Instituto Federal de Brasília
Campus Brasília
Tecnologia em Sistemas para Internet

Ana Luísa Caixeta

ANÁLISE DE TEMAS DE TRABALHOS DE CONCLUSÃO DE CURSO NA REDE FEDERAL: UMA ABORDAGEM BASEADA EM WEB SCRAPING E VISUALIZAÇÃO DE DADOS

Brasília - DF
2025

Ana Luísa Caixeta

ANÁLISE DE TEMAS DE TRABALHOS DE CONCLUSÃO DE CURSO NA REDE FEDERAL: UMA ABORDAGEM BASEADA EM WEB SCRAPING E VISUALIZAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de Tecnologia em Sistemas para Internet do Instituto Federal de Brasília do *Campus* Brasília, como parte da exigência para obtenção do título de tecnólogo.

Orientador: Prof. Msc. Gustavo Henrique Dornelas de Deus
Instituto Federal de Brasília

Brasília - DF
2025

C138 Caixeta, Ana Luísa.

Análise de temas de Trabalhos de Conclusão de Curso na Rede Federal: uma abordagem baseada em web scraping e visualização de dados. / Ana Luísa Caixeta. – Brasília, 2025.

99 f. : il. color.

Orientador: Gustavo Henrique Dornelas de Deus.

Trabalho de conclusão de curso (Graduação) – Instituto Federal de Educação, Ciência e Tecnologia de Brasília, Tecnologia em Sistemas para Internet, 2025.

1. TCC. 2. Web Scraping. 3. Visualização de Dados. 4. Rede Federal de Educação. I. Deus, Gustavo Henrique Dornelas de. (orient.). II. Título.

CDU 004.41:001.8

Elaborado com os dados fornecidos pelo autor.

Ana Luísa Caixeta

ANÁLISE DE TEMAS DE TRABALHOS DE CONCLUSÃO DE CURSO NA REDE FEDERAL: UMA ABORDAGEM BASEADA EM WEB SCRAPING E VISUALIZAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de Tecnologia em Sistemas para Internet do Instituto Federal de Brasília do *Campus* Brasília, como parte da exigência para obtenção do título de tecnólogo.

Aprovado em 8 de Dezembro de 2025

BANCA EXAMINADORA

Prof. Msc. Gustavo Henrique Dornelas de Deus
Instituto Federal de Brasília

Prof. Msc. Fernando Wagner Brito Hortêncio Filho
Instituto Federal de Brasília

Profa. Dra. Sylvana Karla da Silva de Lemos Santos
Instituto Federal de Brasília

Dedico este trabalho à minha família, amigos, ao meu noivo e aos professores, que me ofereceram apoio incondicional e foram essenciais nesta jornada.

AGRADECIMENTOS

A jornada de elaboração deste Projeto do Trabalho de Conclusão de Curso foi marcada por desafios e aprendizados significativos, mas também por um apoio fundamental de diversas pessoas. É com imensa gratidão que expresso meu reconhecimento a todos que contribuíram para a concretização deste projeto.

Aos meus professores que ao longo destes anos, compartilharam conhecimentos e experiências valiosas. Especial reconhecimento ao meu orientador, pela dedicação, paciência e valiosas contribuições que tornaram possível a realização deste trabalho.

À minha família, em especial, meus pais e meus irmãos pelo apoio incondicional, compreensão e incentivo em todos os momentos desta trajetória, sendo meu alicerce e motivação constante.

Ao meu noivo, por ser meu parceiro incansável nesta caminhada. Seu apoio, compreensão, paciência e incentivo diário foram um pilar fundamental, pois tornou os momentos de pressão em leveza.

Aos meus amigos, que, mesmo distantes ou em meio à rotina agitada, sempre encontraram tempo para oferecer uma palavra de encorajamento, um momento de descontração ou um ombro amigo.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho e para minha formação pessoal e profissional, meu sincero agradecimento.

“A ciência nunca resolve um problema sem criar pelo menos dez outros.”
— George Bernard Shaw

RESUMO

Esta monografia aborda a lacuna na sistematização dos temas de TCCs desenvolvidos nos diversos cursos de ensino superior das instituições da Rede Federal de Educação Profissional, Científica e Tecnológica. O objetivo geral é analisar as respectivas temáticas por meio de técnicas de *web scraping* para coleta de dados e desenvolver um *dashboard* interativo para visualização e análise de dados acadêmicos. A metodologia envolveu a implementação de um sistema de *web scraping* assíncrono, que coletou 206.399 metadados de TCCs dos repositórios digitais de 30 instituições. Esses dados foram submetidos a um pipeline de Extração, Transformação e Carga (ETL), que resultou em 81.666 registros validados e estruturados em um modelo *Star Schema*. Subsequentemente, técnicas de Processamento de Linguagem Natural (PLN) e mineração de texto foram aplicadas, onde o algoritmo *Latent Dirichlet Allocation* (LDA) categorizou os trabalhos em 10 tópicos temáticos distintos. A análise dos resultados identificou tendências de crescimento, com destaque para o Tópico 0 "Aprendizagem, Matemática e Revisão"(coeficiente angular de 42,34), e a emergência de termos como "aplicativo"(crescimento de 554,3%). O projeto culminou no desenvolvimento de um *dashboard* interativo em *Streamlit*, que permite a exploração dinâmica dos dados. A pesquisa é classificada como aplicada, descritiva, exploratória, predominantemente quantitativa e técnica. Conclui-se que o artefato desenvolvido é funcional, atende aos requisitos propostos e demonstra que a arquitetura do pipeline é eficaz para transformar dados acadêmicos dispersos em *insights* estratégicos para gestores, pesquisadores e alunos.

Palavras-chave: TCC. Web Scraping. Visualização de Dados. Rede Federal de Educação. Processamento de Linguagem Natural (PLN).

ABSTRACT

This monograph addresses the gap in the systematization of undergraduate thesis topics developed across various higher education programs within Brazil's Federal Network of Professional, Scientific, and Technological Education institutions. The general objective is to analyze these themes using web scraping techniques for data collection and to develop an interactive dashboard for the visualization and analysis of academic data. The methodology involved implementing an asynchronous web scraping system, which collected 206,399 thesis metadata records from the digital repositories of 30 institutions. These data were processed through an Extraction, Transformation, and Loading (ETL) pipeline, resulting in 81,666 validated records structured in a Star Schema model. Subsequently, Natural Language Processing (NLP) and text mining techniques were applied, in which the Latent Dirichlet Allocation (LDA) algorithm categorized the works into ten distinct thematic topics. The analysis of results identified growth trends, highlighting Topic 0 "Learning, Mathematics, and Review" (slope coefficient of 42.34) and the emergence of terms such as "application" (growth of 554.3%). The project culminated in the development of an interactive dashboard in Streamlit, enabling dynamic data exploration. The research is classified as applied, descriptive, exploratory, predominantly quantitative, and technical. It is concluded that the developed artifact is functional, meets the proposed requirements, and demonstrates that the pipeline architecture is effective in transforming dispersed academic data into strategic insights for managers, researchers, and students.

