MONOGRAFIA DE PROJETO FINAL DE GRADUAÇÃO

MÉTODO DE ANÁLISE DE DADOS DOS CADERNOS JUDICIÁRIOS DO DEJT SUPORTADOS POR TÉCNICAS DE OSINT

Bruna Maria de Andrade Augustinho

Natália Andrade Marques

Curso Superior de Engenharia de Redes de Comunicação

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

MONOGRAFIA DE PROJETO FINAL DE GRADUAÇÃO

MÉTODO DE ANÁLISE DE DADOS DOS CADERNOS JUDICIÁRIOS DO DEJT SUPORTADOS POR TÉCNICAS DE OSINT

Bruna Maria de Andrade Augustinho

Natália Andrade Marques

Monografia de Projeto Final de Graduação submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Bacharel em Engenharia de Redes de Comunicação

Banca Examinadora

Dr. Georges Daniel Amvame Nze, EnE/UnB _____
Orientador

Dr. Fábio Lúcio Lopes de Mendonça, EnE/UnB _____
Examinador Interno

M.Sc Bruno Justino Garcia Praciano, EnE/UnB _____
Examinador Externo

FICHA CATALOGRÁFICA

AUGUSTINHO, B.M.A.; MARQUES, N.A.
MÉTODO DE ANÁLISE DE DADOSDOS CADERNOS JUDICIÁRIOS DO DEJT SUPORTADOS POR TÉCNICAS DE OSINT [Distrito Federal] 2023.
xvi, 34 p., 210 x 297 mm (ENE/FT/UnB, Bacharel, Engenharia de Redes de Comunicação, 2023).
Monografia de Projeto Final de Graduação - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. OSINT	2. Kibana
3. DEJT	4. Elasticsearch
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

AUGUSTINHO, B.M.A.; MARQUES, N.A. (2023). *MÉTODO DE ANÁLISE DE DADOSDOS CADERNOS JUDICIÁRIOS DO DEJT SUPORTADOS POR TÉCNICAS DE OSINT*. Monografia de Projeto Final de Graduação, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 34 p.

CESSÃO DE DIREITOS

AUTOR: Bruna Maria de Andrade Augustinho
Natália Andrade Marques
TÍTULO: MÉTODO DE ANÁLISE DE DADOSDOS CADERNOS JUDICIÁRIOS DO DEJT SUPORTADOS POR TÉCNICAS DE OSINT.
GRAU: Bacharel em Engenharia de Redes de Comunicação ANO: 2023

Bruna Maria de Andrade Augustinho
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP: 70919-970 - Brasília-DF - Brasil

Natália Andrade Marques
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP: 70919-970 - Brasília-DF - Brasil

AGRADECIMENTOS

Agradecemos primeiramente a Deus, que permitiu que nossos caminhos se cruzassem durante a trajetória no período de graduação, e que cultivássemos uma amizade ao longo dos anos. É uma grande satisfação chegarmos juntas a este momento e realizarmos este projeto.

Agradecemos também aos nossos professores, em especial aos nossos professores orientadores. Ao Prof. Dr. Robson Albuquerque, expressamos nosso reconhecimento por todo o apoio e por nos fazer acreditar em nossa capacidade. Sua orientação foi fundamental para o desenvolvimento deste projeto. Também gostaríamos de agradecer ao Prof. Dr. Georges Daniel, que esteve conosco ao longo de todos os anos da graduação e nos orientou da melhor forma possível. Sua dedicação e conhecimento foram essenciais para o nosso crescimento acadêmico. Estamos extremamente gratas por termos tido a oportunidade de aprender com vocês.

E por fim, agradecemos às nossas famílias, que nos apoiaram e nos deram motivação para seguir essa caminhada. Eles entenderam os momentos de ausência e nos fortaleceram em todas as situações difíceis. Sua dedicação e apoio foram fundamentais para alcançarmos este ponto em nossas vidas. Somos gratas por todo o amor e suporte que nos proporcionaram.

RESUMO

Diante da problemática de buscar informações em dados governamentais, esse trabalho tem como objetivo aplicar uma metodologia de OSINT para extrair dados dos cadernos do Diário Eletrônico da Justiça do Trabalho. Para esse fim, utilizamos ferramentas como Elasticsearch e Kibana, que permitem realizar buscas inteligentes e promovem uma análise intuitiva das informações extraídas. A utilização da metodologia OSINT oferece uma abordagem eficaz para lidar com a complexidade e o volume de dados presentes nos cadernos do Diário Eletrônico da Justiça do Trabalho. Por meio da aplicação de técnicas adequadas, é possível extrair *insights* relevantes e melhorar a interpretação das informações contidas nesses documentos. Com aplicação dessa metodologia e uso das ferramentas mencionadas, foi possível superar os desafios relacionados à busca e análise de dados nos cadernos do Diário Eletrônico da Justiça do Trabalho, proporcionando uma solução eficiente e intuitiva para a extração de informações relevantes.

Keywords: OSINT, DEJT, Elasticsearch, Kibana;

ABSTRACT

Faced with the problem of seeking information in government data, this work aims to apply an OSINT methodology to extract data from the notebooks of the Electronic Journal of the Labor Justice. To this end, we use tools such as Elasticsearch and Kibana, which allow you to carry out intelligent searches and promote an intuitive analysis of the extracted information. The use of the OSINT methodology offers an effective approach to deal with the complexity and volume of data present in the Labor Justice Electronic Diary. Through the application of appropriate techniques, it is possible to extract relevant insights and improve the interpretation of the information contained in these documents. With the application of this methodology and use of the aforementioned tools, it was possible to overcome the challenges related to the search and analysis of data in the notebooks of the Electronic Journal of the Labor Justice, providing an efficient and intuitive solution for extracting relevant information.

Keywords: OSINT, DEJT, Elasticsearch, Kibana;

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

Siglas

OSINT	*Open Source Intelligence*
DEJT	*Diário Eletrônico da Justiça do Trabalho*
PDF	*Portable Document Format*
TXT	*Text*
VM	*Virtual machine*
DOE	*Diário Oficial do Estado*
TST	*Tribunal Superior do Trabalho*
TRT	*Tribunal Regional do Trabalho*
DO	*Diário Oficial*
DOM	*Diário Oficial Municipal*
DOU	*Diário Oficial da União*
HUMINT	*Human Intelligence*
SIGINT	*Signals Intelligence*
IMINT	*Imagery Intelligence*
DGA	*Dados Governamentais Abertos*

1 INTRODUÇÃO

Com o avanço das Tecnologias da Informação e Comunicação (TICs) e a crescente demanda por transparência nas ações governamentais, ocorreu um progresso significativo na disponibilização aberta de dados governamentais. Inicialmente, eram fornecidos de forma privada, mais tarde evoluíram para a busca em bases de dados. Hoje, essas informações podem ser acessadas em formato bruto, permitindo sua manipulação, filtragem e combinação com outras fontes de informação. Esses dados são conhecidos como Dados Governamentais Abertos (DGA). Esse avanço proporciona oportunidades para a criação de novas aplicações e a geração de conhecimento pela sociedade civil [Vaz, Ribeiro e Matheus 2010].

De forma geral, as instituições públicas têm como princípio constitucional garantir a publicidade de seus atos por meio de uma fonte oficial, que no Brasil pode ser exercida pelos Diários Oficiais (DOs), um tipo de Dado Governamental Aberto. Esses jornais desempenham um papel fundamental ao comunicar à sociedade sobre leis, decretos, instruções normativas, atos de interesse dos servidores da administração pública federal, editais, ineditoriais, avisos e uma variedade de informações relevantes. Além disso, eles também são fonte de notícias sobre concursos públicos, processos administrativos ou judiciais, aberturas de licitações, entre outros.

Mesmo com a implementação de iniciativas governamentais para a disponibilização dos Diários Oficiais em formatos digitais, tornou-se evidente que os métodos tradicionais de busca manual em documentos impressos ou digitais não são mais adequados para a realidade tecnológica em que vivemos. Os dados disponibilizados pelo Governo são extensos e não estão estruturados, além de incluírem um grande volume de informações não padronizadas. Esse formato promove uma deficiência no processo de identificação de informações específicas nessas fontes.

Dessa maneira, iniciativas que utilizam recursos computacionais e estratégias de análise em grandes quantidades de informações podem facilitar o trabalho de quem depende das informações contidas nesses documentos. Ao empregar recursos computacionais, não apenas otimizamos a coleta e análise de informações nos DOs, mas também promovemos a transparência, a acessibilidade e a eficiência na divulgação dos atos e decisões governamentais. Ao disponibilizar meios mais acessíveis e eficazes de pesquisa, essas tecnologias contribuem para fortalecer a relação entre governo e sociedade, facilitando o acesso e o entendimento das informações governamentais.

Além disso, estratégias de análise em larga escala possibilitam identificar padrões, tendências e relações entre as informações contidas nos documentos. Esses *insights* são valiosos para pesquisadores, advogados, empresas e cidadãos em geral, pois proporcionam uma compreensão mais profunda do contexto e auxiliam na interpretação e utilização dessas informações.

Nesse contexto, a aplicação de técnicas de OSINT - *Open Source Intelligence*, ou em português, Inteligência de Fontes Abertas, para extrair, tratar e produzir inteligência com base nos Diários Eletrônicos da Justiça do Trabalho (DEJTs), revela um grande potencial para beneficiar diversas partes interessadas. Ademais, a utilização das técnicas de OSINT nos DEJTs amplia o acesso e a compreensão das informações judiciais, agiliza o processo de pesquisa e contribui para a transparência e a democratização do conheci-

mento jurídico. Essa abordagem pode proporcionar benefícios significativos tanto para profissionais do meio jurídico quanto para o público em geral, ao fomentar uma maior compreensão dos eventos legais e uma maior participação informada no âmbito judicial.

OSINT vem sendo popularizada nos últimos anos devido a facilidade de acesso as fontes de dados, uma vez que tratam-se de informações com um amplo volume de dados disponíveis. Teoricamente, qualquer pessoa com acesso à internet pode utilizar essa prática como um método de pesquisa. Consequentemente, a obtenção de dados por meio de fontes abertas se popularizou, resultando no desenvolvimento de diversas ferramentas para auxiliar na coleta de informações [Tanabe et al. 2022].

Entretanto, é imprescindível enfatizar a relevância do uso de uma metodologia sólida para garantir que o processo seja conduzido de maneira assertiva e eficiente, resultando nos objetivos desejados. Nesse sentido, uma metodologia bem definida também previne oscilações que podem ocorrer quando se depende exclusivamente de ferramentas. Dado que as fontes de dados e as demandas estão em constante atualização, é fundamental que os envolvidos estejam preparados para enfrentar diversas eventualidades.

Em suma, o uso de uma metodologia robusta é um alicerce indispensável para o sucesso do processo, proporcionando maior segurança e eficiência na busca pelos resultados esperados. Ao adotar essa abordagem, os profissionais estarão mais preparados para enfrentar as complexidades do cenário atual, potencializando suas chances de alcançar o êxito em meio às mudanças e desafios do dia a dia [Furuhaug 2019].

Esse trabalho, busca preencher uma lacuna na pesquisa acadêmica e fornecer uma contribuição prática para a área, demonstrando os benefícios e o potencial da aplicação da metodologia OSINT na extração e análise de informações do Diário Oficial da Justiça do Trabalho.

1.1 OBJETIVOS

Esta seção tem como finalidade esclarecer de maneira concisa e precisa os objetivos do trabalho e como o documento será estruturado, proporcionando um melhor entendimento do texto.

