INSTITUTO FEDERAL GOIANO - CAMPUS MORRINHOS

CURSO SUPERIOR DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

João Victor Lemes de Oliveira

Aprendizagem de Ontologias a partir de texto sobre Computação Forense

MORRINHOS - GO

2024

JOÃO VICTOR LEMES DE OLIVEIRA

Aprendizagem de Ontologias a partir de texto sobre Computação Forense

Monografia apresentada ao Curso Superior de Bacharelado em Ciência da Computação do Instituto Federal Goiano – Campus Morrinhos, como requisito parcial para obtenção de título de Bacharel em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Me. Norton Coelho Guimarães

MORRINHOS - GO

2024

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/IF Goiano Campus Morrinhos

O48a Oliveira, João Victor Lemes de.
Aprendizagem de Ontologias a partir de texto sobre Computação Forense. / João Victor Lemes de Oliveira. – Morrinhos, GO: IF Goiano, 2024.
31 f. : il. color.

Orientador: Me. Norton Coelho Guimarães.

Trabalho de conclusão de curso (graduação) – Instituto Federal Goiano Campus Morrinhos, Bacharelado em Ciência da Computação, 2024.

1. Ontologia. 2. Framework (Arquivo de computador). 3. Processamento de linguagem natural (Computação) I. Guimarães, Norton Coelho. II. Instituto Federal Goiano. III. Título.

CDU 004.62

Fonte: Elaborado pela Bibliotecária-documentalista Morgana Guimarães, CRB1/2837

+

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 12/2024 - CCEPTNM-MO/CEPTNM-MO/DE-MO/CMPMHOS/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO

Aos 28 dias do mês de novembro de 2024, às 16 horas, reuniu-se, via Google Meeting, a banca examinadora composta pelos docentes: Norton Coelho Guimarães (orientador), Odilon Fernandes Neto (membro), Leila Roling Scariot da Silva (membro), para examinar o Trabalho de Curso intitulado <Aprendizagem de Ontologias a partir de texto sobre Computação Forense= do estudante **João Victor Lemes de Oliveira** Matrícula nº 2018104201940232 do Curso de Bacharelado em Ciência da Computação do IF Goiano 3 Campus Morrinhos. A palavra foi concedida ao estudante para a apresentação oral do TC, houve arguição do candidato pelos membros da banca examinadora. Após tal etapa, a banca examinadora decidiu pela **APROVAÇÃO** do estudante. Ao ûnal da sessão pública de defesa foi lavrada a presente ata que segue assinada pelos membros da Banca Examinadora.

(Assinado Eletronicamente)

Norton Coelho Guimarães

Orientador

(Assinado Eletronicamente)

Odilon Fernandes Neto

Membro

(Assinado Eletronicamente)

Leila Roling Scariot da Silva

Membro

Observação:

( ) O(a) estudante não compareceu à defesa do TC.

Documento assinado eletronicamente por:

- **Norton Coelho Guimaraes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 28/11/2024 16:48:27.
- **Odilon Fernandes Neto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 28/11/2024 16:54:48.
- **Leila Roling Scariot da Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 28/11/2024 16:58:23.

Este documento foi emitido pelo SUAP em 27/11/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 655626
Código de Autenticação: a3f4e86e5d

INSTITUTO FEDERAL GOIANO
Campus Morrinhos
Rodovia BR-153, Km 633, Zona Rural, SN, Zona Rural, MORRINHOS / GO, CEP 75650-000
(64) 3413-7900

INSTITUTO FEDERAL
Goiano

Repositório Institucional do IF Goiano - RIIF Goiano
Sistema Integrado de Bibliotecas

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor:
João Victor Lemes de Oliveira

Matrícula:
2018104201940232

Título do trabalho:
Aprendizagem de Ontologias a partir de texto sobre Computação Forense

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano:30/12/ 2024

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Morrinhos — Local
30 /12 /2024 — Data

Documento assinado digitalmente
gov.br
JOAO VICTOR LEMES DE OLIVEIRA
Data: 03/12/2024 09:25:29-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

Documento assinado digitalmente
gov.br
NORTON COELHO GUIMARAES
Data: 03/12/2024 09:18:00-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

DEDICATÓRIA

Dedico este a todos aqueles que talvez estão pensando em desistir, que nem ao menos tentam, porque eu entendo o cansaço mental que tudo isso promove, o nervosismo a ansiedade e com isso acabamos deixando sempre para depois e vivendo as nossas vidas como podemos, continue, as coisas vão melhorar quanto mais você tentar.

AGRADECIMENTOS

Agradeço a este a minha mãe, meu irmão e aos meus amigos que sempre me incentivaram e não desistiram de tentar me incentivar e continuar persistindo, agradeço a instituição que faço parte por ser um grande marco na minha carreira, e ao meu orientador pela paciência e pelos auxílios.

RESUMO

A etimologia da palavra ontologia tem raízes no dialeto grego, carregando consigo o conceito de um discurso lógico. Embora essa palavra possua múltiplas definições sob diferentes perspectivas, neste artigo, nosso foco recai exclusivamente sobre a ontologia no contexto da computação. Segundo o dicionário, a ontologia aplicada à computação é uma técnica de organização de informações, representando conhecimento com o objetivo de compartilhar e reutilizar essas informações [ONTOLOGIA, 2020]

Em termos mais específicos, uma ontologia computacional é definida por classes, objetos, referências e relacionamentos entre eles. Em essência, ela serve como a representação formal de um sistema computacional. Alguns exemplos práticos de ontologias incluem a base de uma biblioteca digital integrada, o aprendizado de inteligência artificial e o embasamento de avanços tecnológicos, resultando em máquinas mais habilidosas [GUIMARÃES, 2015].

