CENTRO ESTADUAL DE EDUCAÇÃO TECNOLOGÓGICA PAULA SOUZA
FACULDADE DE TECNOLOGIA DE JUNDIAÍ - "DEPUTADO ARY FOSSEN"
CURSO SUPERIOR DE TECNOLOGIA EM GESTÃO DA TECNOLOGIA DA INFORMAÇÃO

WEB SCRAPING APLICADO A CIBERSEGURANÇA

Alyfer Ricardo Sousa de Oliveira
Andressa de Lima Chiquini
Thais Saturnino de Lima

Jundiaí – SP
2023

Alyfer Ricardo Sousa de Oliveira

Andressa de Lima Chiquini

Thais Saturnino de Lima

WEB SCRAPING APLICADO A CIBERSEGURANÇA

Trabalho de Graduação apresentado à Faculdade de Tecnologia de Jundiaí – “Deputado Ary Fossen” como requisito parcial para a obtenção do título de Tecnólogo em Gestão da Tecnologia da Informação, sob a orientação do Professor Rafael Gross.

Jundiaí – SP

2023

Alyfer Ricardo Sousa de Oliveira

Andressa de Lima Chiquini

Thais Saturnino de Lima

WEB SCRAPING APLICADO A CIBERSEGURANÇA

Trabalho de conclusão de curso apresentado ao curso de Tecnólogo em Gestão da Tecnologia da Informação em 2022, em nível de Graduação, da Faculdade de Tecnologia de Jundiaí (FATEC Deputado Ary Fossen), como requisito parcial à obtenção do certificado Superior de conclusão de curso.

Aprovado em: ____/____/_____ Banca examinadora:

_____

Prof. Rafael Gross (Orientador)

Faculdade de Tecnologia de Jundiaí - FATEC Deputado Ary Fossen

_____

Prof. [titulação] [Nome do 2º examinador]

Faculdade de Tecnologia de Jundiaí - FATEC Deputado Ary Fossen

Dedicamos este trabalho
aos que despertaram
o nosso interesse
no estudo das novas
tecnologias.

AGRADECIMENTOS

Agradecemos primeiramente a Deus que nos mantem de pé e nunca nos abandonou em nossa trajetória, que está conosco em todas as nossas escolhas, mesmo quando obstinamos em fazer as erradas. Além de nos conceder a dádiva da vida.

Agradecemos também às nossas famílias, que nos incentivam e nos acompanham no dia a dia, que nos orientam sempre buscando a melhor direção. E aos nossos amigos que nos auxiliaram durante o curso, com conhecimentos e novidades, com risadas e choros e até mesmo com discussões, mas que foram necessárias para todo o nosso desenvolvimento até então.

Por último, mas não menos importante, agradecemos à FATEC Jundiaí e a todo seu corpo docente, que, academicamente, nos forneceram a base de nossos objetivos profissionais e a avançar barreiras que achávamos impossíveis. Ao nosso orientador que nos incentivou a pesquisar sobre tecnologias recentes e atualizadas. Ao corpo discente, aos alunos e colegas de classe, por todo o apoio nas disciplinas, pelos trabalhos concluídos e pelas risadas dadas, que sustentaram um ambiente leve e de cooperação entre todos. E por último, mas não menos importante, nas pessoas: professores, prestadores de serviços ou palestrantes, agradecemos por todos os conteúdos e ensinamentos de quem passou por nós nesse tempo.

Nosso mais sincero obrigado!

EPÍGRAFE

(A criação bem-sucedida de inteligência artificial seria o maior evento na história da humanidade. Infelizmente, pode também ser o último, a menos que aprendamos a evitar os riscos)

The successful creation of artificial intelligence would be the biggest event in human history. Unfortunately, it could also be the last, unless we learn to avoid the risks.

Stephen Hawking

OLIVEIRA, Alyfer R. Sousa de; CHIQUINI, Andressa de Lima; LIMA, Thais Saturnino de. **Web-scraping aplicado a Cibersegurança. 50** f. Trabalho de Conclusão de Curso de Tecnólogo em Gestão da Tecnologia da Informação. Faculdade de Tecnologia de Jundiaí – “Deputado Ary Fossen”. Centro Estadual de Educação Tecnológica Paula Souza. Jundiaí. 2023.

RESUMO

A raspagem de dados da *Web* é utilizar uma ferramenta que permite coletar automaticamente informações de sites. Embora seja comumente frequentes para coletar informações públicas, é importante observar que o uso de *web scraping* deve ser feito de forma ética e em conformidade com as leis de proteção de dados. Para a maioria das organizações que pretendem ter acesso à informação atual, correta, exata e relevante, e que pretendem também conhecer as tendências e comportamentos dos seus usuários face aos acessos aos seus sites, a implementação de *web scraping* é a ferramenta de eleição que atende bem a todos esses requisitos e, além disso, quando tudo é feito de forma totalmente automática, economiza muito tempo e é mais eficiente e eficaz, economizando também o esforço físico e mental dos profissionais. Porém uma das ameaças de segurança cibernética de hoje é a raspagem da *web*. Uma vez que utilizada com más intenções, não apenas ajuda *hackers* e golpistas a lucrar com o trabalho e o conteúdo de outras empresas, mas também prejudica o desenvolvimento de seus negócios e evoluções.

Palavras-chave: Coleta de dados. *Proteção de dados*. Inteligência Artificial.

OLIVEIRA, Alyfer R. Sousa de; CHIQUINI, Andressa de Lima; LIMA, Thais Saturnino de. **Web-scraping Applied to Cyber Security. 50** p. End-of-course paper in Technologist Degree in Information Technology Management. Faculdade de Tecnologia de Jundiaí – “Deputado Ary Fossen”. Centro Estadual de Educação Tecnológica Paula Souza. Jundiaí. 2023.

ABSTRACT

Web data scraping is utilizing a tool that allows you to automatically collect information from websites. While it is commonly asked to collect public information, it is important to note that the use of web scraping must be done ethically and in compliance with data protection laws. For the majority of organizations that want to have access to current, correct, accurate and relevant information, and that also want to know the trends and behaviors of their users when accessing their websites, the implementation of web scraping is the tool of choice that well to all these requirements and, moreover, when everything is done fully automatically, it saves a lot of time and is more efficient and effective, also saving the physical and mental effort of professionals. But one of today's cybersecurity threats is web scraping. Once used with bad intentions, it not only helps hackers and scammers to profit from other companies' work and content, but also harms their business development and evolutions.

Keywords: Data Collect. *Data Protection*. Artificial intelligence.

LISTA DE ILUSTRAÇÕES

LISTA DE SIGLAS

1. URLs: A sigla *URL* significa: *Uniform Resource Locator*, que é definida como “Localizador Uniforme de Recursos”.
2. HTML: Sigla para *HyperText Markup Language* — Linguagem de Marcação de Hipertexto —, o *HTML* é o componente base da web. Isso quer dizer que ele permite a construção de websites e a inserção de novos conteúdos, como imagens e vídeos, por meio dos hipertextos.
3. CSV: O arquivo *CSV* (valores separados por vírgulas) é um arquivo de texto com formato específico para possibilitar o salvamento dos dados em um formato estruturado de tabela.
4. JSON: Esta sigla é um acrônimo para *JavaScript Object Notation*, ou notação para objeto em *JavaScript*, justamente por ele ser um derivado desta linguagem. Ele define uma formatação para armazenamento, e transferência de dados em formato texto.
5. XML: A sigla significa “*eXtensible Markup Language*” (linguagem de marcação extensível), que é basicamente um formato de arquivo universal usado para criar documentos com dados organizados.
6. ERP: A sigla significa *Enterprise Resource Planning*, ou planejamento de recursos empresariais, é um sistema de gestão integrado que consegue organizar diversas áreas de uma empresa em um só sistema, gerenciando os dados da empresa em um banco de dados único. Isso permite automatizar processos e cria uma visão geral muito mais confiável para a tomada de decisão dos gestores.
7. CPF: O Cadastro de Pessoa Física. Ele é um documento feito pela Receita Federal e serve para identificar os contribuintes. O CPF é uma numeração com 11 dígitos, que só mudam por decisão judicial. O documento é emitido pela Receita Federal.
8. RG: As letras significam Registro Geral. É considerado o documento mais importante do cidadão, pois representa identidade de cada pessoa registrada no Brasil.
9. LGDP: Lei criada com o principal objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural. Também tem como foco a criação de um cenário de segurança jurídica, com a padronização de regulamentos e práticas para promover a proteção aos dados

pessoais de todo cidadão que esteja no Brasil, de acordo com os parâmetros internacionais existentes.

10. EDM: Em sua essência é sigla para *Electronic Document Management* – ou seja, nada mais é do que a sigla em inglês para o nosso conhecido *GED* (Gestão Eletrônica de Documentos). Se trata dos processos e tecnologias otimizadas para gerir dados, informações e documentos por meio digital.

11. APIs: *Application Programming Interface* (Interface de Programação de Aplicação). No contexto de APIs, a palavra Aplicação refere-se a qualquer software com uma função distinta. A interface pode ser pensada como um contrato de serviço entre duas aplicações.

SUMÁRIO

1. INTRODUÇÃO

O *Web scraping* é uma ferramenta que permite a coleta de dados automatizada de *websites*. “[...] Essa técnica possibilita a aquisição de grandes quantidades de dados em tempo reduzido, permitindo assim análises e estudos para desenvolvimento de modelos de inteligência artificial. [...]” (SOUZA, 2022). A utilização da ferramenta pode ser realizada para transformar dados em informação inteligente para gerar *insights* de negócio, qualificar a prospecção de novos clientes ou avaliar a reputação da marca ou de um produto, dentre outros diversos métodos legais para encontrar informações em grande escala, como documentos na base do governo. Embora seja comumente usado para coletar informações públicas, é importante destacar que o uso de *web scraping* deve ser feito com ética e em conformidade com as leis de proteção de dados.

A problemática sobre o tema possibilita que indivíduos e profissionais de empresas obtenham acesso a informações públicas sobre qualquer pessoa, tais como números de documentos, telefone celular, endereço de e-mail e senhas, fotos de perfis em redes sociais, idade e gênero, bem como outras informações confidenciais. Contudo, é importante salientar que a natureza das informações coletadas depende da fonte acessada pela ferramenta, e algumas pessoas podem utilizar essa técnica de forma maliciosa.

Além disso, é válido destacar que as redes sociais são frequentemente alvo de invasões, e que tais ataques podem permitir que os invasores tenham acesso a postagens públicas e a todo tipo de conteúdo que esteja disponível publicamente. Com o acesso concedido pela plataforma digital, como *Instagram*, *Facebook* e *Twitter*, é possível descobrir o número correto de seguidores, pessoas que estão sendo seguidas, bem como o nível de engajamento do usuário na rede social e os *links* acessados e compartilhados.

E como base nas indagações remete-se a seguinte hipótese pela crescente popularidade do *web scraping* como uma ferramenta de coleta de dados automatizada tem levantado diversas questões acerca de seus potenciais benefícios e desafios para as empresas. Entre as principais indagações suscitadas, destacam-se: quais são as

vantagens do uso de *web scraping* na obtenção de dados para tomada de decisões empresariais? E quais são os benefícios de implementar medidas de cibersegurança para mitigar os riscos associados ao uso dessa ferramenta? Essas questões são cruciais para avaliar a viabilidade e a sustentabilidade do uso da *web scraping* em diferentes contextos empresariais, levando em conta não apenas seus potenciais benefícios, mas também seus riscos e desafios inerentes.

Para a maioria das organizações que desejam ter acesso a informações atuais, corretas, precisas, e relevantes e também desejam saber sobre as preferências e comportamentos de seus usuários de acordo com os acessos em seus sites, a implementação de *web scraping* é a opção de ferramenta que atende bem todos esses quesitos, e além de tudo é feito de forma totalmente automática economizando muito tempo, e sendo mais eficiente e eficaz que a coleta de dados manual economizando assim também o esforço dos profissionais.

