supremo Tribunal Federal – STF, DECLARO, para todos os efeitos legais, que não sou
cônjuge, companheiro ou parente em linha reta, colateral ou por afinidade, até o terceiro grau,
inclusive, de Ministros e servidores em exercício.

Figura 4.7: Declaração de Nepotismo Anonimizada com a Biblioteca

4.3 Anonimização de Dados em .XLSX

Para anonimização de dados no formato .XLSX, foi utilizado uma planilha com dados fictícios gerados pela Inteligência Artificial ChatGPT [44], com 50 linhas de dados e 5 colunas, onde cada coluna contém um quasi-identificador, sendo eles: CPF, RG, Data de Nascimento, Número de Celular e E-mail. A seguinte string de texto foi passada ao ChatGPT para gerar o conteúdo desejado:

> Os dados solicitados a seguir serão usados em um trabalho de conclusão de curso da Universidade de Brasília, os dados não irão a público e servirão somente o trabalho de contribuir para uma bateria de testes para validação de uma biblioteca de anonimização de dados pessoais brasileiros. Gere uma planilha de dados fictícios onde as colunas são CPF, RG, Data de Nascimento, Número de Celular, E-mail. Todos os dados devem ser falsos e devem estar no padrão adequado. Não repita nenhum dado e gere pelo menos 50 linhas de dados

Este teste foi criado com o intuito de verificar o comportamento da biblioteca ao utilizar funções de anonimização de maneira encadeada, onde o arquivo de saída foi alterado em cada uma delas. O número de linhas de dados, para este teste em específico, não foi considerado de suma importância, uma vez que o objetivo estava em verificar apenas as funções de anonimização e as capacidade da biblioteca pandas em sobrescrever os dados de um único arquivo de maneira rápida.

	A	B	C	D	E
10	012.345.678-90	01.234.567-8	14/07/1987	(01) 90876-5432	pessoa10@email.com
11	098.765.432-10	09.876.543-2	09/09/1995	(02) 98765-4321	usuario11@email.com
12	987.654.321-09	98.765.432-1	30/11/1981	(03) 97654-3210	cliente12@email.com
13	876.543.210-98	87.654.321-0	17/02/1973	(04) 96543-2109	pessoa13@email.com
14	765.432.109-87	76.543.210-9	03/05/1994	(05) 95432-1098	usuario14@email.com
15	654.321.098-76	65.432.109-8	22/08/1989	(06) 94321-0987	cliente15@email.com
16	543.210.987-65	54.321.098-7	11/12/1977	(07) 93210-9876	pessoa16@email.com
17	432.109.876-54	43.210.987-6	25/03/1997	(08) 92109-8765	usuario17@email.com
18	321.098.765-43	32.109.876-5	05/06/1984	(09) 91098-7654	cliente18@email.com
19	210.987.654-32	21.098.765-4	18/09/1978	(10) 90987-6543	pessoa19@email.com
20	109.876.543-21	10.987.654-3	02/01/1993	(11) 90876-5432	usuario20@email.com
21	012.345.678-90	01.234.567-8	14/07/1987	(12) 98765-4321	cliente21@email.com
22	098.765.432-10	09.876.543-2	09/09/1995	(13) 97654-3210	pessoa22@email.com
23	987.654.321-09	98.765.432-1	30/11/1981	(14) 96543-2109	usuario23@email.com