Ao adentrarmos o campo da computação forense, percebemos que ela está intrinsecamente ligada aos avanços tecnológicos do cotidiano. Essa disciplina é altamente precisa na identificação, coleta, exame e análise de dados, facilitando a extração de evidências criminais de computadores e mídias de armazenamento .[FILHO, 2016]

Por sua vez, conseguimos provar o suficiente que em determinado conjunto de palavras como artigos, por exemplo, podemos obter mais de 1000 palavras derivadas de um tema específico utilizando as ontologias, mesmo sendo um processo demorado até a organização de seus resultados.

Palavras-chave: Ontologias de domínio, Framework, Computação Forense, RDF, OWL, W3C, Processamento de Linguagem Natural (NLP)

ABSTRACT

The etymology of the word ontology has roots in the Greek dialect, carrying with it the concept of logical discourse. Although this word has multiple definitions under different perspectives, in this article, our focus is exclusively on ontology in the context of computing. According to the dictionary, ontology applied to computing is a technique for organizing information, representing knowledge with the goal of sharing and reusing that information [ONTOLOGIA, 2020].

In more specific terms, a computational ontology is defined by classes, objects, references, and relationships between them. Essentially, it serves as the formal representation of a computational system. Some practical examples of ontologies include the foundation of an integrated digital library, artificial intelligence learning, and the basis for technological advancements, resulting in more skilled machines [GUIMARAES,2015].

As we delve into the field of forensic computing, we realize that it is intrinsically linked to the technological advances of everyday life. This discipline is highly precise in the identification, collection, examination, and analysis of data, facilitating the extraction of criminal evidence from computers and storage media.[FILHO,2016]

In turn, we were able to prove enough that in a given set of words such as articles, for example, we can obtain more than 1000 words derived from a specific topic using ontologies, even though it is a time-consuming process to organize the results.

Keywords: Domain Ontologies, Framework, Computer Forensics, RDF, OWL, W3C, Natural Language Processing (NLP)

SUMÁRIO

INTRODUÇÃO **1**
1. OBJETIVOS **3**
1.1 OBJETIVO GERAL **3**
1.2 OBJETIVOS ESPECÍFICOS **3**
1.2.1 IMPORTÂNCIA DA COMPUTAÇÃO FORENSE 3
1.2.2 VERSATILIDADE DAS ONTOLOGIAS 3
2 BIBLIOGRAFIA CORRELATA **4**
2.1 FERRAMENTAS AUXILIARES **4**
2.2 LINGUAGENS **4**
2.3 BANCO DE DADOS **5**
3 REFERENCIAL TEÓRICO **6**
3.1 CONCEITOS INICIAIS NECESSÁRIOS **6**
3.2 ONTOLOGIA **6**
3.4 APRENDIZAGEM DE ONTOLOGIAS **6**
3.5 PROCESSAMENTO DE LINGUAGEM NATURAL **7**
3.6 RDF **9**
3.7 OWL **9**
3.8 TÉCNICAS DE ESTATÍSTICA **9**
3.9 COMPUTAÇÃO FORENSE **10**
4 METODOLOGIA **11**
5 RESULTADOS E DISCUSSÕES **12**
5.1 CÁLCULOS SOBRE OS RESULTADOS 14
5.2 REPRESENTAÇÃO DOS RESULTADOS EM HISTOGRAMA 15
6 CONCLUSÃO **15**
6.1 APRENDIZAGEM 16
6.2 DIFICULDADES 16
6.3 TRABALHOS FUTUROS 16
REFERÊNCIAS BIBLIOGRÁFICAS **17**

LISTA DE ILUSTRAÇÕES

Figura 01 - Etapas de aprendizagem de Ontologias 20
Figura 02 - Etapas processamento de linguagem natural utilizando AZURE 21
Figura 03 - Etapas do processo de computação forense 23
Figura 04 - Grafo gerado 26
Figura 05 - Grafo gerado escala reduzida 27
Figura 06 - Distribuição do tamanho das palavras 28

LISTA DE ABREVIATURAS E SIGLAS

RDF	*Resource Description Framework* (Estrutura de Descrição de Recursos)
OWL	*Web Ontology Language* (Linguagem de Ontologia da Web)
IA ou IAs	Inteligência Artificial
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
OCR	*Optical Character Recognition* (Reconhecimento Óptico de Caracteres)
W3C	*World Wide Web Consortium* (Consórcio da World Wide Web)
PDF	*Portable Document Format* (Formato de Documento Portátil)
TXT	Extensão de arquivo de texto
R	Linguagem de programação estatística
C#	Linguagem de programação

INTRODUÇÃO

Ontologia de domínio é o estudo da estrutura e organização do conhecimento em um determinado domínio ou área específica. Ela é frequentemente utilizada em inteligência artificial e mineração de dados para ajudar a compreender e organizar informações complexas. Uma ontologia de domínio é uma representação formal de conceitos e relações em um determinado domínio. Ela é composta por uma coleção de conceitos e relações hierárquicas entre eles, que são usadas para modelar o conhecimento e fornecer uma estrutura para a representação e inferência de informações [GRUBER,1993].

Existem várias abordagens diferentes para o desenvolvimento de ontologias de domínio, incluindo ontologias estruturadas manualmente, ontologias geradas automaticamente a partir de texto livre e ontologias semiautomáticas que combinam técnicas automáticas e manuais.