1.1.1 OBJETIVO GERAL

O objetivo deste trabalho é utilizar a metodologia OSINT para extrair informações do Diário Eletrônico da Justiça do Trabalho. A proposta é abordar a dificuldade encontrada na localização de informações relevantes nesses documentos, explorando o potencial de OSINT como método para superar esse desafio.

1.1.2 OBJETIVOS ESPECÍFICOS

Para delimitar e alcançar a conclusão esperada citada no objetivo geral, os seguintes objetivos específicos nortearão a consecução do objetivo deste estudo, conforme descrito abaixo:

- Estruturar o direcionamento do ciclo de OSINT e seus resultados esperados;
- Planejar e executar as atividades de coleta;

- Demonstrar as etapas de processamento dos dados após a coleta;
- Visualização e análise dos dados processados;
- Citar o uso de OSINT na obtenção de informações através de Fontes;
- Através do estudo de caso, validar os dados e informações que foram obtidas nas pesquisas dentro dos Diários Eletrônicos da Justiça do Trabalho por meio de aplicação de técnicas de OSINT;
- Citar como essas informações são úteis com base na análise dos dados e produção de inteligência;

1.2 ORGANIZAÇÃO DO RELATÓRIO

Este documento será estruturado da seguinte forma:

- O primeiro capítulo se refere a introdução, que tem cunho principal de realizar uma abordagem descritiva das motivações, primeiros estudos e conceitos e objetivos do trabalho;
- O segundo capítulo se refere a fundamentação teórica e tem cunho principal realizar um estudo mais a fundo bem como relacionar os trabalhos correlatos que serviram como base teórica necessária para toda a concepção deste estudo;
- O terceiro capítulo é dedicado a descrição da metodologia utilizada para a concepção do estudo, com a descrição das etapas realizadas;
- O quarto capítulo retrata uma visão detalhada e abrangente dos resultados alcançados no trabalho;
- O quinto capítulo apresenta a conclusão, fornecendo uma síntese da proposta do estudo e destacando os pontos mais relevantes abordados ao longo do trabalho, assim como os resultados obtidos ;
- O sexto e último capítulo visa apresentar sugestões de trabalhos futuros relacionados a este estudo, mas não sendo este o limitante para novas contribuições.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Este capítulo tem como objetivo apresentar a fundamentação teórica necessária para compreender os assuntos discutidos no trabalho. Além disso, busca relacionar alguns estudos relevantes com as temáticas escolhidas, a fim de aprimorar e aprofundar os tópicos abordados.

2.1 OPEN SOURCE INTELLIGENCE

A criação de ferramentas de inteligência para aprimorar a tomada de decisões tem sido uma prática antiga na esfera governamental. Essa busca por informações estratégicas é uma meta comum tanto da diplomacia pública, que visa melhorar a compreensão e a comunicação com o público externo, quanto da espionagem, cujo foco é obter *insights* cruciais para a segurança nacional e os interesses do Estado [Jota et al. 2022].

É importante ressaltar que a utilização de informações públicas e amplamente disponíveis em relatórios de inteligência já existia anteriormente. No entanto, com o advento da Revolução da Informação, essa abordagem ganhou novos contextos, apresentando desafios e oportunidades inéditas para o campo da inteligência.

Nesse contexto, a adoção da *Open Source Intelligence* (OSINT) torna-se extremamente relevante no que tange a extração de informações a partir de fontes governamentais. O potencial impacto proporcionado por esta técnica no âmbito de tomada de decisões estratégicas é notável, proporcionando uma vantagem competitiva significativa para aqueles que sabem utilizá-la adequadamente.

De forma geral, OSINT pode ser definida como "coleta, processamento e correlação de informações públicas de fontes de dados abertos, como mídia de massa, redes sociais, fóruns, blogs, dados públicos governamentais e publicações ou dados comerciais". Através da aplicação de técnicas avançadas de coleta e análise, a OSINT possibilita a expansão contínua do conhecimento sobre o alvo em questão [Pastor-Galindo et al. 2020].

Além disso, OSINT permite que as agências de inteligência complementem suas fontes tradicionais de informação, como fontes classificadas e confidenciais, agregando uma perspectiva mais diversificada e atualizada sobre os acontecimentos globais. Isso possibilita uma tomada de decisão mais informada e fundamentada.

O trabalho [Pastor-Galindo et al. 2020] trata-se de um estudo abrangente que fornece uma visão perspicaz de OSINT no contexto da cibersegurança. O artigo discute as vantagens de usar essa tecnologia, ao mesmo tempo em que destaca suas limitações. Ele explora uma variedade de técnicas e serviços relacionados à OSINT, oferecendo um recurso valioso para pesquisadores e profissionais da área [Pastor-Galindo et al. 2020].

O estudo enfatiza a importância da metodologia na realização de investigações eficientes e apresenta várias ferramentas que podem aprimorar a eficácia das práticas de OSINT. Ele também aborda o futuro da OSINT, considerando questões como o uso indevido da tecnologia, a filtragem de dados irrelevantes e as implicações éticas e de privacidade associadas ao seu uso. Servindo como um importante referencial para este trabalho, direcionando o entendimento e enfatizando a importância do assunto.

2.2 MÉTODOS PARA OSINT

A utilização de ferramentas por si só não possui a capacidade de produzir inteligência, podendo tornar-se obsoleta com o tempo e/ou sofrer interferências durante sua fase de coleta devido a eventuais mudanças que a fonte de dado pode sofrer. Nesse sentido, assim como qualquer outra forma de inteligência, OSINT segue uma metodologia precisa, e o processo de produção de inteligência pode ser descrito como uma sequência cíclica de etapas [Tanabe et al. 2022].

De forma geral, as fases de uma busca via OSINT não são fixas e podem intercalar entre si, exigindo um planejamento durante todas as etapas, além de um planejamento geral a fim de se alcançar um objetivo especifico. Essas fases são definidas pela Figura 2.1.

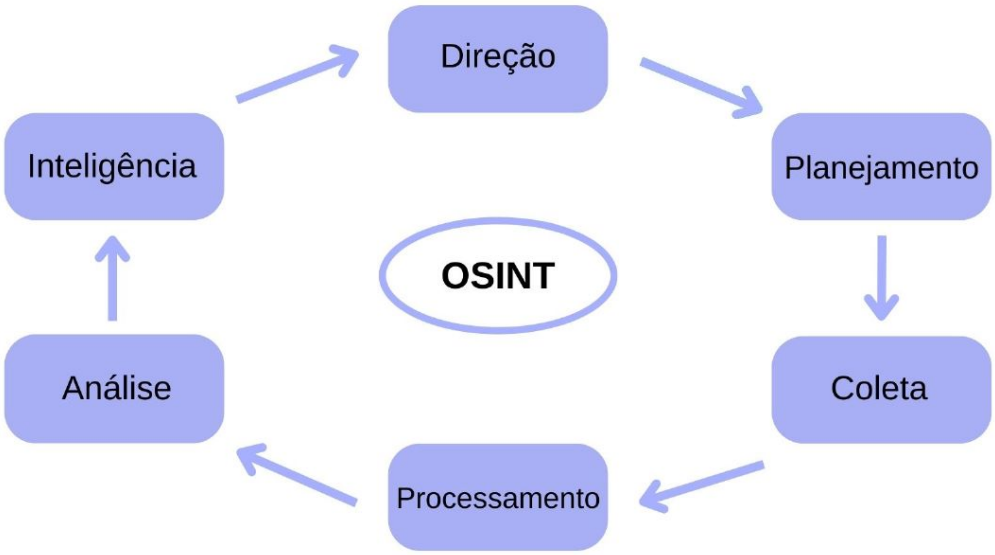

Figura 2.1: Metodologia básica para OSINT. Fonte: Autores

A fase inicial do processo de construção de inteligência usando OSINT é a definição dos requisitos, nesta são estabelecidas as informações relevantes e a criticidade da situação. Essa etapa também orienta os investigadores sobre o objeto da investigação, fornecendo diretrizes sobre as fontes a serem utilizadas.

As fontes podem ser classificadas em abertas e fechadas. As fontes abertas são aquelas de livre acesso, que não impõem obstáculos para obtenção de dados e conhecimento, incluindo mídia de massa, redes soci-

ais, fóruns, blogs, dados públicos do governo e publicações ou dados comerciais. [Williams e Blum 2018]

Já as fontes fechadas, referem-se a fontes de informação nas quais o acesso aos dados é restrito ou protegido. Diferentemente das fontes abertas, fontes fechadas requerem autorização ou privilégios especiais para acessar suas informações. Isso pode incluir documentos classificados, bancos de dados restritos, sistemas internos de uma organização, informações proprietárias ou dados pessoais protegidos por leis de privacidade. Vale ressaltar que a utilização dessas fontes para a maioria dos fins, seja de segurança, ou acadêmico, não é aceito [Tomás 2019]. Portanto, para que a metodologia de OSINT seja aplicada é necessário que os dados sejam oriundos de fontes abertas.

No contexto da coleta de informações, existem disciplinas especializadas conhecidas como "disciplinas de inteligência", que se concentram na coleta de informações por meio de métodos específicos, como HUMINT (inteligência de fontes humanas), SIGINT (interceptação e decodificação de sinais eletromagnéticos) e IMINT (inteligência obtida por meio de imagens de reconhecimento, principalmente imagens de satélite). A Inteligência de Fontes Abertas pode ser definida como um compêndio dessas informações que estão publicamente disponíveis de alguma forma [Williams e Blum 2018].

Existem três formas básicas de coleta de informações por meio do OSINT, conhecidas como coleta de informações passivas, coleta de informações ativas e coleta de informações semi-passivas:

1. Coleta de informações passivas: baseia-se na não detecção da coleta, ou seja, o alvo da investigação via OSINT não deve perceber que está sendo investigado. Essa tática é de alta dificuldade e baixa precisão, pois não é possível enviar qualquer tipo de tráfego para a organização-alvo, limitando-se, muitas vezes, à coleta de informações armazenadas que podem estar desatualizadas.

2. Coleta de informações ativas: envolve o mapeamento ativo da infraestrutura de rede, sendo que o alvo tem conhecimento de que seus dados estão sendo coletados e analisados. Essa técnica pode ser descrita como um reconhecimento e mapeamento inicial para a realização de possíveis atividades de teste de penetração.

3. Coleta de informações semi-passivas: nesse tipo de coleta, não são executados *portscans* ou *crawlers* em nível de rede, a busca é limitada a metadados em documentos e arquivos publicados. O perfil do alvo é traçado a partir de buscas que aparecem como tráfego de comportamento normal na internet.

No entanto, é importante ressaltar que o mero acesso aos dados não possui utilidade intrínseca; é necessário interpretá-los adequadamente para extrair fatos relevantes por meio de uma análise aprofundada. Após a fase de coleta, segue-se a etapa de análise dos dados filtrados, na qual diversas técnicas podem ser empregadas com o objetivo de obter uma compreensão mais completa e precisa. Nesse sentido, destacam-se as seguintes técnicas como relevantes:

- Análise léxica: os dados brutos devem ser examinados para extrair informações concretas do texto. É essencial aplicar processos de tradução para o idioma usado na investigação OSINT e filtrar o ruído que não agrega valor à investigação.

- Análise semântica: com o objetivo de compreender os dados, os algoritmos de processamento de linguagem natural (PNL) são amplamente utilizados. Além disso, as técnicas de análise de sentimento

permitem contextualizar postagens subjetivas ou opiniões para classificar o estado emocional do autor (por exemplo, positivo, negativo ou neutro). Dessa forma, o processo de descoberta da verdade aborda a tarefa desafiadora de resolver conflitos em dados de várias fontes que defendem posições opostas sobre o mesmo assunto.