Para a coleta de dados ser feita é preciso seguir um passo a passo do processo de raspagem de dados, o primeiro passo é fazer a identificação do site a ser coletado, logo depois é feita a coleta de *URLs* de todas as páginas de onde os dados serão extraídos, então é feita uma solicitação a esses *URLs* para se obter o *HTML* da página analisada, após essa parte do processo os localizadores serão utilizados para encontrar os dados dentro do *HTML*. Após toda a extração dos dados, os arquivos devem ser salvos em formatos estruturados, geralmente sendo utilizados os tipos de arquivos *CSV* ou *JSON*, após a finalização do processo, o arquivo já pode ser analisado pelos especialistas de acordo com a finalidade de cada organização.

Como já citado essa técnica, permite e dá acesso a todo tipo de informação, então o *web scraping* é usado de maneira ilegal para ataques e crimes cibernéticos, mas quando se tem um conhecimento amplo desta ferramenta alinhado ao uso da cibersegurança, que é a prática que protege tudo o que envolve a internet e o ambiente digital, é garantido aos usuários um acesso seguro e confiável, mantendo a confidencialidade e a integridade de todas as suas informações pessoais. Desta maneira, é mais fácil de se evitar e combater esses ataques, impedindo que as violações ocorram melhorando a segurança cibernética da organização.

Se o principal objetivo tem por finalidade garantir a integridade e fidelidade das informações, então o problema maior não é o acesso a este tipo de informações, mas sim no que se pode fazer com elas. Sobre os vazamentos, têm-se que:

> Para evitar novos vazamentos, as plataformas têm bloqueado a coleta de dados por robôs e lançado novas opções de privacidade. Mas como nem sempre é possível evitar as ações de *hackers* e bots, o usuário deve tomar o máximo cuidado para que suas informações não caiam em mãos erradas. (TECMUNDO, 2022).

Apesar das plataformas tentarem, esse bloqueio ainda é muito falho e é neste momento que entra a Cibersegurança, que faz a proteção de computadores, servidores, dispositivos móveis, redes e dados, e os sistemas eletrônicos das pessoas/organizações contra esses tipos de ameaças e ataques sendo entre eles os principais: o ataque ou terrorismo cibernético e o crime virtual.

Dentre muitas ameaças à segurança cibernética, uma das mais atuais é o *web scraping*, que se resume a ser um processo de utilização de robôs para extrair rapidamente grandes quantidades de dados de um site, salvando as informações para uso pessoal. A técnica é um tanto controversa: embora empresas legítimas a utilizem, o *web scraping* também é utilizado de forma ilegal, por exemplo, para surrupiar conteúdo protegido por direitos autorais e criar ofertas de um produto ou serviço para concorrência desleal.

Ele também permite que hackers e golpistas lucrar com o trabalho e o conteúdo de outras empresas. o que é danosa aos seus negócios. Esses bots geralmente são programados para imitar o comportamento de usuários usuais. Isso dificulta a detecção e o bloqueio. E essa técnica está se tornando cada vez mais acessível e barata.

Esta pesquisa teve como objetivo analisar o uso do web-scraping em seus diferentes aspectos, colocando-o como objeto de estudo dentro do âmbito da cibersegurança. Para tanto, foram pesquisados mecanismos e ferramentas eletrônicas para analisar e validar os níveis de segurança oferecidos mediante ameaças encontradas no mundo cibernético. Visando atingir o Objetivo Principal, alguns Objetivos Específicos são requeridos:

Encontrar e analisar soluções que tenham a capacidade de realizar a análise e proteção de variadas formas que possam ser previstas, como por exemplo: Inspeção das interações de usuário e navegador; validar integridade do dispositivo de uso; detectar imitações, ataques falhos; dentre outros, de maneira que não afete a maneira da aplicação; como melhorar e ampliar a proteção de dados do usuário nas plataformas, de maneira que seu acesso e utilização sejam limpos; analisar a quantidade de usuários e suas respectivas formas de segurança de dados, quais métodos utilizam para que sua segurança seja fortificada; realizar uma discussão dos resultados obtidos.

2. REFERENCIAL TEÓRICO

Cada vez mais os números de usuários na internet aumentam, fazendo com que os acessos também aumentam. É perceptível que o ambiente digital está se expandindo cada vez mais e tomando espaço entre as pessoas. Todos estão tentando se adequar, inovar, e usar as novas ferramentas e técnicas do meio tecnológico para obterem sucesso entre o mundo e um mercado competitivo. Os acessos não servem apenas para benefício do usuário, mas também para o governo e organizações que desejam se manter no mercado e precisam saber o que atingem o seu público. Para isso, é necessário realizar análises e comparativos que costumam ser criteriosos, com dados e informações verdadeiras e de fontes seguras. Para que estes dados sejam extraídos, o acesso a grande quantidade de informações é indispensável. Mas afinal, o que seria a extração, qual o conceito de coleta de dados?

> A coleta de dados é um processo utilizado para captar informações geradas pelas pessoas (ou por processos) e que servirão de insumos para planejar estratégias para o negócio. Esses dados podem ser coletados em plataformas específicas para coletas, formulários, sites e outras metodologias. (ALINE OLIVEIRA, 2022).

O *Web Scraping* é uma técnica amplamente utilizada e ao mesmo tempo pouco conhecida no Brasil. Mas, dentro do âmbito da Cibersegurança, seu monitoramento se torna de extrema importância, como ressaltado por MITCHEL (2019, .p 11):

> [...] Os web scrapers podem acessar lugares que as ferramentas de pesquisa tradicionais não conseguem. Uma pesquisa no Google por “voos mais baratos para Boston” resultará em uma grande quantidade de anúncios publicitários e sites populares para busca de voos. O Google sabe apenas o qe esses sites dizem em suas páginas de conteúdo, mas não os resultados exatos de várias consultas fornecidas a de uma aplicação de busca de voos. [...]

No entanto, um *Web Scraper* bem desenvolvido pode colocar em um gráfico o custo de um voo para *Boston* ao longo do tempo para uma variedade de sites e informar qual é o melhor momento para comprar uma passagem.

Para utilizar o *web scraping* de forma segura, é importante lembrar que mexer com dados exige extrema atenção, pois pode violar algum direito da Lei Geral de Proteção de Dados (LGPD). De acordo com a Lei de Proteção de Dados:

> A LGPD regulamentará qualquer atividade que envolva utilização de dados pessoais, inclusive nos meios digitais, por pessoa natural ou jurídica, no território nacional ou em países onde estejam localizados os dados. LGPD (2023).

A Cibersegurança zela por essas informações que são extraídas dos *websites*, para que não sejam utilizadas de forma maliciosa. Revela-se assim um questionamento, que se abre mediante a uma possibilidade de análise do fato, que foi definido por um autor:

> [...] enquanto negócios legítimos a utilizam, o *web scraping* também é usado ilegalmente para, por exemplo, roubar conteúdos protegidos por direitos autorais e gerar cotações de um produto ou serviço para uma concorrência desleal. PUGA (2017).

3. ANÁLISE E A COLETA DE DADOS

O princípio inicial é o de que para a formação de uma análise, é necessário realizar a atividade de pesquisa, utilizando ferramentas e sites que possibilitam a coleta e análise dos dados que forem fornecidos para que seja possível a formação de um

conjunto de informações. É uma prática fundamental para realizar as análises de comportamento, preferências, estatísticas etc., o que permitirá desenvolver campanhas de marketing ou até mesmo desenvolver campanhas mais profundas de conscientização.

As informações são valiosas, já que é com elas que as empresas podem conhecer mais a opinião de seu público-alvo, seu mercado e seus resultados, além de reconhecer a sua imagem vista de fora. Para isso, inicia-se a coleta de dados e suas respectivas análises.

Devido à sociedade atual viver na era digital, os dados são uma verdadeira mina de ouro em questão de material para uma coleta produtiva. Mas não basta apenas captar esse conteúdo, é importante considerar algumas condições para que seus resultados sejam coerentes com a pesquisa e o objetivo final da análise: é importante saber onde buscá-lo, como estruturá-lo e onde será aplicado, considerando as ferramentas que serão utilizadas no decorrer da captação dos conteúdos.

A coleta de dados é um processo de captação de conteúdo estratégico, que pode ser realizada através de ferramentas próprias para isso, formulários, questionários, dentre outros *softwares* ou aplicativos que retenham essas informações.

Esse processo ocorre com o objetivo de garantir que as empresas reconheçam a fundo como anda o mercado, a percepção dos consumidores, resultado de setores que a abrangem e o desempenho do negócio de modo geral. Apesar de aparentar ser um evento realizar a coleta de dados, ela deve ser tratada como uma atividade rotineira, assim os dados sempre estarão atualizados e não haverá um acúmulo de informação (sejam elas ultrapassadas ou em grandes quantidades, de maneira desnecessária).

Devido aos dados, é possível otimizar ações, planejar campanhas mais adequadas, reconhecer como o consumidor vê a empresa e, consequentemente, agir de forma estratégica as melhorias necessárias para sempre atrair e ampliar seu público. (FERREIRA, 2020).

Por estarmos vivendo a era digital, os dados estão por toda a parte, com o acesso à internet na palma de nossas mãos, são gerados a partir de um simples toque entre consumidor e um aparelho conectado à rede.

3.1. TIPOS DE ANÁLISES

Além de entender a importância da análise de dados, é necessário organizar as ideias para que tenham em mente a transformação que pode ocorrer devido a uma boa gestão dela, com resultados considerativos, como: aumento da colaboração do time, decisões mais assertivas e eficiência para os processos comerciais, sem que haja desperdício ou escassez no mercado.

3.1.1. ANÁLISE DESCRITIVA

É uma análise baseada em fatos, ou seja, este tipo de avaliação é feito a partir de resultados concretos que foram obtidos. Exemplos:

- Relatórios;
- Segmentação e controles;
- Análise de negócio;
- Aplicação métrica;
- Avaliação de resultados

3.1.2. ANÁLISE PREDITIVA

Este é o modelo estimado entre as análises, já que seu objetivo é a previsão de cenários futuros com base na análise de padrões revelados pelos dados, um adendo é de que os cenários variam de acordo com as condições determinadas. Exemplo:

- Haverá o ingresso de concorrentes no mercado: a análise não será capaz de te dizer quando o concorrente iniciará as atividades. Em contrapartida, te ajudará prever o que pode ocorrer quando iniciar com base em situações anteriores.

3.1.3. ANÁLISE PRESCRITIVA

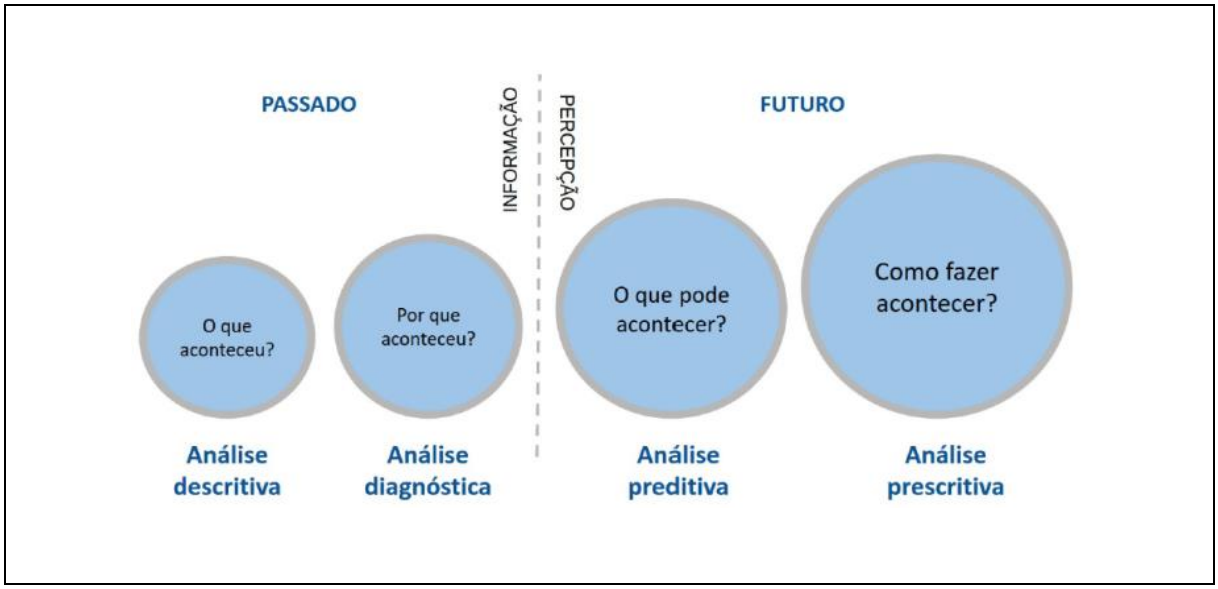
A análise prescritiva é uma continuação após a análise preditiva, já que se baseia em uma recomendação posterior a uma potencial previsão. Este modelo se baseia em realizar projeções com o direcionamento para obter o melhor resultado a partir das possibilidades das situações preditivas.

