Python para Análise de Dados em Segurança Cibernética Utilizando Regex

Priscila de Sousa Silva [1, 2]
orcid.org/0009-0000-5801-0090

Emmanuel Andrade de Barros Santos [1, 3, 4, 5, 6]
orcid.org/0000-0002-1786-0934

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: pss@poli.br

[2]Graduação em Engenharia de Elétrica com Ênfase em Controle e Automação, Escola Politécnica de Pernambuco, Recife, Brasil.

[3]Graduação em Engenharia Elétrica/Eletrônica, Universidade Federal de Pernambuco, Recife, Brasil.

[4]Mestrado em Engenharia Elétrica, Universidade Federal de Pernambuco, Recife, Brasil.

[5]Especialização em MBA em Gerenciamento de Projetos na visão do PMI, Faculdade Estácio do Recife, Recife, Brasil.

[6]Doutorado em Engenharia Elétrica, com Ênfase em Eletrônica, Universidade Federal de Pernambuco, Recife, Brasil.

DOI: 10.25286/repa.v10i1.2508

Como citar este artigo pela NBR 6023/2018: Priscila de Sousa Silva; Emmanuel Andrade de Barros Santos. Python para Análise de Dados em Segurança Cibernética Utilizando Regex. Revista de Engenharia e Pesquisa Aplicada, v.10, n. 1, p. 33-42, 2025

RESUMO

Com a vigência da Lei Geral de Proteção de Dados, a LGPD, o Brasil conta hoje com leis que buscam garantir a Segurança dos dados de consumidores de serviços. Hoje as empresas são obrigadas a seguir a LGPD no que diz respeito à coleta dos dados, manipulação e tratamento, compartilhamento, armazenamento, caso contrário podem sofrer punições. Propõe-se o desenvolvimento de um projeto para auxiliar equipes responsáveis pela proteção dos dados de uma empresa no controle do compartilhamento de informações consideradas sensíveis. Em estudos baseados na LGPD foi desenvolvido um script, utilizando a plataforma PyCharm tendo como base a linguagem Python, e foram aplicadas Expressões Regulares para identificação de padrões. Como resultado foi obtido um código que faz análises em e-mails enviados por usuários, assim como também em anexos contidos nestes e-mails para que caso sejam encontrados padrões de dados que podem ser considerados sensíveis, os administradores da empresa estejam cientes e possam tomar as ações necessárias.

PALAVRAS-CHAVE: Segurança Cibernética; Lei Geral de Proteção de Dados; Expressões Regulares; Segurança da Informação; Violação de Dados; Vazamento de Dados.

ABSTRACT

With the General Data Protection Law, LGPD, Brazil now has laws that guarantee data security for service consumers. Today companies are obliged to follow the LGPD with regard to data collection, handling and treatment, sharing, storage, otherwise they may be punished.
It is proposed to develop a script to help teams responsible for protecting a company's data in controlling the sharing of information considered sensitive.
In studies based on the LGPD, a script was developed, using the PyCharm platform based on the Python language, and Regular Expressions were applied to identify patterns. As a result, a code wasobtained that analyzes e-mails sent by users, as well as attachments contained in these e-mails so that if data patterns are found that can be considered sensitive, company administrators are aware and can take the necessary actions.

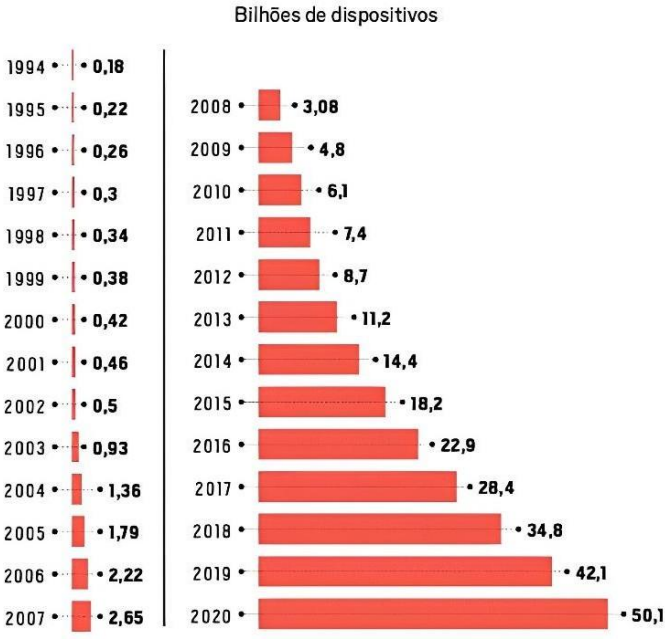
KEY-WORDS: Cibersecurity; General Data Protection Law; Regular expression; Information Security; Data Leak; Data Breach.