- Análise geoespacial: os dados coletados de redes sociais, eventos, sensores ou endereços IP podem ser analisados a partir de uma perspectiva baseada na localização. O uso de mapas ou gráficos facilita a representação e compreensão dos dados, bem como a extração de conexões significativas entre incidentes ou pessoas.

- Análise de mídias sociais: permite que os pesquisadores realizem análises aprofundadas dos usuários. Nesse cenário, a análise de dados sociais permite traçar perfis comportamentais e/ou interesses sobre um determinado assunto ou autor.

- Análise de dados governamentais: envolve a exploração sistemática e a interpretação das informações contidas nos veículo oficiais de divulgação governamental.

Por fim, são estabelecidas estratégias visando alcançar o objetivo final, que pode variar desde a otimização das estratégias de marketing de uma empresa até a identificação dos responsáveis por atividades maliciosas ou a obtenção de informações relevantes para embasar estratégias políticas. Nesse sentido, os objetivos do Open Source Intelligence são diversificados e abrangem uma ampla gama de possibilidades.

2.2.1 CICLOS DE INTELIGÊNCIA

As etapas do ciclo básico podem ser explicadas de maneia sucinta na seguinte citação, "Em síntese, a Atividade de Inteligência teria que saber o que o seu usuário precisa, coletar e analisar informações para transformá-las em Inteligência, e então, entregar esse produto para o usuário."[Tanabe 2023]. Todavia, por se tratar de um método flexível onde cada agência o desenvolve a seu critério, o ciclo de Inteligência também é descrito de diversas formas pela literatura especializada.

O trabalho de [Tanabe et al. 2022] apresenta uma proposta de desenvolvimento de um método para OSINT que se destaca por sua abordagem lógica, coerente e fundamentada em valores, princípios éticos, metodológicos e profissionais. De forma geral, o trabalho tem como objetivo melhorar as propostas de metodologia de OSINT que são muitas vezes apresentadas apenas como versões do ciclo genérico de Inteligência. Nesse sentido, o presente projeto utiliza-se da junção de duas metodologia para trazer uma proposta simplificada e objetiva no desenvolvimento da arquitetura de forma a alcançar uma análise eficiente dos dados governamentais.

O primeiro é o Ciclo Básico de Inteligência adotado pela Organização do Tratado do Atlântico Norte (OTAN), composto por quatro fases: Direção, Coleta, Processamento e Disseminação, este pode ser visto na Figura 2.2.

Figura 2.2: Ciclo de Inteligência (adotado pela OTAN). Fonte: [Tanabe 2023]

O segundo, o ciclo da Doutrina dos Estados Unidos da América (EUA), visto na Figura 3.1. Este, apresenta um maior nível de complexidade, ainda de acordo com [Tanabe 2023] e é composto por 6 etapas as quais podem ser explicadas da seguinte maneira:

- Planejamento e direção: Estabelecimento das solicitações do usuário, e planejamento das atividades que serão realizadas pelo profissional para realizar o trabalho de Inteligência solicitado;
- Coleta: Coletar as informações/dados necessários para a realização do trabalho solicitado na etapa anterior;
- Processamento e Exploração: Tratamento e transformação dos dados obtidos em um formato legível e funcional para a criação do produto final;
- Análise e Produção: Examinar as informações, combinar e tratar para que as mesmas tenham a utilização necessária no produto final.
- Disseminação: Entrega do produto final ao usuário.
- Avaliação: Realizar avaliação de cada etapa do ciclo para aperfeiçoar as etapas individualmente e refinar o ciclo, de acordo com as solicitações do usuário.

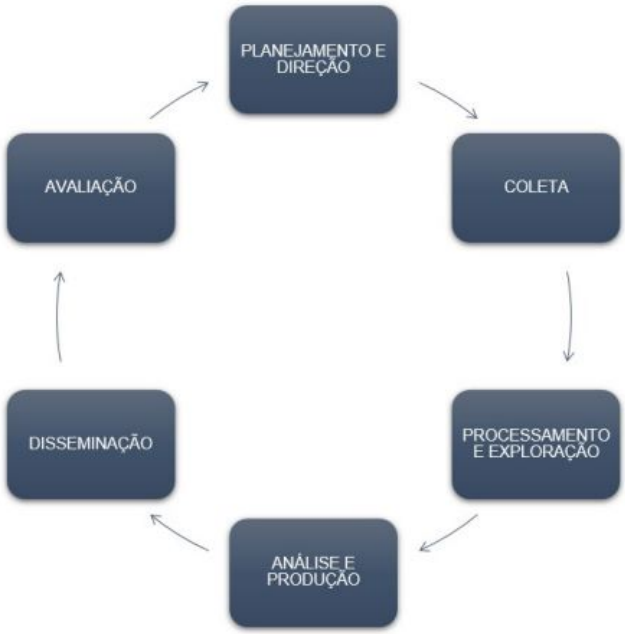

Figura 2.3: Ciclo de Inteligência (adaptado da Doutrina dos EUA). Fonte:[Tanabe 2023]

2.3 DESAFIOS E ABORDAGENS NA ANÁLISE DE DADOS NÃO ESTRUTURADOS

A implementação bem-sucedida de uma metodologia OSINT possui um conhecimento completo sobre quais dados serão utilizados como fonte para o estudo. Para obter resultados eficazes, as informações usadas como base da pesquisa devem ser tratadas, estudadas e processadas de forma adequada. O cuidado na seleção e preparação dos dados contribui significativamente para a qualidade e precisão dos resultados obtidos.

A extração de informações a partir de uma fonte de dado não estruturada é um processo que inicia com o mapeamento do conteúdo apresentado, passa pela coleta e finaliza na análise para então transformá-los em informações compreensíveis. Neste trabalho, a fonte de dados consiste em dados não estruturados, ou seja, informações que não seguem uma organização específica [Simões 2022].

Dados não estruturados não possuem uma estrutura bem definida e isso dificulta todo o processo de análise [Simões 2022]. Neste trabalho, a fonte de dados é composta por uma grande quantidade de PDFs, extensos e sem padronização. Os diários que serão utilizados como fonte são publicados pelo Diário da Justiça Eletrônico e contêm atos judiciais do Poder Judiciário Estatal. Os documentos têm garantia de autenticidade e confiabilidade, sendo publicados todos os dias úteis por todos os tribunais regionais de todas as regiões do Brasil, incluindo o Tribunal Superior do Trabalho. Em média, são 25 documentos por dia, sendo que cada caderno pode chegar a ter dezenas milhares de páginas.

O manuseio desse tipo de dado apresenta desafios significativos em várias frentes. O tempo de processamento é um fator crítico, exigindo recursos computacionais robustos e podendo atrasar o processo. O

armazenamento também se torna uma questão complexa, uma vez que se trata de um conjunto de dados em constante crescimento, e cada documento tem um tamanho considerável. A extração de *insights* é outro desafio, pois a análise de um texto tão extenso dificulta a identificação de padrões e a obtenção de informações relevantes. Além disso, a inconsistência de formato é uma questão pertinente, uma vez que os dados não são padronizados, exigindo abordagens diferenciadas para a extração de informações úteis.

O trabalho destaca como um estudo baseado em metodologia pode conduzir a resultados eficazes, especialmente garantindo o principal objetivo de capacitar a extração, filtragem, análise e busca simultânea de informações em todos os documentos disponibilizados pelo DJE. Isso possibilitará a obtenção de informações relevantes e o desenvolvimento de inteligência por meio dessa fonte de dados.

2.3.1 PROCESSAMENTO DE DADOS DE FONTES OFICIAIS DO GOVERNO

Os Diários Oficiais são instrumentos de publicação governamental amplamente utilizados em muitos países ao redor do mundo publicados sob a responsabilidade das autoridades governamentais e portanto, considerados fontes confiáveis e autênticas.

Esses são compostos por uma variedade de documentos que podem incluir leis, regulamentos, decretos, portarias, editais de licitação, nomeações de cargos públicos, comunicados e outros documentos relevantes. Essas informações são normalmente publicadas em uma periodicidade regular, fornecendo uma fonte contínua de dados para análise.

No contexto de OSINT, a análise dos Diários Oficiais apresentam alguns desafios específicos. Primeiramente, a quantidade de informações disponíveis pode ser volumosa e requer técnicas adequadas de coleta e processamento. Além disso, os documentos podem estar em formatos variados, exigindo a conversão e organização dos dados para facilitar a análise. A complexidade da linguagem jurídica e a presença de termos técnicos também podem dificultar a compreensão e a extração de informações. Esta monografia apresenta uma possível solução para essas problemáticas discutida na sessão 3.

Em [Alles 2018], o autor explora a utilização de uma arquitetura para extrair dados do Diário Oficial da União, empregando ferramentas que reconhecem entidades nomeadas na fonte de dados selecionada. O trabalho ressalta a importância de compreender e familiarizar-se com os dados com os quais se está trabalhando, bem como entender sua estrutura. O embasamento teórico deste projeto desempenhou um papel crucial na elaboração do trabalho proposto nesse documento.

Os conhecimentos adquiridos auxiliaram na fase de coleta, proporcionando direcionamento na coleta e tratamento dos dados, uma vez que se tratava de uma fonte de dados semelhante. Além disso, o trabalho teve relevância na fase de análise, pois ter acesso às informações e conhecer o formato do texto auxiliou em uma análise mais precisa.

2.4 ELASTICSEARCH

O ElasticSearch é um banco de dados orientado a documentos, amplamente utilizado como ferramenta para busca e análise de dados distribuído, gratuito e aberto para todos os tipos de dados, incluindo textuais,

numéricos, geoespaciais, estruturados e não estruturados.

Desenvolvido com base no Lucene - mecanismo de busca de texto com alto desempenho, o Elasticsearch pode funcionar como um banco de dados, entretanto, não é possível realizar transações no sentido típico, ou seja, não há como reverter um documento uma vez que foi enviado, e não há como enviar um grupo de documentos e ter todos ou nenhum deles indexados. Todavia existe uma funcionalidade de modo a garantir a durabilidade das operações.

A visibilidade das alterações é controlada quando um índice é atualizado, que, por padrão, ocorre uma vez por segundo e a cada *shard*. É importante destacar a relevância dos índices no ElasticSearch. De acordo com a documentação da ferramenta, "um indice é como um 'banco de dados' em um banco de dados relacional e tem um mapeamento que define vários tipos. De forma sucinta, um índice é um espaço de nome lógico que mapeia para um ou mais *shards* principais e pode ter zero ou mais *shards* de réplica"[Elastic 2023].

Para exemplificar, pode-se dizer que um índice presente em um banco de dados relacional são compostos por uma coleção de tabelas, e estas armazenam as informações relevantes. Quanto aos *shards*, podemos defini-los como instâncias de um índice, que funciona como um mecanismo de pesquisa independente. Cada *shard* indexa e manipula consultas para uma parte dos dados em um *cluster* do ElasticSearch. Podemos considerar um *shard* como uma unidade autônoma que contém uma fração dos dados em um índice [Elastic 2023].

3 METODOLOGIA E ARQUITETURA PROPOSTA

Este capítulo pretende descrever a arquitetura proposta e vincular os processos realizados com a metodologia utilizada. Nesta parte da monografia é possível verificar todo o processo de instalação, configuração e implementação das ferramentas e bibliotecas utilizadas para a realização do projeto.