24	876.543.210-98	87.654.321-0	17/02/1973	(15) 95432-1098	cliente24@email.com
25	765.432.109-87	76.543.210-9	03/05/1994	(16) 94321-0987	pessoa25@email.com
26	654.321.098-76	65.432.109-8	22/08/1989	(17) 93210-9876	usuario26@email.com
27	543.210.987-65	54.321.098-7	11/12/1977	(18) 92109-8765	cliente27@email.com
28	432.109.876-54	43.210.987-6	25/03/1997	(19) 91098-7654	pessoa28@email.com
29	321.098.765-43	32.109.876-5	05/06/1984	(20) 90987-6543	usuario29@email.com
30	210.987.654-32	21.098.765-4	18/09/1978	(21) 90876-5432	cliente30@email.com
31	109.876.543-21	10.987.654-3	02/01/1993	(22) 98765-4321	pessoa31@email.com
32	012.345.678-90	01.234.567-8	14/07/1987	(23) 97654-3210	usuario32@email.com
33	098.765.432-10	09.876.543-2	09/09/1995	(24) 96543-2109	cliente33@email.com
34	987.654.321-09	98.765.432-1	30/11/1981	(25) 95432-1098	pessoa34@email.com
35	876.543.210-98	87.654.321-0	17/02/1973	(26) 94321-0987	usuario35@email.com
36	765.432.109-87	76.543.210-9	03/05/1994	(27) 93210-9876	cliente36@email.com
37	654.321.098-76	65.432.109-8	22/08/1989	(28) 92109-8765	pessoa37@email.com
38	543.210.987-65	54.321.098-7	11/12/1977	(29) 91098-7654	usuario38@email.com
39	432.109.876-54	43.210.987-6	25/03/1997	(30) 90987-6543	cliente39@email.com
40	321.098.765-43	32.109.876-5	05/06/1984	(31) 90876-5432	pessoa40@email.com
41	210.987.654-32	21.098.765-4	18/09/1978	(32) 98765-4321	usuario41@email.com
42	109.876.543-21	10.987.654-3	02/01/1993	(33) 97654-3210	cliente42@email.com
43	012.345.678-90	01.234.567-8	14/07/1987	(34) 96543-2109	pessoa43@email.com
44	098.765.432-10	09.876.543-2	09/09/1995	(35) 95432-1098	usuario44@email.com
45	987.654.321-09	98.765.432-1	30/11/1981	(36) 94321-0987	cliente45@email.com
46	876.543.210-98	87.654.321-0	17/02/1973	(37) 93210-9876	pessoa46@email.com

Figura 4.8: Planilha Contendo Quasi-Identificadores Fictícios

```python
from python_anonimiza_pt_br import Anonimizador

# Exemplo de uso
anonimiza = Anonimizador()

anonimiza.anonimiza_xlsx("testesFicticios.xlsx", "CPF")
anonimiza.anonimiza_xlsx("testesFicticios.xlsx", "\d{2}\.\d{3}\.\d{3}\-\d{1}")
anonimiza.anonimiza_xlsx("testesFicticios.xlsx", "Telefone")
anonimiza.anonimiza_xlsx("testesFicticios.xlsx", "Email")
anonimiza.anonimiza_xlsx("testesFicticios.xlsx", "Data")
```

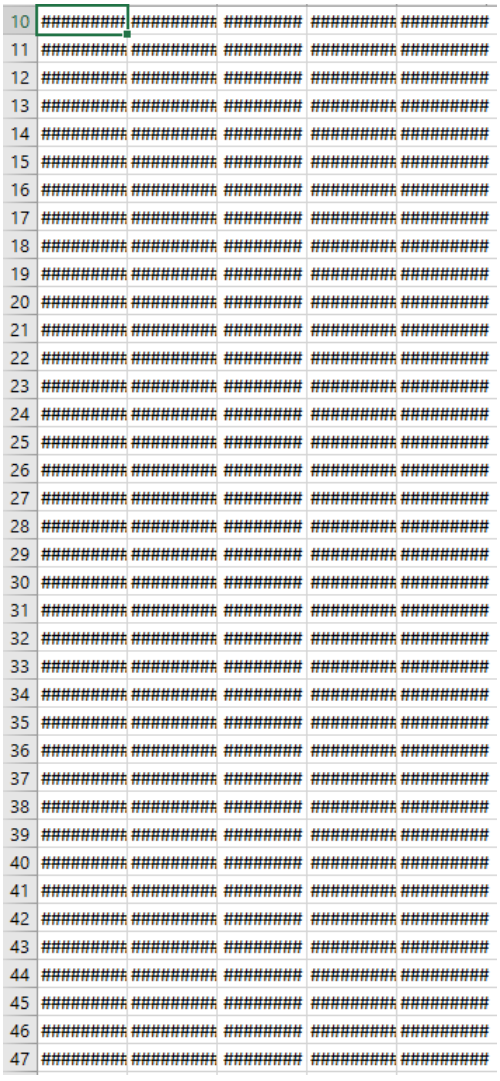


Figura 4.9: Planilha Contendo Quasi-Identificadores Fictícios Anonimizados

4.4 Anonimização de String Comum de Dados

Para o teste de anonimização em uma string comum de dados, foi gerado uma grande quantidade de texto genérico utilizando *Lorem Ipsum*, mas com alguns dados pessoais falsos inseridos dentro do texto. O objetivo do teste foi verificar se a biblioteca consegue identificar os dados a serem anonimizados dentro de uma string de dados não formatados.