Uma das principais vantagens da ontologia de domínio é a capacidade de fornecer uma representação consistente e precisa do conhecimento em um determinado domínio. Isso pode ser útil para aplicações como recuperação de informações, inferência automática e geração de perguntas e respostas. Além disso, uma ontologia de domínio pode ser compartilhada e utilizada em várias aplicações e sistemas, o que ajuda a garantir a consistência do conhecimento em todo o sistema [GUIMARÃES, 2015].

Desenvolver ontologias de domínio também traz desafios. Garantir precisão e completude é um dos principais obstáculos, especialmente em domínios complexos e dinâmicos. A criação e manutenção de ontologias pode ser complexa, especialmente quando lidamos com áreas em constante mudança.

É possível ter ontologias de domínio para o estudo da computação forense. A computação forense é uma área que envolve a coleta, análise e apresentação de provas digitais em um ambiente legal. Ela é amplamente utilizada em investigações criminais e civis para ajudar a recuperar e analisar dados digitais.

Quando é necessário realizar as análises forenses de algum documento e ou arquivo é necessário a extração de textos, dentro da computação forense existe um termo bastante forte e mencionado que é a mineração de textos sendo um processo altamente necessário para transformar textos não estruturados em um formato mais

adequado para sua leitura, tentando sempre identificar padrões que são significativos a fim de conseguir desvendar [NASSIF, 2011].

Uma vez estabelecida a ontologia de domínio para a computação forense, torna-se mais fácil extrair textos relevantes das evidências digitais e realizar a mineração de textos para identificar padrões, tendências ou informações importantes para o caso em questão. Com a estruturação adequada dos dados através da ontologia, é possível automatizar o processo de análise textual, facilitando o trabalho dos profissionais envolvidos na investigação [PALMEIRA, 2007].

Além disso, as ontologias de domínio também podem ser utilizadas para representar as normas e regulamentações legais relacionadas à coleta, armazenamento e apresentação de provas digitais, permitindo que os investigadores possam seguir as diretrizes legais e evitar problemas legais.

No entanto, é importante notar que o desenvolvimento de ontologias de domínio para computação forense pode ser desafiador devido à natureza complexa e dinâmica da computação forense e das regulamentações legais. Além disso, a precisão e a completude de uma ontologia de domínio deve ser cuidadosamente verificada para garantir a confiabilidade das informações e evitar problemas legais.

Ao final será possível unir todo o resultado gerado a fim de conseguir uma demonstração mais clara e concisa do que foi gerado e assim pode-se provar que as ontologias de domínio podem ser eficientes o necessário para construir aquilo que se deseja, unindo diversos temas e por mais que desconexo seja, sempre há uma forma de unir com suas derivantes.

1. OBJETIVOS

1.1 OBJETIVO GERAL

Busca-se avaliar como as ontologias podem ser utilizadas para representar conceitos, relações e contextos específicos no âmbito da computação forense, aprimorando a eficiência e a precisão das análises.

1.2 OBJETIVOS ESPECÍFICOS

1.2.1 IMPORTÂNCIA DA COMPUTAÇÃO FORENSE

A computação forense desempenha um papel crucial no meio criminalístico investigativo, permitindo a coleta e análise de evidências digitais de maneira admissível em tribunais. Essa disciplina combina ciência da computação com forense legal, possibilitando a identificação de práticas ilícitas e crimes cibernéticos. Pretende-se explorar como a modelagem ontológica pode contribuir para a organização, interpretação e extração de informações relevantes em investigações criminais.

1.2.2 VERSATILIDADE DAS ONTOLOGIAS

A flexibilidade das ontologias é notável, pois sua aplicação não se limita à área de computação. Elas podem ser adaptadas e utilizadas em diversos domínios, transcendendo as fronteiras da tecnologia. Essa versatilidade permite que ontologias sejam empregadas em áreas variadas como por exemplo na computação forense.

2 BIBLIOGRAFIA CORRELATA

2.1 FERRAMENTAS AUXILIARES

O Protégé, um framework de código aberto, permite a edição de ontologias. Ele é amplamente empregado na construção de sistemas inteligentes baseados em conhecimento. Além disso, o *Protégé* conta com o suporte de uma comunidade robusta de usuários acadêmicos, governamentais e corporativos, que o aplicam em áreas tão diversas quanto biomedicina, comércio eletrônico e modelagem organizacional. [GUIMARÃES,2015], oferece suporte aos padrões OWL 2 e RDF do W3C, além de ser uma plataforma flexível para prototipagem rápida e desenvolvimento de aplicativos

A ferramenta Sabença é voltada para o aprendizado semi-automatizado de ontologias a partir de textos, extrair conhecimento estruturado e por meio desta representar em forma de ontologias com o objetivo de facilitar a organização e a interpretação de informações em diversos domínios [GUIMARÃES,2015].

2.2 LINGUAGENS

Para conseguir fazer a extração de textos é necessário alguma linguagem para conseguir tal feito, tais como RDF e OWL.

O RDF (Resource Description Framework) é um padrão da *World Wide Web* para representar informações sobre recursos, oferecendo uma estrutura para expressar relacionamentos e descrever recursos de maneira legível por máquina [LASSILA & SWICK, 1999].

O OWL (Web Ontology Language) é uma linguagem destinada à definição e instalação de ontologias na Web, permitindo a inclusão de descrições de classes, suas propriedades e relacionamentos [GUIMARÃES, 2015].

O Prolog é uma linguagem de programação lógica comumente utilizada para trabalhar com ontologias, possibilitando a introspecção da base de dados em tempo de execução e o acesso às estruturas de dados, tabelas e ligação de variáveis tanto em nível de programação quanto de meta-programação [STERLING, 1986].

O Ontoprolog, por sua vez, é uma linguagem formal textual utilizada para representar e deduzir problemas reais com base em ontologias de domínio e de fundamentação [ARAUJO & LIMA-MARQUES, 2017].