Keywords: TCC. Web Scraping. Data Visualization. Federal Education Network. Natural Language Processing (NLP).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Interfaces de Programação de Aplicações
APL	Arranjo Produtivo Local
ARIMA	*Autoregressive Integrated Moving Average*
BI	*Business Intelligence*
BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag-of-Words*
CEFETs	Centros Federais de Educação Tecnológica
CGU	Controladoria-Geral da União
CPU	*Central Processing Unit*
CSS	*Cascading Style Sheets*
CSV	Valores Separados por Vígula
DSR	*Design Science Research*
DOM	*Document Object Model*
DW	*Data Warehouse*
ETL	Extração, Transformação e Carga
GRU	*Gated Recurrent Unit*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
IDF	*Inverse Document Frequency*
IFB	Instituto Federal de Brasília
IFs	Institutos Federais
IFSP	Instituto Federal de São Paulo
I/O	*Input/Output*
JSON	*JavaScript Object Notation*

KPI	*Key Performance Indicator*
LDA	*Latent Dirichlet Allocation*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MER	Modelo Entidade-Relacionamento
NFD	Normalização *Form Canonical Decomposition*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
OLAP	*Online Analytical Processing*
OLTP	*Online Transaction Processing*
ORM	*Object-Relational Mapping*
PLN	Processamento de Linguagem Natural
PTCC	Projeto de Trabalho de Conclusão de Curso
RAM	*Random Access Memory*
RF	Requito Funcional
RFEPCT	Rede Federal de Educação Profissional, Científica e Tecnológica
RI	Recuperação de Informação
RMSE	*Root Mean Squared Error*
RNF	Requisito Não Funcional
RNN	*Recurrent Neural Networks*
RU	Requisto de Usuários
SPA	*Single Page Application*
SQL	*Structured Query Language*
TCC	Trabalho de Conclusão de Curso
TCCs	Trabalhos de Conclusão de Curso
TF	*Term Frequency*

TF-IDF	*Term Frequency-Inverse Document Frequency*
UC	Casos de Uso
UFPB	Universidade Federal da Paraíba
URL	*Uniform Resource Locator*
UTFPR	Universidade Tecnológica Federal do Paraná
VSM	*Vector Space Model*
WWW	*World Wide Web*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

O cenário educacional brasileiro tem passado por transformações significativas nas últimas décadas, especialmente no que se refere à formação profissional e tecnológica. Em 2008, a Rede Federal foi criada pela Lei nº 11.892/2008. Ela é composta pelos Institutos Federais de Educação Ciência e Tecnologia (IFs), pela Universidade Tecnológica Federal do Paraná (UTFPR), pelos Centros Federais de Educação Tecnológica (CEFETs) do Rio de Janeiro e de Minas Gerais, e pelo Colégio Pedro II (Brasil, 2008). Esta estrutura representa uma das maiores expansões da educação profissional e tecnológica da história do país, que abrange todo o território nacional com mais de 600 unidades distribuídas entre as diferentes instituições da rede federal (MEC, 2024).

A produção acadêmica gerada nessas instituições, em particular, os Trabalhos de Conclusão de Curso (TCCs), constitui um importante indicador das áreas de interesse, competências desenvolvidas e tendências de pesquisa dentro da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT). Nesse contexto, é fundamental compreender que a pesquisa nos Institutos Federais deve transcender a mera produção científica, ao integrar-se ao processo formativo dos estudantes e ao contribuir para o desenvolvimento local e regional (Pacheco, 2011).

No cenário contemporâneo de proliferação de informações, o conceito de Dados Abertos (*Open Data*) se consolidou como um pilar fundamental para a transparência, a inovação e o desenvolvimento social em escala global. Segundo a *Open Knowledge Foundation (2025)*, dados são considerados abertos quando qualquer pessoa possui liberdade para usar, reutilizar e redistribuir, desde que, no máximo, haja a exigência de atribuição da autoria e de compartilhamento pela mesma licença. Essa definição estabelece os princípios fundamentais de acesso irrestrito e liberdade de uso, que caracterizam o movimento de dados abertos e formam as bases conceituais de iniciativas governamentais e institucionais em todo o mundo.

No contexto governamental brasileiro, essa abordagem se formalizou por meio do Decreto nº 8.777/2016 (Brasil, 2016), que institui a Política de Dados Abertos no âmbito do Poder Executivo Federal, ao estabelecer diretrizes para a publicação de dados governamentais de forma estruturada e acessível. A disponibilização de dados abertos representa um instrumento essencial para o fortalecimento da governança democrática, ao possibilitar que cidadãos, pesquisadores e organizações da sociedade civil tenham acesso e realizem a análise de informações anteriormente restritas ou fragmentadas. Esse processo favorece o controle social, a criação de novos serviços digitais e o avanço do conhecimento científico, fundamentado em evidências empíricas.

Para que o potencial dos Dados Abertos alcance sua plena exploração, a análise de dados se apresenta como uma disciplina essencial. Essa área tem como foco a examinação,

transformação e modelagem de dados, com o propósito de descobrir informações úteis e apoiar a tomada de decisões. Conforme Sharda *et al.* (2019), a análise de dados corresponde ao processo de gerar recomendações práticas fundamentadas em informações obtidos a partir de dados históricos, por meio da combinação de tecnologia e técnicas estatísticas. De forma complementar, o *Business Intelligence* (BI) se caracteriza como um conjunto de estratégias e tecnologias que possibilita o acesso interativo e a manipulação de dados, o que capacita gestores a realizar decisões mais embasadas. Seu objetivo principal consiste em converter dados em informações, capazes de subsidiar decisões e, consequentemente, ações estratégicas (Sharda *et al.*, 2019).

1.1 Problema

Apesar da importância estratégica da produção acadêmica da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPT), existe uma lacuna significativa no que se refere a um panorama consolidado e sistematizado dos temas de TCCs desenvolvidos em todas as instituições que compõem essa rede. A dispersão dessas informações em repositórios institucionais distintos dos IFs, CEFETs, UTFPR e Colégio Pedro II, com diferentes estruturas e padrões de organização, dificulta uma visão abrangente das tendências de pesquisa e dos focos de interesse acadêmico na rede federal.

Diante desse cenário, a questão que norteia essa pesquisa é: Quais são as principais temáticas abordadas nos Trabalhos de Conclusão de Curso (TCCs) da Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil e como esses temas podem ser categorizados e visualizados para fornecer achados acadêmicos relevantes?

1.1.1 Objetivo geral

A proposta deste trabalho consiste em analisar os temas dos Trabalhos de Conclusão de Curso (TCCs) produzidos na (RFEPCT), por meio de técnicas de *web scraping* para coleta de dados e do desenvolvimento de um *dashboard* interativo para visualização e análise de *insights* acadêmicos.

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

- Implementar um sistema de *web scraping* para coletar metadados de TCCs (título, autor, professor orientador, curso, instituição, ano, resumo, palavras-chave) disponíveis nos repositórios digitais das instituições integrantes da Rede Federal de Educação, com a ressalva de que a presença e a completude desses dados variam entre elas.
- Utilizar técnicas de Processamento de Linguagem Natural (PLN) e mineração de texto para limpar, padronizar e categorizar os temas dos TCCs.

- Realizar análises quantitativas para identificar a frequência, distribuição e evolução temporal dos temas de pesquisa, além de comparar tendências entre diferentes instituições da Rede Federal (IFs, CEFETs, UTFPR e Colégio Pedro II).
- Desenvolver um *dashboard* interativo que permita a exploração dinâmica dos dados, por meio de recursos como filtros por período, instituição e área temática, e visualizações gráficas que facilitem a identificação de temáticas e tendências, lacunas e oportunidades de pesquisa.

1.2 Estrutura do Trabalho de Conclusão de Curso (TCC)

1.2.1 Classificação da Pesquisa

Quanto à finalidade, esta pesquisa caracteriza-se como aplicada, uma vez que busca gerar conhecimento prático e uma ferramenta concreta para análise da produção acadêmica da Rede Federal de Educação Profissional (RFEPCT), Científica e Tecnológica (Gil, 2022). Em relação aos objetivos, classifica-se como descritiva e exploratória: descritiva ao buscar caracterizar e quantificar os temas de TCCs da Rede Federal, e exploratória por investigar padrões em um conjunto amplo e heterogêneo de dados acadêmicos (Richardson, 2017).

Quanto à abordagem, adota-se uma perspectiva quantitativa predominante, devido à coleta e análise de grande volume de dados, com elementos qualitativos na categorização e interpretação semântica dos temas (Creswell, 2014). Sob a perspectiva dos procedimentos técnicos, a pesquisa combina aspectos documentais (análise de TCCs como documentos), bibliográficos (revisão de literatura) e técnicos (desenvolvimento de sistemas de *web scraping*, banco de dados e *dashboard*).

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Extração, transformação e carga (ETL) e modelagem dimensional

2.1.1 Extração, transformação e carga (ETL)

O processo de extração, transformação e carga (ETL) constitui um dos componentes fundamentais da arquitetura de dados, pois permite que informações provenientes de diferentes fontes sejam coletadas, tratadas para garantir qualidade e consistência dos dados e, por fim, armazenadas em *data warehouse* (DW). Segundo Kimball e Caserta (2004), o ETL vai além de apenas transportar os dados, ele agrega valor ao reduzir erros e inconsistências, documentar medidas de confiabilidade, registrar o fluxo das transações, integrar informações provenientes de diferentes fontes e estruturar os dados de forma que possam ser efetivamente utilizados pelos usuários finais.