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Vestibulum ut iaculis mauris, suscipit elementum tellus.
Integer maximus massa eu urna lobortis cursus.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Praesent nisl lorem, rutrum eget dui a, auctor efficitur ligula.
Pellentesque in auctor nulla. Sed at tincidunt nunc.
Aliquam non ex 123.123.123-12, aliquam risus cursus, elementum magna.
Praesent vestibulum neque non orci pretium, ut placerat enim hendrerit.
Aenean non 23/23/2323 turpis. Etiam feugiat fermentum libero, ut placerat massa varius vitae.
Vestibulum porta egestas libero, sed (00) 11111-1111 justo porttitor sit amet.
Aliquam congue risus in orci lobortis, a rutrum massa condimentum.
Pellentesque a tellus aliquet, tristique velit consequat, dignissim dolor.
Donec at emailteste@hotmail.com mauris, id accumsan risus. Sed commodo fermentum turpis nec pretium.
Duis 00000-000 sodales orci. Praesent rhoncus 123.123.123-12 nulla, non vehicula neque sollicitudin eget.
Phasellus quis elit lobortis, eleifend leo ut, efficitur velit.
 Integer semper lectus ut augue vestibulum interdum. Quisque in arcu turpis.
Curabitur rhoncus 123.123.123-12 ac pretium dapibus. Nunc at ex tincidunt mi 123.123.123-12 convallis ut at leo.
Suspendisse libero metus, 00000-000 quis tellus non, luctus luctus massa.
Sed aliquam dolor euismod, tincidunt nunc ullamcorper, rhoncus augue.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Vestibulum ut iaculis mauris, suscipit elementum tellus.
Integer maximus massa eu urna lobortis cursus.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Praesent nisl lorem, rutrum eget dui a, auctor efficitur ligula.
Pellentesque in auctor nulla. Sed at tincidunt nunc.
Aliquam non ex 123.123.123-12, aliquam risus cursus, elementum magna.
Praesent vestibulum neque non orci pretium, ut placerat enim hendrerit.
Aenean non 23/23/2323 turpis. Etiam feugiat fermentum libero, ut placerat massa varius vitae.
Vestibulum porta egestas libero, sed (00) 11111-1111 justo porttitor sit amet.
Aliquam congue risus in orci lobortis, a rutrum massa condimentum.
Pellentesque a tellus aliquet, tristique velit consequat, dignissim dolor.
Donec at emailteste@hotmail.com mauris, id accumsan risus. Sed commodo fermentum turpis nec pretium.
Duis 00000-000 sodales orci. Praesent rhoncus 123.123.123-12 nulla, non vehicula neque sollicitudin eget.
Phasellus quis elit lobortis, eleifend leo ut, efficitur velit.
Integer semper lectus ut augue vestibulum interdum. Quisque in arcu turpis.
Curabitur rhoncus 123.123.123-12 ac pretium dapibus. Nunc at ex tincidunt mi 123.123.123-12 convallis ut at leo.
Suspendisse libero metus, 00000-000 quis tellus non, luctus luctus massa.
Sed aliquam dolor euismod, tincidunt nunc ullamcorper, rhoncus augue.
```

Figura 4.10: Texto Genérico Não Anonimizado

```
loremIpsum = anonimiza.anonimiza_string(loremIpsum, "CPF")
loremIpsum = anonimiza.anonimiza_string(loremIpsum, "Telefone")
loremIpsum = anonimiza.anonimiza_string(loremIpsum, "Email")
loremIpsum = anonimiza.anonimiza_string(loremIpsum, "Data")
loremIpsum = anonimiza.anonimiza_string(loremIpsum, "CEP")

with open("resultado.txt", 'w') as arquivo_txt:
    arquivo_txt.write(loremIpsum)
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Vestibulum ut iaculis mauris, suscipit elementum tellus.
Integer maximus massa eu urna lobortis cursus.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Praesent nisl lorem, rutrum eget dui a, auctor efficitur ligula.
Pellentesque in auctor nulla. Sed at tincidunt nunc.
Aliquam non ex #########, aliquam risus cursus, elementum magna.
Praesent vestibulum neque non orci pretium, ut placerat enim hendrerit.