Para conseguir organizar o resultado foi utilizado a linguagem de programação da Microsoft o C# onde foi possível automatizar o processo de leitura e análise das palavras obtidas.

E por final foi utilizado uma linguagem de estatística chamada R com foco maior em programação funcional, é uma linguagem completa e bastante complexa de se mexer, realiza cálculos e também desenha gráficos sobre os dados estatísticos.

2.3 BANCO DE DADOS

Se a geração de textos segue uma estrutura específica, torna-se essencial armazenar os dados resultantes. Nesse contexto, é crucial utilizar um banco de dados adequado para essa finalidade.

Os Bancos de dados orientados a grafos, como Neo4J, são ideais para representar e consultar eficientemente as relações complexas presentes em ontologias.

Os Bancos de dados NoSQL oferecem flexibilidade e escalabilidade, especialmente úteis para grandes volumes de dados [ALEXANDRE, 2014].

Os Bancos de dados de triplas RDF, projetados especificamente para armazenar e consultar dados no formato de triplas RDF, comumente utilizado na representação de ontologias [GUIMARÃES, 2015].

Os Bancos de dados relacionais, embora não sejam a primeira opção devido à sua estrutura tabular, podem ser adequados em casos de ontologias simples ou quando já há uma infraestrutura relacional estabelecida.

3 REFERENCIAL TEÓRICO

3.1 CONCEITOS INICIAIS NECESSÁRIOS

As ontologias de domínio, que encapsulam o conhecimento de um campo específico, são estruturas conceituais essenciais. No caso da computação forense, elas desempenham um papel fundamental. Estas representações são vitais para a gestão do conhecimento, pois proporcionam um vocabulário unificado para todos os participantes envolvidos [FILHO, 2016].

A computação forense envolve a aplicação de técnicas de investigação e análise para coletar e preservar evidências de dispositivos específicos, com o objetivo de torná-las admissíveis em contextos judiciais. Assim, a ontologia desempenha um papel crucial na definição e organização dos conceitos e relações relevantes ao domínio da computação forense, facilitando a comunicação eficaz e a compreensão entre os profissionais envolvidos na área [COSTA, 2014].

3.2 ONTOLOGIA

Suas discussões só foram surgir na computação com a necessidade de criar algumas representações principiológicas no desenvolvimento de IAs, o que foi criado fóruns e tópicos de discussões entre a comunidade computacional a fim de debater esses assuntos e em algumas áreas da modelagem de dados tem sido um ótimo exemplo de aplicação de ontologias. Os modelos iniciais de modelagem de dados fornecem um conjunto confuso de modelagem abstrata que não eram estruturados e não poderiam ser usados para tal [GUIZZARDI, 2008].

3.4 APRENDIZAGEM DE ONTOLOGIAS

Para iniciar, é imperativo dispor de um ambiente semi-automático ou automático para o desenvolvimento de ontologias. No caso de uma abordagem semi-automática, é necessário reunir arquivos, documentos e artigos pertinentes ao tema selecionado. Já na abordagem automática, requer-se um mecanismo de busca altamente eficiente para extrair informações tangíveis para o aprendizado. Posteriormente, é essencial adotar uma linguagem de ontologia facilmente

recuperável por dispositivos computacionais. Este processo pode ser demorado, dependendo da metodologia escolhida para a obtenção do material relevante, e pode envolver tanto a criação de novas ontologias quanto o aprimoramento das já existentes [GUIMARÃES, 2015], na Figura 01 visualizamos este processo.

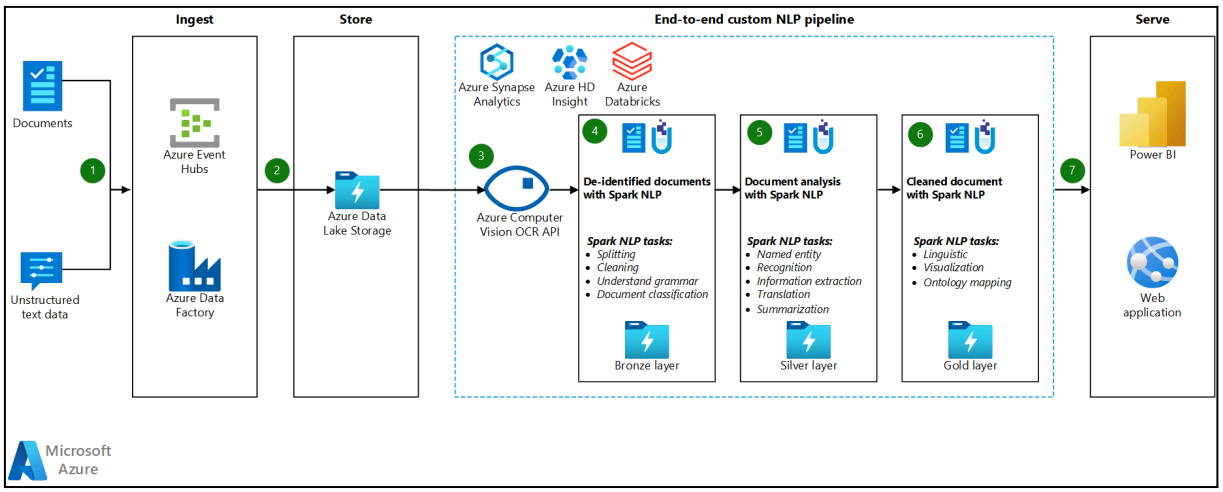
Figura 01 - Etapas de aprendizagem de Ontologias

Fonte: Vinente, Julie & Barbosa, Ademilson & Ponte, Márcio & Pixiúba, Carla & Lima, Celson. (2018). Modelo de atualização de bases de conhecimento: Um Estudo de Caso ONTO-AmazonTimber.