Por fim, espera-se que as informações contidas neste capítulo possibilitem o entendimento completo da implantação das tecnologias presentes no projeto.

3.1 METODOLOGIA

Como já enunciado anteriormente, OSINT é composto por uma metodologia flexível e portanto, adapta-se de acordo com a necessidade especifica da organização, desde que se mantenham os requisitos básicos para aplicação dessa técnica. Sendo assim, na análise de requisitos realizada para construção desse estudo, ficou evidente que a utilização de apenas uma das metodologias não seria aplicável para a realização do projeto.

Dessa forma, a metodologia implementada no projeto uniu dois dos ciclos base de inteligência, o Ciclo Básico de Inteligência adotado pela Organização do Tratado do Atlântico Norte (OTAN), e o ciclo da Doutrina dos EUA, ambos apresentados na seção 2.2.1. Essa adaptação foi necessária uma vez que este trabalho possui finalidade acadêmica e portanto são descartadas as fases de disseminação e avaliação pois para a perfeita aplicação dessas, seriam necessários um cliente ou usuário final.

Conforme evidenciado na Figura 3.1, a metodologia adotada pôde ser compreendida em 4 etapas: Definição requisitos, Coleta, Processamento e Análise, especificadas da seguinte forma:

- Definição de requisitos: Nesta etapa foi definido o problema a ser solucionado, no caso a busca de dados nos Diários oficiais da Justiça do Trabalho. Dessa forma, foi definida a base de dados como o site do DEJT (Diário Eletrônico da Justiça do Trabalho).
- Coleta: A coleta dos dados foi realizada a partir de um robô para fazer o *Download* de todos os Diários do dia em formato PDF.
- Processamento: A fase de processamento é a etapa onde os arquivos PDF são convertidos em arquivos de texto com a finalidade de facilitar a manipulação dos dados, nesta etapa também é realizada a indexação dos arquivos no ElasticSearch.
- Análise: Esta etapa é realizada a busca e visualização dos dados de interesse, e é feita pela plataforma Kibana.

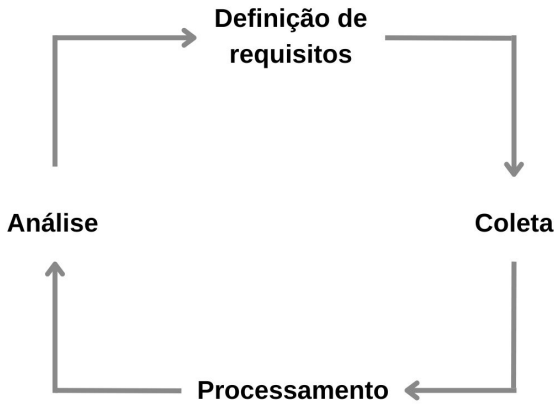

Figura 3.1: Metodologia proposta. Fonte: autores

3.2 DEFINIÇÃO DE REQUISITOS

Nesta etapa, a fase de definição de requisitos tem como objetivo estabelecer os critérios necessários para a busca de informações através de OSINT, determinar as fontes de dados a serem utilizadas e as estratégias de coletas que serão aplicadas para então produzir inteligência.

3.2.1 REQUISITOS DA FONTE DE DADOS

A fonte de dados selecionada consiste nos cadernos eletrônicos judiciários da Justiça do Trabalho. Esses cadernos estão divididos em 25 documentos distintos, cada um correspondente a 24 regiões e um documento do Tribunal Superior do Trabalho (TST). Cada região engloba um ou mais estados, conforme ilustrado na Figura 3.2.

Tribunais Regionais do Trabalho:

Tribunal Regional do Trabalho da 1ª Região (TRT1) (RJ)
Tribunal Regional do Trabalho da 2ª Região (TRT2) (SP / Grande São Paulo e Baixada Santista)
Tribunal Regional do Trabalho da 3ª Região (TRT3) (MG)
Tribunal Regional do Trabalho da 4ª Região (TRT4) (RS)
Tribunal Regional do Trabalho da 5ª Região (TRT5) (BA)
Tribunal Regional do Trabalho da 6ª Região (TRT6) (PE)
Tribunal Regional do Trabalho da 7ª Região (TRT7) (CE)
Tribunal Regional do Trabalho da 8ª Região (TRT8) (AP e PA)
Tribunal Regional do Trabalho da 9ª Região (TRT9) (PR)
Tribunal Regional do Trabalho da 10ª Região (TRT10) (DF e TO)
Tribunal Regional do Trabalho da 11ª Região (TRT11) (AM e RR)
Tribunal Regional do Trabalho da 12ª Região (TRT12) (SC)
Tribunal Regional do Trabalho da 13ª Região (TRT13) (PB)
Tribunal Regional do Trabalho da 14ª Região (TRT14) (AC e RO)
Tribunal Regional do Trabalho da 15ª Região (TRT15) (SP / Interior e Litoral Norte e Sul)
Tribunal Regional do Trabalho da 16ª Região (TRT16) (MA)
Tribunal Regional do Trabalho da 17ª Região (TRT17) (ES)
Tribunal Regional do Trabalho da 18ª Região (TRT18) (GO)
Tribunal Regional do Trabalho da 19ª Região (TRT19) (AL)
Tribunal Regional do Trabalho da 20ª Região (TRT20) (SE)
Tribunal Regional do Trabalho da 21ª Região (TRT21) (RN)
Tribunal Regional do Trabalho da 22ª Região (TRT22) (PI)
Tribunal Regional do Trabalho da 23ª Região (TRT23) (MT)
Tribunal Regional do Trabalho da 24ª Região (TRT24) (MS)

Figura 3.2: Lista de regiões. Fonte: [Tribunais]

No processo de mapeamento, serão identificadas as entidades específicas que serão alvo da busca, realizadas por meio de diferentes critérios, como números de processos, nomes das pessoas envolvidas no caso, suas funções, datas de publicação, portarias, nomes de desembargadores, entre outros. Esses elementos são utilizados para identificar as informações úteis nos diários eletrônicos da justiça do trabalho, conforme visto na Figura 3.3.

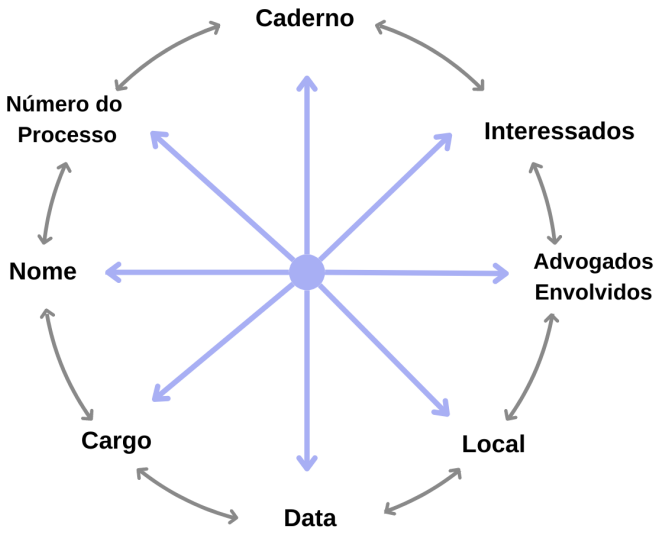

Figura 3.3: Mapeamento da Informação. Fonte: Adaptada - [Pastor-Galindo et al. 2020]

Como produto final espera-se que as informações especificadas para cada tipo de informação resulte em uma saída que também poderá ser utilizada como entrada para outro processo, conforme metodologia proposta por [Pastor-Galindo et al. 2020], permitindo a identificação de cada informação entre diferentes etapas.

Essa abordagem de encadeamento de técnicas proporciona a geração contínua de novos dados à medida que cada etapa é executada. Dessa forma, é possível obter informações adicionais e aprimorar a qualidade e a abrangência dos resultados ao longo do processo.

3.2.2 REQUISITOS DA COLETA

Para garantir uma coleta eficiente, é essencial definir os requisitos necessários, como a disponibilidade dos dados no portal da Justiça do Trabalho que só estão disponíveis em dias úteis. Importante ressaltar que os documentos disponíveis na data atual referem-se sempre aos eventos do dia anterior. Com isso, sugere-se o desenvolvimento de um robô capaz de acessar automaticamente o site do DEJT diariamente, preencher os campos obrigatórios e efetuar o *download* dos dados desejados.

O site onde os dados estão disponíveis permite o *download* dos arquivos em PDF após o preenchimentos dos campos de "Data de Início", "Data de Fim", "Tipo de Caderno"e "Orgão", conforme mostrado na Figura 3.4.

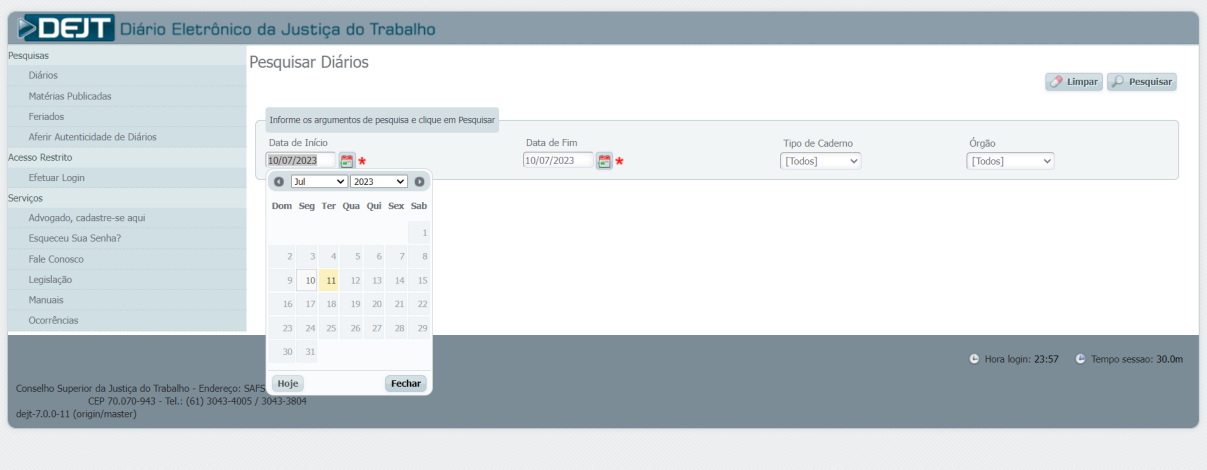

Figura 3.4: Site para download dos Diários Eletrônicos. Fonte: Justiça do Trabalho

3.2.3 REQUISITOS DO PROCESSAMENTO

Percebe-se que os dados são disponibilizados em formato PDF em um número consideravelmente grande de páginas, podendo ultrapassar 20.000 (vinte e mil) páginas a depender do dia e da região do caderno. Dessa forma, o processamento de arquivos em PDF trazem desafios, tais quais:

- Estrutura não padronizada: Arquivos PDF podem ser criados de várias maneiras, e a estrutura interna pode variar significativamente entre diferentes PDFs. Isso torna difícil desenvolver uma abordagem única para extrair informações de todos os documentos PDF.
- Falta de semântica: O conteúdo em um PDF geralmente não possui informações semânticas bem definidas, como *tags* ou marcadores específicos para cada tipo de dado. Isso torna difícil identificar e extrair automaticamente informações importantes, como títulos, subtítulos, tabelas ou dados específicos.
- Dificuldade de pesquisa: PDFs são frequentemente usados como formatos de apresentação ou distribuição de documentos finais e não possuem um mecanismo simples para identificação de uma determinada informação.
- Tamanho do arquivo: Documentos PDF grandes e complexos podem exigir muito processamento computacional para extrair dados com eficiência.