Aenean non ######### turpis. Etiam feugiat fermentum libero, ut placerat massa varius vitae.
Vestibulum porta egestas libero, sed ######### justo porttitor sit amet.
Aliquam congue risus in orci lobortis, a rutrum massa condimentum.
Pellentesque a tellus aliquet, tristique velit consequat, dignissim dolor.
Donec at ######### mauris, id accumsan risus. Sed commodo fermentum turpis nec pretium.
Duis ######### sodales orci. Praesent rhoncus ######### nulla, non vehicula neque sollicitudin eget.
Phasellus quis elit lobortis, eleifend leo ut, efficitur velit.
 Integer semper lectus ut augue vestibulum interdum. Quisque in arcu turpis.
Curabitur rhoncus ######### ac pretium dapibus. Nunc at ex tincidunt mi ######### convallis ut at leo.
Suspendisse libero metus, ######### quis tellus non, luctus luctus massa.
Sed aliquam dolor euismod, tincidunt nunc ullamcorper, rhoncus augue.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Vestibulum ut iaculis mauris, suscipit elementum tellus.
Integer maximus massa eu urna lobortis cursus.
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Praesent nisl lorem, rutrum eget dui a, auctor efficitur ligula.
Pellentesque in auctor nulla. Sed at tincidunt nunc.
Aliquam non ex #########, aliquam risus cursus, elementum magna.
Praesent vestibulum neque non orci pretium, ut placerat enim hendrerit.
Aenean non ######### turpis. Etiam feugiat fermentum libero, ut placerat massa varius vitae.
Vestibulum porta egestas libero, sed ######### justo porttitor sit amet.
Aliquam congue risus in orci lobortis, a rutrum massa condimentum.
Pellentesque a tellus aliquet, tristique velit consequat, dignissim dolor.
Donec at ######### mauris, id accumsan risus. Sed commodo fermentum turpis nec pretium.
Duis ######### sodales orci. Praesent rhoncus ######### nulla, non vehicula neque sollicitudin eget.
Phasellus quis elit lobortis, eleifend leo ut, efficitur velit.
Integer semper lectus ut augue vestibulum interdum. Quisque in arcu turpis.
Curabitur rhoncus ######### ac pretium dapibus. Nunc at ex tincidunt mi ######### convallis ut at leo.
Suspendisse libero metus, ######### quis tellus non, luctus luctus massa.
Sed aliquam dolor euismod, tincidunt nunc ullamcorper, rhoncus augue.
```

Figura 4.11: Texto Genérico Anonimizado

4.5 Discussão dos Resultados

A análise dos resultados obtidos durante o teste foi de extrema importância, uma vez que sua funcionalidade e confiabilidade pôde ser verificada. Esta seção visa interpretar esses resultados e fornecer uma visão mais aprofundada do que foi observado.

4.5.1 Desempenho Geral

Os testes realizados demonstraram que a biblioteca possui uma eficiência notável na anonimização de dados encadeada, como foi possível perceber principalmente nos testes realizados com documentos .XLSX. Os testes em outros tipos de arquivos também se mostraram satisfatórios, uma vez que a biblioteca foi capaz de anonimizar e retornar as string de texto sem complicações. Entretanto, notou-se uma demora na anonimização dos dados de va-

cinação, onde a execução do programa levou 48.78 segundos. Futuras otimizações serão realizadas na tentativa de diminuir o tempo de execução.

4.5.2 Robustez e Confiabilidade

A biblioteca demonstrou robustez ao lidar com uma variedade de tipos de dados e situações de teste, não tendo problemas com a opção de inserção manual de uma expressão regular para identificar dados pessoais não suportados por padrão, aumentando ainda mais as possibilidades de anonimização em diferentes contextos. A biblioteca, durante os testes, realizou a anonimização em todos os dados selecionados de forma impecável, de forma que não pudessem ser recuperados, garantindo a segurança e inviolabilidade dos quasi-identificadores apresentados.

4.5.3 Comparação com Outras Soluções

Ao comparar a biblioteca com outras já existente, foi possível perceber um desempenho comparável em diversos aspectos, entretanto, este desempenho é superior quando se trata de anonimização de dados pessoais brasileiros.