1 INTRODUÇÃO

Em estudo realizado pelo laboratório de inteligência e ameaças da Fortinet, empresa de soluções em segurança cibernética, no primeiro semestre de 2022, o número de ataques cibernéticos no Brasil foi de aproximadamente 31,5 bilhões. Esse número significa um aumento de 94% em comparação ao primeiro semestre de 2021, que houveram 16,2 bilhões de registros. Ainda de acordo com este estudo, o Brasil é o segundo país da América Latina com mais ataques cibernéticos em 2022, ficando atrás apenas do México, com 85 bilhões de tentativas **[1]**.

Por mais que os ataques de segurança cibernética ocorram desde o surgimento das tecnologias, os dados mostram que estas tentativas de invasão e obtenção de benefícios de forma ilícita se intensificou com a pandemia da Covid-19, pois o número de pessoas trabalhando de forma remota aumentou, consequentemente, aumentou também o número de dispositivos conectados com acesso à dados empresariais (Figura 1), deixando as empresas mais vulneráveis **[2]**. Dependendo do ataque, as consequências podem ser desastrosas, como por exemplo, perda, roubo, bloqueio de dados, instalação de softwares mal-intencionados, entre outros.

Figura 1 - Crescimento exponencial do número de dispositivos conectados durante os anos



Fonte: New Signature, Julius Braer (2020)

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Pilares da Segurança da Informação

O grande objetivo da Segurança da informação é a proteção de dados e existem três pilares que dão suporte às estratégias que visam essa proteção: a confidencialidade, integridade e disponibilidade **[3]**.

A disponibilidade busca garantir que a informação ou serviço esteja sempre disponível no momento que se deseja utilizá-la. Este pilar busca garantir que os usuários possam acessar os dados em tempo integral, sem qualquer interrupção.

A integridade busca garantir que os dados estejam preservados, consistentes e confiáveis durante todo o ciclo de vida deles.Visando cumprir este pilar, os dados não podem ser alterados ou deletados de forma não autorizada.

A confidencialidade garante que os dados estejam acessíveis apenas a usuários selecionados e estejam protegidos contra acessos não autorizados.

2.2 Lei Geral de Proteção de Dados (LGPD)

As empresas, atualmente, enfrentam um grande desafio que é o tratamento de todos os dados dos quais está em posse, sejam estes de funcionários, clientes ou fornecedores. Ninguém quer ter sua privacidade violada, mesmo que seja apenas alguns dados considerados "simples".
Esta privacidade é prevista na Constituição Federal e no Código Civil:

> Art. 5, X, Constituição Federal de 1988: "são invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação;" **[4]**.

> Art.21, Código Civil: "A vida privada da pessoa natural é inviolável, e o juiz, a requerimento do interessado, adotará as providências necessárias para impedir ou fazer cessar ato contrário a esta norma." **[5]**.

A privacidade de dados foca nos direitos das pessoas, nos fins da coleta e do processamento dos dados, nas preferências de privacidade e na forma

DOI: 10.25286/repa.v10i1.2508

como as instituições fazem o controle dos dados pessoais de seus proprietários.

Visando regulamentar a proteção e a privacidade deles, foi sancionada em 14 de agosto de 2018 a LGPD (Lei Geral de Proteção de Dados), baseada na GDPR (*General Data Protection Regulation*), regulamentação em vigor na União Europeia com função similar. Apesar de ter sido sancionada em 2018, a LGPD entrou em vigor dois anos depois, no dia 18 de setembro de 2020.

2.2.1 Classificação dos Dados

Tomando como base a GDPR da União Europeia, no Art. 5°, a LGPD também separa os dados quanto ao seu tipo.

Os dados pessoais são informações relacionadas à pessoa natural identificada ou identificável. Informações que tornam possível a identificação da pessoa, como: Nome, Endereço, CPF, endereço de IP, fotos, placa de carro, número de matrícula, número de passaporte, outros.

Já os dados pessoais sensíveis são dados pessoais "sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dados referentes à saúde ou à vida sexual, dados genéticos ou biométricos, quando vinculados a uma pessoa natural." **[6]**. Os dados pessoais sensíveis devem ser tratados de maneira ainda mais cautelosa, justamente pelo fato de serem dados ainda mais reservados e íntimos do titular. Se estes dados forem mal utilizados, podem causar discriminação e segregação, causando prejuízos ao titular.

Dados anonimizados são os que sozinhos não conseguem fazer a identificação de um indivíduo. Eles são usados na realização de estudos e a estes dados não são aplicadas as exigências da LGPD.

2.2.2 Ciclo de vida dos dados

O ciclo de vida dos dados é a ordem que representa o caminho deles dentro da companhia, desde o momento da coleta até o momento da eliminação ou quando são arquivados.

As etapas do ciclo de vida dos dados podem ser resumidas em:

Captação ou coleta dos dados, processamento, análise, compartilhamento, armazenamento, reutilização e eliminação **[7]**.

2.2.3 Titular, Operador e Controlador

Como se trata de uma legislação que tem como objetivo estabelecer uma cultura de proteção à dados, temos então três grupos envolvidos que podem ser encontrados, descritos no Art. 5 da Lei: o titular, o operador e o controlador **[8]**.