Portanto, concluiu-se que o processamento dessa quantidade de informações a partir de arquivos em PDF seria inviável. Assim, tornou-se necessário converter os arquivos PDF para um formato de dados tratável. A opção escolhida foi utilizar arquivos do tipo texto, o que demandaria o desenvolvimento de um *script* capaz de realizar essa conversão de forma precisa, preservando a estrutura original da apresentação dos dados. Esse desafio adicional representa uma importante etapa para o êxito do trabalho proposto.

3.2.4 REQUISITOS DA ANÁLISE

Todas as etapas anteriores são fundamentais para chegarmos à etapa final do ciclo apresentado - a análise. É nessa etapa que a inteligência pode ser extraída.

Cada uma dessas etapas desempenha um papel importante ao transformar dados complexos em informações tratáveis e capazes de serem analisadas. Com base nisso, sugere-se a apresentação e disponibilização de algumas informações relevantes por meio de *dashboards*, enquanto outras podem ser acessadas por meio de pesquisas, de acordo com a necessidade individual do pesquisador.

3.3 ARQUITETURA PROPOSTA

Sequencialmente, após mapear e apresentar os requisitos como base para a pesquisa e produção de inteligência, foi desenvolvida a arquitetura a ser implementada, conforme ilustrado na Figura 3.5. As etapas que compõem essa arquitetura serão detalhadas a seguir, dando continuidade ao ciclo de inteligência apresentado como metodologia.

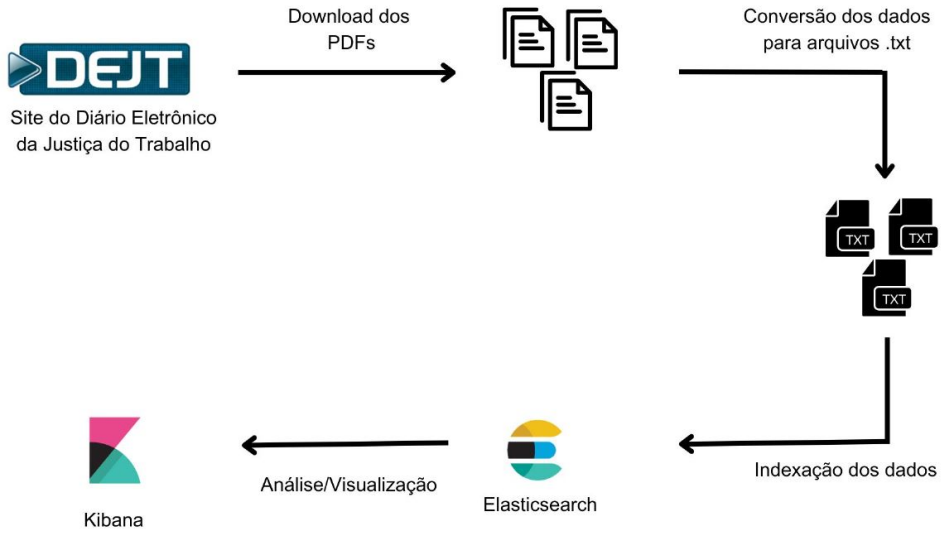

Figura 3.5: Arquitetura proposta. Fonte: autores

3.4 COLETA

Para a coleta, criou-se um robô que desempenha uma função semelhante a de um *crawler* - algoritmo utilizado para extrair e analisar dados de *websites*. A logica do robô é detalhada de acordo com a Figura 3.6.

Figura 3.6: Logica do robô de coleta dos cadernos judiciários. Fonte: autores

Inicialmente o robô deve acessar o site onde os diários estão disponíveis, preencher os campos conforme definidos em 3.2.2, setar o local onde os dados serão armazenados e então realizar o download dos arquivos PDFs.

Para que o robô pudesse preencher os campos disponíveis, foi necessário acessar o código fonte do site e identificar o *id* e o *name* das classes de cada um dos campos, dessa forma, o robô pôde simular a ação do *mouse* e preencher os campos corretamente, conforme mostrado na Figura 3.4.

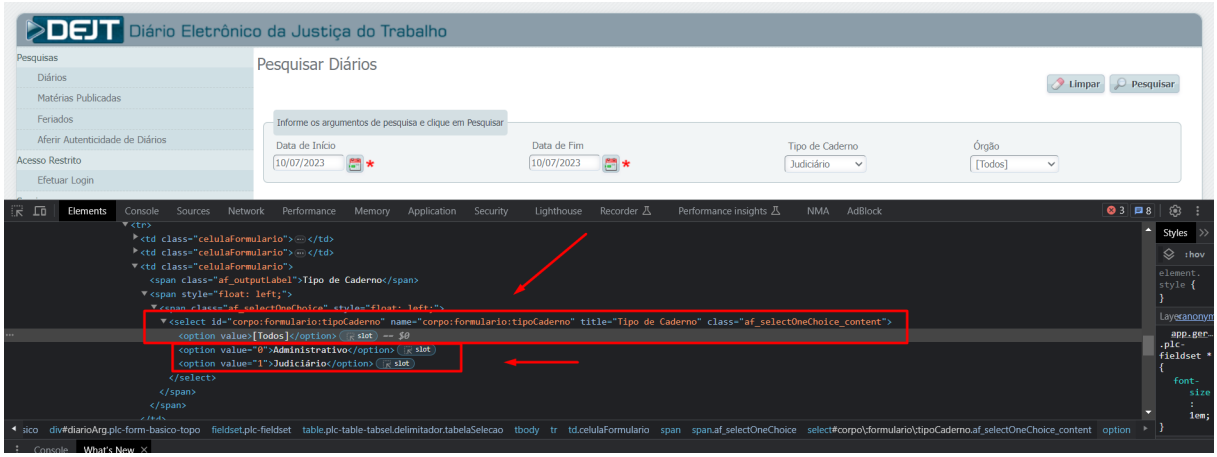

Figura 3.7: Código fonte do site da Justiça do Trabalho. Fonte: Justiça do Trabalho

Dessa forma, definimos os valores dos nomes das classes e seus respectivos *ids* conforme evidenciado na Tabela 3.1:

Por fim, vale ressaltar que o robô apresentado funciona especificamente para essa fonte de dados, não

Tabela 3.1: Tabela definição dos campos que deverão ser preenchidos pelo robô proposto na Figura 3.7

Campo	**name**	**Id**
Data de Início	corpo:formulario:dataIni	corpo:formulario:dataIni
Data de Fim	corpo:formulario:dataFim	corpo:formulario:dataFim
Tipo de Caderno	corpo:formulario:tipoCaderno	0 ou 1

devendo ser utilizado para outras fontes.

3.5 PROCESSAMENTO

Para o processamento dos dados também foi preciso criar uma robô para conversão dos arquivos PDFs em tipo texto, a logica pode ser vista na Figura 3.8.

Figura 3.8: Logica do robô de conversão de PDF para txt. Fonte: autores

Primeiramente o robô deve ler todos os arquivos da pasta, chamar a função de conversão e então salvar os novos arquivos convertidos em uma nova pasta. Após a conversão, os dados são inseridos no banco de dados orientado a documentos onde estarão disponíveis para análise.

3.6 ANÁLISE

Na fase de análise, é de suma importância que os dados estejam prontamente disponíveis para acesso e que seja viável realizar pesquisas por informações contidas nos cadernos.

Após identificação das informações requeridas, torna-se viável produzir inteligência e gerar gráficos e tabelas para uma visualização mais eficaz. Essa capacidade de análise aprofundada e representação visual contribui para uma compreensão mais completa dos dados.

3.7 FERRAMENTAS UTILIZADAS

Nesta seção, serão apresentadas e definidas as ferramentas utilizadas no projeto para a implementação da proposta de arquitetura. Essas ferramentas desempenharam um papel fundamental na execução do projeto, proporcionando recursos e funcionalidades específicas necessárias para a consumação da metodologia, que será melhor apresentada na seção 3.

3.7.1 Oracle VM VirtualBox

Oracle VM VirtualBox é um software de virtualização que permite ampliar a capacidade de execução de um computador hospedeiro, independentemente do seu Sistema Operacional base (Windows, Mac OS, Linux ou Oracle Solaris), permitindo a execução de vários Sistemas Operacionais simultaneamente. Isso é possível através do uso de instâncias virtuais, que são sistemas operacionais independentes, sejam eles idênticos ou diferentes, que operam em recursos dedicados de um único dispositivo de forma emulada. Esses recursos são limitados pela disponibilidade de hardware, como espaço em disco, capacidade de processamento e memória RAM [VirtualBox 2022].

Para este projeto, usaremos o Virtual Box para hospedar a ferramenta que será utilizada para visualização dos dados extraídos do Diários Oficiais Eletrônicos da Justiça do Trabalho. O virtualizador, o Oracle VM VirtualBox permite a execução de simultâneos sistemas operacionais utilizando-se do hardware hospedeiro da aplicação. Esta virtualização pode ser feita mediante a inserção de imagem de disco (Imagens ISO (extensão de arquivo de imagem obtido através da antiga ISO9660, comumente utilizada para padrões de CD-ROM), aqui será utilizada uma imagem do sistema operacional Linux - do Ubuntu 20.05.6.

Desta forma, o Oracle VM VirtualBox se torna a ferramenta importante para esse trabalho, sendo o mantenedor de toda a estrutura virtualizada que permite a hospedagem da pilha ELK (*Elasticsearch, Logstash e Kibana*), discutido a seguir.

3.7.2 ELK Stack

O Elastic Stack, também conhecido como ELK Stack, é um conjunto de projetos de código aberto, composto pelo Elasticsearch, Logstash e Kibana. Ele é amplamente utilizado para armazenar, processar e visualizar dados de diversos tipos e formatos de maneira segura.

- O Elasticsearch é um mecanismo de armazenamento e análise de dados distribuído, que oferece recursos avançados de busca e análise de dados em tempo real. Ele permite armazenar dados centralmente e realizar buscas rápidas e precisas, além de oferecer recursos analíticos poderosos [Elastic 2023].

- O Logstash ingere, transforma e envia dados dinamicamente, independentemente do formato ou complexidade. Ele pode derivar estrutura de dados não estruturados, decifrar coordenadas geográficas de endereços IP, anonimizar ou excluir campos confidenciais e facilitar o processamento geral. [Elastic 2023]

- O Kibana, por sua vez, é uma interface de visualização de dados gratuita e de código aberto. Ele permite aos usuários criar painéis interativos e dashboards para visualizar e explorar os dados armazenados no Elasticsearch. Com o Kibana, é possível criar gráficos, tabelas, mapas e outros tipos de visualizações personalizadas para analisar e monitorar os dados de forma intuitiva [Elastic 2023].

No contexto do projeto apresentado neste relatório, o Elasticsearch na versão 7.17.11 será adotado como o mecanismo de armazenamento e indexação das informações extraídas dos Diários Eletrônicos da Justiça do Trabalho. A escolha desse sistema se baseia em sua capacidade de centralizar os dados após transformação de arquivos PDF para arquivos de texto, que serão posteriormente analisados e visualizados por meio do Kibana - versão 7.17.11.

Para realizar a coleta e o envio dos dados ao Elasticsearch, optou-se por não utilizar o Logstash. Em vez disso, será desenvolvido um script em Python, aproveitando a API do Elasticsearch, que oferece uma biblioteca compatível com a linguagem Python. Essa abordagem proporcionará maior flexibilidade e controle sobre o processo de coleta e envio dos dados, permitindo uma integração mais direta e personalizada com as necessidades específicas do projeto.