A fase de extração refere-se à identificação, acesso e coleta de diferentes fontes, como sistemas transacionais, bancos de dados relacionais, arquivos estruturados e não estruturados, APIs (*Interfaces de Programação de Aplicações*) e serviços *web*. Kimball e Caserta (2004) ressaltam os desafios encontrados durante o desenvolvimento da etapa de extração que estão relacionados à conectividade, performance e integridade dos dados. Consequentemente, torna-se necessário desenvolver sistemas de acesso robustos, capazes de lidar com diversos formatos, protocolos e restrições de segurança.

Em seguida, na fase de transformação, Inmon (2002) destaca a necessidade do processamento ser realizado em uma área de *staging* que fique fora do *data warehouse*, local onde os dados são submetidos a operação de limpeza, padronização, enriquecimento, agregação e aplicação de regras de negócio. Essa etapa é considerada a mais complexa do processo de ETL, visto que deve converter dados brutos em informações estruturadas que atendam aos requisitos analíticos da organização com intuito de garantir a qualidade e conformidade com os padrões estabelecidos.

Por fim, a fase de carga compreende a inserção dos dados transformados no sistema de destino. Conforme, Kimball e Ross (2013) é de suma importância serem considerados aspectos tais como *performance*, integridade referencial, controle de versões, estratégias de atualização, com destaque quanto a escolha do tipo de carga ou completa ou incremental devido a possibilidade de impactar o desempenho e a disponibilidade do sistema aos usuários.

Sendo assim, a implementação de processos ETL pode aderir diferentes abordagens arquiteturais e ferramentas de desenvolvimento. Inmon (2002) propõe uma abordagem *top-down*, na qual ferramentas tradicionais de bancos de dados relacionais são adaptadas às necessidades de um *data warehouse* corporativo, com ênfase na normalização e na integração empresarial. Em contrapartida, Kimball e Ross (2013) defendem uma metodolo-

gia *bottom-up*, voltada para o desenvolvimento de *data marts* com esquemas estrela (*star schema*), que prioriza agilidade, rapidez na entrega e foco nas necessidades do usuário final.

2.1.2 Modelagem Dimensional

A modelagem dimensional é reconhecida, muitas vezes, como a principal técnica de projeto de bancos de dados para *data warehouses* e sistemas de *Business Intelligence* (BI). Diferentemente da modelagem transacional, utilizada em sistemas OLTP (*Online Transaction Processing*) que possui foco na eficiência de operações de inserção, atualização e exclusão de dados para garantir consistência e integridade transacional, a modelagem dimensional é usada em sistemas OLAP (*Online Analytical Processing*), construídos para otimizar consultas complexas, sumarizações e análises multidimensionais. Conforme Kimball e Ross (2013), seu objetivo é apresentar os dados de forma clara e estruturada, de modo a facilitar a interpretação e o acesso às informações para apoiar decisões rápidas e fundamentadas.

Um modelo dimensional é composto por dois tipos principais de tabelas, as fato e as de dimensão. As tabelas fato centralizam as métricas e medidas referentes aos eventos de uma organização, elas são caracterizadas por apresentarem grande volume de dados, visto que cada linha corresponde a um evento específico. Por outro lado, as tabelas de dimensão fornecem o contexto descritivo para os fatos e armazenam atributos textuais que permitem responder às principais perguntas analíticas de "quem, o quê, onde, quando, por quê e como" (Kimball; Ross, 2013).

2.1.2.1 Star Schema (Esquema Estrela)

O *star schema* é um modelo de implementação da modelagem dimensional cujo nome se deve à sua disposição visual semelhante a uma estrela, formada por uma tabela fato central cercada por tabelas de dimensão. Nesta arquitetura, as chaves estrangeiras presentes na tabela fato referenciam diretamente as chaves primárias das respectivas tabelas de dimensão, de tal forma que estabelece uma arquitetura de junção simples e desnormalizada (Kimball; Ross, 2013). Essa configuração de relacionamentos diretos, sem níveis intermediários de hierarquia, confere ao modelo simplicidade estrutural o que facilita tanto a compreensão por usuários de negócio quanto a otimização de consultas analíticas.

A característica mais distintiva do *star schema* é a desnormalização proposital de suas tabelas de dimensão. Em modelos transacionais normalizados, os atributos dimensionais são distribuídos em diversas tabelas para reduzir redundâncias e otimizar operações de inserção, atualização e exclusão de dados. No entanto, no *star schema*, esses mesmos atributos contextuais são agrupados em uma única tabela de dimensão, o que facilita o acesso e a consulta aos dados. Segundo Kimball e Ross (2013), essa desnormalização representa uma grande vantagem no *data warehousing*, pois simplifica a estrutura de consultas. Essa característica é especialmente importante para a adoção de ferramentas de

self-service, pois permite que analistas formulem questões de negócio complexas sem a necessidade de escrever consultas com várias junções.

2.2 *Web scraping*

O *web scraping*, também denominado raspagem de dados da web, consiste no processo automatizado de extração de informações de *websites* (Broucke; Baesens, 2018). Esta técnica emprega softwares especializados, conhecidos como *scrapers* ou *bots*, que simulam o comportamento de navegação humana para coletar dados específicos de páginas web. O propósito fundamental desta abordagem é converter dados não estruturados ou semiestruturados, como é o caso de documentos HTML *(HyperText Markup Language)*, em formatos estruturados e processáveis, tais como estruturas de banco de dados, de modo a viabilizar a análise e utilização em processos analíticos.

A terminologia *web scraping*, como é conhecida atualmente, representa a evolução de práticas de extração de dados. Embora o *web scraping* não seja um termo novo, no passado essa prática era mais conhecida como *screen scraping*, mineração de dados, *web harvesting* ou variações similares (Broucke; Baesens, 2018).

O *screen scraping* referia-se originalmente à extração de dados exibidos em terminais de computador, uma técnica amplamente utilizada antes da padronização de APIs (*Application Programming Interfaces*) como método de integração entre sistemas. Com o surgimento da *World Wide Web* (WWW), a técnica foi adaptada ao ambiente web e o consenso geral atualmente favorece o termo *web scraping*, embora expressões como *web harvesting* e *web data extraction* ainda sejam encontradas na literatura técnica (Mitchell, 2018).

Os sistemas de *web scraping* recorre principalmente a duas estratégias de coleta de dados, em que a primeira realiza requisições HTTP (Hypertext Transfer Protocol) diretas aos servidores web, o que inclui o consumo de APIs públicas ou privadas, e a segunda utiliza automação de navegadores para simular o comportamento de usuários reais por meio de código.

A abordagem de requisição HTTP direta constitui o método mais fundamental e eficiente de extração de dados web. Nesta técnica, o *scraper* envia uma requisição HTTP, geralmente dos tipos GET ou POST, diretamente ao servidor web, que retorna o conteúdo bruto da página em formatos como HTML, XML ou JSON. No contexto de APIs (*Application Programming Interfaces*), esse processo é consideravelmente facilitado, pois essas interfaces fornecem dados em formatos estruturados de modo que a extração seja mais direta e confiável (Mitchell, 2018).

Em contrapartida, a automação de navegador é necessária quando o conteúdo é carregado dinamicamente via *JavaScript client-side*, como ocorre em *Single Page Applications* (SPAs). Nestes casos, requisições HTTP diretas são insuficientes, o que demanda ferramentas como *Selenium* (Selenium, 2025), *Playwright* (Playwright, 2025) ou *Puppeteer*

(Puppeteer, 2025) para controlar navegadores como *Chrome* ou *Firefox*. O navegador automatizado executa o *JavaScript*, renderiza a página completa e, somente então, o *scraper* extrai os dados do DOM renderizado. Embora mais robusta para sites dinâmicos, essa abordagem consome significativamente mais tempo de processamento, CPU e memória (Mitchell, 2018).

2.2.1 Abordagens de coleta de dados na Web

Ao implementar *scrapers* que serão responsáveis por realizar múltiplas requisições, é de suma importância que seja considerada a estratégia de execução, visto que a abordagem escolhida pode impactar diretamente o desempenho na obtenção das informações desejadas e na utilização de recursos computacionais. Nesse contexto, destacam-se duas abordagens principais de gerenciamento, a coleta síncrona e a assíncrona.