O titular dos dados é a pessoa física dona da informação, a quem se refere os dados pessoais. De forma geral, toda pessoa física é um titular de dados.

O controlador dos dados é a empresa ou responsável por coordenar e definir como o dado pessoal deve ser tratado, desde o momento da coleta até o momento da eliminação. De forma simples, é quem lida diretamente com os dados e, uma de suas funções principais é garantir a transparência e a comunicação que precisa ser feita com o titular.

O operador dos dados é a empresa ou responsável por realizar o tratamento de dados em nome do controlador. O operador atua como parceiro técnico e deve respeitar a política e as regras impostas pelo controlador. O controlador é responsável pelo operador, sendo necessárias orientações para que tudo ocorra dentro das determinações da lei. Em caso de situações de danos que podem ter sido causados pelo operador, o controlador pode responder por isto.

Um exemplo simples são as empresas de *call center*, elas atuam como operadora de dados e os tratam de acordo com o que foi definido pelo controlador, baseado na lei.

2.3 Autoridade Nacional de Proteção de Dados (ANPD)

A função de fiscalizar, monitorar e regulamentar o cumprimento da LGPD fica com a Autoridade Nacional de Proteção de Dados, conhecida como ANPD. Com essa fiscalização, o crescimento da cultura de privacidade e proteção é agilizado, pois as empresas querem estar em adequação à Lei.

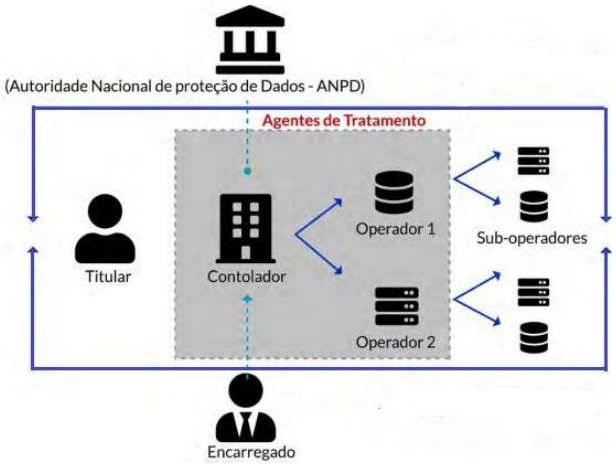
A ANPD atua em todo território nacional, é um órgão da administração pública federal e foi criada pela Lei 13.853/2019. Para que seu papel na sociedade seja cumprido, este órgão tem poder de elaboração de guias orientativos de boas práticas e normas para auxílio dos controladores nos meios de conformidade da LGPD, além de qualificado na aplicação de multas e sanções, de acordo com o Art. 52 da Lei **[9]**.

2.4 DPO ou Encarregado de dados

Uma figura importante citada no Art. 5° VIII é o encarregado, mas conhecido como DPO (data

protection officer). Este especialista é definido como o responsável por atuar como um canal de comunicação entre o titular dos dados, o controlador e a ANPD, e também monitora as empresas para assegurar que estejam em conformidade com as boas práticas de acordo com a LGPD, exercendo os requisitos de governança de gestão de riscos e também de privacidade e proteção de dados **[10]**.

Figura 2: Entidades envolvidas com a LGPD.



Fonte: CertiProf (2021)

2.5 Consequências do não cumprimento da LGPD

As companhias que tem posse dos dados dos clientes ou funcionários e não está em conformidade com a Lei Geral de Proteção de Dados estão passíveis de sanções e punições devido ao não cumprimento da Lei. Estas sanções administrativas começaram a ser aplicadas a partir de 1° de agosto de 2021. São elas:

Advertência: É a primeira medida da ANPD caso seja cometida alguma irregularidade por parte de uma empresa. Determina um prazo para que a empresa que cometeu a infração possa regularizara situação de acordo com a LGPD.

Multas: Após a advertência não surtir efeito, a empresa que segue em não conformidade pode receber uma multa simples que pode chegar à 2% do valor de faturamento da empresa (limitada a R$ 50 milhões por infração) ou multas diárias também limitadas à 50 milhões de reais. O valor da multa vai depender da gravidade da infração e dos danos que foram causados.

Publicitação da infração: Após confirmado que a empresa não está em conformidade com a lei, a ANPD também pode solicitar que a empresa assuma publicamente o vazamento e/ou a violação dos dados.

Bloqueio de dados: A empresa pode ter os dados bloqueados até que a regularização seja finalizada. Dessa forma, as atividades da empresa que precisam destas informações para serem realizadas precisam ser pausadas e de certo modo, afeta a produtividade.

Eliminação de dados: Acontece em casos mais graves e a punição é a eliminação dos dados coletados que estão no sistema da empresa.

Não apenas consequências financeiras, descumprir a LGPD põe em risco a imagem da empresa, a confiança dos clientes, a confiança da marca, a reputação da empresa no mercado, a desvalorização das ações da empresa, entre outros.