A utilização do Elasticsearch em conjunto com o Python fornece uma solução eficiente e escalável para o armazenamento, indexação e análise dos dados extraídos, ao mesmo tempo em que simplifica o processo de desenvolvimento do sistema.

3.7.3 Python e bibliotecas

Python é uma linguagem de programação de alto nível, interpretada e de propósito geral, amplamente utilizada em diversos domínios acadêmicos. Sua popularidade crescente deve-se às suas características que promovem uma sintaxe clara e legível, além de oferecer suporte a uma vasta gama de bibliotecas e *frameworks*, tornando-a adequada para uma ampla variedade de aplicações [Python 2023]. Além disso, possui uma ampla gama de bibliotecas especializadas, para o desenvolvimento desse projeto utilizou-se a versão 3.11 do Python e as seguintes bibliotecas:

- *Playwright* é uma ferramenta de automação de testes e automação de multi-navegador. Ela usa uma abordagem baseada em protocolo, aproveitando o mecanismo interno de automação dos navegadores, em vez de emular a interação do usuário [Playwright 2023].

  Nesse projeto, ela foi utilizada na automação do processo de download dos Diários Oficiais em

formato PDF diretamente do site do Diário Eletrônico da Justiça do Trabalho <(https://dejt.jt.jus.br/dejt/f/n/diariocon)>. Vale salientar que de acordo com a disponibilidade dos diários, os downloads são realizados para o dia anterior (D-1), por exemplo, os diários referentes ao dia atual estarão disponíveis para download somente no dia seguinte.

- *os* é um módulo incorporado do Python que fornece uma interface para interagir com o sistema operacional em que o código Python está sendo executado. Ela oferece um conjunto de funções e métodos que permitem realizar tarefas relacionadas ao sistema operacional, como manipulação de arquivos, acesso ao ambiente, gerenciamento de processos entre outros [Python 2023]. Nesse sentido, com o auxílio da biblioteca, tornou-se possível manipular os arquivos de forma mais eficiente, simplificando as operações relacionadas aos documentos dos Diários Oficiais.

- *pdfminer* é uma biblioteca em Python utilizada para extrair e transformar texto e metadados de arquivos PDF. Ela fornece uma maneira eficiente de analisar e processar documentos em formato PDF, permitindo a extração de informações relevantes para uso em projetos de processamento de texto, mineração de dados e análise de dados. [PyPI 2023]. Aqui ela desempenhou um papel fundamental na conversão dos Diários Oficiais da Justiça do Trabalho, disponíveis em formato PDF, para arquivos de texto. Essa conversão foi realizada com o objetivo de melhorar a indexação dos documentos no Elasticsearch e facilitar o tratamento dos dados extraídos.

- *Elasticsearch-Py* é uma biblioteca oficial do Elasticsearch projetada especificamente para a linguagem de programação Python. Sua finalidade é fornecer uma interface de programação que permita interagir com um cluster do Elasticsearch. Essa biblioteca desempenha um papel crucial ao simplificar a comunicação e a execução de várias operações com o Elasticsearch por meio de uma API dedicada [ElasticPy 2023]. No contexto deste trabalho, o Elasticsearch-Py foi utilizado para indexar os arquivos previamente convertidos em texto. Essa etapa de indexação é fundamental, pois possibilita um armazenamento eficiente dos documentos no Elasticsearch. Essa abordagem otimiza a pesquisa e recuperação de informações posteriormente, facilitando o acesso aos documentos com base em critérios específicos.

- *Glob*: módulo para listar todos os nomes de caminhos correspondentes a um padrão especificado de acordo com as regras usadas pelo shell [Tecnologia 2023], utilizada aqui para ler todos os arquivos presentes nas pastas pós transformação dos arquivos texto.

É importante ressaltar que a utilização dessas bibliotecas no projeto permitiu automatizar tarefas, otimizar a extração e o tratamento dos dados dos Diários Oficiais, além de possibilitar uma indexação mais eficiente no Elasticsearch.

4 TESTES E RESULTADOS

Nessa seção, aprofunda-se na implementação da arquitetura abordada na seção 3.3, além de promover testes em busca dos resultados esperados, conforme descrito na seção anterior. Para garantir a validade e a robustez dos testes, foram conduzidos experimentos, utilizando um cenário fictício e um conjuntos de dados representativos.

Os resultados obtidos revelaram um desempenho promissor da arquitetura, com ganhos significativos em termos de processamento e análise dos dados do DEJT. Dessaforma, foir possível mensurar o impacto positivo da abordagem proposta em comparação com métodos tradicionais de pesquisa, destacando sua superioridade e vantagens distintas.

4.1 COLETA

O robô desenvolvido para a coleta dos dados foi escrito utilizando a linguagem *python*. Para o acesso a aplicação web do DEJT onde estão disponíveis os Diários Eletrônicos, adicionou-se um período para o carregamento e preenchimento dos campos entre cada ação que o robô deveria executar. Essa etapa foi necessária para garantir que a página seja carregada por completo antes do robô começar o processo de coleta dos arquivos.

```
page.goto('https://dejt.jt.jus.br/dejt/f/n/diariocon', timeout=80_000)
sleep(6)
page.locator("span input[name='corpo:formulario:dataIni']").fill(date)
sleep(6)
page.locator("span input[name='corpo:formulario:dataFim']").fill(date)
sleep(6)
page.select_option("span select[name='corpo:formulario:tipoCaderno']","1")
sleep(6)
```

Figura 4.1: Definição de tags para realizar o *Download* dos Diários. Fonte: autores

Inicialmente foi necessário estabelecer as datas de coleta como sendo sempre do dia anterior. Essa consideração é fundamental para garantir a integridade e a atualidade dos dados coletados. Dessa forma, assegura-se que os dados extraídos correspondam aos diários eletrônicos mais recentes disponíveis para análise e classificação.

Sendo assim, nota-se pela Figura 4.2 que após as importações das bibliotecas *Sleep* e *playwright*, o robô chama a função <open_page()> para iniciar a coleta. Entretanto, primeiramente ele define a data de coleta de forma a definir o dia atual e subtrair uma unidade conforme mencionado acima, vide Figura 4.2.

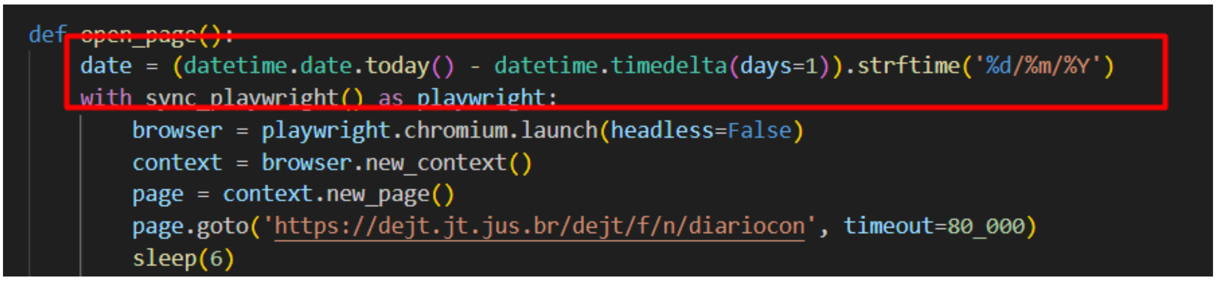

Figura 4.2: Definição das datas do script. Fonte: autores

Vale salientar que para esse trabalho utilizaremos apenas os cadernos judiciários definidos com o id 1, e o download será realizado para todos os órgãos preenchido como *default* pelo próprio site do DEJT. Sendo assim, o *script* foi construído de acordo a Figura 4.1.

Por fim, de acordo com a Figura 4.3 o robô da início ao download dos arquivos e salva todos os arquivos em uma pasta na maquina virtual finalizando a fase de coleta,

Figura 4.3: Exemplo de execução do robô. Fonte: autores

4.2 PROCESSAMENTO

Na fase de processamento, também foi desenvolvido um *script* em Python. Nessa etapa, o primeiro passo consistiu na conversão dos arquivos em formato PDF para arquivos em formato de texto. Essa etapa é fundamental para simplificar o processo de indexação e tratamento dos dados coletados.

O *script* em Python implementado para essa finalidade utiliza bibliotecas apropriadas já citadas na Seção 3.7.3, que fornecem funcionalidades específicas para a extração e conversão de texto a partir dos arquivos PDF. Essas bibliotecas possibilitam a extração precisa do texto contido nos documentos, mantendo a estrutura e a formatação adequadas.

Ao realizar a conversão dos arquivos PDF para texto, abre-se a possibilidade de explorar e utilizar efetivamente os dados coletados nos próximos passos do processo. Essa etapa é essencial para garantir a qualidade e a acessibilidade dos dados, contribuindo para a eficiência e eficácia das análises e tratamentos subsequentes.

A Figura 4.4 apresenta o código desenvolvido para facilitar o processo de conversão dos arquivos. Para realizar a conversão dos arquivos PDF, foi utilizada a biblioteca pdfminer. Essa biblioteca permitiu percorrer a pasta onde os PDFs são armazenados após o download, realizar a conversão e salvar os novos arquivos em formato de texto em uma nova pasta.

```
folderpdf =  r'/home/vboxuser/TCC/Diarios//'
foldertxt =  r'/home/vboxuser/TCC/Diarios_txt//'

for file_name in os.listdir(folderpdf):
    from pdfminer.high_level import extract_text
    text = extract_text(folderpdf + file_name)
    arquivo = open(foldertxt + file_name[:-4] , "a")
    arquivo.write(text)
#    os.remove(folderpdf + file_name)
```

Figura 4.4: Script de donversão dos PDFs em arquivos txt. Fonte: autores

A Figura 4.5 ilustra os arquivos pré e pós conversão, do lado esquerdo é possível visualizar o caderno de número Nº3758/2023 ainda em formato PDF, já do lado direto observa-se o caderno de mesmo número já convertido para o formato texto. Nota-se ainda pelo exemplo que não houve perda de informações no processo de conversão dos arquivos garantindo a integridade das informações contidas nesses documentos.