A coleta síncrona, ou sequencial, representa o modelo de programação mais tradicional para *web scraping*. Nesta abordagem, as requisições são executadas uma de cada vez, ou seja, o *script* realiza uma requisição, aguarda a resposta do servidor, processa os dados recebidos e somente então inicia a próxima operação. O principal gargalo dessa estratégia reside no fato de que operações de rede constituem operações de I/O (*Input/Output*) essencialmente lentas, nas quais a maior parte do tempo de execução é consumida por ter que aguardar respostas dos servidores enquanto o processador permanece ocioso. Embora simples de implementar e depurar, o modelo síncrono é considerado ineficiente e impraticável para coletas de dados em larga escala (Mitchell, 2018).

A coleta assíncrona opera através de um modelo de execução não-bloqueante, ela utiliza um *event loop* (laço de eventos), o *scraper* pode iniciar uma operação de I/O, como uma requisição HTTP, e, ao invés de aguardar a resposta, registra a tarefa pendente e prossegue imediatamente para as próximas requisições (Mitchell, 2018). Dessa forma, o *script* consegue gerenciar centenas ou milhares de conexões simultâneas de forma eficiente. Quando um servidor retorna uma resposta, o *event loop* notifica o sistema e direciona o processador para processar aquela resposta específica, de modo que o tempo de latência de I/O de múltiplas requisições seja sobreposto e mantém tanto o processador quanto a rede consistentemente ocupados (Mitchell, 2018).

A principal diferença entre as duas abordagens está no gerenciamento do tempo de latência de I/O. Na coleta síncrona, o comportamento bloqueante resulta em um tempo total de execução equivalente, na melhor das hipóteses, à soma dos tempos individuais de cada requisição. Em contraste, a coleta assíncrona, não-bloqueante, permite que múltiplas requisições permaneçam em ação simultaneamente, o que reduz bastante o tempo total de coleta, que passa a ser limitado não pela serialização das operações, mas sim pelo desempenho da rede disponível, pela capacidade de resposta do servidor de destino e pelo número de conexões concorrentes que o *scraper* consegue gerenciar eficientemente.

Para projetos de *web scraping* que envolvem a coleta de dezenas, centenas ou

milhares de páginas, a abordagem assíncrona oferece ganhos de desempenho significativos. Em um cenário hipotético onde cada requisição individual demanda 1 segundo para completar, um *scraper* síncrono necessitaria de aproximadamente 16,7 minutos para processar 1.000 páginas, enquanto um *scraper* assíncrono que gerencia 100 conexões concorrentes poderia completar a mesma tarefa em cerca de 10 segundos. Essa diferença consolida a abordagem assíncrona como a estratégia preferencial para extração de dados em larga escala.

2.3 Mineração de texto e Processamento de Linguagem Natural

O aumento do volume de dados textuais disponíveis digitalmente nas últimas décadas, provenientes de redes sociais, websites, artigos científicos e diversas outras fontes, aliado aos avanços em *software* e *hardware*, trouxe à tona novos desafios e oportunidades para a extração automática de informações. Neste contexto, surgiu a mineração de texto (*text mining*) como uma área fundamental dedicada à descoberta de padrões, tendências e conhecimento utéis a partir de grandes coleções de dados textuais (Aggarwal; Zhai, 2012).

O Processamento de Linguagem Natural (PLN) constitui a base tecnológica sobre a qual a mineração de texto se fundamenta. Trata-se de um campo interdisciplinar que combina linguística computacional, inteligência artificial e aprendizado de máquina para capacitar computadores a compreender, interpretar e gerar linguagem humana de forma automática, o que fornece os métodos e algoritmos necessários para processar texto em suas múltiplas dimensões, sintática, semântica e pragmática, de modo que máquinas extraiam significado de dados linguísticos (Jurafsky; Martin, 2024).

A relação entre mineração de texto e PLN é cooperativa visto que o PLN atua como fundação teórica por meio do desenvolvimento de técnicas fundamentais para análise linguística em múltiplas camadas, tokenização, análise morfológica, sintática e semântica, enquanto a mineração de texto opera como dimensão aplicada com a utilização dessas técnicas em escala industrial para extrair padrões e conhecimento de grandes volumes de dados não estruturados. Esta combinação possibilita que avanços teóricos em PLN sejam imediatamente aproveitados em aplicações práticas, enquanto os desafios reais da mineração motivam novas pesquisas em PLN, de tal modo que ocorre um ciclo contínuo de inovação (Manning *et al.*, 2008).

2.3.1 Pré-processamento textual

Dados textuais em seu estado bruto são chamados de ruidosos e não podem ser diretamente interpretados por algoritmos de *machine learning*. O pré-processamento é uma etapa crítica que visa limpar e padronizar o texto, para convertê-lo em um formato mais adequado para análise (Manning *et al.*, 2008). Este processo envolve um *pipeline* de tarefas sequenciais que transformam o texto original em uma representação estruturada e viável

computacionalmente.

A primeira etapa consiste na normalização, que padroniza o texto para reduzir variações superficiais que não agregam valor semântico. Dessa forma, a técnica mais comum é a conversão para minúsculas (*lowercase*), que visa garantir que palavras como "Texto", "texto" e "TEXTO" sejam tratadas como uma única entidade (Jurafsky; Martin, 2024). Outras formas de normalização incluem a remoção de acentuação, expansão de contrações e correção ortográfica, a depender dos objetivos da análise.

A remoção de ruído elimina elementos textuais que não contribuem para o significado, mas que podem interferir no processamento algorítmico. Pontuação, caracteres especiais, números, URLs (Localizador Uniforme de Recursos), menções em redes sociais e *hashtags* são frequentemente removidos ou tratados separadamente (Feldman; Sanger, 2007). Esta etapa deve ser cuidadosamente ajustada ao contexto, em análise de sentimentos, por exemplo, pontuações como "!" e "?" podem carregar informação emocional relevante e devem ser preservadas ou codificadas adequadamente.

A tokenização é o processo de segmentar o texto contínuo em unidades básicas de análise, denominadas *tokens*, que geralmente correspondem a palavras individuais (Manning *et al.*, 2008). Embora pareça trivial, a tokenização apresenta desafios significativos, como tratar palavras compostas ("guarda-chuva"), contrações ("dele"), expressões multipalavra ("*machine learning*") e emojis. Dessa forma, a qualidade da tokenização impacta diretamente todas as etapas subsequentes do pipeline de análise (Jurafsky; Martin, 2024).

Por fim, a remoção de *stopwords* consiste na filtragem de palavras de alta frequência e baixo valor semântico, como artigos ("o", "a"), preposições ("de", "para"), conjunções ("e", "mas") e pronomes ("ele", "sua"). Estas palavras aparecem em praticamente todos os documentos e raramente contribuem para a distinção de tópicos ou padrões significativos (Manning *et al.*, 2008). Entretanto, em tarefas como modelagem de linguagem ou análise sintática, *stopwords* podem ser essenciais e sua remoção inadequada. A decisão de remover ou preservar *stopwords* deve considerar o objetivo específico da aplicação (Feldman; Sanger, 2007).

O conjunto dessas etapas de pré-processamento não segue uma regra específica, pois cada aplicação pode exigir ajustes específicos, como a análise de sentimentos pode beneficiar-se da preservação de pontuação enfática, enquanto a classificação de documentos técnicos pode requerer a manutenção de números e siglas. Portanto, a experimentação e validação empírica são fundamentais para determinar a configuração ideal do pipeline de pré-processamento para cada contexto (Aggarwal; Zhai, 2012).

2.3.2 Vetorização de texto

Após o pré-processamento, o texto limpo e tokenizado ainda não pode ser diretamente utilizado por algoritmos de *machine learning*, que operam exclusivamente com dados numéricos. A vetorização de texto é o processo de transformar dados textuais em

representações numéricas, vetores, que preservam informações relevantes do texto original enquanto permitem operações matemáticas e estatísticas (Aggarwal; Zhai, 2012). Esta conversão é fundamental para possibilitar que algoritmos possam calcular similaridades, identificar padrões e realizar predições sobre documentos textuais.