Por ser um modelo de constante atualização, é comum ser acompanhado por tecnologias de algoritmos e inteligências artificiais. As ferramentas ampliam as análises e nos apresentam padrões diferenciados e percepções de objetivos organizacionais, assim como limitações no processo e/ou fatores de influência.

3.1.4. ANÁLISE DIAGNÓSTICA

Assim como a análise descritiva, a análise diagnóstica é referente a algo que já ocorreu, mas tem como objetivo encontrar relações de causa e efeito para destrinchar um acontecimento, como não é uma atividade simples de ser realizada, o método se baseia em probabilidades.

Figura 1 - Análise Preditiva ou Prescritiva? Sua empresa precisa de ambos



Fonte: MEDIUM, 2019.

4. AUTOMATIZAÇÃO DA COLETA DE DADOS

Apesar de a coleta de dados ser efetiva, muitas das vezes a realização de sua atividade é extensa, cansativa e repetitiva, ela contém diversos objetivos eficientes para um melhor desenvolvimento dentro da empresa, mas, por ser realizada de maneira manual tem como consequência alguns problemas, como a não padronização, tempo de validação de informações, alguns pontos que passam despercebidos, dentre outros.

Com o fato de que hoje as informações são dinâmicas e constantes, para as empresas, é crucial a sua confiabilidade e atualizações constantes, então a automatização é um meio de realizar esse processo. Iniciando pela busca de um *software* que interaja com um ERP utilizado na empresa, já reduz o tempo de trabalho em uma atividade do funcionário, já que a partir dali “um simples clique na tela” já realiza toda a comunicação que, ora antes, levasse até dias para poder acontecer. Tendo uma atualização em tempo real, a eficiência operacional, visibilidade e a rentabilidade dos trabalhos se tornam maiores e melhores devido a uma integração total entre os funcionários.

4.1. VANTAGENS

Como já apresentado, o tempo de processo será reduzido em muitos aspectos, mas as vantagens não se resumem a apenas isso, conta também com:

- Facilita o acesso à informação;
- Reduz custos de negócio;
- Maior volume de produção;
- Padronização.

4.2. DESVANTAGENS

Naturalmente, quando estamos a adotar uma determinada tecnologia ou promover uma mudança mais disruptiva, sempre irão surgir os fatores negativos referente a implementação, como:

- Investimento inicial;
- Possível inflexibilidade na execução de processos;
- Erros e negligências que afetam o processo e o interrompem;
- Erros em alta escala.

5. WEB SCRAPING

Cada vez mais o volume de informações necessárias para construir uma estratégia de *marketing* que leve às organizações ao mercado aumenta gradativamente. O *web scraping* é utilizado para coletar dados de páginas *web*, transformando-os em dados semiestruturados e em textos simples.

Geralmente o processo é todo automático. Existem variações de ferramentas e linguagens de programação para a conversão desses dados, transformando os dados em formatos como *CSV*, *XML*, *JSON*, mas, também pode ser feito manualmente. Um exemplo simples de comando é o famoso "copiar e colar". Com ele, o usuário extrai toda informação que ele deseja de forma manual. As diferentes ferramentas de *web scraping* automáticas tem níveis de complexidade e isso depende da exigência do usuário e quais os parâmetros cadastrados para fazer o processo.

Conforme observado no tópico anterior, a automatização reduz o tempo do processo de coleta dos dados, produzindo em maior quantidade com facilidade. Dentro deste âmbito da coleta dos dados e geração de insights para uma empresa, o *web scraping* se instaura criando *bots* automatizados que coletam os dados de forma rápida e precisa. O uso dessa técnica pode ser responsável por extrair uma grande quantidade de dados e transformá-los em informação, gerando estatísticas que farão parte da criação de estratégias inteligentes.

6. LEI GERAL DE PROTEÇÃO DE DADOS (LGPD)

A LGPD foi a lei criada durante o governo de Michel Temer em agosto de 2018, mas começou a ter efeitos jurídicos somente em setembro 2020. Ela é responsável por proteger e realizar o tratamento (coleta, utilização, compartilhamento e armazenamento) dos dados pessoais sejam eles públicos ou privados.

Dados pessoais são aqueles referentes a identificação de uma pessoa, dados esses utilizados geralmente em cadastros quando é preenchido os campos de informações pessoais, ou seja, dados pessoais como por exemplo o CPF, RG, data de nascimento, endereço, nome, dados bancários etc. Existe um cuidado especial com alguns tipos de dados sendo eles considerados dados pessoais sensíveis e não podem ser disponibilizados de qualquer maneira ou a qualquer pessoa, e são eles: dados relacionados a religião, etnia e origem racial, opinião política, dados biométricos, relacionados a vida sexual, saúde ou genética.

A proteção dos dados pessoais é de extrema importância pois seu tratamento deve ser feito apenas com autorização e consentimento explicito do titular desses dados, por exemplo quando uma pessoa tem acesso aos dados bancários de outra com o intuito de usar para aplicar golpes, fazer compras solicitar empréstimos ou cartões é algo que gera muitos transtornos ao titular essas atitudes se adequam ao uso indevido e violação de dados, mas quando o titular decide ir até uma agencia bancária ou até mesmo acessar um banco digital e decide de livre e espontânea vontade fornecer suas informações para pedir limite, cartões realizar o financiamento é um tipo de tratamento correto desses dados.

Lembrando que a LGPD é uma lei válida em território nacional, para a proteção dos dados tratados no Brasil. A lei entrou em ação para a proteção dos dados e gerar atenção maior a quem fornece e a quem recebe essas informações. Qual seria então a relação do *Web Scraping* é a LGPD? Lembra-se que o *Web Scraping* é uma ferramenta de raspagem de dados, portanto as vezes durante este processo de coleta incluem alguns dados pessoais, mas se a coleta é feita sem consentimento entra na violação da LGPD gerando as devidas punições como multas podendo ser diárias ou até mesmo de 2% sobre o faturamento da empresa(caso a empresa tenha violado) e também advertências, bloqueio ao acesso dos dados, e caso seja um profissional que trabalha com tratamento de dados, será proibido parcialmente ou totalmente de exercer sua função.

6.1. DADOS SENSÍVEIS E NÃO SENSÍVEIS

Dados são um conjunto de valores brutos que levam ao receptor informações, podendo ser elas, confidenciais ou públicas. Os dados sensíveis são aqueles que devem ser tratados com confidencialidade, pois de acordo com gov.br (2021), o tratamento dos dados sigilosos precisa ter um fim definido. Estes transportam informações que não podem ser passados a terceiros sem o consentimento prévio do indivíduo o qual ele pertence.

Em paralelo a isso, em conformidade com a Lei de Proteção de Dados Pessoais (Art. 5º, X), o tratamento é realizado com dados que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração.

Dentro destes conhecimentos retirados de websites públicos, é possível realizar uma análise ética e coerente do tratamento destas informações e como elas devem ser usadas corretamente para um *web scraping* ético que não infrinja as leis e normas da utilização dos dados.

6.2. CIBERSEGURANÇA

Quando falamos em cibersegurança, estamos tratando de forma geral como a proteção das informações cibernéticas será feita. Essa prática assume os dados com sigilo e descrição, visando evitar possíveis invasões e o furto indesejado de informações. “Um dos focos das ameaças atuais de cibersegurança é o *web scraping*, o processo de usar *bots* para extrair grandes quantidades de dados de um site rapidamente, salvando as informações para uso pessoal. [...]” (PUGA, 2017).

A estratégia traz consigo uma dualidade de objetivos em sua extração, sendo elas voltadas tanto para o lado do roubo de informações, quanto para a sua utilização de forma consciente e ética. Desta forma, dados sensíveis podem ser obtidos do banco de dados do *Host* por meio da leitura e raspagem de *websites* mal estruturados e desprotegidos, ou até mesmo, dados públicos e não sensíveis podem ser manipulados a fim de gerar *insights* e possibilitar a sua aplicação no ambiente corporativo.

7. ESTUDOS DE CASO

A seguir serão apresentados dois estudos de casos para demonstrar e comparar sobre o uso correto e incorreto do *web scraping*.

7.1. APLICAÇÃO DO WEB SCRAPING: ANÁLISE EXPERIMENTAL

Para demonstração de aplicação do *scraping*, uma biblioteca da Linguagem *Python* chamada *SCRAPY* foi selecionada como objeto principal de estudo. Diante disso, para que o caso seja válido, faz-se necessário realizar uma breve abordagem conceitual sobre o assunto.

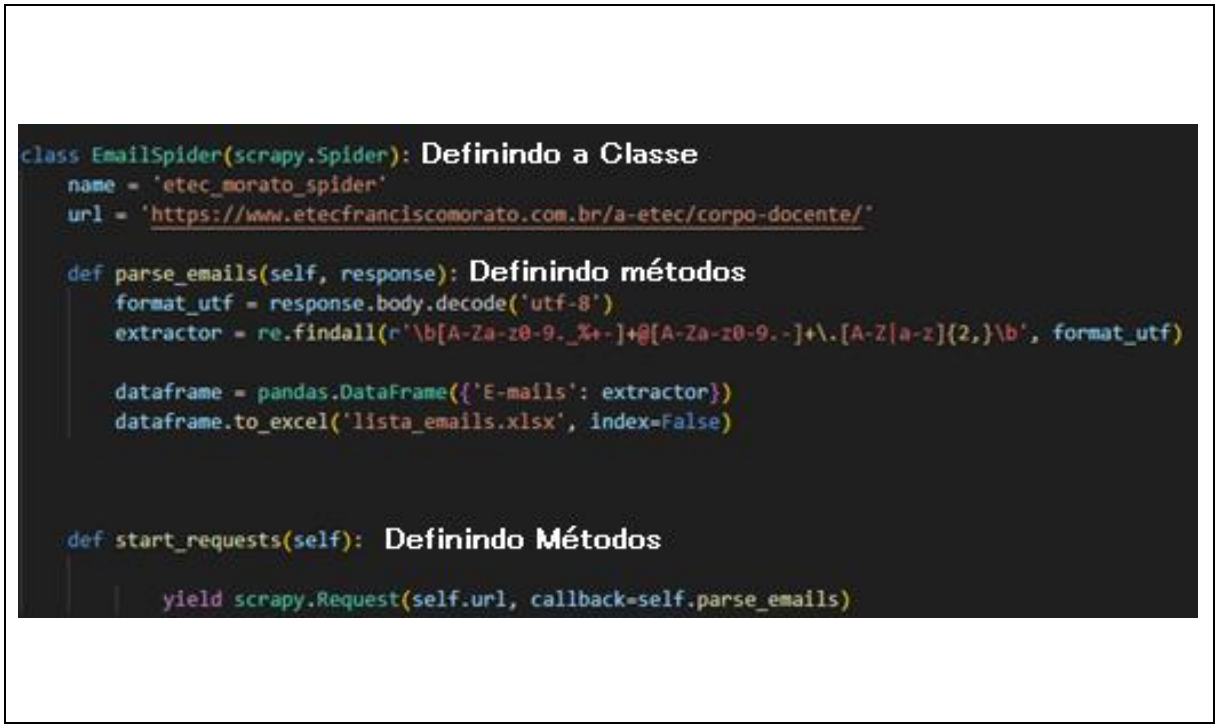
Trazendo para perto o conceito da Linguagem, a página oficial afirma que *Python* é a linguagem da programação que permite trabalhar de forma rápida e integrar sistemas mais efetivamente. *PYTHON* (2019, tradução nossa).[1] Dentro dela, existem as bibliotecas, sendo elas “pacotes”, que contém funções, classes e métodos definidos de forma prévia. Elas auxiliam ao desenvolvedor para que ele utilize somente o necessário e não precise definir mais funcionalidades, deixando o código mais “limpo” (mais leve e visualmente bonito). As funções podem ser definidas dentro da linguagem como blocos de código que facilitam para que ele fique mais organizado e não precise ser reescrito diversas vezes. (LOPES, 2022). Similares a estas, de acordo com *PYTHON*, “Um método é uma função que “pertence” a um objeto instância”. Por fim, as classes definem um tipo de objeto e são utilizadas para criar instâncias deste objeto. Partindo destes pontos, é possível explorar ainda mais as bibliotecas do *Python*, indicando cada uma delas para assim obter maior entendimento do código que será apresentado. As Figuras 2, 3 e 4 indicam corretamente como são definidos cada um dos conceitos apresentados.