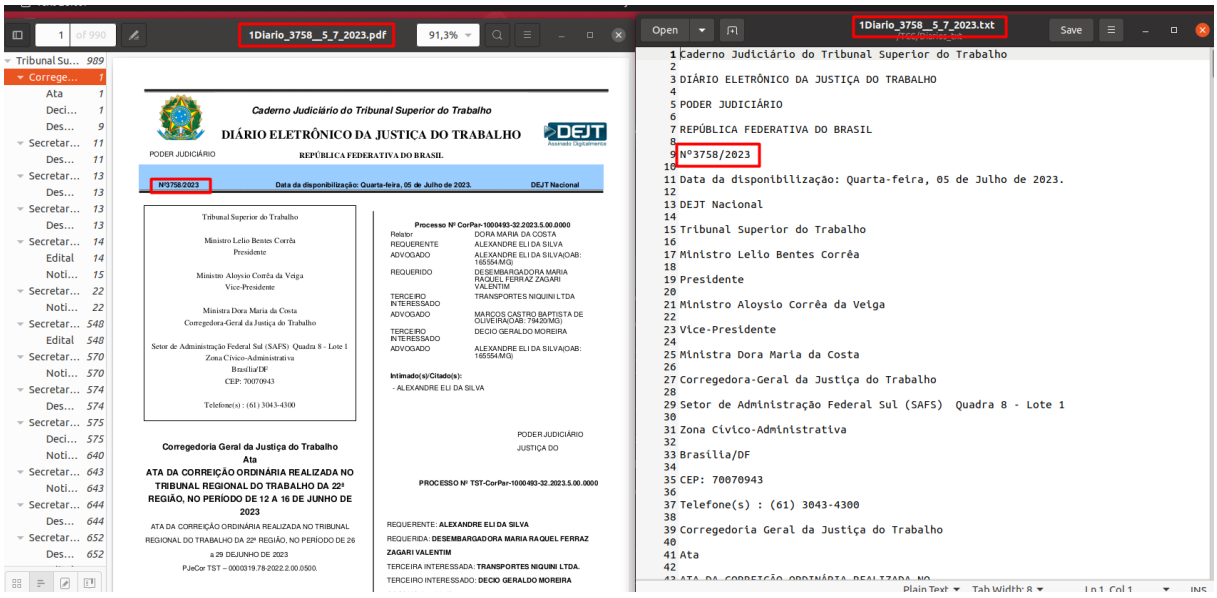

Figura 4.5: Exemplo de arquivos convertidos. Fonte: autores

Após a etapa de conversão segue-se para a etapa de indexação dos arquivos de texto no Elasticsearch, para isso utilizou-se a biblioteca Elaticsearch PY, citada na seção 3.7.3.

Conforme a Figura 4.6, o *script* estabelece a conexão com o servidor elasticsearch que está hospedado localmente na maquina virtual rodando a imagem do Ubuntu 20.04. Nota-se que o elasticsearch está hospedado no endereço <http://localhost:5601>. Após estabelecer a conexão o script chama a função de indexação que varre a pasta dos arquivos do tipo de texto, e indexa no *cluster* do elastic dentro do *index* também criado no código e nomeado por "diários".

```
    # Connect to Elasticsearch
es = Elasticsearch(hosts=['http://localhost:9200'])

print(es.info().body)

# Index a text file
def index_text_files(folder_path, index_name):
    file_paths = glob.glob(folder_path + '/*.txt')

    for file_path in file_paths:
        with open(file_path, 'r', encoding='latin-1') as file:
            text_content = file.read()

        document = {
            'content': text_content
        }

        # Index the document
        es.index(index=index_name, body=document)

        print(f"Text file '{file_path}' indexed successfully in '{index_name}'.")

# Usage example
index_name = 'diarios'
folder_path = '/home/vboxuser/TCC/Diarios_txt/'
```

Figura 4.6: Script para indexação dos dados. Fonte: autores

Posteriormente, realizou-se a criação de um *index* também nomeado como "diarios"no Kibana para que a fase de análise, presente na próxima seção, pudesse ser realizada.

4.3 ANÁLISE

Essa fase final da metodologia proposta envolve o uso do Kibana, que desempenhou um papel crucial ao simplificar as etapas de análise dos dados. O Kibana oferece recursos avançados, como a criação de *dashboards* personalizados, que possibilitam a visualização e análise dos dados de forma intuitiva e interativa.

Por meio do Kibana, os dados previamente indexados no Elasticsearch podem ser facilmente explorados e visualizados seguindo a proposta discutida na fase de requisitos. Através da criação de *dashboards*, é possível selecionar e combinar diferentes tipos de visualizações, como gráficos, tabelas e mapas, de acordo com as necessidades específicas do projeto.

A Figura 4.7 ilustra os dados indexados que poderão ser visualizados no Kibana. Essa visualização permite que os usuários tenham uma compreensão mais clara dos dados coletados e possam realizar análises mais aprofundadas.

Figura 4.7: Dados disponíveis no Kibana. Fonte: autores

O Kibana facilita a interação com os dados indexados, oferecendo recursos de filtragem, agrupamento e análise. Além disso, ele também permite a definição de painéis interativos que podem ser compartilhados com outras partes interessadas, tornando a comunicação e a colaboração mais eficazes durante o processo de análise.

Assim, com o uso do Kibana como parte integrante da metodologia proposta, foi possível obter *insights* valiosos e tomar decisões embasadas em informações concretas. A capacidade de criar *dashboards* personalizados no *Kibana* impulsionou a eficiência e a eficácia do processo de análise dos dados coletados.

Como forma de dar continuidade ao processo de análise criou-se um ambiente de visualização contendo a base dos cadernos judiciáios indexados no Kibana. Gráficos interativos trazem uma melhor melhoria na identificação do conteúdo dos cadernos, conforme visto na Figura 4.8.

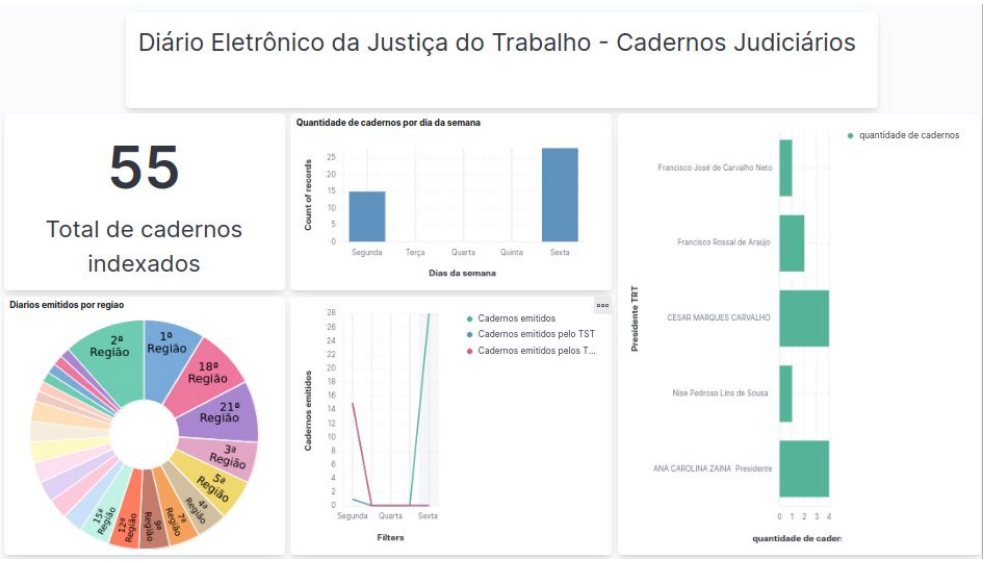

Figura 4.8: Dashboard criado a partir dos diários indexados. Fonte: autores

O gráfico apresentado na Figura 4.9 exemplifica a indexação dos diários por dias na semana, vale ressaltar que há dias da semana que não há dados disponíveis para download e, nesses dias, os campos ficam vazios. Além disso, só há diários disponíveis em dias úteis, excluindo finais de semanas e feriados.

Figura 4.9: Gráfico de indexação por dia na semana. Fonte: autores

A Figura 4.10 exemplifica os diários separados por região, cada região corresponder a um ou mais estados.

Figura 4.10: Gráfico por região. Fonte: autores

Vale ressaltar que o *dashboard* foi criado com o intuito de aprimorar e facilitar a visualização dos documentos indexados de forma geral. Ele permite ter uma noção mais clara da quantidade de publicações de diários em diferentes regiões, além de validar se os cadernos de todas as regiões mantêm um padrão de publicação. Além disso, o *dashboard* também é útil para validar a indexação desses documentos.

Como última fase do testes e resultados, foi adotada uma situação fictícia com o nome completo não

exposto por motivos de preservação, privacidade e segurança das entidades envolvidas. O primeiro passo consistiu em utilizar o nome da pessoa como parâmetro e realizar uma busca em todo o conteúdo de todos os documentos indexados no Elasticsearch. Os documentos que continham o nome do individuo procurado foram retornados como resultado da busca. O resultado pode ser visto na Figura 4.11.

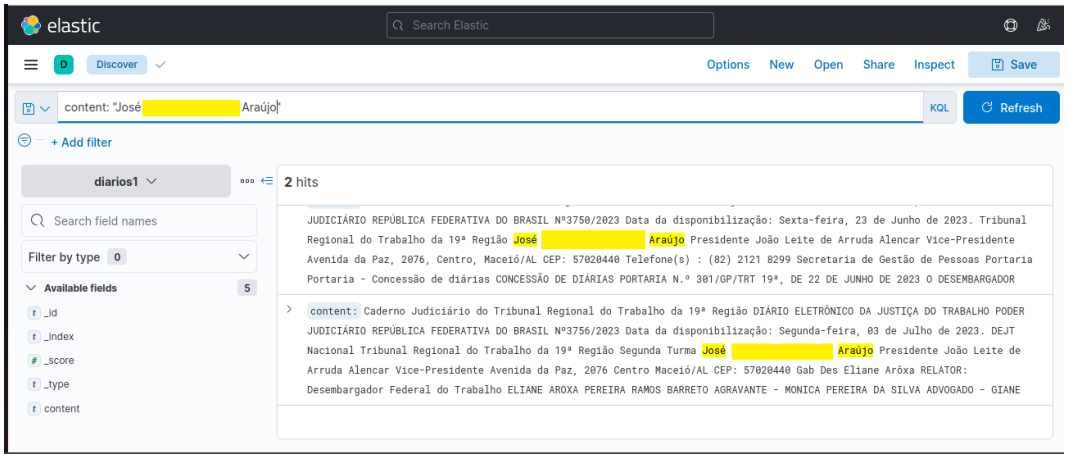

Figura 4.11: Exemplo de busca por nome em dados indexados. Fonte: autores

Conforme evidenciado na Figura 4.12, através do nome em questão, é possível verificar em um dos cadernos indexados a presença dos seguintes dados:

- Número do processo: 3756/2023
- Data de Divulgação: 03 de Julho de 2023
- Região: 19º Região
- Cargo: Presidente
- Outras pessoas envolvidas no processo: João (segundo nome grifado por motivos de segurança).
- Local: Maceió - AL

Ressalta-se que conforme proposto na seção 3 e explicado na Figura 3.3, se o valor da entrada for alterado para qualquer um dos valores de saída mencionados, será possível identificar as mesmas informações, uma vez que estão lincadas entre si por fazerem parte do mesmo processo.

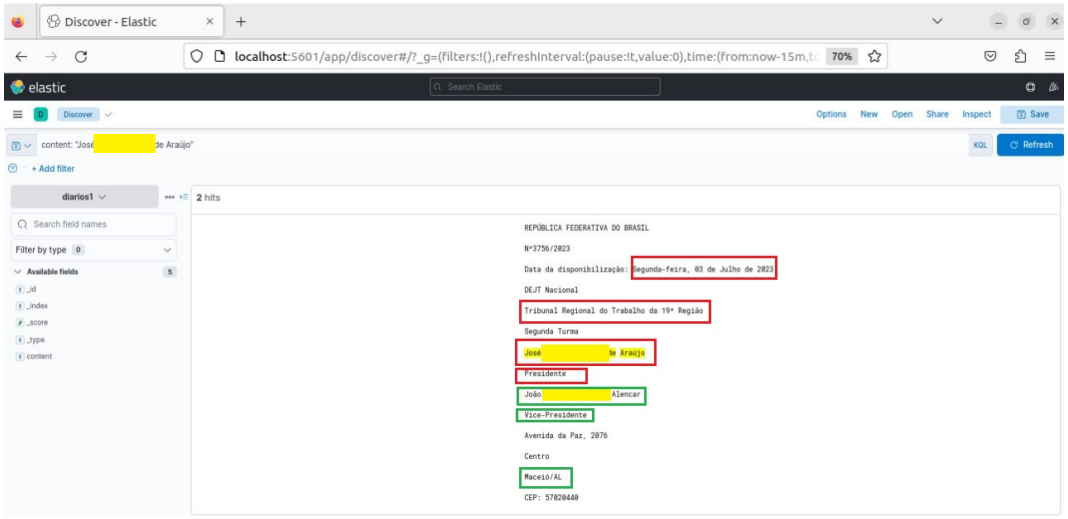

Figura 4.12: Exemplo de extração de informações. Fonte: autores

Por fim, o exemplo nos mostra com sucesso que é possível coletar informações valiosas provenientes das análises realizadas e então produzir inteligencia de acordo com a necessidade de cada possível usuário.