3.5 PROCESSAMENTO DE LINGUAGEM NATURAL

Para extrair texto do tema escolhido, é essencial empregar um processo de linguagem natural. Esse processo visa facilitar a compreensão entre humanos e máquinas, com o propósito de ler, decifrar e compreender o conteúdo. Enquanto os computadores lidam com zeros e uns, o processamento de linguagem natural permite que os humanos entendam essas informações de forma mais acessível. No contexto da ontologia de domínio, esse processamento envolve a extração e organização de textos não estruturados [FALCÃO, 2021], pode-se ilustrar um pouco de seu processo (Figura 2).

Figura 02 - Etapas processamento de linguagem natural utilizando AZURE



Fonte: (Processamento de linguagem natural personalizado em larga escala (NLP)).

Em resumo, primeiro é necessário obter material que não está estruturado, os Hubs de Eventos do Azure e o Azure Data Factory coletam dados de texto não (é necessário alguma ferramenta para coleta destes textos) estruturados e os armazenam como arquivos no Azure Data Lake Storage, seguindo uma estrutura de diretório alinhada aos requisitos de negócios. A API da Pesquisa Visual Computacional do Azure, utilizando OCR, consome esses dados e os registra na camada bronze de uma arquitetura de lakehouse. Nesta camada, o Spark NLP realiza o pré-processamento do texto, incluindo tarefas como divisão, correção ortográfica e análise gramatical. A classificação de documentos também é feita nesta etapa, antes de passar os dados para a camada prata.

Na camada prata, o Spark NLP executa análises mais complexas, como reconhecimento de entidade nomeada e sumarização. Os resultados podem ser gravados na camada ouro em certas arquiteturas. Finalmente, na camada ouro, análises linguísticas avançadas são realizadas para fornecer insights sobre dependências de linguagem e facilitar a visualização dos rótulos. Os usuários podem consultar esses dados textuais como um quadro de dados e visualizar os resultados através do Power BI ou aplicativos web [MICROSOFT, 2024].

O importante deste processo é entender de como funciona, mesmo que seja utilizado outras ferramentas como descrito na imagem , é importante coletar os

dados, armazenar estes dados e por fim interpretá-los utilizando alguma linguagem que consiga organizar um texto não estruturado para um texto totalmente estruturado [MICROSOFT, 2024].

3.6 RDF

Trata-se de um padrão estabelecido pelo World Wide Web Consortium (W3C) que permite a interpretação de recursos na web e suas inter-relações. Esse padrão utiliza a estrutura de triplos, composta por um sujeito, um predicado e um objeto, para representar de forma abrangente a totalidade da web. Ele encontra aplicação significativa no processamento de linguagem natural, na construção de grafos de conhecimento e é fundamental para o desenvolvimento de sistemas de inteligência artificial que demandam uma compreensão mais ampla e hierarquias complexas [GUIZZARDI, 2008].

3.7 OWL

Esta é uma linguagem especializada para ontologias que oferece vocabulários adicionais e uma estrutura mais elaborada para representar conhecimento complexo sobre recursos e materiais adquiridos. Ela possibilita a definição de classes, propriedades e suas relações, ampliando as possibilidades de inferência e lógica semântica necessárias [GUIMARÃES, 2015].

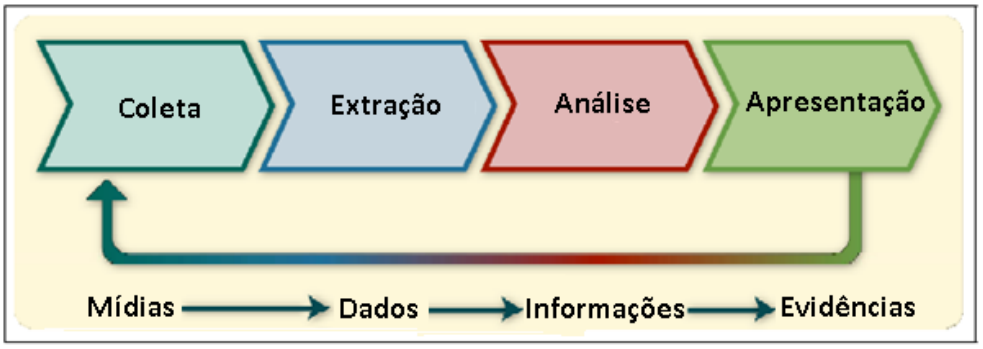
3.8 TÉCNICAS DE ESTATÍSTICA

Para interpretar os dados gerados como resultados, é essencial incorporar uma variedade de técnicas da estatística, a fim de compreender a extensão dos materiais adquiridos ao longo da pesquisa. Para obter um valor aproximado da pesquisa é necessário incluir meios matemáticos a fim de conseguir demonstrar os resultados gerados e se foram realmente úteis para isso foi necessário utilizar uma linguagem de programação baseada em estatística, chamada “R”, com ela é capaz de calcular a média, mediana, moda e desvio padrão dos resultados e ainda é capaz de fazer uma geração visual do resultado [WICKHAM, 2019].

3.9 COMPUTAÇÃO FORENSE

É um ramo utilizado em análises predominantemente no contexto criminalístico. Esse campo abrange uma variedade de análises, incluindo sensoriais e digitais, entre outras fontes, para determinar pistas através de dados que requerem uma análise aprofundada de dispositivos computacionais. Para lidar com esse processo, é crucial adquirir o material que contenha pistas definidas como evidências criminais e determinar sua origem. Portanto, as ontologias foram aplicadas para gerar textos específicos para a computação forense, estabelecendo uma ponte entre essas duas áreas distintas [BEGOSSO, 2010].