[1] No original: Python is a programming language that lets you work quickly and integrate systems more effectively.

Figura 2 - Importando bibliotecas

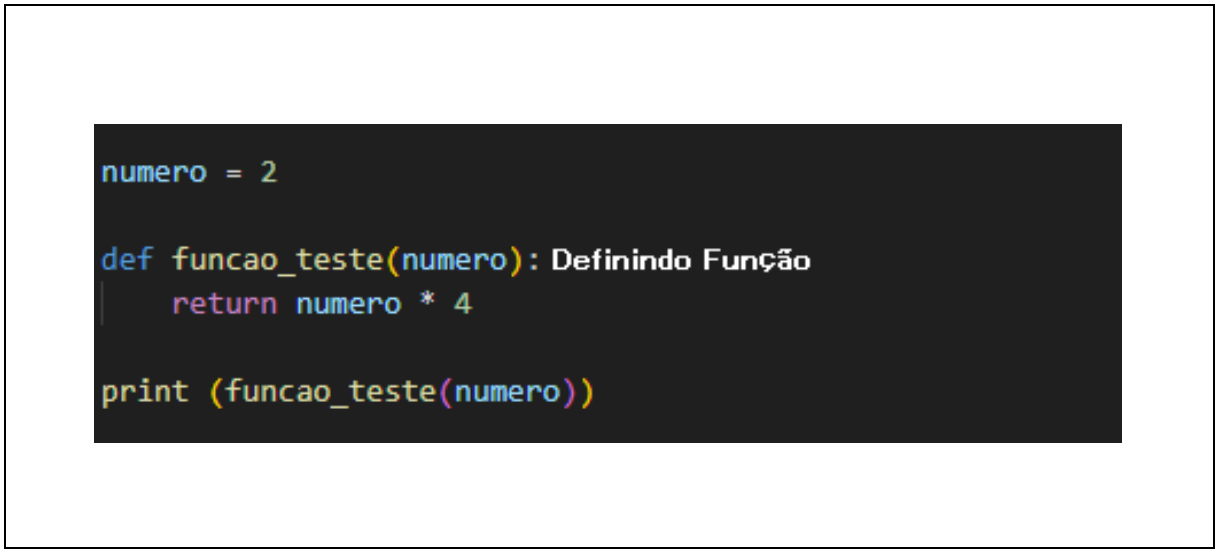
Fonte: Elaborado pelo Autor (2023).

Figura 3 - Definindo Classes e Métodos



Fonte: Elaborado pelo Autor (2023).

Figura 4 - Definindo Funções



Fonte: Elaborado pelo Autor (2023).

O *SCRAPY* é um *framework* bastante utilizado do *python*, que se configura por ser um código aberto e próprio para extração de dados de forma rápida e simples de *websites* (SCRAPY, 2023). O PANDAS é uma biblioteca *Open Source*, que permite ao *Python* analisar dados e manipulá-los de diversas formas (PANDAS, 2023). O RE também é uma biblioteca *Open Source*, responsável por manipular expressões regulares de diversas formas, podendo filtrar e encontrar padrões nelas. Nesta demonstração, será utilizada a biblioteca *SCRAPY*, juntamente às bibliotecas *PANDAS* e *RE*.

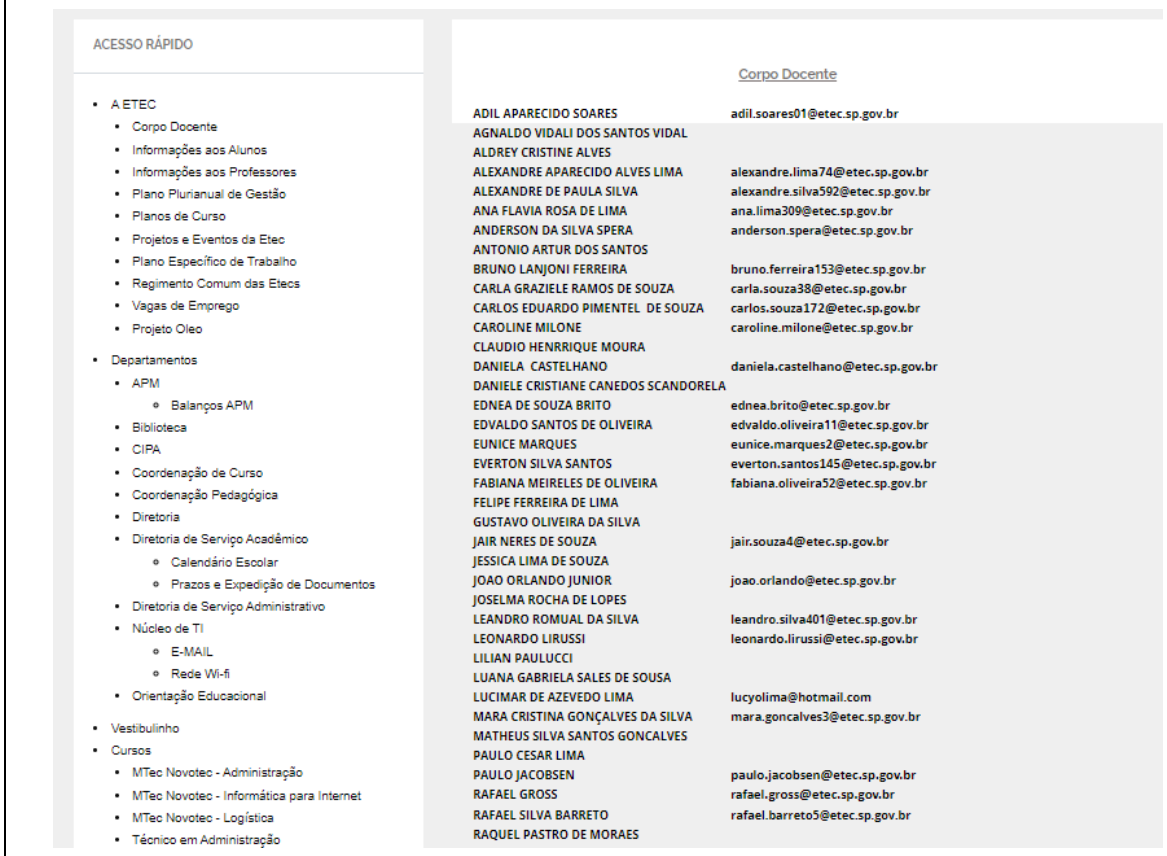
Dentro da biblioteca *SCRAPY*, existem os *spiders*. Um *spider* é definido como uma classe do *SCRAPY* que define como será realizada a extração, quais sites serão raspados e quais dados deseja-se obter, sendo ativado por meio do *Crawl*. O *Crawl* é uma funcionalidade usada para iniciar os *Spiders* e realizar a extração das informações desejadas de um *website*. O *Spider* também contém uma função *Callback* que permite analisar o conteúdo da página (SCRAPY, 2023).

Para realizar a extração de dados, foram escolhidos dois *websites*. Cada *website* terá um *spider,* cujas informações serão indicadas a seguir.

7.1.1. SPIDER: ETEC MORATO

A “Figura 6” indica a estrutura do código que foi elaborada pelos integrantes do grupo, tomando como base o seu objetivo que é: “Extrair os e-mails presentes na página Corpo Docente”, conforme indicado na “Figura 5”.

Figura 5 - Website Etec Morato



Fonte: ETEC FRANCISCO MORATO, 2023.

Figura 6 – Spider: Etec Morato

```
import re
import scrapy
import pandas

class EmailSpider(scrapy.Spider):
    name = 'etec_morato_spider'
    url = 'https://www.etecfranciscomorato.com.br/a-etec/corpo-docente/'

    def parse_emails(self, response):
        format_utf = response.body.decode('utf-8')
        extractor = re.findall(r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b', format_utf)

        dataframe = pandas.DataFrame({'E-mails': extractor})
        dataframe.to_excel('lista_emails.xlsx', index=False)



    def start_requests(self):

            yield scrapy.Request(self.url, callback=self.parse_emails)
```

Fonte: Elaborado pelo autor (2023).

Dentro do *Spider*, foi realizada a importação das três bibliotecas: RE, *SCRAPY* e *PANDAS*. A classe *EmailSpider* será utilizada, e dentro dela, será chamado o argumento *Spider*, que está dentro da biblioteca *scrapy*. Portanto, *Spider* é uma classe do *framework*.

Seguindo pela classe, há o nome do *Spider*, definido por *name* e a *url* de onde os dados serão extraídos. O nome e a *URL* foram definidos como tipo *string*.

Nota: A *URL* também poderia ser definida como uma lista. Desta forma, dentro da função *start_requests*, haveria uma modificação, adicionando um laço de repetição (*for*) para realizar a *request* a cada "*url*" inclusa na lista.

Dentro da classe, dois métodos são definidos. São eles: *parse_emails* e *start_requests*. O método *parse_emails* será chamado pelo método *start_requests*.

Antes de iniciar, o método *start_requests* utiliza o atributo *self* para identificar as variáveis definidas no início da classe. A partir do momento em que é chamado *“yield scrapy.Request”*, inicia-se a *request*. O *scrapy* fica responsável por realizar a *request HTTP* para a *URL* especificada e o *yield* realiza o *return* desta *request*, como objeto de resposta, que será atribuído automaticamente ao atributo *response*. Este atributo será consumido no método *parse_emails*.

Dentro do método *parse_emails*, utiliza-se os atributos *self* e *response*. O atributo *response* será atribuído a uma variável, chamada de *format_utf*. Esta variável é responsável por formatar o conteúdo para UTF-8. Este fato é necessário para que os dados sejam traduzidos de forma legível, podendo assim tratá-los.