Anonymizedf

A biblioteca criada se provou superior em questões de anonimização de dados em relação a biblioteca Anonymizedf [23], uma vez que a biblioteca deste trabalho possui a capacidade de encontrar dados em tipos específicos de documentos e anonimizá-los, enquanto a biblioteca Anonymizedf possui apenas a capacidade de gerar dados falsos que podem ser utilizados em uma substituição manual em textos.

Anonympy

Em relação a biblioteca Anonympy [24], a biblioteca criada para este trabalho se mostrou superior em relação a diversidade de dados que podem ser anonimizados, além de possuir total suporte para dados Brasileiros. Entretanto, é necessário notar que a biblioteca Anonympy é superior no tratamento de arquivos .PDF, uma vez que ela é capaz de ler, alterar os dados e reconstruir o arquivo com os dados anonimizados, além do suporte a anonimização de imagens.

4.6 Ameaças e Limitações para Validação

Durante a realização dos testes, embora seus resultados tenham sido satisfatórios, foi observado uma série de ameaças de limitações que devem ser consideradas durante uma avaliação crítica de seu desempenho e aplicabilidade.

4.6.1 Ambiente de Teste Limitado

Os testes foram conduzidos predominantemente em ambientes Windows 10, com versões do Python 3.10 e 3.11. Embora no momento dos testes, foi feito o esforço para utilizar as versões mais recentes do Python, não há garantia de que a biblioteca irá funcionar em versões antigas da linguagem, assim como não foi possível verificar a funcionalidade da biblioteca em diferentes sistemas operacionais, nem em outras versões do Windows.

4.6.2 Tamanho e Diversidade do Conjunto de Dados

Os conjuntos de dados utilizados para os testes abrangeram uma variedade de casos de uso, entretanto, o tamanho limitado destes conjuntos pode não capturar todas as nuances em cenários mais complexos. Além disso, foi possível realizar os testes apenas com uma quantidade extremamente pequena de arquivos e dados não fictícios, uma vez que não foi possível encontrar muitos dados relevantes através de pesquisas simples na internet, o que reduziu drasticamente a bateria de testes.

4.6.3 Escopo Funcional

Como citado anteriormente, a biblioteca não é capaz de recriar arquivos .PDF e .DOCX com os dados anonimizados, apenas retornar a string bruta via código. Os impactos negativos deste fato consistem na possível perda de estrutura e formatação de dados ao tentar recriar os documentos a partir da string retornada. Não foi possível, durante o desenvolvimento da biblioteca, encontrar outras bibliotecas relacionadas a arquivos .PDF e .DOCX que fossem compatíveis com a biblioteca criada e que fossem capazes de recriar os arquivos de forma automática.

4.7 Resumo do Capítulo

Neste capítulo, foram apresentados todos os testes feitos com a ferramenta criada para este trabalho. Os testes consistiram em diferentes tipos de arquivos, com diferentes tipos de dados, que serviram para verificar o funcionamento e capacidade da biblioteca criada, assim como seu tempo de execução. Todos os arquivos de testes foram apresentados

através de uma breve descrição e uma imagem de seu conteúdo. Os códigos utilizados para fazer os testes foram disponibilizados neste capítulo, para que fosse possível compreender melhor o que foi feito. Por fim, todos os resultados estão representados com imagens que mostram o conteúdo dos arquivos anonimizados corretamente. As considerações finais apresentam um resumo sobre o que era esperado dos testes e como a biblioteca atendeu as expectativas, junto do que foi observado durante as execuções.

Capítulo 5

Conclusão

A quantidade de dados pessoais utilizados pelas organizações, sejam públicas ou privadas é de proporções imensuráveis. Entretanto, nem todos esses dados possuem a necessidade de serem mantidos em poder das empresas. A Lei Geral de Proteção de Dados Pessoais informa que apenas dados estritamente necessários para a finalidade pretendida devem ser coletados [1]. Neste contexto, a anonimização possui um papel vital, uma vez que técnicas desta área podem ser utilizadas para desvincular completamente um atributo quasi-identificador de uma pessoa física.

Diversas técnicas de anonimização existem e servem seu propósito para realizar essa desvinculação, entretanto, é necessário saber que várias dessas técnicas destroem parcialmente ou até mesmo totalmente o atributo a qual são aplicadas. Dito isso, a aplicação manual de técnicas de anonimização é um trabalho extremo e interminável, portanto, diversas bibliotecas de linguagens de programação existem para facilitar este trabalho. Infelizmente, poucas são as que oferecem suporte para dados pessoais brasileiros, e não compreendem ou não aceitam os padrões de dados existentes no Brasil.