Figura 03 - Etapas do processo de computação forense



Fonte: (Adaptado de Kent et al, 2006)

Em resumo, de início é feito um isolamento da área e é feito a coleta de evidências garantindo ao máximo sua integridade, etiqueta-se as evidências, depois é feito a extração desse material. Dependendo do tipo de material que é manuseado se é extraído, filtrado ou apenas identificado, em seguida é feito a análise minuciosa, reconstruindo todo o ocorrido e nesta etapa é fundamental documentar tudo o que se encontra e por fim é apresentado essa análise para dar prosseguimento com a investigação.

4 METODOLOGIA

Durante o processo de coleta de material para a geração de texto, foi-se pesquisado por informações em revistas educacionais, artigos, documentos, sites e blogs relevantes para complementar a fundamentação teórica. Utilizando recursos como o Portal de Periódicos da CAPES e o periódico científico "Em Questão", publicado pelo Programa de Pós-graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul, é uma fonte eletrônica de acesso aberto, com contribuições teóricas e empíricas de autores nacionais e internacionais.

A pesquisa abrangeu também documentos relacionados à computação forense e ontologias de domínio, incluindo material empírico para a produção de textos sobre computação forense. Ao todo, foram identificados cerca de 8 artigos de grande relevância para a obtenção de informações não estruturadas. Além disso, foi introduzido conceitos básicos de estatística para evidenciar a extensão e a estrutura do material obtido e também adicionando artigos visuais para melhor clareza.

Posteriormente, utilizamos o “*Framework* Sabença" [GUIMARÃES, 2015] para construir a Ontologia de domínio sobre Computação Forense a partir dos textos selecionados..

5 RESULTADOS E DISCUSSÕES

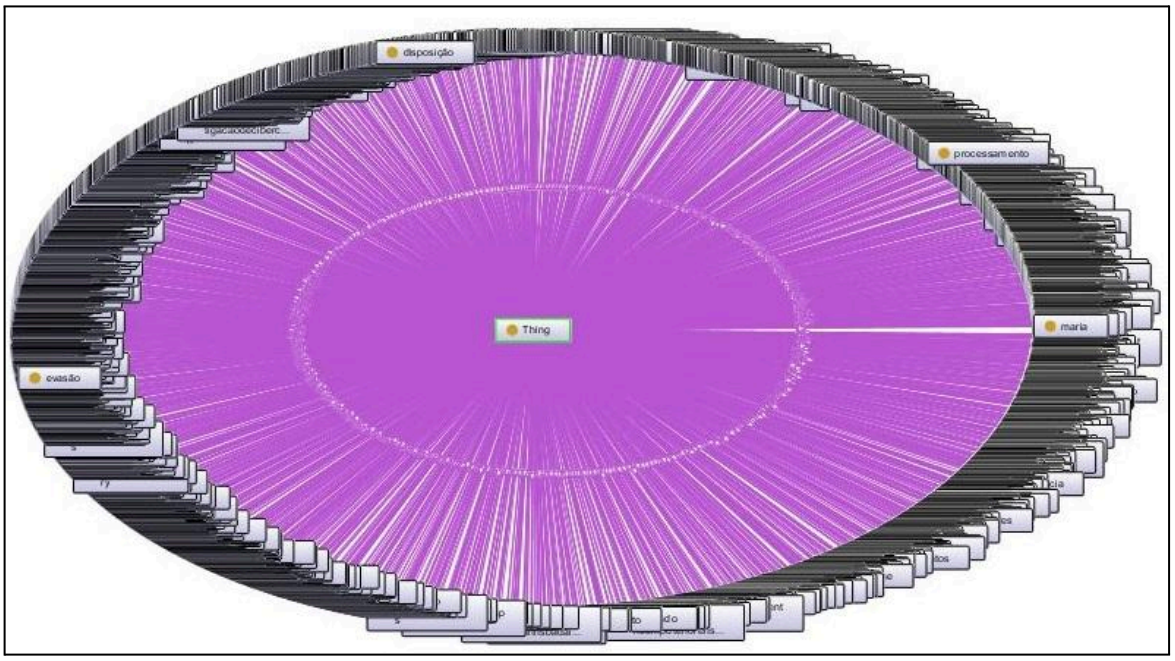
O alfabeto foi obtido através do framework Sabença, com ele foi possível obter muitos resultados e com uma análise em apenas 8 artigos em PDF [GUIMARÃES, 2015]. Envolvendo apenas o tema de computação forense é possível visualizar diversas palavras de nosso cotidiano.

Como resultados temos:

1. Com todo o material foi possível obter 2122 palavras relacionadas com a computação forense entre elas palavras na língua inglesa e na língua portuguesa.
2. Foram gerados 22 arquivos após a obtenção do alfabeto, tais quais os arquivos PDFs em formato TXT e alguns desses arquivos foram etiquetados, também é gerado as suas palavras de parada (*Stop Words*).

Na Figura 4, pode-se ver um grafo gerado pela Ferramenta Protegè e como ficou a extração de palavras tendo como base a computação forense, como foram bastante palavras é difícil de visualizar assim uma por uma, mas tendo como base computação forense, isso é fascinante de pensar, porque foi-se de um tema para um variado exemplar de palavras, tendo diversas ramificações e os mesmo estão unidos por um simples “pai”, algumas fazem mais sentido que outras estarem ali, mas todas fazem parte do mesmo círculo de geração de palavras.