Utilizando esta variável, podemos definir o *extrator*, que utilizará a biblioteca *RE*, chamando sua classe *findall*. Para filtrar e buscar somente por e-mails válidos, houve a necessidade da utilização de uma expressão regular, que possui as seguintes características:

- \b permite que o e-mail seja interpretado para que o e-mail seja considerado como uma palavra completa, iniciando o e-mail. No fim da expressão, ela define o fim do e-mail.
- [A-Za-z0-9._%+-] permite o uso de caracteres maiúsculos, minúsculos, bem como números e caracteres especiais (define a primeira parte do e-mail).
- [A-Za-z0-9.-] permite o uso de caracteres maiúsculos, minúsculos, bem como números, ponto e o sinal de menos (Define o meio, que vem após o “@”).
- .\ para que o “.” seja interpretado de forma literal, ele precisa ser acompanhado por uma barra invertida.
- [A-Z|a-z]{2,} por fim, o final da expressão permite o uso de dois ou mais caracteres maiúsculos ou minúsculos para representar a extensão utilizada após o e-mail (Define o final do e-mail, após o ponto).

Após passada a variável *format_utf* e executada a filtragem com o *extractor*, criamos uma variável *dataframe*, que chamará a classe *Dataframe* da biblioteca *PANDAS*. Esta classe criará o *dataframe* com a coluna “*E-mails*” e conteúdo *extractor* (conteúdo extraído, formatado e filtrado para trazer somente e-mails). Na próxima linha, o

dataframe será convertido para *excel*, tendo como *index* igual a falso. Ou seja, a coluna não será incluída no índice de dados.

Ao executar o código por meio do *crawl*, uma planilha será criada, com os e-mails organizados em uma tabela, como na “Figura 7”.

Figura 7 – Planilha Etec Morato

E-mails
adil.soares01@etec.sp.gov.br
alexandre.lima74@etec.sp.gov.br
alexandre.silva592@etec.sp.gov.br
ana.lima309@etec.sp.gov.br
anderson.spera@etec.sp.gov.br
bruno.ferreira153@etec.sp.gov.br
carla.souza38@etec.sp.gov.br
carlos.souza172@etec.sp.gov.br
caroline.milone@etec.sp.gov.br
daniela.castelhano@etec.sp.gov.br
ednea.brito@etec.sp.gov.br
edvaldo.oliveira11@etec.sp.gov.br
eunice.marques2@etec.sp.gov.br
everton.santos145@etec.sp.gov.br
fabiana.oliveira52@etec.sp.gov.br
jair.souza4@etec.sp.gov.br
joao.orlando@etec.sp.gov.br
leandro.silva401@etec.sp.gov.br
leonardo.lirussi@etec.sp.gov.br
lucyolima@hotmail.com
mara.goncalves3@etec.sp.gov.br
paulo.jacobsen@etec.sp.gov.br
rafael.gross@etec.sp.gov.br
rafael.barreto5@etec.sp.gov.br
ricardo.jesus10@etec.sp.gov.br
rodrigo.silva901@etec.sp.gov.br
gordramos@yahoo.com.br
rodrigo.souza240@etec.sp.gov.br
rogerio.silva138@etec.sp.gov.br
sergio.fernandes01@etec.sp.gov.br
tatiane.cardoso6@etec.sp.gov.br
wagner.vieira8@etec.sp.gov.br

Fonte: Elaborado pelo autor (2023).

7.1.2. SPIDER: FATEC SANTANA DO PARNAÍBA

A “Figura 9” indica a estrutura do código que foi elaborada pelos integrantes do grupo, tomando como base o seu objetivo que é: “Extrair os cursos presentes na página Inicial”, conforme indicado na “Figura 8”.

Figura 8 - Website Fatec Santana

Fonte: FATEC SANTANA DO PARNAÍBA, 2023.

FIGURA 9 - Spider Fatec Santana

```
import scrapy
import pandas

class CourseSpider(scrapy.Spider):
    name = 'fatec_santana'
    url = 'https://www.fatecsdp.edu.br/'

    def parse_courses(self, response):
            courses = response.xpath('/html/body/div[3]/div/div[1]/div[3]/div/div/div[2]')
            path_dt = response.xpath('/html/body/div[3]/div/div[1]/div[3]/div/div/div[2]/dl/dt')
            data = []

            len_dt = len(path_dt)

            for scrapy in courses:
                for x in range (1,len_dt + 1):
                    value_title = scrapy.xpath(f'//dl/dt[{x}]/a/text()').get(),
                    value_course = scrapy.xpath(f'//dl/dt[{x}]/span/text()').get()
                    data.append ({
                        'Título': value_title,
                        'Nível do Curso': value_course
                    })

            dataframe = pandas.DataFrame(data)
            dataframe.to_excel('lista_cursos.xlsx', index=False)



    def start_requests(self):

            yield scrapy.Request(self.url, callback=self.parse_courses)
```

Fonte: Elaborado pelo autor (2023).

Como no último *spider*, a importação das bibliotecas foi realizada. Para este código, utilizou-se o *scrapy* e o *pandas*. Nele, também foram definidos o nome e a *URL*, portando a estrutura é parcialmente similar ao anterior.

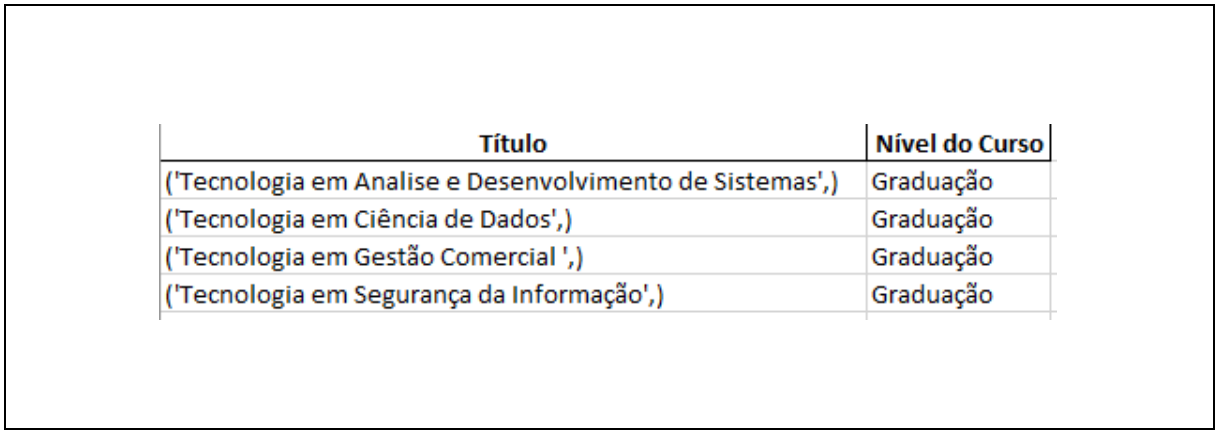
O que o difere do outro, é que neste caso, há filtro no *response*, com base no *xpath*. A variável *courses* navega pelas *div*, buscando o *xpath* indicado. Esta variável será utilizada para encontrar as listas de termos (*dt*) do *html*. Já a variável *path_dt* navega até o elemento <*dt*>, para que a variável *len_dt* possa usá-la como base para identificar a quantidade de elementos <*a*> e <*span*> existentes no elemento selecionado.

Após definir *data* como um *array*, definimos outra variável para verificar o comprimento de *path_dt*. Em seguida, definimos um *loop "for scrapy in courses"*, que é responsável por iterar sobre elementos selecionados por *xpath*. Dentro deste loop, outro *loop* é definido. Desta vez, *"for x in range (1, len_dt + 1)"*. Essa expressão define que *x* será

um *range* que irá de 1 até *len_dt* + *1*. O valor final é somado em 1 para que ele seja iterado corretamente, pois geralmente, o valor final não é incluído no intervalo, dentro de uma iteração.

Dentro deste *loop*, *value_title*, executará um *get()* no *xpath* definido, usando como base o *xpath* em *scrapy* (proveniente de *courses*). Ainda dentro do *loop*, ele adiciona os itens na lista, conforme definido. Por fim, é criado o *dataframe*, como explicado no código do tópico anterior. Em conclusão, há o retorno em uma planilha. Conforme “Figura 10”.

Figura 10 - Planilha Fatec Santana



Título	Nível do Curso
('Tecnologia em Analise e Desenvolvimento de Sistemas',)	Graduação
('Tecnologia em Ciência de Dados',)	Graduação
('Tecnologia em Gestão Comercial ',)	Graduação
('Tecnologia em Segurança da Informação',)	Graduação

Fonte: Elaborado pelo autor (2023).

7.2. EXPOSIÇÃO DE DADOS EM REDES SOCIAIS: CASO FACEBOOK

No ano de 2021, uma rede social amplamente conhecida pela conexão e interação de pessoas sofreu uma extração de dados de aproximadamente 1,5 bilhão de usuários. Dados mostram que o *scraper* divulgou a venda dos dados em um fórum, cujas informações continham dados como o nome, sexo do cliente, e-mail, número, entre outras constatadas pelo vendedor como “novas”. (WELIVESECURITY, 2021).

Segundo Zoltan (2022), os comerciantes haviam dito que as contas que detinham os dados não foram comprometidas (Fato que foi comprovado tecnicamente pela plataforma). Desta forma, os dados obtidos por raspagem não eram comprometedores, e sim públicos.

Embora dados como estes não sejam considerados tão sensíveis, por serem disponibilizados de forma pública, é cabível neste caso uma análise da plataforma e das possibilidades de utilização dos dados coletados.

O Facebook é uma plataforma que lida com uma alta quantidade de dados, podendo eles ser públicos ou não. Dentro da plataforma, o usuário consegue personalizar o que se torna público e o que é privado. Logo, o investimento em segurança da informação naturalmente deve ser alto, e crescer a cada ano, prevenindo assim ataques à base de dados.

A extração de dados do Facebook por *Scraping* (Agindo de acordo com as diretrizes) pode ser realizada por meio da raspagem de dados considerados como públicos, disponíveis para todos os usuários, sem exceção. É importante ressaltar que para a plataforma, o conceito de público é definido como:

- Informações exibidas na interface ao serem pesquisadas por um usuário, bem como acessadas por jogos, sites, aplicativos e até APIs do Facebook.

Atualmente, a aplicação web já conta com informativos que identificam o que é uma extração de dados e como prevenir que estes dados sejam extraídos sem a sua autorização. Embora não consigam prevenir todas as tentativas, a equipe interna EDM fica responsável por dificultar o processo de coleta de dados e impedir o lucro dos extratores. Para isso, um dos processos que ela executa é deixar a técnica menos atrativa, detectando padrões, bloqueando-os e limitando o número de extrações que podem ser feitas em um período determinado. Dentro desta estratégia, eles também analisam o comportamento dos extratores, punindo suas contas e utilizando as informações coletadas pela equipe para aprimoramento de seus sistemas (FACEBOOK, 2023).