O modelo *Bag-of-Words* (BoW) é a abordagem mais simples e intuitiva para vetorização textual, onde cada documento é representado como um vetor de frequências em que cada dimensão corresponde a uma palavra única do vocabulário e o valor indica quantas vezes aquela palavra aparece no documento (Manning *et al.*, 2008). Por exemplo, em um vocabulário com "mineração", "texto", "dados" e a frase "mineração de texto e mineração de dados" seria representada como [2, 1, 1]. Embora simples e computacionalmente eficiente, o BoW apresenta limitações significativas, pois ignora completamente a ordem das palavras, perde informações contextuais e sintáticas, e trata todas as palavras com igual importância, de modo que frases como "o algoritmo processa os dados" e "os dados processam o algoritmo" tenham representações idênticas apesar de significados distintos (Feldman; Sanger, 2007).

Para superar essas limitações, o TF-IDF (*Term Frequency–Inverse Document Frequency*) propõe um esquema de ponderação que representa de forma mais adequada a importância relativa dos termos, que combina a frequência do termo em um documento (TF) com a sua raridade na coleção completa (IDF) (Salton; Buckley, 1988). O peso TF-IDF de um termo t em um documento d, considerando uma coleção de documentos D, é definido como

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D), \quad \text{onde} \quad \text{IDF}(t, D) = \log\left(\frac{N}{df(t)}\right)$$

em que N representa o número total de documentos na coleção D e $df(t)$ corresponde ao número de documentos que contêm o termo t. Assim, termos que aparecem em muitos documentos recebem pesos menores, enquanto termos mais raros e específicos recebem pesos mais altos, de modo que a representação textual seja mais informativa para tarefas como recuperação de informação e classificação de documentos (Manning *et al.*, 2008).

Tanto BoW quanto TF-IDF geram representações dispersas de alta dimensionalidade, onde o número de dimensões equivale ao tamanho do vocabulário e a maioria dos valores é zero, já que cada documento contém apenas uma pequena fração do vocabulário total (Aggarwal; Zhai, 2012). Apesar de não capturarem relações semânticas entre palavras, essas técnicas permanecem amplamente utilizadas devido à sua simplicidade, interpretabilidade e eficácia em diversas tarefas de mineração de texto, que servem de base para o desenvolvimento de técnicas mais sofisticadas, como *embeddings* de palavras e modelos de linguagem contextualizados (Jurafsky; Martin, 2024).

2.3.3 Modelagem de tópicos

A modelagem de tópicos é uma técnica de aprendizado não supervisionado que visa descobrir automaticamente temas latentes em coleções de documentos textuais, a partir de análise de padrões de ocorrência simultânea de palavras sem necessidade de dados rotulados ou intervenção humana (Blei *et al.*, 2003). A premissa fundamental é que documentos são compostos por misturas de tópicos, onde cada tópico representa uma distribuição de probabilidades sobre palavras do vocabulário. Por exemplo, em artigos científicos, o algoritmo poderia identificar tópicos como "inteligência artificial" com palavras "algoritmo", "aprendizado", "rede neural" ou "biologia molecular" com "proteína" ou "célula", o que revela a estrutura temática oculta e permite extrair *insights* de volumes massivos de texto de forma eficiente.

O *Latent Dirichlet Allocation* (LDA) é um dos algoritmos mais populares para modelagem de tópicos, caracterizado como um modelo probabilístico generativo que assume cada documento como uma mistura de tópicos e cada tópico como uma distribuição de probabilidades sobre palavras (Blei *et al.*, 2003). Formalmente, o LDA modela cada documento como uma distribuição multinomial sobre K tópicos, onde K é definido pelo usuário, e cada tópico como uma distribuição multinomial sobre as palavras do vocabulário. Para gerar um documento, o processo inicia com a escolha de uma distribuição de tópicos a partir de uma distribuição de *Dirichlet*, que é uma distribuição de probabilidade sobre vetores de proporções não negativas que somam 1. Em seguida, para cada palavra do documento, seleciona-se um tópico com base nessa distribuição e, a partir do tópico escolhido, uma palavra é selecionada conforme a distribuição de palavras associada ao tópico.

Na prática, o objetivo do LDA é inferir as distribuições latentes de tópicos que melhor explicam os dados observados através de métodos como *Variational Bayes* ou *Gibbs Sampling* (Blei *et al.*, 2003). O funcionamento baseia-se em padrões de coocorrência de palavras, ou seja, termos que aparecem frequentemente juntos tendem a ser agrupados no mesmo tópico como, por exemplo, "paciente", "diagnóstico", "tratamento" e "sintoma" formariam um tópico médico. A qualidade dos tópicos identificados depende do número de tópicos especificado (K), do tamanho e da qualidade do corpus e dos hiperparâmetros α e β, que regulam a dispersão das distribuições. A interpretabilidade dos tópicos é avaliada com base nas palavras mais prováveis de cada um.

Apesar de sua elegância matemática e eficácia prática, o LDA possui limitações importantes, pois assume a chamada hipótese *bag-of-words*, segundo a qual a ordem das palavras não importa, requer que o número de tópicos seja especificado a priori, e pode produzir tópicos de difícil interpretação (Blei, 2012). Variações e extensões do LDA foram desenvolvidas para endereçar essas limitações, o que inclui modelos hierárquicos, dinâmicos e supervisionados, de tal modo a demonstrar a flexibilidade e relevância continuada desta abordagem. A aplicabilidade da modelagem de tópicos abrange organização e sumarização de grandes acervos documentais, descoberta de tendências em mídias sociais, análise de

feedback de clientes, recomendação de conteúdo e identificação de pesquisas emergentes em literatura científica.

2.4 Recuperação de Informação e Similaridade Textual

A Recuperação de Informação (RI) é uma área fundamental da ciência da computação dedicada à busca, localização e ranqueamento de informações relevantes em grandes coleções de documentos textuais não estruturados (Manning *et al.*, 2008). Diferentemente dos sistemas de banco de dados tradicionais, que operam sobre dados estruturados por meio de consultas precisas, os sistemas de RI lidam com consultas em linguagem natural e documentos textuais, nos quais a relevância assume caráter subjetivo e probabilístico. Assim, o desafio central da RI consiste em determinar quais documentos são mais relevantes para satisfazer uma necessidade de informação expressa por uma consulta dentro de uma determinada coleção de documentos (Baeza-Yates; Ribeiro-Neto, 2011).

Os sistemas de RI são onipresentes no cotidiano digital e englobam mecanismos de busca na web, bibliotecas digitais, sistemas de busca em intranets corporativas, *e-discovery* em documentos jurídicos, busca em repositórios científicos e sistemas de recomendação de conteúdo (Croft *et al.*, 2015). Nesse contexto, o processo típico de recuperação de informação envolve três componentes principais. Primeiramente, a indexação processa e organiza os documentos para permitir buscas eficientes. Em seguida, a recuperação identifica documentos candidatos que correspondem à consulta, enquanto o ranqueamento ordena os resultados de acordo com a relevância estimada. Por fim, a eficácia é tradicionalmente avaliada por métricas como precisão, que representa a proporção de documentos recuperados que são relevantes, e a revocação, que indica a proporção de documentos relevantes efetivamente recuperados, e medidas combinadas, como *F1-score* e *Mean Average Precision* (Manning *et al.*, 2008).

2.4.1 Modelo Vetorial

O Modelo Vetorial (VSM), proposto por Salton *et al.* (1975), introduziu uma representação geométrica para documentos e consultas, de modo a revolucionar a área de recuperação de informação. Neste modelo, tanto documentos quanto consultas são representados como vetores em um espaço multidimensional, onde cada dimensão corresponde a um termo distinto do vocabulário e os valores dos vetores representam a importância de cada termo (Baeza-Yates; Ribeiro-Neto, 2011). Por exemplo, em um vocabulário com três termos, "mineração", "texto", "dados", um documento poderia ser representado como o vetor [0.5, 0.8, 0.3], o que indica diferentes graus de relevância para cada termo.