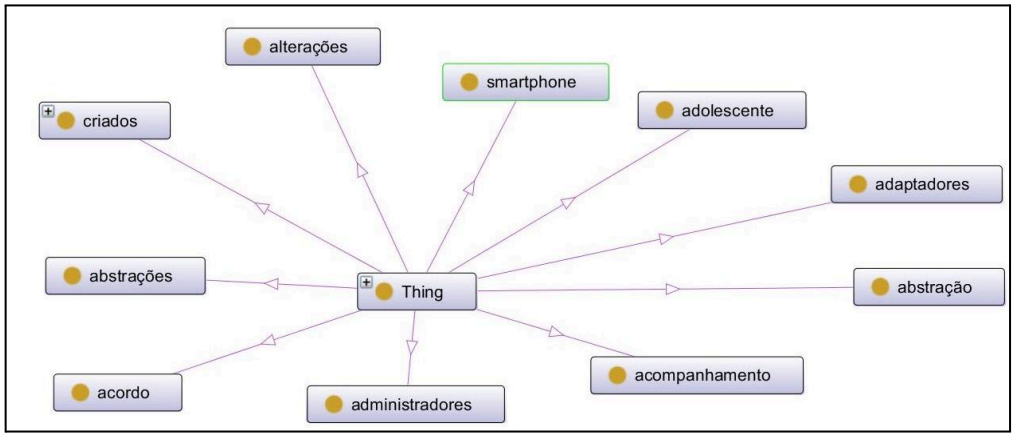
Figura 04 - Grafo gerado



Fonte: elaborado pelo autor (2024)

Na Figura 5 apresentamos uma pequena fração do grafo, assim é visto de uma maneira mais clara de como foram gerados os resultados, a palavra “thing” sendo o tema central do todo e criado várias ramificações, também é possível visualizar que a palavra “criados” possui alguma ramificação, e é notável que cada um deles possa conter mais ramificações, sem limite máximo abrindo as portas para infinitas possibilidades, depende muito do material conquistado e dos parâmetros que se utiliza para geração de texto.

Figura 05 - Grafo gerado escala reduzida



Fonte: elaborado pelo autor (2024)

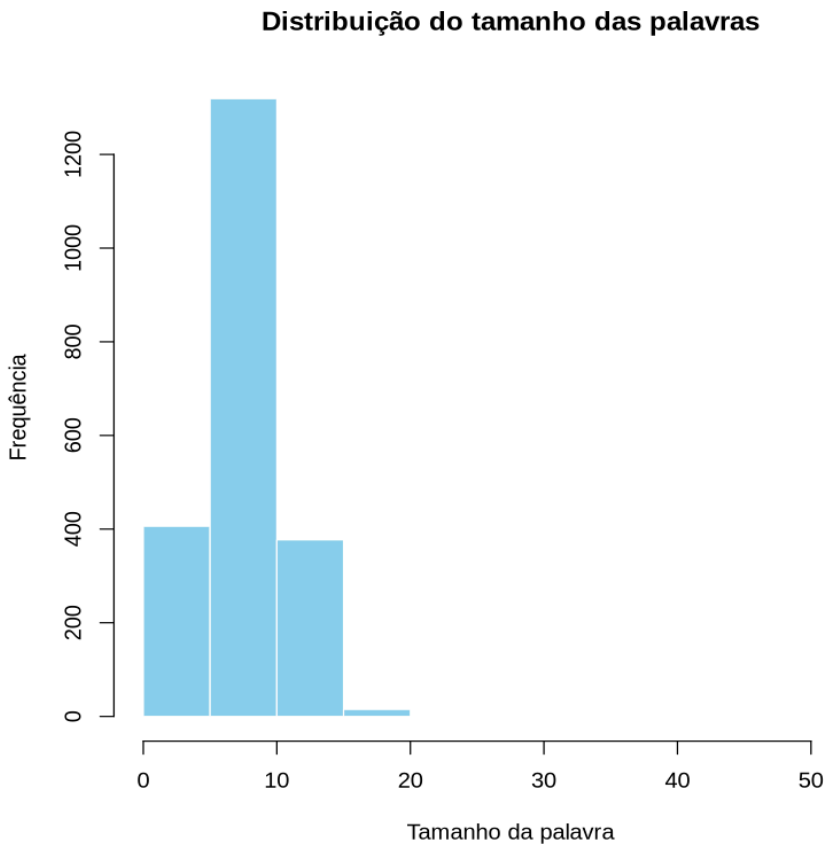
5.1 CÁLCULOS SOBRE OS RESULTADOS

Em média, se houver um cenário que cada artigo contém a mesma quantidade de páginas é possível obter 265 palavras por artigo com a utilização do *framework Sabença,* com isso podemos gerar um grafo com base nos dados brutos retirados, o grafo é uma representação visual de todo o contexto, incluindo palavras derivadas do tema, e palavras que também possuam derivadas de outras.

Com a obtenção de 2122 palavras foi possível visualizar que nenhuma delas é encontrada repetição, dando ainda mais ênfase na geração dessas palavras, porque assim sabe-se que são palavras cada vez distintas, algumas com seus plurais, algumas em suas versões na língua inglesa.

5.2 REPRESENTAÇÃO DOS RESULTADOS EM HISTOGRAMA

Figura 06 - Distribuição do tamanho das palavras



Fonte: elaborado pelo autor (2024)

Este histograma mostra a distribuição de frequência dos comprimentos de palavras na lista. O eixo x representa o comprimento da palavra, e o eixo y mostra a frequência de palavras com esse comprimento.