7.3. VOCÊ CUIDA DOS SEUS DADOS? (FORMULÁRIO GOOGLE)

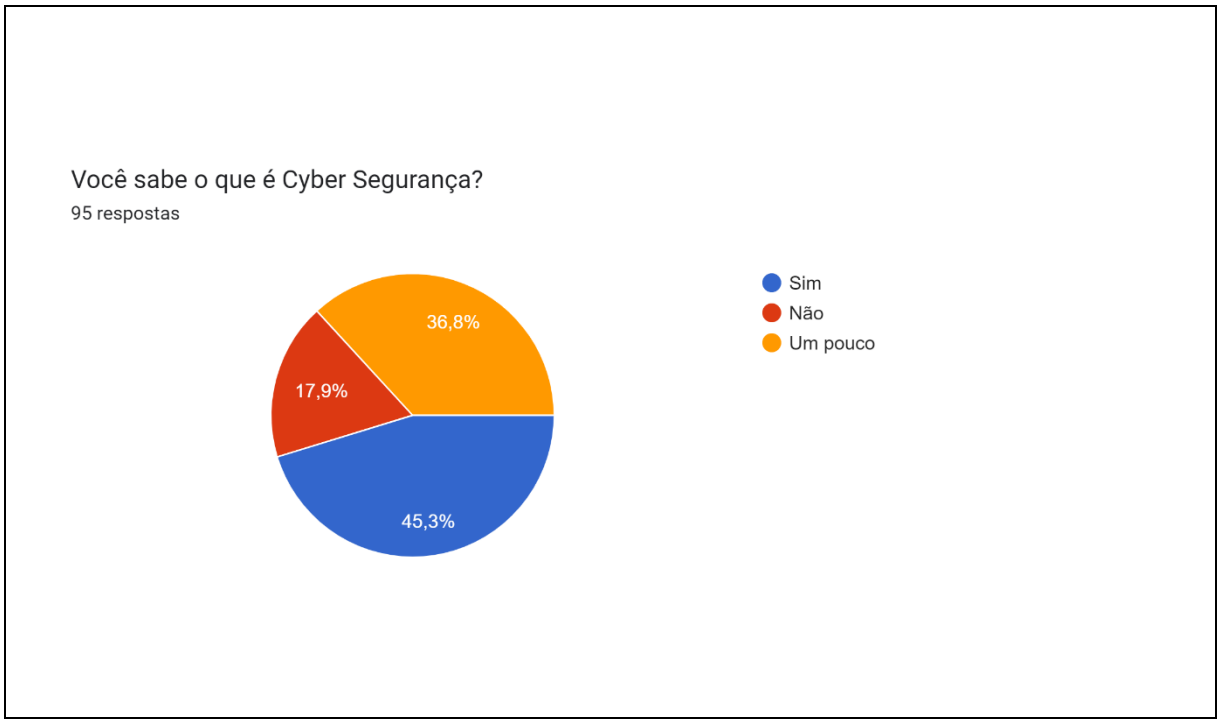
O formulário foi desenvolvido com o objetivo de reconhecer se a população, como um todo, tem conhecimento da importância da proteção de dados online e, de maneira simples, questionar os métodos de seguranças que elas aplicam, junto com indicações de como melhorar sua própria segurança.

Contamos com a colaboração de 95 participantes, desde pessoas que não tem muito conhecimento sobre a internet, até pessoas que trabalham justamente com Cibersegurança. O link do Google Forms foi enviado pelas redes sociais *Whatsapp*, *Instagram* e *Facebook*.

Como afirma Moran: “O professor tem um grande leque de opções metodológicas, de possibilidades de organizar sua comunicação com os alunos, de introduzir um tema, de trabalhar com os alunos presencial e virtualmente, de avaliá-los” (MORAN, 2000, p. 30).

A primeira pergunta focou em avaliar se as pessoas sabiam o que é a Cibersegurança.

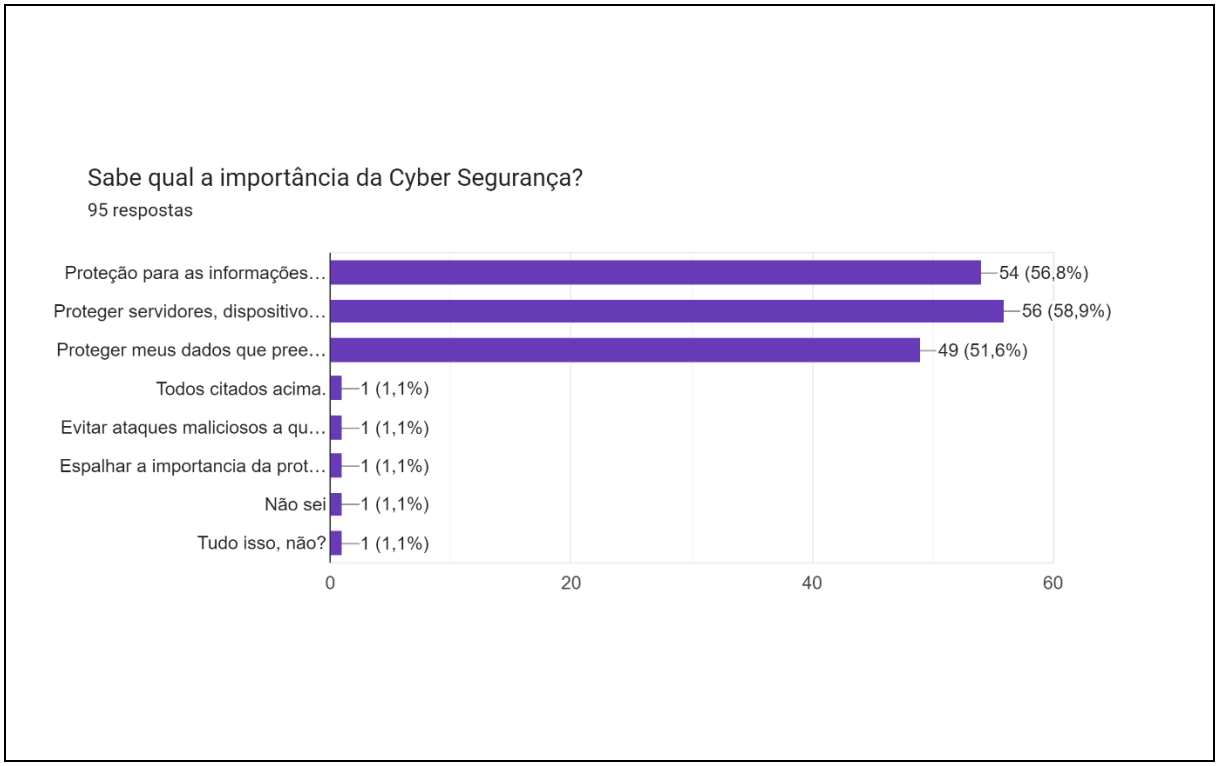
Gráfico 1: Você sabe o que é Cibersegurança?



Fonte: Dados da Pesquisa.

Percebeu-se que dos 95 participantes, 43 deles reconhecem o que é a Cibersegurança, 17 não conhecem e 35 distinguem suas particularidades, selecionando a opção “um pouco”, sendo respectivamente: 45,3% sim, 17,9% não e 36,8% um pouco.

Gráfico 2: Sabe qual a importância da Cibersegurança?

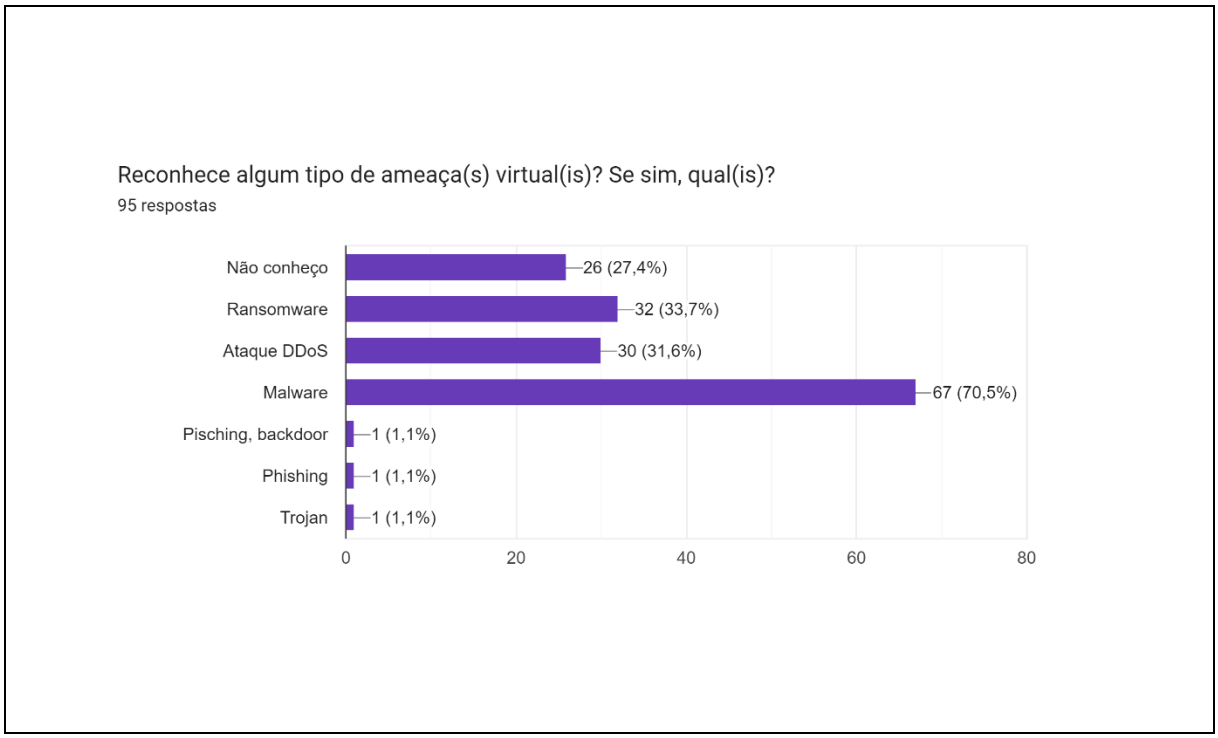


Fonte: Dados da Pesquisa.

Nesse tópico, apresentamos 3 exemplificações do que acham que a Cibersegurança faz e demos a oportunidade de citar novos exemplos. No caso, a 3ª opção (proteger meus dados contra sites de compras) não era totalmente correta, já que a função não

se remete a proteger seus dados inseridos em sites de compra online, consequentemente, as respostas nos surpreenderam pela facilidade em “influência”.

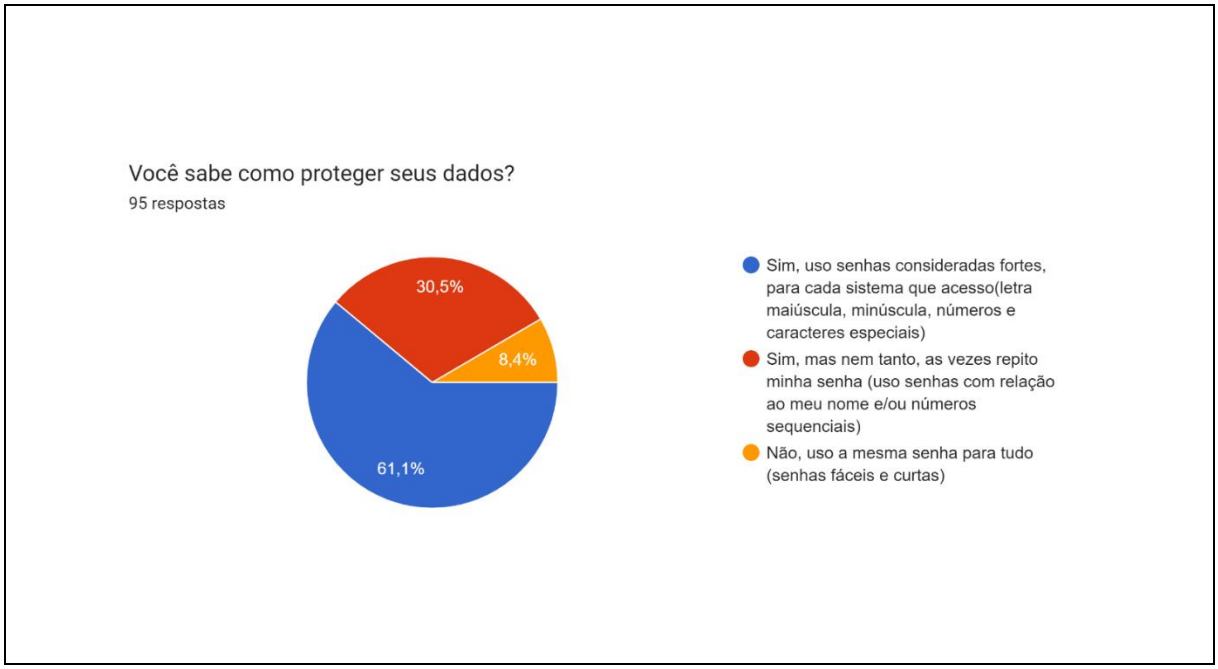
Gráfico 3: Reconhece algum tipo de ameaça(s) virtual(is)? Se sim, qual(is)?