A principal inovação do VSM é transformar o problema de recuperação de informação em álgebra linear, de modo que encontrar documentos relevantes equivale a localizar vetores de documentos próximos ao vetor da consulta no espaço vetorial (Manning *et al.*,

2008). Essa abordagem oferece várias vantagens, o que inclui ranqueamento parcial, robustez a variações terminológicas e uma base matemática sólida para calcular similaridade. Os pesos dos termos nos vetores são geralmente calculados com esquemas como TF-IDF, que ponderam a importância ao se considerar tanto a frequência local no documento quanto a raridade na coleção (Salton; Buckley, 1988).

2.4.2 Métricas de Similaridade

Uma vez que documentos e consultas estão representados como vetores no modelo vetorial, torna-se necessário estabelecer uma forma de medir a proximidade ou similaridade entre eles. Dessa forma, surge a similaridade cosseno como a métrica mais amplamente utilizada para esta finalidade em aplicações de recuperação de informação e mineração de texto (Manning *et al.*, 2008). Esta métrica calcula o cosseno do ângulo entre dois vetores, de modo que seja produzido um valor entre -1 e 1, ou entre 0 e 1 para vetores com componentes não-negativos, como é típico em representações TF-IDF.

A similaridade cosseno entre dois vetores A e B é definida como

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

$A \cdot B$ representa o produto escalar dos vetores, e $\|A\|$ e $\|B\|$ correspondem às suas magnitudes ou normas euclidianas (Baeza-Yates; Ribeiro-Neto, 2011). De maneira conceitual, a similaridade cosseno mede o quanto dois vetores apontam na mesma direção, independentemente de suas magnitudes. Sendo assim, dois documentos com vocabulário similar mas tamanhos diferentes terão alta similaridade cosseno, onde um valor próximo a 1 indica vetores muito similares, quase paralelos, enquanto um valor próximo a 0 indica vetores ortogonais, sem similaridade alguma.

A escolha da similaridade cosseno em vez de outras métricas como distância euclidiana é estratégica, haja vista que ela normaliza automaticamente pelo tamanho dos documentos, de modo que documentos longos e curtos sejam comparáveis de forma justa (Manning *et al.*, 2008). Por exemplo, um documento de 10.000 palavras e outro de 100 palavras sobre o mesmo tema terão alta similaridade cosseno, mesmo que suas frequências absolutas de termos sejam muito diferentes. Na prática, quando vetores são representados a partir da utilização de pesos TF-IDF, a similaridade cosseno captura efetivamente a sobreposição de termos importantes e discriminativos entre documentos, de modo que sirva como uma medida robusta de relevância semântica (Salton; Buckley, 1988).

2.5 Análise de Tendências e Machine Learning Aplicado

2.5.1 Análise de Séries Temporais

Séries temporais consistem em conjuntos de observações sequenciais indexadas no tempo, cuja ordem cronológica contém informações sobre a dinâmica do fenômeno

estudado (Box *et al.*, 2015). Ao contrário de dados transversais, que consideram observações independentes, séries temporais apresentam dependência temporal, de modo que valores próximos tendem a ser correlacionados e o passado influencia o presente e o futuro. Consequentemente, a análise de séries temporais busca identificar padrões subjacentes, o que inclui tendências de longo prazo, sazonalidade com repetições regulares, ciclos de longo prazo sem periodicidade fixa e componentes irregulares ou ruído aleatório (Chatfield, 2003).

Além disso, em mineração de texto e análise de linguagem, séries temporais surgem naturalmente a partir de dados como frequência de menções de termos, volume de publicações, evolução de sentimentos em redes sociais, tendências de buscas e popularidade de hashtags (Atefeh; Khreich, 2015). Dessa forma, a análise dessas séries permite identificar temas em ascensão ou declínio, avaliar impactos de eventos específicos no discurso público e acompanhar a evolução de conceitos científicos ao longo do tempo. Enquanto técnicas tradicionais, como decomposição sazonal, suavização exponencial e modelos de autorregressão integrada de médias móveis ARIMA (*Autoregressive Integrated Moving Average*), extraem componentes temporais e permitem previsões, métodos modernos, o que inclue redes neurais recorrentes RNN (*Recurrent Neural Networks*) como redes de memória de longo prazo LSTM (*Long Short-Term Memory*) e unidades recorrentes fechadas GRU (*Gated Recurrent Unit*) capturam dependências complexas de longo prazo (Hyndman; Athanasopoulos, 2018).

2.5.2 Modelagem Preditiva

Modelagem preditiva refere-se ao uso de técnicas estatísticas e algoritmos de Machine Learning para construir modelos que estimam valores futuros ou desconhecidos com base em padrões aprendidos de dados históricos (Kuhn; Johnson, 2013). No contexto de séries temporais, o objetivo consiste em prever observações futuras a partir de informações passadas e, quando disponível, de variáveis externas relacionadas. O processo de modelagem preditiva envolve várias etapas que incluem seleção e engenharia de *features* ou variáveis preditoras, escolha do algoritmo apropriado, treinamento do modelo com dados históricos, validação através de técnicas como validação cruzada temporal e, finalmente, avaliação do desempenho preditivo com métricas como RMSE (Root Mean Squared Error) ou MAE (Mean Absolute Error) (Hyndman; Athanasopoulos, 2018).

A Regressão Linear Simples constitui um dos modelos preditivos mais fundamentais e interpretáveis para análise de tendências. Este modelo assume uma relação linear entre uma variável independente, geralmente o tempo t, e uma variável dependente y, expressa pela equação

$$y = \beta_0 + \beta_1 t + \varepsilon,$$

em que β_0 representa o intercepto, β_1 o coeficiente angular ou taxa de mudança da tendência, e ε o erro aleatório (Montgomery *et al.*, 2012). O objetivo consiste em estimar os

parâmetros β_0 e β_1 que melhor se ajustam aos dados observados, geralmente por meio do método dos mínimos quadrados ordinários, que minimiza a soma dos quadrados dos resíduos. Uma vez estimados, esses parâmetros permitem projetar valores futuros por meio da extrapolação da linha de tendência.

2.5.3 Detecção de Termos Emergentes

A detecção de termos emergentes é uma técnica especializada de mineração de texto que identifica palavras-chave, conceitos ou tópicos cuja frequência de uso cresce significativamente ao longo do tempo, o que indica tendências emergentes, novos desenvolvimentos tecnológicos ou mudanças no discurso público (Kleinberg, 2003). Essa análise é particularmente valiosa em inteligência competitiva, monitoramento de tendências científicas, análise de mercado e vigilância de mídias sociais, pois permite que organizações e pesquisadores identifiquem de forma proativa tópicos em ascensão antes que se tornem comuns.

A abordagem mais direta para detecção de termos emergentes consiste em calcular a frequência relativa de cada termo em janelas temporais consecutivas, como meses, trimestres ou anos, e identificar aqueles que exibem crescimento acelerado (Atefeh; Khreich, 2015). Técnicas mais sofisticadas incluem o modelo de *burst detection* proposto por Kleinberg (2003), que trata a ocorrência de termos como processos probabilísticos e identifica explosões estatisticamente significativas na frequência de uso, o que permite distinguir crescimento genuíno de flutuações aleatórias. Outras abordagens comparam distribuições temporais por meio de testes estatísticos, como o teste qui-quadrado para comparação entre períodos, calculam taxas de crescimento relativo ou aplicam técnicas de detecção de anomalias para identificar desvios relevantes da frequência esperada (FUNG *et al.*, 2005).

Além da frequência bruta, métodos avançados consideram também características contextuais dos termos. Entre essas características, destacam-se a novidade, que se refere a palavras inexistentes em períodos anteriores; a ressonância, que diz respeito a termos que se difundem rapidamente entre diferentes fontes ou comunidades; e, por fim, a persistência, que indica termos cujo crescimento se mantém por vários períodos, em vez de ocorrer apenas em picos isolados (ALLAN *et al.*, 1998). Dessa forma, a combinação da análise de séries temporais com técnicas de processamento de linguagem natural torna possível identificar termos emergentes isolados, além de detectar tópicos compostos em ascensão, analisar o contexto semântico de uso e mapear redes de termos que surgem simultaneamente.

2.6 Visualização de Dados e *Dashboards* Interativos

A visualização de dados e os *dashboards* interativos são componentes essenciais para comunicar *insights* de mineração de texto e análise de tendências, pois transformam

resultados analíticos complexos em representações gráficas intuitivas e exploráveis (Cairo, 2016).