O comprimento médio das palavras é de cerca de 8 caracteres, com uma mediana de 8 caracteres também. Sugerindo que a distribuição é relativamente simétrica. E o comprimento de palavra mais comum é de 7 caracteres, o que é próximo da média e da mediana. E o seu desvio padrão é cerca de 3 caracteres indica uma dispersão moderada nos comprimentos de palavras em torno da média.

No histograma, podemos ver que a maioria das palavras na lista têm comprimentos entre 5 e 12 caracteres, com um pico em torno de 7 a 8 caracteres.

6 CONCLUSÃO

6.1 APRENDIZAGEM

Desde o processo de obtenção do material para conseguir realizar a monografia, foi-se descoberto que as ontologias podem facilitar a aprendizagem de qualquer tema podendo ser relacionado e também foi possível obter conhecimento sobre a computação forense e deixar um interesse sobre essa área que tanto é importante e os cálculos estatísticos foram de grande ajuda para a demonstração de maneira mais exata de todo o resultados obtido

6.2 DIFICULDADES

Em relação a dificuldades, pode se dizer que conseguir material empírico o suficiente para conseguir entender ou manipular as informações não foi uma tarefa muito fácil e por ser um trabalho bastante pesado em termos técnicos e grande fonte de arquivos encontrados na linguagem inglesa. Em principal podemos colocar a maior dificuldade como organizar os dados e também conseguir provas o suficiente para definir sua eficácia

6.3 TRABALHOS FUTUROS

Pode se realizar a automatização do processo de obtenção de material não estruturado como artigos, documentos e ou informações que sejam cabíveis e úteis para utilização de ontologias em um determinado tema, pensando em termos de inteligência artificial seria ideal que fosse possível inserir algum *prompt* para obter uma gama maior de documentos para que sejam usados para gerar uma ontologia.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEXANDRE, Joel. **NoSQL como suporte à análise de dados não-normalizados e de grande volume** [Em linha]. In InforAbERTA 2014. Jornadas de Informática, 4, Porto, 2014 - "A tecnologia digital ao serviço do ensino da informática em regime de e-learning". Editado por Jorge Morais; Leonel Morgado; Vitor Rossio. Porto : Universidade Aberta, 2014. p. 1-28.

ALMEIDA, B. M. **Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia**. Belo Horizonte: UFMG, 2014.

ARAUJO, Lauro César; LIMA-MARQUES, Mamede. **Ontoprolog: uma linguagem para especificação de discursos sobre ontologias**. Perspectivas em Ciência da Informação, Belo Horizonte, v. 22, n. 2, p. 188-210, abr./jun. 2017. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362017000200188&lng=en&nrm=iso>. Acesso em: 25 jan. 2018. doi: http://dx.doi.org/10.1590/1981-5344/2532.

BEGOSSO, H. R. **Computação forense. Assis: Fundação Educacional do Município de Assis** - FEMA, 2010.

COSTA, D. P. R.; GARCIA, R. **Princípios e análises da computação forense. In: ETIC 2014 - Encontro de iniciação científica**. 2014. ISSN 21-76-8496.

ELEUTÉRIO, P. M. D. S.; MACHADO, M. P. **Desvendando a Computação forense**. São Paulo: Novatec Editora, 2010. ISBN 978-85-7522-260-7.

FALCÃO, L. C. de J.; LOPES, B.; SOUZA, R. R. **Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI**. Em Questão, Porto Alegre, v. 28, n. 1, p. 13–34, 2021. DOI: 10.19132/1808-5245281.13-34. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/111323. Acesso em: 9 abr. 2024.

FILHO, L. B. V. E. **Ontologia aplicada nos processos de computação forense.** Brasília: Universidade de Brasília – Departamento de engenharia elétrica – PPGENE.DM – 634/16, 2016.

GRUBER, T. R. **A translation approach to portable ontology specifications**. Knowledge Acquisition, 5(2), 199-220, 1993.

GUIMARÃES, C. Norton., SABENÇA - **Um arcabouço computacional baseado na aprendizagem de ontologias a partir de textos**. Dissertação (Mestrado) – Mestrado em Ciência da Computação, INF, UFG, Goiânia, 2015.

GUIZZARDI, G.; HALPIN, T. **Ontological foundations for conceptual modelling**. Applied Ontology, v. 3, n. 1-2, p. 1–12, 2008.

LASSILA, O.; SWICK, R. **Resource Description Framework (RDF): model and syntax specification**, W3C Recommendation, 1999. Disponível em: <http://www.w3.org/TR/REC-rdf-syntax>.

MICROSOFT. **Processamento De Linguagem Natural Personalizado Em Grande Escala** - Azure Architecture Center. 4 Jan. 2024. **learn.microsoft.com**. Available at: https://learn.microsoft.com/pt-br/azure/architecture/ai-ml/idea/large-scale-custom-natural-language-processing. Accessed on: 17 Jun. 2024.

NASSIF, C. D. F. L. **Técnicas de agrupamento de textos aplicadas à computação forense**. Brasília: PPGENE.DM - 094 A/11, 2011.

ONTOLOGIA. In: DICIO, **Dicionário Online de Português**. Porto: 7Graus, 2020. Disponível em: <https://dicionario.priberam.org/ontologia>. Acesso em: 21/07/2021.

PALMEIRA, E.; FREITAS, F. **Ontologias detalhadas e classificação de texto: uma união promissora**. Maceió: Centro Federal de Educação Tecnológica de Alagoas - UNED/Pln, 2007.

Sterling, Leon; Shapiro, Ehud (1986). **The Art of Prolog. Advanced Programming Techniques**. Cambridge: The MIT Press. 437 páginas. ISBN 0-262-19250-0.

WICKHAM, Hadley. **R para Data Science**. Rio de Janeiro: Alta Books, 2019.