2.6 A segurança dos Dados é responsabilidade de todos

Ao analisar casos de exposição de dados, quase sempre é possível identificar condutas de segurança inadequadas dos funcionários das empresas. Isto traz a seguinte pergunta: Seriam as pessoas o alvo mais fácil para atacar? Precisamos então entender a diferença entre vazamento e violação de dados.

A violação de dados ou *Data Breach*, acontece quando uma organização sofre algum tipo de incidente de segurança que tem como resultado a violação da disponibilidade, confidencialidade ou integridade dos dados, ou seja, uma fonte externa consegue violar o sistema por um ataque cibernético.

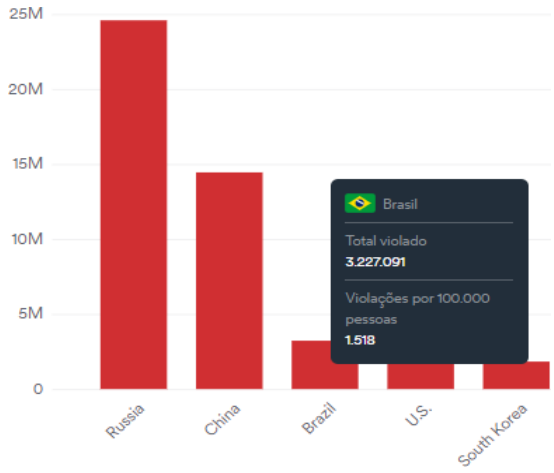
Já o vazamento de dados ou *Data Leak*, acontece quando pessoas não autorizadas tem acesso à dados devido a erros internos. Este tipo de erro pode levar a violação dos dados, roubo de credenciais ou instalação de programas mal-intencionados. Estes conceitos podem parecer confusos, pois os criminosos podem usar informações resultantes de vazamentos de dados para iniciar violações de dados em massa.

De acordo com informações publicadas pela empresa SurfShark **[11]**, no primeiro semestre de 2022 o Brasil ficou entre os três países mais afetados com incidentes cibernéticos de violação e/ou vazamento de dados, como visto na figura (3).

Figura 3 - Países mais afetados com a violação de dados no 1° Trimestre de 2022.

Fonte: Surfshark [11]

DOI: 10.25286/repa.v10i1.2508



Como humanos, estamos suscetíveis a cometer falhas, mas falando de segurança, um erro considerado pequeno pode levar a uma exposição gigante de dados. De acordo com um estudo realizado em mais de 5.000 empresas no mundo, pela *Kaspersky Lab* e *B2B International* (empresas de Segurança e Tecnologia da Informação, respectivamente), 46% dos ataques relacionados à segurança cibernética resultaram de descuido ou falta de treinamento dos usuários **[12]**.

As empresas precisam garantir que os sistemas de segurança cibernética estejam cada vez mais atualizados e funcionem bem.

É importante deixar claro que nenhuma empresa pode evitar que os ataques ocorram, mas pode evitar que estes ataques possam ser bem-sucedidos. Todos os colaboradores da empresa devem ter conhecimento das medidas utilizadas para proteger os dados e compartilhar esta responsabilidade.

3 MATERIAIS E MÉTODOS

É possível utilizar a tecnologia ao nosso favor, seja para ajudar a proteger ambientes virtuais, realizar detecção de e-mails de phishing ou até mesmo usar as linguagens de programação para auxiliar na verificação do cumprimento da LGPD **[13]**.

De acordo com todos os fatos citados, o presente trabalho tem como objetivo apresentar o desenvolvimento de um sistema que visa realizar análises de dados utilizando *REGEX*, para identificar em e-mails enviados pelos colaboradores de uma empresa o possível compartilhamento de dados que podem ser considerados infrações pela LGPD.

No projeto foi usada a linguagem de programação em Python, utilizando como IDE o software PyCharm.

3.1 Expressões Regulares (REGEX)

As expressões regulares, REGEX, são padrões que podem ser usados para coletar e identificar informações em textos, encontrar padrões em *strings* (conjuntos de caracteres de texto dentro de um código), filtrar elementos, dividir *strings*, fazer remoção de caracteres, entre outros **[14]**.

As expressões regulares podem ser escritas em qualquer linguagem de programação.Um exemplo é o número de CPF, é possível criar uma expressão regular para identificá-lo, pois nele há uma sequência de números que precisam estar num formato específico, com alguns blocos separados por pontos.

Para utilizar o REGEX em Python, é necessário a instalação da biblioteca re **[15]**.
Existe uma lista de componentes e significados que possibilitam as operações e a criação dos padrões **[16]**.
Alguns dos componentes podemos ver na Tabela 1.
Por exemplo, podemos validar na equação (1) palavras como abacaxi, amora ou ameixa dentro de um texto:

$$(abacaxi) \mid (amora) \mid (ameixa) \quad \textbf{(1)}$$

Onde temos:

() – agrupando as palavras dentro do marcador;

| - indicando a possibilidade de escolha, funcionando como o operador lógico OR.