Fonte: Dados da Pesquisa.

O *Malware* é realmente o tipo de invasão mais comentada, tanto nas redes quanto em cursos e adversos, mas a quantidade de pessoas que reconheceram os demais perigos (inclusive aqueles não citados como base) foram altos, o que traz certo conforto em observar esses resultados.

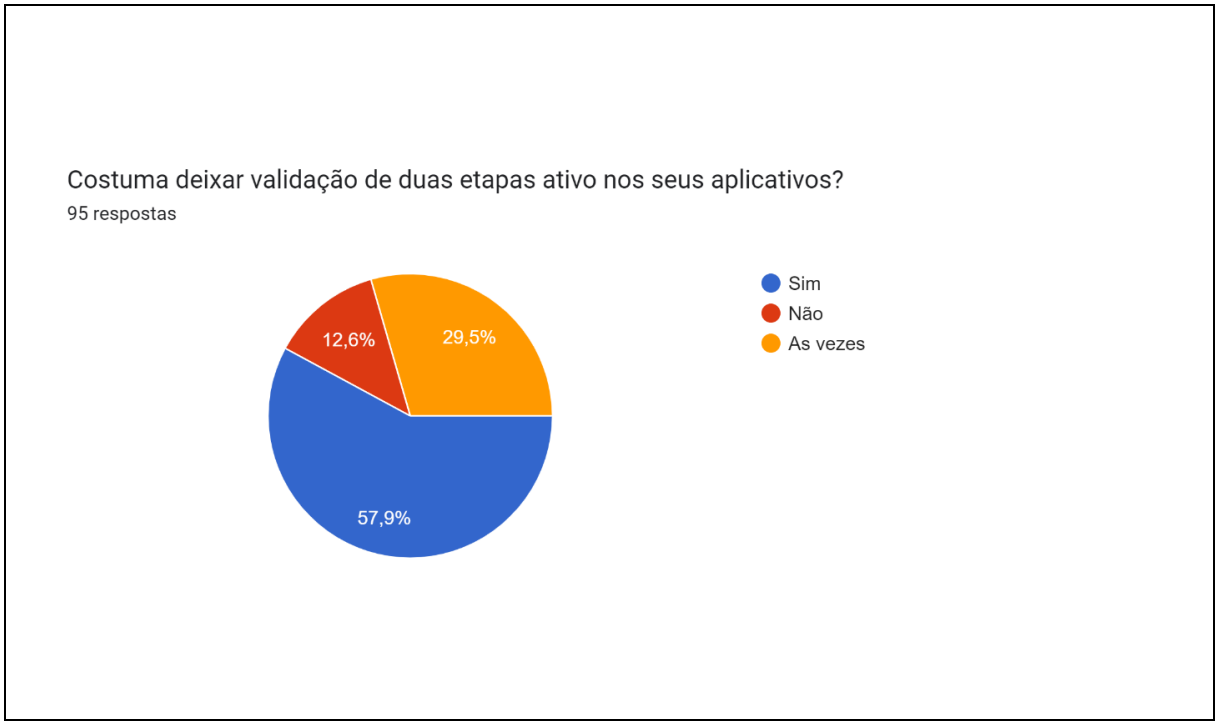
Gráfico 4: Você sabe como proteger seus dados?



Fonte: Dados da Pesquisa.

Conforme avaliado, os participantes, em grande parte, reconhecem a importância de ter senhas difíceis e com características que não se remetem as nossas informações pessoais. Mesmo de que nem em todas as ocasiões elas sejam utilizadas.

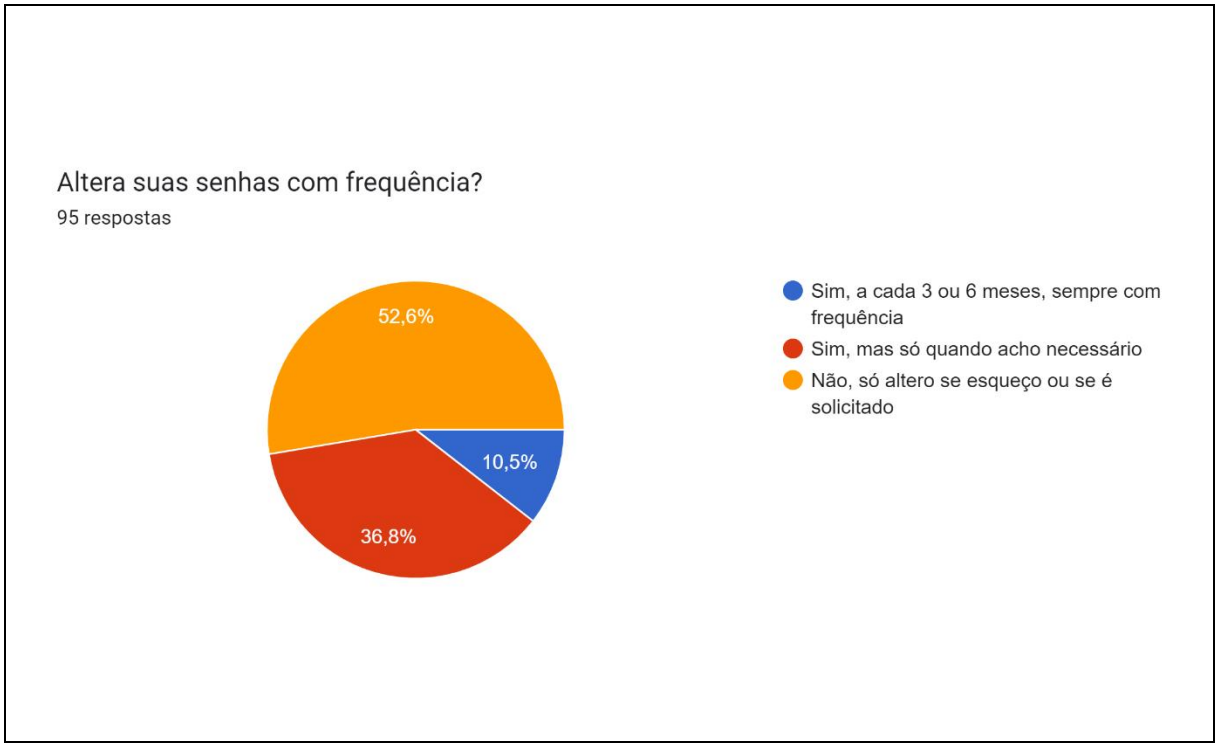
Gráfico 5: Costuma deixar validação de duas etapas?



Fonte: Dados da Pesquisa.

Se há a disponibilidade da validação em duas etapas, 45,3% das pessoas optam por deixá-las ativas, o que é uma barreira além para evitar vazamento de informações indesejadas. Nesse ponto, vale ressaltar que nem todas as redes obtém essa função de validação.

Gráfico 6: Altera suas senhas com frequência?



Fonte: Dados da Pesquisa.

Os resultados desse gráfico foram os mais alarmantes quanto a pesquisa, pois apesar de os participantes terem consciência da importância da Segurança da Informação, aqui eles pecam quanto a rotina de validações. 52,6% dos pesquisados não tem o costume de alterar suas senhas com frequência, nem sequer quando conveniente, alteram a senha apenas se acham necessário.

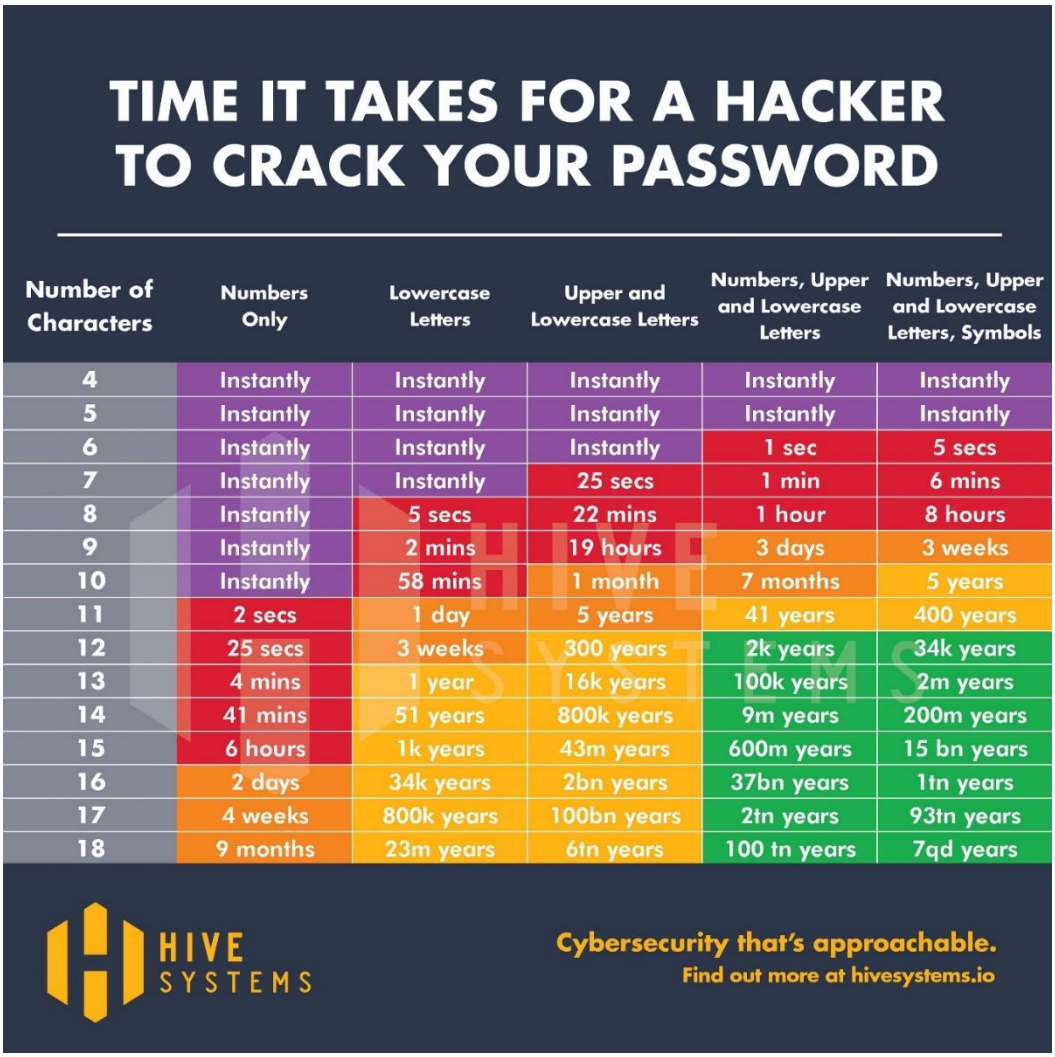
Mesmo que sua senha seja difícil de ser descoberta, a alteração frequente dela é uma dificuldade além de um *hacker* invadir sua máquina e ter acesso a suas informações, a alteração frequente não é necessária quando se trata de aplicações mais singelas, porém é importante que ela seja regular, de maneira que sempre estejam atualizadas.

Lembrando que existem algumas dicas de composição de senhas:

- Não obter informações pessoais na senha (como data de aniversário, senhas de cartões, bancos, seu nome, nome de sua mãe, etc.);
- Não obter sequencias ou repetições de números e/ou letras (como 123 ou GGG como composição;

- Adquirir o hábito de sempre inserir caracteres especiais. Exemplos: (,!,@,#,$,*,). ;
- Adquirir o hábito de inserir letras maiúsculas e minúsculas. Exemplos: MvTmjSunP.