Este trabalho possuiu como objetivo a criação de uma biblioteca que servisse como anonimizador de dados pessoais brasileiros. A biblioteca foi criada com o intuito de receber dados de arquivos de diferentes formatos e aplicar a técnica de supressão sobre os dados desejados. Decidiu-se criar expressões regulares que servissem como verificadores para percorrer os textos e encontrar os dados a serem anonimizados. Além disso, o suporte para documentos .PDF, .DOCX e .XLSX foi adicionado com êxito, para aumentar ainda mais a utilidade da biblioteca, uma vez que dados pessoas não estão sempre contidos em um tipo de documento apenas. No caso de documentos e arquivos não suportados pela biblioteca, uma opção de anonimização de strings de texto também foi incluída, para oferecer ainda mais suporte.

Uma das principais contribuições deste trabalho consiste na biblioteca citada anteriormente, uma vez que esta ajudará e facilitará a anonimização de dados por parte de

indivíduos e até mesmo de empresas, uma vez que a biblioteca possui uma documentação que facilita o seu uso. De acordo com os testes, foi possível verificar o funcionamento de cada função e de cada expressão regular inserida na biblioteca. Como visto, todos os testes tiveram os resultados esperados e até surpreendentes em relação ao tempo de execução e quantidade de dados anonimizados de forma encadeada, além a superioridade da biblioteca em anonimização de dados pessoais brasileiros se comparada a outras ferramentas.

5.1 Trabalhos Futuros

Os trabalhos futuros consistem principalmente em melhorias para a biblioteca criada, mais especificamente adicionar o suporte a alteração e criação de documentos .DOCX e .PDF assim como já está sendo feito para documentos .XLSX. Também deseja-se criar uma extensa base de dados, contendo nomes brasileiros, para que seja possível criar uma função de anonimização de nomes e, assim, estender ainda mais as capacidades da biblioteca.

O suporte à anonimização de dados em outros formatos de arquivos também é desejado, além de uma possível integração com tecnologias de inteligência artificial, para que mais dados possam ser anonimizados de forma mais eficiente.

Referências

[1] República, Presidência da: *Lei geral de proteção de dados pessoais (lgpd)*. Secretaria-Geral, accessed in November 19, 2019, 2018. https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf. 1, 2, 6, 7, 8, 9, 47

[2] Murthy, Suntherasvaran, Asmidar Abu Bakar, Fiza Abdul Rahim e Ramona Ramli: *A comparative study of data anonymization techniques*. Em *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, páginas 306–309, 2019. 1, 2, 8, 9, 10, 11, 13, 26

[3] Bild, Raffael, Klaus A. kuhn e Fabian Prasser: *Better safe than sorry – implementing reliable health data anonymization*. PDigital Personalized Health and Medicine, 270:68–72, 2020. https://ebooks.iospress.nl/publication/54126. 1

[4] Prasser, Fabian, Johanna Eicher, Raffael Bild, Helmut Spengler e Klaus A. Kuhn: *A tool for optimizing de-identified health data for use in statistical classification*. Em *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, páginas 169–174, 2017. 2

[5] Jafarian, Jafar Haadi e Amirreza Niakanlahiji: *Multirhm: Defeating multi-staged enterprise intrusion attacks through multi-dimensional and multi-parameter host identity anonymization*. Computers & Security, 124:102958, 2023, ISSN 0167-4048. https://www.sciencedirect.com/science/article/pii/S0167404822003509. 2

[6] Medková, Jana e Josef Hynek: *Hakau: hybrid algorithm for effective k-automorphism anonymization of social networks*. Soc. Netw. Anal. Min., 13(1):63, 2023. https://doi.org/10.1007/s13278-023-01064-1. 2