Enquanto isso, as técnicas de mineração de texto extraem padrões a partir de dados não estruturados, ao passo que as visualizações traduzem esses achados em formatos acessíveis, o que possibilita uma compreensão rápida e uma exploração interativa dos resultados. Além disso, a integração de *dashboards* com *frameworks* modernos e formatos de dados otimizados favorece a criação de ecossistemas dinâmicos, nos quais os usuários podem interagir com as visualizações por meio de filtros e seleções, para que assim obtenham respostas em tempo real (Few, 2006).

2.6.1 Princípios da Visualização de Dados

A visualização de dados é definida como o processo de representar informações e conhecimento através de gráficos, diagramas e outras formas visuais, com o objetivo de comunicar *insights* complexos de forma clara, eficaz e acionável (Cairo, 2016). Diferentemente da simples produção de gráficos decorativos, a visualização de dados fundamenta-se em princípios da percepção visual humana, teoria da informação e design gráfico para maximizar a compreensão e minimizar distorções ou ambiguidades na interpretação dos dados (Tufte, 2001). Os principais objetivos da visualização de dados são facilitar a exploração de informações para identificar padrões desconhecidos, validar hipóteses por meio de representações visuais claras, comunicar resultados analíticos a diferentes públicos e auxiliar na tomada de decisões com base em evidências visuais objetivas e fundamentadas.

Princípios fundamentais orientam a criação de visualizações eficazes. Por conseguinte, o princípio da integridade gráfica estabelece que a representação visual deve ser proporcional às quantidades numéricas representadas com o intuito de evitar distorções que induzam interpretações errôneas (Tufte, 2001). Além disso, o princípio da maximização da razão dados-tinta recomenda remover elementos gráficos supérfluos, concentrando a atenção nos dados. Já a clareza exige que visualizações sejam facilmente compreendidas, por meio da utilização de codificações visuais adequadas como posição, comprimento, cor e forma, conforme o tipo de dado e a mensagem a ser transmitida (Few, 2006). Adicionalmente, visualizações devem respeitar limitações da percepção humana, como o número limitado de cores distinguíveis simultaneamente, a dificuldade de comparar áreas ou volumes, e a eficácia superior de codificações posicionais em relação a codificações por cor ou tamanho (Cairo, 2016).

A escolha do tipo de visualização apropriada depende da natureza dos dados e dos objetivos analíticos. Por exemplo, séries temporais são efetivamente representadas por gráficos de linha, que revelam tendências e sazonalidades, enquanto comparações entre categorias se beneficiam de gráficos de barras ou colunas. Além disso, distribuições estatísticas são bem capturadas por histogramas ou *box plots*, e relações entre variáveis numéricas podem ser exploradas através de gráficos de dispersão (Few, 2006). No caso de

dados textuais e resultados de mineração de texto, visualizações especializadas também são relevantes, pois nuvens de palavras permitem analisar frequências de termos, gráficos de rede evidenciam relações entre conceitos, mapas de calor facilitam a visualização de matrizes de similaridade e gráficos de bolhas mostram a evolução temporal de tópicos (Cairo, 2016).

2.6.2 Dashboards Analíticos

Dashboards analíticos são painéis visuais que consolidam e exibem métricas, indicadores-chave de desempenho (KPIs) e visualizações diversas em uma única interface integrada, de modo a fornecer uma visão abrangente e em tempo real do estado de um sistema, processo ou fenômeno de interesse (Few, 2006). Diferentemente de relatórios estáticos tradicionais, *dashboards* enfatizam a apresentação simultânea de múltiplas informações relacionadas, o que permite aos usuários compreenderem contextos complexos através da justaposição visual de diferentes métricas e suas relaçoes. Com isso, a eficácia de um *dashboard* reside em sua capacidade de responder rapidamente às questões mais importantes dos usuários, destacar exceções ou anomalias que requerem atenção e facilitar o monitoramento contínuo de processos críticos.

A interatividade constitui um elemento distintivo e fundamental dos *dashboards* modernos, pois transforma visualizações passivas em ferramentas exploratórias dinâmicas (Murray, 2017). Entre os recursos interativos destacam-se filtros, que permitem aos usuários segmentar dados por dimensões específicas, como período temporal, categoria ou região geográfica, e seleções dinâmicas, em que a escolha de um elemento em uma visualização atualiza automaticamente outras visualizações relacionadas, de modo a revelar conexões entre diferentes aspectos dos dados. Além disso, o recurso de *drill-down* possibilita navegar de visões agregadas para níveis mais detalhados de granularidade, enquanto ferramentas de zoom e *pan* permitem explorar regiões específicas de interesse em gráficos densos. Dessa forma, essa interatividade capacita os usuários finais a conduzirem análises *ad hoc* sem necessidade de conhecimento técnico avançado ou dependência de analistas de dados para cada questão que surge (Few, 2006)

A arquitetura de *dashboards* eficazes segue princípios de design orientado ao usuário. Nesse sentido, informações mais críticas devem ocupar posições privilegiadas no layout, tipicamente o quadrante superior esquerdo, onde o olhar naturalmente inicia a leitura. Além disso, visualizações relacionadas devem ser agrupadas espacialmente para facilitar comparações. Do mesmo modo, o uso consistente de cores, escalas e formatos ao longo do *dashboard* reduz a carga cognitiva e acelera a interpretação (Cairo, 2016). Ainda, *dashboards* devem equilibrar densidade informacional com clareza para evitar tanto sobrecarga visual por excesso de elementos quanto interfaces minimalistas que omitem informações relevantes. No contexto de mineração de texto, *dashboards* podem integrar visualizações de frequências de termos ao longo do tempo, distribuições de tópicos descobertos, métricas de

qualidade dos modelos, alertas sobre termos emergentes e comparações entre diferentes segmentos analisados (Murray, 2017).

2.7 Trabalhos Correlatos

A análise de dados educacionais tem se consolidado como uma importante ferramenta para compreensão de tendências acadêmicas e apoio à gestão institucional. Os trabalhos correlatos apresentados nesta seção demonstram diferentes abordagens para o tratamento e análise de informações educacionais, ao incluir estudos bibliométricos sobre produção acadêmica, aplicação de metodologias de BI em dados educacionais e iniciativas de transparência institucional. Essas contribuições fornecem o embasamento teórico necessário para o desenvolvimento da presente pesquisa.

2.7.1 Uma análise das áreas temáticas do trabalho de conclusão de curso (TCC) em Ciências Contábeis Campus I da UFPB no quadriênio 2016.1-2019.1

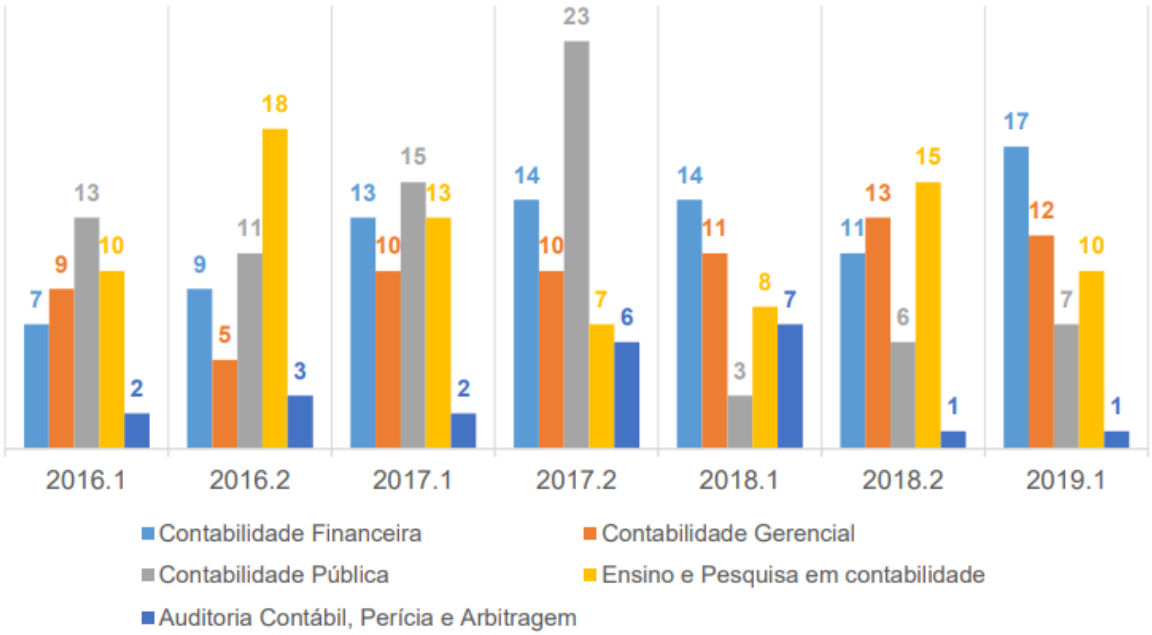
A monografia de Parnaíba (2020) teve como objetivo identificar as áreas temáticas mais recorrentes nos Trabalhos de Conclusão de Curso (TCCs) do tipo monografia, no curso de Ciências Contábeis do Campus I da Universidade Federal da Paraíba (UFPB), no período de 2016.1 a 2019.1. O estudo possui caráter descritivo, quantitativo, bibliográfico, bibliométrico e censitário, e abrange um universo de 336 (TCCs). A coleta dos dados ocorreu por meio da base de dados da Comissão de TCC do curso de Ciências Contábeis da UFPB.