Figura 11 - Reflexão



Number of Characters	Numbers Only	Lowercase Letters	Upper and Lowercase Letters	Numbers, Upper and Lowercase Letters	Numbers, Upper and Lowercase Letters, Symbols
4	Instantly	Instantly	Instantly	Instantly	Instantly
5	Instantly	Instantly	Instantly	Instantly	Instantly
6	Instantly	Instantly	Instantly	1 sec	5 secs
7	Instantly	Instantly	25 secs	1 min	6 mins
8	Instantly	5 secs	22 mins	1 hour	8 hours
9	Instantly	2 mins	19 hours	3 days	3 weeks
10	Instantly	58 mins	1 month	7 months	5 years
11	2 secs	1 day	5 years	41 years	400 years
12	25 secs	3 weeks	300 years	2k years	34k years
13	4 mins	1 year	16k years	100k years	2m years
14	41 mins	51 years	800k years	9m years	200m years
15	6 hours	1k years	43m years	600m years	15 bn years
16	2 days	34k years	2bn years	37bn years	1tn years
17	4 weeks	800k years	100bn years	2tn years	93tn years
18	9 months	23m years	6tn years	100 tn years	7qd years

Fonte: TECMUNDO, 2021.

Para finalizar a pesquisa, um quadro foi inserido, com base nas informações do ano de 2022, onde apresentam (simbolicamente) o tempo de que uma pessoa conseguiria invadir suas informações para realizar coleta de dados pessoais, conforme apresentado na imagem acima.

8. CONCLUSÃO

Ao desenvolver a pesquisa ficou perceptível a necessidade de manter a atualização, inovação e proteção de dados, quanto em etapas de validações de seguranças. Muitas pessoas têm acesso a tecnologia, mas não tem o conhecimento de quais ameaças existem quando se está exposto na internet, ou muito menos como ocorre a proteção dos dados ou como utilizá-la contra as ameaças, conforme demonstra os resultados do relatório desenvolvido com perguntas abertas diretamente ao público.

Como observado o *Web Scraping* é uma ferramenta de raspagem de dados que traz muitos benefícios para o usuário, para profissionais da área de TI que trabalham com a ferramenta, e para as organizações que necessitam desses profissionais. Com a raspagem automática de dados é possível reduzir custos na empresa, permitindo manipular um volume muito maior de dados em menos tempo, padronizando o processo de coleta/raspagem de dados, e poupando esforços manuais de seus profissionais.

Uma raspagem automatizada além de trazer dados de forma rápida, traz dados cada vez mais precisos e atualizados, e se tratando de atualização no mercado de trabalho, quem se mantém mais atualizado e dentro do mercado tem mais chances de se destacar entre este meio. As empresas produzem é precisam analisar uma grande quantidade de dados todos os dias, portanto o *Web scraping* é uma ótima ferramenta para atender a essa demanda de dados, e está sendo bastante utilizada no âmbito empresarial.

Sendo assim, é essencial ter extrema atenção ao mexer com dados, pois nem qualquer pessoa/profissional podem ter acessos a esse tipo de informação quando se

trata de dados dos tipos sensíveis, quando acessados sem a devida permissão é considerado invasão, ataque ou até mesmo um crime cibernético.

Nesse sentido saber detalhadamente sobre como é realizado o processo *Web scraping,* permite que as áreas voltadas a segurança de informação, expandam seus conhecimentos e estudos sobre a ferramenta, criando assim tecnologias melhores e padrões de segurança mais eficazes contra esse tipo de uso e acesso indevido de informações. Por consequência a invasão/compartilhamento ou vazamento de dados mesmo sendo algo extremamente negativo, é necessário para a evolução da cibersegurança, e para proteger os usuários de ameaças existentes e de possíveis novos ataques.

REFERÊNCIAS

ALINE OLIVEIRA. **Coleta De Dados: Quais São Os Métodos e Como Fazer?**. 2022. Disponível em: <https://mindminers.com/blog/coleta-de-dados-como-fazer/>. Acesso em: 5 abr. 2023.

CACHE, G. **O que é Web Scraping? Para iniciantes - GoCache.** Disponível em: <https://www.gocache.com.br/seguranca/o-que-e-web-scraping-para-iniciantes/>. Acesso em: 21 nov. 2022.

FERREIRA, Kellison. **Coleta de dados: o que é, ferramentas e como fazer no marketing?** Disponível em: <https://rockcontent.com/br/blog/coleta-de-dados/>. Acesso em: 2 de fev. 2023.

COMO. **Scraping: entenda como terceiros acessam dados que você expõe sem saber**. Disponível em: <https://www.techtudo.com.br/listas/2020/09/scraping-entenda-como-terceiros-acessam-dados-que-voce-expoe-sem-saber.ghtml>. Acesso em: 1 dez. 2022.

ETEC FRANCISCO MORATO. **Corpo Docente**. Disponível em: <https://www.etecfranciscomorato.com.br/a-etec/corpo-docente/>. Acesso em: 15 mai. 2023.

Facebook: Ajuda. **O que é extração de dados e o que posso fazer para proteger as minhas informações no Facebook**. Disponível em: <https://www.facebook.com/help/463983701520800/?helpref=uf_share>. Acesso em: 12 de mai. 2023.

FATEC SANTANA DO PARNAÍBA. **Início**. Disponível em: < https://www.fatecsdp.edu.br/>. Acesso em: 16 mai. 2023.

FERRI, Edsel. **Análise Preditiva ou Prescritiva? Sua empresa precisa de ambos**. Disponível em: <https://medium.com/@edselferri/an%C3%A1lise-preditiva-ou-prescritiva-sua-empresa-precisa-de-ambos-63b49caf09cf>. Acesso em: 22 mai. 2019.

LGPD. **Principais objetivos da nova lei.** 2023? Disponível em: <https://www.lgpdbrasil.com.br/o-que-muda-com-a-lei/>. Acesso em: 1 abr. 2023.

LOPES, Erickson. **FUNÇÕES EM PYTHON.** Disponível em: < https://pythonacademy.com.br/blog/funcoes-em-python>. Acesso em: 23 mai. 2023.

MESTRE, P. **A automatização de processos: benefícios e desvantagens**. Disponível em: <https://pt.primaverabss.com/pt/blog/automatizacao-de-processos/>. Acesso em: 6 fev. 2023.

MITCHELL, Ryan. **Web Scraping com Python: Coletando dados na Web moderna**. Segunda Edição. p 11. ed. [S. l.]: Novatec Editora LTDA., 2019.

PANDAS. Disponível em: <https://pandas.pydata.org/>. Acesso em: 17 mai. 2023.

PYTHON. **Regular expression operations**. Disponível em: <https://docs.python.org/3/library/re.html>. Acesso em: 17 mai. 2023.

PYTHON. Welcome to Python.org. Disponível em: < https://www.python.org/>. Acesso em: 23 mai. 2023.

PYTHON. Classes. Disponível em: < https://docs.python.org/pt-br/3/tutorial/classes.html#:~:text=Um%20m%C3%A9todo%20%C3%A9%20uma%20fun%C3%A7%C3%A3o,remove%2C%20sort%2C%20entre%20outros.>. Acesso em: 23 mai. 2023.

SCRAPY. Disponível em: <https://scrapy.org/>. Acesso em: 15 maio 2023.

SCRAPY**. Scrapy 2.9 documentation**. Disponível em: <https://docs.scrapy.org/en/latest/topics/spiders.html>. Acesso em: 15 maio 2023.

SIMOVA. **ENTENDA MAIS SOBRE A AUTOMAÇÃO DA COLETA DE DADOS**. Disponível em: <https://www.simova.com.br/post/entenda-mais-sobre-a-automa%C3%A7%C3%A3o-da-coleta-de-dados>. Acesso em: 6 fev. 2023.

SOUZA, L. F. **Web scraping: coleta de dados automatizada - Aquarela Analytics.** Disponível em: <https://www.aquare.la/web-scraping-coleta-de-dados-automatizada/>. Acesso em: 21 nov. 2022.

Web scraping: como proteger os negócios dessa ameaça silenciosa? | PUGA, Rafael. Disponível em: <https://www.securityreport.com.br/destaques/web-scraping-como-proteger-os-negocios-dessa-ameaca-silenciosa/#.Y4qNV3bMLIU>. Acesso em: 02 dez. 2022.

TECMUNDO. Web scraping: conheça a técnica de coleta de dados. Disponível em: <https://www.tecmundo.com.br/internet/215525-web-scraping-conheca-tecnica-coleta-dados.htm>. Acesso em: 29 nov. 2022.

TECMUNDO. Quanto tempo leva para um hacker descobrir a sua senha? | Disponível em: <https://www.tecmundo.com.br/seguranca/210443-tempo-leva-hacker-descobrir-senha.htm>. Acesso em: 2 mai. 2023.

Web scraping: o que é, como funciona e para que serve? - Netrin. Disponível em: <https://netrin.com.br/web-scraping-o-que-e-como-funciona/>. Acesso em: 21 nov. 2022.

WELIVESECURITY. **Criminosos vendem dados de 1,5 bilhão de usuários do Facebook coletados por meio de scraping**. Juan Manuel Harán, 2021. Disponível em: <https://www.welivesecurity.com/br/2021/10/06/criminosos-vendem-dados-de-15-bilhao-de-usuarios-do-facebook-coletados-por-meio-de-scraping/> Acesso em: 8 mai. 2023.

ZOLTAN, Miklos. **Web Scrapers alegam possuir e vender dados pessoais de 1,5 bilhão de usuários do Facebook em um fórum de hackers.** Privacity Affairs**,** 2022. Disponível em: <https://www.privacyaffairs.com/facebook-data-sold-on-hacker-forum/> Acesso em: 12 de mai. 2023.

Anexo A- Relatório *CopySpider*

CopySpider
https://copyspider.com.br/

Versão do CopySpider: 2.2.0
Relatório gerado por: alyfercd@hotmail.com
Modo: web / detailed

Arquivos	Termos comuns	Similaridade
TCC-FatecJdi-WebScraping.docx X https://www.tecmundo.com.br/internet/215525-web-scraping-conheca-tecnica-coleta-dados.htm	102	1,14
TCC-FatecJdi-WebScraping.docx X https://blog.betrybe.com/tecnologia/scraping-como-fazer	155	1,13
TCC-FatecJdi-WebScraping.docx X https://rockcontent.com/br/blog/web-scraping	61	0,65
TCC-FatecJdi-WebScraping.docx X https://canaltech.com.br/seguranca/o-que-e-web-scraping	51	0,60
TCC-FatecJdi-WebScraping.docx X https://www.passeidireto.com/arquivo/97856282/modelo-tcc	54	0,43
TCC-FatecJdi-WebScraping.docx X https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html	18	0,19
TCC-FatecJdi-WebScraping.docx X https://genius.com/Stephen-hawking-on-the-threat-of-artificial-intelligence-annotated	16	0,15
TCC-FatecJdi-WebScraping.docx X https://impactforbreakfast.com/privacy-policy	6	0,06
TCC-FatecJdi-WebScraping.docx X http://www.fatecsdp.edu.br	4	0,05
TCC-FatecJdi-WebScraping.docx X https://www.fatecsdp.edu.br/a-fatec-santana	4	0,05
Arquivos com problema de download		
https://www.reference.com/world-view/moral-little-red-riding-hood-3c0f20738e2ece43?utm_content=params%3Ao%3D740005%26ad%3DdirN%26qo%3DserpIndex&ueid=f9f29335-01a4-47ad-993b-0f47c6496d56	Não foi possível baixar o arquivo. É recomendável baixar o arquivo manualmente e realizar a análise em conluio (Um contra todos). - Erro: Parece que o documento não existe ou não pode ser acessado. HTTP response code: 403 - Server returned HTTP response code: 403 for URL: https://www.reference.com/world-view/moral-little-red-riding-hood-3c0f20738e2ece43?utm_content=params%3Ao%3D740005%26ad%3DdirN%26qo%3DserpIndex&ueid=f9f29335-01a4-	