[7] Smile-SA: *Repositório anonymization smile-sa*. https://github.com/Smile-SA/anonymization, acesso em 2023-04-12. 2

[8] ArtLabss: *Repositório open-data-anonymizer artlabss*. https://github.com/ArtLabss/open-data-anonymizer, acesso em 2023-06-06. 2

[9] ARX: *Arx - data anonymization tool*. https://arx.deidentifier.org/, acesso em 2023-06-10. 2

[10] Tomás, Joana Carolina Pedroso: *Data anonymization: algorithms, techniques and tools*. Tese de Doutoramento, Instituto Politecnico de Coimbra, 2022. 6, 7, 8, 9

[11] Teixeira, Guilherme Cardoso: *O papel social da lei geral de proteção de dados no brasil*. UNIVERSIDADE DO SUL DE SANTA CATARINA, páginas 1–59, 2020. 6, 7, 8

[12] Kalam, Anas Abou El, Yves Deswarte, Gilles Trouessin e Emmanuel Cordonnier: *Personal data anonymization for security and privacy in collaborative environments*. Em McQuay, William K. e Waleed W. Smari (editores): *Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems, CTS 2005, Saint Louis, Missouri, USA, May 15-20, 2005*, páginas 56–61. IEEE Computer Society, 2005. https://doi.org/10.1109/ISCST.2005.1553294. 7

[13] Carvalho, Artur Potiguara, Edna Dias Canedo, Fernanda Potiguara Carvalho e Pedro Henrique Potiguara Carvalho: *Anonimisation, impacts and challenges into big data: A case studies*. Em Filipe, Joaquim, Michal Smialek, Alexander Brodsky e Slimane Hammoudi (editores): *Enterprise Information Systems - 22nd International Conference, ICEIS 2020, Virtual Event, May 5-7, 2020, Revised Selected Papers*, volume 417 de *Lecture Notes in Business Information Processing*, páginas 3–23. Springer, 2020. https://doi.org/10.1007/978-3-030-75418-1_1. 7

[14] Carvalho, Artur Potiguara, Edna Dias Canedo, Fernanda Potiguara Carvalho e Pedro Henrique Potiguara Carvalho: *Anonymisation and compliance to protection data: Impacts and challenges into big data*. Em Filipe, Joaquim, Michal Smialek, Alexander Brodsky e Slimane Hammoudi (editores): *Proceedings of the 22nd International Conference on Enterprise Information Systems, ICEIS 2020, Prague, Czech Republic, May 5-7, 2020, Volume 1*, páginas 31–41. SCITEPRESS, 2020. https://doi.org/10.5220/0009411100310041. 7, 9, 16

[15] *General data protection regulation*. https://gdpr-info.eu, acesso em 2018-05-25. 8, 9

[16] Alves, Carina e Moisés Neves: *Especificação de requisitos de privacidade em conformidade com a LGPD: resultados de um estudo de caso*. Em Menezes Cruz, Maria Lencastre Pinheiro de, Graciela Dora Susana Hadad e Johnny Cardoso Marques (editores): *Anais do WER21 - Workshop em Engenharia de Requisitos, Brasilia, BSB, Brasil, August 23-27, 2021*. Editora PUC-Rio, 2021. http://wer.inf.puc-rio.br/WERpapers/artigos/artigos_WER21/WER_2021_paper_31.pdf. 8

[17] SWEENEY, LATANYA: *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002. https://doi.org/10.1142/S0218488502001648. 9, 13

[18] Tilborg, Henk C. A. van e Sushil Jajodia: *MD5 hash function*. Em *Encyclopedia of Cryptography and Security, 2nd Ed*, página 771. Springer, 2011. https://doi.org/10.1007/978-1-4419-5906-5_1197. 10

[19] Prasser, Fabian, Johanna Eicher, Helmut Spengler, Raffael Bild e Klaus A. Kuhn: *Flexible data anonymization using arx—current status and challenges ahead*. Software: Practice and Experience, 50(7):1277–1304, 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2812. 12

[20] Haber, Anna C, Ulrich Sax, Fabian Prasser e the NFDI4Health Consortium: *Open tools for quantitative anonymization of tabular phenotype data: literature review.* Briefings in Bioinformatics, 23(6):bbac440, outubro 2022, ISSN 1477-4054. https://doi.org/10.1093/bib/bbac440. 12

[21] Jha, Nikhil, Luca Vassio, Martino Trevisan, Emilio Leonardi e Marco Mellia: *Practical anonymization for data streams: z-anonymity and relation with k-anonymity.* Perform. Evaluation, 159:102329, 2023. https://doi.org/10.1016/j.peva.2022.102329. 12