5 CONCLUSÃO

Evidencia-se portanto que a partir do método e da arquitetura proposta, este estudo permitiu explorar o potencial da metodologia OSINT para a pesquisa e análise de dados nos diários eletrônicos da Justiça do Trabalho. A dificuldade na localização de informações relevantes nesses documentos foi abordada, e a utilização de técnicas de OSINT emergiu como uma solução promissora para superar esse desafio.

Através da construção de um robô especifico para essa fonte de dado, foi possível realizar o *download* dos diários diretamente do site do DEJT, permitindo a obtenção dos dados necessários para nossa análise. Em seguida, realizamos o processo de conversão dos arquivos PDF para arquivos de texto (.txt) e, por meio do uso de *dashboards* intuitivos no Kibana, pode-se visualizar os dados de forma clara e facilitar a identificação das informações relevantes para os interessados.

No entanto, durante o desenvolvimento do projeto, enfrentou-se uma dificuldade significativa no processo de conversão de PDFs para arquivos de texto (.txt), devido à demora no tempo de conversão e à alta exigência de processamento. Essa limitação pode afetar a eficiência e o desempenho do método OSINT em projetos futuros que envolvam a análise de grandes volumes de dados dos diários eletrônicos da Justiça do Trabalho.

Apesar desse desafio, os resultados obtidos até o momento são promissores. A aplicação da metodologia OSINT permitiu transformar os dados brutos coletados em inteligência acionável, proporcionando uma compreensão mais aprofundada e uma identificação mais precisa dos dados relevantes para os usuários.

6 TRABALHOS FUTUROS

Neste capítulo, serão expostas algumas direções e possibilidades para trabalhos futuros relacionados ao tema abordado neste trabalho. Com base nas atividades realizadas e nas limitações encontradas durante a execução desse estudo, manifestam-se oportunidades para o aprimoramento da contribuição para o estudo das metodologias de OSINT e o uso de ferramentas para auxiliar na obtenção de Inteligencia a partir de Fontes de dados abertos.

6.0.1 Paralelização da conversão dos Diários

Durante a fase de processamento dos dados, especificamente na etapa 3 do ciclo proposto neste estudo e representada na Figura 3.1 como a conversão de dados de PDF para TXT, nos deparamos com um desafio significativo relacionado ao tempo de conversão dos documentos. Diariamente, uma média de 25 documentos é disponibilizada, variando em quantidade de páginas, podendo variar desde um documento com poucas páginas até documentos extensos com mais de 10000 páginas.

Essa variação no tamanho dos documentos resultou em tempos de conversão bastante divergentes. Alguns arquivos podiam ser convertidos em apenas um segundo, enquanto outros demandavam vários minutos, o que se tornou uma preocupação crítica para o trabalho em questão, uma vez que lidamos com a indexação desses arquivos diariamente.

Imaginemos um cenário hipotético em que temos 30 arquivos a serem convertidos, e cada um deles leva uma hora para ser processado. Teríamos então um total de 30 horas apenas na fase de conversão. Uma situação como essa inviabilizaria o trabalho e apresentaria riscos à integridade dos dados.

É fundamental abordar essa questão de maneira eficaz para garantir a fluidez e a eficiência do processo de conversão. Uma solução possível seria explorar abordagens de processamento paralelo, distribuindo a conversão entre várias unidades de processamento simultaneamente. Isso ajudaria a reduzir o tempo necessário para converter os documentos extensos e agilizaria todo o processo de indexação.

Além disso, a busca por ferramentas especializadas ou soluções de terceiros que ofereçam uma conversão mais rápida e eficiente também pode ser considerada. Essas soluções podem ser integradas ao fluxo de trabalho existente, melhorando significativamente o desempenho e mitigando o risco de atrasos e impactos negativos na integridade dos dados.

6.0.2 Capacidade de Armazenamento

Outro desafio identificado durante a realização do projeto foi a limitação de armazenamento das máquinas utilizadas. O Elasticsearch foi configurado em um ambiente local, juntamente com o *script* de obtenção de dados e conversão. Isso resultou na necessidade de armazenar os dados processados, os dados não processados e o banco de dados em uma única máquina.

Essa configuração restrita limitou a capacidade de busca e dificultou a ampliação do volume de do-

cumentos indexados. Seria mais vantajoso para o projeto ter uma capacidade de armazenamento maior, permitindo indexar um maior número de documentos. Quanto mais dados forem indexados, mais informações relevantes podem ser obtidas.

Para superar esse problema, seria recomendado considerar soluções de armazenamento escaláveis e distribuídas. Por exemplo, o uso de um *cluster* Elasticsearch distribuído, com diferentes nós dedicados a funções específicas, como armazenamento de dados brutos, armazenamento de dados processados e banco de dados, permitiria aumentar significativamente a capacidade de armazenamento e a escalabilidade do sistema. Além disso, seria importante avaliar a utilização de serviços de armazenamento em nuvem, para armazenar os dados de forma escalável e eficiente.

6.0.3 Machine learning

Durante a concepção do projeto, a ideia de utilizar aprendizado de máquina para aprimorar a implementação da metodologia já foi considerada. No entanto, ao longo da realização do trabalho, ficou evidente que a incorporação dessa tecnologia traria melhorias significativas para o projeto.

Como mencionado na dissertação de mestrado de Vanderlei Jandir Alves, na seção 2, a definição de entidades nomeadas desempenha um papel crucial no processo de busca de dados. Nesse sentido, uma proposta para trabalhos futuros é a aplicação de técnicas de aprendizado de máquina para estruturar os dados em categorias, como Data, Caderno, Número de Processo e Pessoa. Essa abordagem facilitaria a organização e estruturação dos dados.

A utilização do aprendizado de máquina permitiria que o sistema automaticamente identificasse e classificasse as informações relevantes nos documentos, o que otimizaria significativamente o processo de busca e recuperação de dados. Além disso, a capacidade de aprendizado e adaptação do modelo de *machine learning* permitiria aprimorar continuamente a precisão e a eficiência do sistema.

Dessa forma, a aplicação de técnicas de aprendizado de máquina representaria um passo promissor para avançar ainda mais a metodologia proposta. Essa abordagem aumentaria a capacidade de estruturação dos dados e melhoraria a experiência de busca, permitindo a recuperação de informações mais relevantes e precisas.

REFERÊNCIAS BIBLIOGRÁFICAS

Alles 2018 ALLES, V. J. Construção de um corpus para extrair entidades nomeadas do diário oficial da união utilizando aprendizado supervisionado. 2018.

Elastic 2023 ELASTIC. *Centralize, transform stash your data*. 2023. <https://www.elastic.co/logstash/>. (Accessed on 10/07/2023).

Elastic 2023 ELASTIC. *The heart of the free and open Elastic StackThe heart of the free and open Elastic Stack*. 2023. <https://www.elastic.co/elasticsearch/>. (Accessed on 10/07/2023).

Elastic 2023 ELASTIC. *Turn data into results, response and resolution*. 2023. <https://www.elastic.co/kibana/>. (Accessed on 10/07/2023).

ElasticPy 2023 ELASTICPY. *Elasticsearch Python Client.* 2023. <https://www.elastic.co/guide/en/elasticsearch/client/python-api/current/index.html>. (Accessed on 10/07/2023).

Furuhaug 2019 FURUHAUG, R. A. *Open Source Intelligence Methodology*. Dissertação (Mestrado) — School of Computer Science and Informatics, University College Dublin, 2019.

Jota et al. 2022 JOTA, L. M. G. et al. A informação como elemento de difusão de poder no espaço cibernético: o uso da inteligência de fontes abertas (osint) no conflito entre rússia e ucrânia. Florianópolis, SC, 2022.

Pastor-Galindo et al. 2020 PASTOR-GALINDO, J.; NESPOLI, P.; MÁRMOL, F. G.; PÉREZ, G. M. The not yet exploited goldmine of osint: Opportunities, open challenges and future trends. *IEEE Access*, IEEE, v. 8, p. 10282–10304, 2020.

Playwright 2023 PLAYWRIGHT. *Playwright enables reliable end-to-end testing for modern web apps.* 2023. <https://playwright.dev/>. (Accessed on 10/07/2023).

PyPI 2023 PYPI. *Project description - PDFMiner.* 2023. <https://pypi.org/project/pdfminer/>. (Accessed on 10/07/2023).

Python 2023 PYTHON. *os — Miscellaneous operating system interfaces.* 2023. <https://docs.python.org/3/library/os.html>. (Accessed on 10/07/2023).

Python 2023 PYTHON. *Python is powerful... and fast; plays well with others; runs everywhere; is friendly easy to learn; is Open.* 2023. <https://www.python.org/about/>. (Accessed on 10/07/2023).

Simões 2022 SIMÕES, R. A. M. *A importância dos Dados Estruturados, Não Estruturados e Semiestruturados os desafios da sua utilização nas organizações brasileiras*. Dissertação (B.S. thesis) — Brasil, 2022.

Tanabe 2023 TANABE, R. Proposta de um método para inteligência de fontes abertas: valores e princípios para uma atividade ética e profissional. 2023.

Tanabe et al. 2022 TANABE, R.; OLIVEIRA-ALBUQUERQUE, R. de; SILVA-FILHO, D. da; SILVA, D. Alves-da; COSTA-GONDIM, J.-J. Osint methods in the intelligence cycle. In: SPRINGER. *International Conference on Computer Science, Electronics and Industrial Engineering (CSEI)*. [S.l.], 2022. p. 42–54.

Tecnologia 2023 TECNOLOGIA, A. *Unix style pathname pattern expansion*. 2023. <https://docs.python.org/3/library/glob.html#module-glob>. (Accessed on 11/07/2023).

Tomás 2019 TOMÁS, H. F. D. A. *iKNOW–Sistema distribuído de intelligence em fontes abertas*. Tese (Doutorado), 2019.

Tribunais TRIBUNAIS. <https://www.cnj.jus.br/poder-judiciario/tribunais/>. (Accessed on 10/07/2023).

Vaz, Ribeiro e Matheus 2010 VAZ, J. C.; RIBEIRO, M. M.; MATHEUS, R. Dados governamentais abertos e seus impactos sobre os conceitos e práticas de transparência no brasil. *Cadernos ppg-au/ufba*, 2010.

VirtualBox 2022 VIRTUALBOX, O. V. *Chapter1.First Steps*. 2022. <https://www.virtualbox.org/manual/ch01.html#virtintro>. (Accessed on 04/13/2022).

Williams e Blum 2018 WILLIAMS, H.; BLUM, I. *Definindo inteligência de código aberto de segunda geração (OSINT) para a empresa de defesa*. [S.l.], 2018.

Williams e Blum 2018 WILLIAMS, H. J.; BLUM, I. *Defining second generation open source intelligence (OSINT) for the defense enterprise*. [S.l.]: Rand Corporation Santa Monica, 2018.