No exemplo da equação (2), utilizamos a máscara de *REGEX* abaixo para identificação de um CEP:

$$((\hat{}(\backslash d\{5\}) - (\backslash d\{3\})\$)|(\backslash d\{8\})) \quad \textbf{(2)}$$

Onde temos:

^ - marcando o início da string (também pode ser usado como sinal de negação, quando vem como primeiro sinal em uma lista);
\d{5} – indicando cinco dígitos, variando de 0 a 9;
\d{3} – indicando três dígitos, variando de 0 a 9;
$ - indicando que o padrão está no final da linha;
\d{8} - indicando 8 dígitos, variando de 0 a 9.

Alguns exemplos de aplicações das expressões regulares são: análise de texto na colaboração de um sistema de agendamento de missões espaciais de uma tripulação **[17]**, sistema de análise de documentos e reconhecimento de código chinês usando Regex **[18]**, expressões regulares em

conjunto com bibliotecas Python usadas para extração de conteúdos de mídias sociais **[19]**.

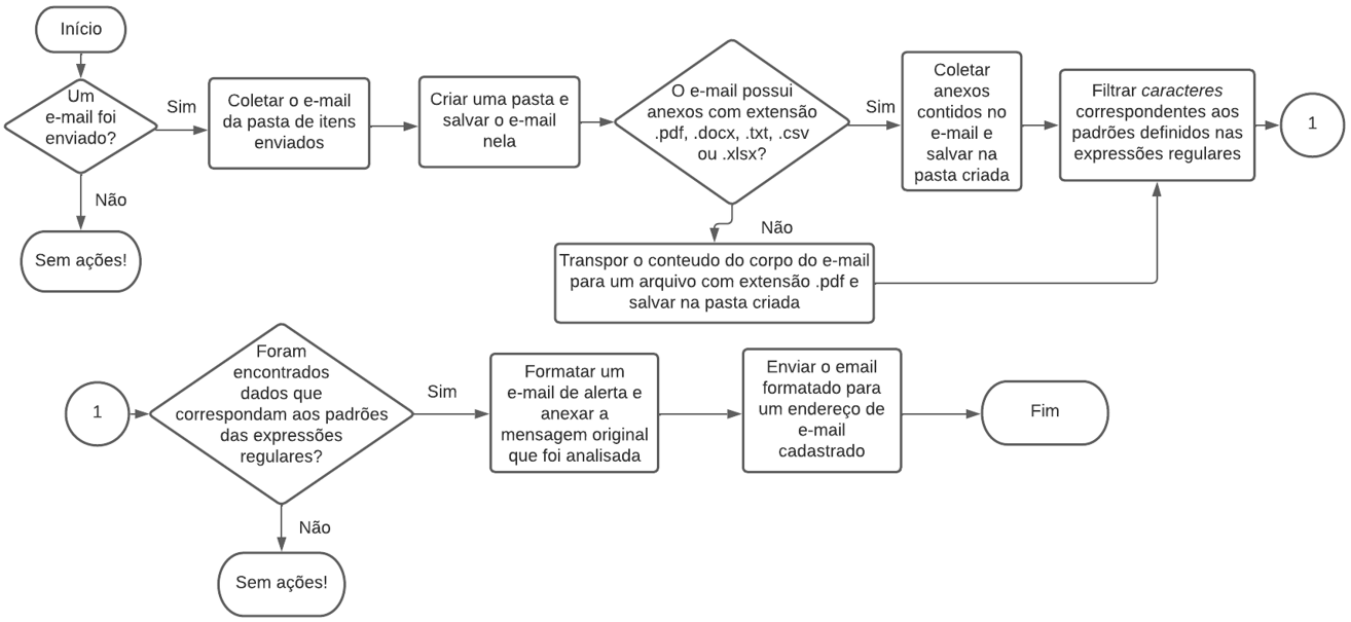
Tabela 1 - Alguns elementos para criar um REGEX

Elementos	Utilização	Padrão	Exemplo
[] e ()	Indicam um conjunto de *caracteres*	[ae] (ab)	"a" em "c**a**s**a**" "ab" em "**ab**acaxi"
.	Indica qualquer *caracter*, funciona como um coringa	Lu.a	"Lu**n**a" ou "Lu**v**a" em "Lu**.**a"
\	Indica a eliminação do significado especial de um caracter	Lu\.a	"Lu**.**a" e não "Lu**v**a" ou "Lu**n**a"
^	Indica a busca por elementos no início de uma *string*	^a ^[bo]	"a" em "**a**brir" "bo" em "**bo**la"
$	Indica a busca por elementos no fim de uma *string*	9$ [15]$	"9" em "1**9**" "15" em "17**15**"
|	Indica a busca por um padrão ou por outro padrão	Lu(a|na|z)	"Lua" e "Luna" em "**Luna** está na **Lua**"
d	Indica um caractere numérico equivalente à [0-9]	\d	"4" em "**4** = IV"
{}	Indica repetições de *caracteres* ou de uma sequência de *caracteres*	d{3}	"123" em "A**123**"

Fonte: Autoria Própria (2022)

3.2 Abordagem proposta

Figura 4 - Fluxograma de funcionamento



Fonte: Autoria Própria (2022

O projeto foi desenvolvido para que o script realizasse uma extração e análise de dados em e-mails e tomasse ações de acordo com o que foi estabelecido.