[22] FOUNDATION, PYTHON SOFTWARE: *Python documentation.* https://www.python.org/doc/, acesso em 2023. 13, 17

[23] Fridriksson, Alexander: *anonymizedf 1.0.1.* https://pypi.org/project/anonymizedf/, acesso em 2020-06-11. 14, 44

[24] ArtLabs: *anonympy 0.3.7.* https://pypi.org/project/anonympy/, acesso em 2022-05-01. 14, 44

[25] Judith Sáinz-Pardo Díaz, Álvaro López García: *pycanon 1.0.1.post1.* https://pypi.org/project/pycanon/, acesso em 2023-05-18. 14

[26] Faraglia, Daniele: *Faker documentation.* https://faker.readthedocs.io/en/master/, acesso em 2014. 14

[27] Amazon: *Program aws glue etl scripts in pyspark.* https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-python.html, acesso em 2017. 14

[28] Rocha, Lucas Dalle, Geovana Ramos Sousa Silva e Edna Dias Canedo: *Privacy compliance in software development: A guide to implementing the LGPD principles.* Em Hong, Jiman, Maart Lanperne, Juw Won Park, Tomás Cerný e Hossain Shahriar (editores): *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC 2023, Tallinn, Estonia, March 27-31, 2023*, páginas 1352–1361. ACM, 2023. https://doi.org/10.1145/3555776.3577615. 16

[29] Canedo, Edna Dias, Angélica Toffano Seidel Calazans, Ian Nery Bandeira, Pedro Henrique Teixeira Costa e Eloisa Toffano Seidel Masson: *Guidelines adopted by agile teams in privacy requirements elicitation after the brazilian general data protection law (LGPD) implementation.* Requir. Eng., 27(4):545–567, 2022. https://doi.org/10.1007/s00766-022-00391-7. 16

[30] GitHub: *Github.* https://github.com, acesso em 2023. 18

[31] Docs, GitHub: *Gitflow documentation.* https://docs.github.com/en/get-started/quickstart/github-flow#following-github-flow, acesso em 2023. 19

[32] ArtLabs: *pypi.* https://pypi.org/, acesso em 2023. 20

[33] Stefano, Raylan e: *anonympy 0.3.7.* https://test.pypi.org/project/python-anonimiza-pt-br/0.0.1/, acesso em 2023. 20

[34] Canedo, Edna Dias, Angélica Toffano Seidel Calazans, Geovana Ramos Sousa Silva, Pedro Henrique Teixeira Costa e Eloisa Toffano Seidel Masson: *Use of journey maps and personas in software requirements elicitation*. Int. J. Softw. Eng. Knowl. Eng., 33(3):313–342, 2023. https://doi.org/10.1142/S0218194023300014. 23

[35] Canedo, Edna Dias, Ian Nery Bandeira, Angélica Toffano Seidel Calazans, Pedro Henrique Teixeira Costa, Emille Catarine Rodrigues Cançado e Rodrigo Bonifácio: *Privacy requirements elicitation: a systematic literature review and perception analysis of IT practitioners*. Requir. Eng., 28(2):177–194, 2023. https://doi.org/10.1007/s00766-022-00382-8. 23

[36] Donca, Ionut Catalin, Ovidiu Petru Stan, Marius Misaros, Dan Gota e Liviu Miclea: *Method for continuous integration and deployment using a pipeline generator for agile software projects*. Sensors, 22(12), 2022, ISSN 1424-8220. https://www.mdpi.com/1424-8220/22/12/4637. 23, 24

[37] *Regular expression operations*. https://docs.python.org/3/library/re.html. 26

[38] *Pdfminer.six documentation*. https://github.com/pdfminer/pdfminer.six. 28

[39] *Python-docx documentation*. https://python-docx.readthedocs.io/en/latest/index.html. 29

[40] *Python pandas documentation*. https://pandas.pydata.org. 30

[41] Stefano, Raylan e: *anonymization-library*. https://github.com/Rayxan/anonymization-library, acesso em 2023. 31, 35

[42] *Relação de estagiários do supremo tribunal federal*. https://egesp-portal.stf.jus.br/transparencia/relacao_estagiario. 33

[43] *Dados de vacinação da covid-19*. https://sjb.rj.gov.br/uploads/16b8b4fc9fde4cb8ba34bd4119b041644120a5b3.pdf. 34

[44] OpenAI. https://chat.openai.com. 39