Uma abordagem similar foi desenvolvida por Stephen **[20]**. O projeto desenvolvido por ele foi implantado numa plataforma de processamento de e-mails para extrair conteúdo de texto de mensagens de e-mail 24 horas por dia e 7 dias por

DOI: 10.25286/repa.v10i1.2508

semana, com o objetivo de melhorar o serviço de filtragem de e-mails.

Neste projeto, o código precisa fazer uma conexão com o Microsoft Outlook (encontrado localmente na máquina do colaborador) e, ao identificar o último e-mail localizado no *mailbox* de Itens enviados, realizar a análise no e-mail encontrado. Para esta ação ser realizada, foi utilizada a biblioteca *Win32com.client*.

Foi utilizada outra biblioteca (*Fpdf*) para salvar o conteúdo do corpo do e-mail em um arquivo com extensão .pdf e realizar a leitura do arquivo (estes arquivos foram salvos em uma pasta criada na máquina do usuário). Porém, não somente no corpo do e-mail, os dados também podem estar inseridos em anexos. O projeto foi desenvolvido para também ler algumas extensões de anexos, como docx (word), csv e xlsx (*excel*), txt (bloco de notas) e pdf (*Portable Document Format*).

Após lidas todas as informações e armazenadas em variáveis, através da biblioteca *re* e da função *compile*, foram criados os padrões das expressões regulares para que fossem compilados antes de realizar as buscas nas variáveis criadas.

Utilizamos então a função *Search*, ainda da mesma biblioteca, para examinar os caracteres das variáveis com o objetivo de encontrar o padrão de expressões regulares que corresponde ao que foi pré-estabelecidos no código, dados que podem identificar um titular e ferir a LGPD em caso de incidentes de segurança cibernética.

Como ação, ao encontrar algum dado que corresponda aos padrões, um endereço de e-mail configurado no *script* (um profissional, que pode ser um administrador de segurança ou conformidade) receberá um alerta informando que o usuário em questão pode estar violando a Lei Geral de Proteção de Dados.

O alerta que será recebido inclui também o e-mail original anexado, para uma melhor análise por parte do administrador.

Caso se trate de fato de uma violação da Lei, a ação ficará por conta do administrador de acordo com as normas da empresa. Este *script* ajudará a entender se há algum usuário que não está agindo conforme a Lei, e possibilita que alguma ação venha a ser tomada antes que ocorra algum incidente de vazamento e/ou violação de dados.

Para o desenvolvimento deste projeto foram utilizadas as bibliotecas encontradas na tabela 2:

Tabela 2 - Bibliotecas utilizadas no projeto

Biblioteca	Motivo da Utilização
Win32com.client	Utilizada para fazer integração com o Microsoft Outlook e enviar e-mails
Pathlib	Utilizada para manipular caminhos dos arquivos
Fpdf	Utilizada para transformar o corpo do e-mail em um arquivo .pdf
Re	Utilizada para criar e validar as expressões regulares
PyPDF2	Utilizada para ler arquivos com extensão .pdf
Os	Utilizada para obter o diretório atual dos arquivos
Docxetxt	Utilizada para ler arquivos com extensão .docx
Pandas	Utilizada para ler arquivos com extensões .xlsx, .csv, .txt

Fonte: Autoria Própria (2022)

4 RESULTADOS E DISCUSSÃO

Após finalizado o projeto, foram realizados testes que ajudaram a comprovar que as expressões regulares podem ser aliadas no cumprimento da Lei Geral de Proteção de Dados.

Os testes foram realizados em:

- E-mails **sem** informações sensíveis: Sem anexos e com anexos (extensões .pdf, .docx, .xlsx, .csv e .txt).
- E-mails **com** informações sensíveis:

Sem anexos e com anexos (extensões .pdf, .docx, .xlsx, .csv e .txt).

Nas figuras (5) e (6) podemos ver alguns exemplos dos testes que foram realizados em e-mails com anexos que possuíam dados considerados sensíveis:

Figura 5: Exemplo de tabela em .xlsx analisada nos testes

Matrícula	Nome	Sexo	Idade	CPF
00015	Guilherme Maciel	MASC	25	11.222.333-44
00013	Felipe Lopes	MASC	35	123.456.789-00
00019	Davi Lucas	MASC	19	987.654.321-00
00035	Emmanuele Prado	FEM	22	500.400.300-20
00021	Fernando Silva	MASC	42	999.888.777-88

Fonte: Autoria própria (2022)

Figura 6: Exemplo de documento em .pdf analisado nos testes

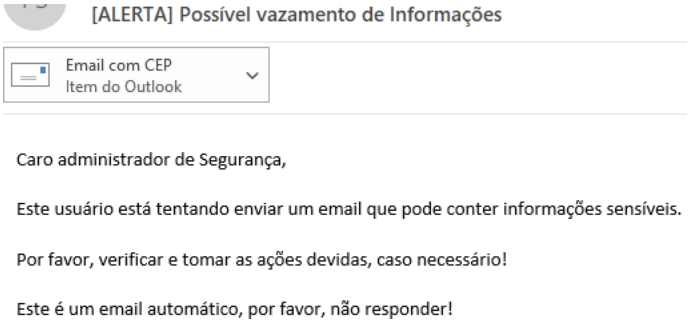
Fonte: Autoria própria (2022)

O comportamento do script quando encontrados *caracteres* compatíveis com os padrões das expressões regulares foi acionar o "gatilho" de envio de um e-mail de alerta.

É importante ressaltar que quando a informação passa pelas validações da expressão regular, é verificado apenas o padrão encontrado, logo não é confirmada a veracidade e/ou legitimidade dos dados.

Conforme configurado, foi então recebido em um endereço de e-mail, uma mensagem sobre uma possível violação dos dados, como visto na figura (8).

Figura 7: *Exemplo de e-mail de alerta recebido*



Fonte: *Autoria própria*

A mensagem recebida pelo administrador, foi enviada do endereço do próprio usuário e nela contém o e-mail original que causou o alerta, para que fossem feitas as análises necessárias por parte do responsável.

Como dito anteriormente, o sistema atua na máquina do usuário, mais especificamente ao fazer uma conexão com o Outlook presente na máquina dele, o que possibilita toda análise das mensagens e o envio dos emails.

A ideia inicial do projeto não é gerar bloqueios nos e-mails enviados pelos usuários (tendo em vista que falsos-positivos podem ocorrer e gerar crises na comunicação), mas gerar um alerta sobre o que está sendo compartilhado pelos usuários para evitar surpresas desagradáveis.

Em e-mails que não possuem anexos, foi analisado apenas o corpo do e-mail, que foi convertido em um arquivo com extensão .pdf. Após realizada esta conversão, foi então realizada a busca pelos padrões neste arquivo .pdf.

Em e-mails que possuem anexos, primeiro foi realizada a conversão do corpo do e-mail e a análise no arquivo convertido, e depois analisados todos os arquivos anexados ao e-mail, não tendo limite de quantidade de anexos e/ou quantidade de páginas dos arquivos.

Com êxito, as análises realizadas em e-mails que possuíam informações como CEP, CPF, CNPJ, expressões ligadas a etnia, religião e sexualidade foram captados pelas expressões regulares.

Os e-mails que não possuíam informações sensíveis também performaram bem, não executando nenhuma ação, como programado.

Apesar dos ótimos resultados obtidos, foi percebido que a aplicação desenvolvida não consegue identificar padrões quando estes estão

DOI: 10.25286/repa.v10i1.2508

inseridos em imagens (estando elas dentro de anexos ou como anexos), logo, neste cenário as expressões regulares não conseguiram analisar as informações do e-mail em sua totalidade.

Dos cinquenta testes realizados, em apenas dez deles o resultado não aconteceu como esperado. Nestes dez testes malsucedidos, os dados a serem encontrados estavam inseridos em imagens (figura 8) e o script desenvolvido não realizou o reconhecimento de textos dentro delas.

Para que esta análise seja bem-sucedida, são necessários estudos adicionais e o uso de outras bibliotecas, para que a inspeção dos padrões feitos com REGEX sejam aplicados com sucesso.

Figura 8 - Exemplo de imagem em .png anexada a um email utilizada nos testes

Fonte: Autoria própria

Para que o modelo desenvolvido neste artigo seja utilizado em uma empresa real, se faz necessário o aprimoramento deste modelo, com a utilização de uma biblioteca que execute a leitura e faça o reconhecimento dos textos em imagens ou então o uso de alguma outra ferramenta complementar que seja capaz de realizar as análises diretamente nas imagens.

5 TRABALHOS FUTUROS

Como dito anteriormente, a utilização das expressões regulares em análises de e-mails pode ser útil para estar em conformidade com a LGPD. O modelo apresentado pode ser utilizado desde que nos e-mails enviados não haja imagens, pois os padrões de leitura são aplicados em strings.

Recomenda-se para trabalhos futuros a incorporação de um modelo que realize a digitalização de imagens e que consiga transformar o conteúdo delas em strings, para que as expressões regulares consigam atuar diretamente nestas informações e concluir a análise com precisão total, dispensando o uso de uma ferramenta adicional.

6 CONCLUSÃO

A aplicação desenvolvida neste projeto se mostrou eficiente em grande parte dos testes realizados, emitindo os alertas, como esperado quando os padrões foram identificados em e-mails, sendo necessário um aprimoramento para abranger mais extensões de arquivos nas análises.

A ferramenta pode ser útil aos responsáveis por Segurança da Informação em ambientes empresariais, pois permite a rápida ciência do possível compartilhamento proibido de dados, evitando maiores transtornos posteriormente. Com a finalização deste trabalho, empresas que não possuem iniciativas de proteção de dados podem ser beneficiadas.

Espera-se que exista a conscientização na forma como os dados são compartilhados, pois dependendo dos dados e da exposição, os danos podem ser gigantes.

REFERÊNCIAS

[1] FORTINET. **FortiGuard Labs**. 2022. Disponível em: <https://encr.pw/fortinet-ataques-ciberneticos>. Acesso em: 5 jun. 2022.

[2] ÉPOCA. **Época Negócios**. 2020. Disponível em: <https://acesse.one/epoca-crescimentodispositivos>Acesso em:8 jun. 2022.

[3] PESSOA, Raimundo Alan Matos. **Um estudo de caso sobre a gestão da segurança da informação em uma instituição financeira.** 2012. Artigo (Bacharelado em Ciência da Computação) - Universidade Estadual Do Sudoeste Da Bahia, [S. l.], 2012.

[4] BRASIL. Casa Civil. **Constituição da República Federativa do Brasil de 1988,** Artigo 5, Inciso X. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm>. Acesso em: 10 jul. 2022.

[5] BRASIL. Casa Civil. Artigo 21, **Lei n° 10406**, de 10 de janeiro de 2002. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm> Acesso em: 9 jun. 2022.

[6] BRASIL. Secretaria Geral. **Lei Nº 13.709,** de 14 de Agosto de 2018, 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 11 ago. 2022.

[7] Sant'Ana, Ricardo. (2017). **Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação.** Pesquisa Brasileira em Ciência da Informação e Biblioteconomia. 12. 10.22478/ufpb.1981-0695.2017v12n1.34194.

[8] SILVA, Francisco Thiago; Pereira, Manoel Phelipe Xavier; SILVA, Danilo Lima. **VAZAMENTO DE DADOS E A ANÁLISE DA LEI 13.709/18 NO BRASIL, 2022**.

[9] MELLO, A. P. .; MIRAMONTES, G. C. . **LGPD: agentes De Tratamento, Resposável E ANPD.** Cadernos Jurídicos da Faculdade de Direitode Sorocaba, [S. l.], v. 3, n. 1, p. 73–80, 2022. Disponível em: https://www.fadi.br/revista/index.php/cadernosjuridicos/article/view/88. Acesso em: 2 set. 2022.

[10] SOUSA, Sérgio Lopes de. **Lei Geral De Proteção De Dados:** Aspectos Da Titularidade De Dados E A Importância De Um Data Protection Officer Em Uma Instituição De Ensino. 2020. Artigo (Bacharelado em Direito) - PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS, [S. l.], 2020.

[11] SURFSHARK. **SurfShark**. 2022. Disponívelem: <https://surfshark.com/research/data-breach-monitoring>. Acesso em: 14 out. 2022.

[12] KASPERSKY. **Kaspersky Daily**. 2021? Disponível em: <https://www.kaspersky.com/blog/the-human-factor-in-it-security/>. Acesso em: 14 set. 2022.

[13] M. Pandey and V. Ravi, "**Detecting phishing e-mails using text and data mining,"** *2012 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2012, pp. 1-6, doi: 10.1109/ICCIC.2012.6510259.

[14] JARGAS, Aurelio Marinho. Livro Online, **Expressões Regulares - Guia de Consulta Rápida.**Editora Novatec, 2001. Disponível em: <https://aurelio.net/regex/guia/>. Acesso em: 24 out. 2022.

[15] PYTHON. **re — Regular expression operations** — Python 3.7.2 documentation. Disponível em: <https://docs.python.org/3/library/re.html>. Acesso em: 15 jun. 2022.

[16] CUNHA, José Antônio. **Desmistificando Expressões Regulares**. Congresso Iniciação Científica do CEFET-RN, 2005. Acesso em: 13 jun. 2022.

[17] J. Li, J. Xing and J. Li, "**Text analysis technology in crew collaboration scheduling system for space missions**," *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, China, 2014, pp. 43-47, doi: 10.1109/ITAIC.2014.7065002.

[18] J. Zhang and H. Yao, "**A Chinese document parsing and code recognition system using Regex and SVM**," *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, 2021, pp. 1860-1864, doi: 10.1109/IAEAC50856.2021.9390975.

[19] S. Thivaharan., G. Srivatsun. e S. Sarathambekai., "**A Survey on Python Libraries Used for Social Media Content Scraping**," *2020 International Conference on Smart Electronics and Communication (ICOSEC)* , Trichy, India, 2020, pp. 2020.9215357.

[20] S. Sun, "**Stably extracting text contents from email messages with Python**," *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, London, UK, 2009, pp. 199-203, doi: 10.1109/ICADIWT.2009.5273961.

DOI: 10.25286/repa.v10i1.2508