Os resultados apontaram que a área de Contabilidade Financeira representou a maior proporção, com 25% dos trabalhos, seguida de Ensino e Pesquisa em Contabilidade (24%) e Contabilidade Pública (23%). Por outro lado, Auditoria Contábil, Perícia e Arbitragem apresentou a menor incidência, com apenas 7% das monografias analisadas.

O estudo também trouxe a distribuição dos TCCs por gênero dos autores, com leve predominância masculina, com 171 trabalhos elaborados por homens em comparação aos 165 produzidos por mulheres. Além disso, a análise revelou que a maioria dos trabalhos possui, em média, três palavras-chave, com destaque para “Ciências Contábeis” como o termo mais frequente (24 ocorrências), seguido de “Educação Financeira” e “Contabilidade” (19 ocorrências cada).

A Figura 1 apresenta um gráfico extraído da monografia da Parnaíba (2020). Essa imagem foi incluída por representar de forma objetiva a distribuição das monografias segundo as áreas temáticas e sua evolução temporal. A estrutura do gráfico evidencia a concentração de temas em determinados períodos e permite identificar possíveis mudanças nos focos de interesse dos estudantes ao longo do tempo, de modo a oferecer uma base comparativa relevante entre os dados observados.

Figura 1 – **Distribuição das monografias segundo a área temática e por período letivo**



Fonte: Parnaíba (2020)

2.7.2 Aplicação das metodologias de Business Intelligence para análise dos dados abertos governamentais do Instituto Federal de Brasília

A monografia de Campos (2021) teve como objetivo realizar uma análise exploratória dos dados abertos do Instituto Federal de Brasília (IFB), por meio da aplicação de metodologias de *Business Intelligence* (BI) para facilitar a compreensão das informações e subsidiar a tomada de decisões de forma mais assertiva.

Para isso, adotou a metodologia *Design Science Research* (DSR), que tem como foco a construção e avaliação de artefatos. Como resultado, o trabalho desenvolveu um painel interativo (*dashboard*), apoiado em processos de Extração, Transformação e Carga (ETL) e em modelagem dimensional, conforme a abordagem *bottom-up* proposta por Kimball. As ferramentas utilizadas incluíram o *Power BI*, *Power Query* e a linguagem M, aplicadas na manipulação e na visualização dos dados.

O projeto teve como foco a análise de dados públicos do IFB, foram considerados indicadores como permanência e êxito, além do número de alunos matriculados, certificados e diplomados, com informações extraídas do Portal Brasileiro de Dados Abertos. Apesar dos desafios, especialmente da limitação dos dados brutos, que impediu a realização de novos cálculos, o estudo resultou em um processo de *Business Intelligence* (BI) devidamente documentado, que contemplou todas as etapas, desde a coleta até a entrega de visualizações de qualidade.

A Figura 2 exibe uma tela de *dashboard* retirada da monografia de Campos (2021). Essa imagem foi selecionada por ilustrar de forma clara a funcionalidade de análise da evolução temporal de dados, recurso fundamental para identificar tendências e padrões

ao longo do tempo. A visualização temporal, como a apresentada, demonstra o potencial de ferramentas interativas para transformar dados brutos em informações acessíveis e relevantes para o processo decisório em contextos educacionais e institucionais.

Figura 2 – **Página Evolução Temporal**

Ano	BACHARELADO	ESPECIALIZACAO (LATO SENSU)	FORMACAO CONTINUADA	FORMACAO INICIAL	LICENCIATURA	MESTRADO PROFISSIONAL	TÉCNICO	TECNOLOGIA	Total
2009	0	2	690	0	0	0	948	0	**1640**
2010	0	224	3336	1592	66	0	1506	154	**6878**
2011	0	218	3952	1224	238	0	4210	310	**10152**
2012	0	208	4810	4612	922	0	9898	542	**20992**
2013	0	234	5130	4404	1384	0	18022	862	**30036**
2014	128	92	5754	2158	1658	0	16996	1120	**27906**
Total	**2436**	**2804**	**48378**	**28224**	**23058**	**142**	**172856**	**20962**	**298860**

Fonte: Campos (2021)

2.7.3 IFB em Números

A plataforma “IFB em Números” é uma iniciativa do Instituto Federal de Brasília (IFB) voltada à transparência pública, que organiza e disponibiliza dados e informações institucionais de forma acessível e estruturada. Lançada em agosto de 2015, a plataforma foi atualizada em junho de 2020, com a adição de novos módulos e a ampliação da quantidade de dados e informações disponíveis. Atualmente, o sistema conta com sete módulos principais, e abrange áreas como Ensino, Orçamento, Servidores, Ouvidoria, Pesquisa, Extensão e Mundo do Trabalho. A iniciativa foi reconhecida em 2016 com o 4º Concurso de Boas Práticas da CGU , na categoria "Promoção de transparência ativa e/ou passiva", o que a consolidou como referência em transparência pública (IFB, 2025).

A Figura 3 apresenta a interface da página de pesquisas da plataforma IFB em Números, que disponibiliza dados institucionais de forma visual e interativa. Essa tela exemplifica uma das funcionalidades da plataforma, ao demonstrar o uso de recursos gráficos que facilitam a exploração e a compreensão das informações por parte de gestores, pesquisadores e demais interessados. A visualização dos dados fortalece a transparência institucional e oferece uma referência útil para projetos que tenham como objetivo a análise e a divulgação de informações acadêmicas e administrativas.

Figura 3 – **IFB em Números - Página Pesquisa**

Fonte: IFB em Números (2025)

2.7.4 Comparativo entre os Trabalhos Correlatos

A análise dos trabalhos correlatos permite identificar convergências e divergências em relação ao sistema proposto neste trabalho, com destaque para as contribuições específicas e o diferencial representado pelo *dashboard* interativo na análise das temáticas dos TCCs da Rede Federal de Educação Profissional, Científica e Tecnológica.

O estudo de Parnaíba (2020) compartilha o objetivo central de identificar áreas temáticas recorrentes em trabalhos de conclusão de curso, por meio da metodologia quantitativa e descritiva similar à proposta nesta pesquisa. Ambos os trabalhos focam na análise de metadados dos TCCs e incluem palavras-chave, distribuição por períodos e categorização temática, para demonstrar a relevância acadêmica desse tipo de investigação. Em contrapartida, o trabalho de Parnaíba (2020) apresenta limitações significativas de escopo, pois restringiu-se a apenas uma única instituição (UFPB), um único curso (Ciências Contábeis) e um período específico (2016.1 a 2019.1), além de utilizar coleta manual de dados através da base da Comissão de TCC. Essa abordagem, embora válida, limita a generalização dos resultados e a replicabilidade do estudo.

A pesquisa de Campos (2021) apresenta maior proximidade metodológica e tecnológica com o sistema proposto, uma vez que também desenvolve um *dashboard* interativo para análise de dados educacionais, por meio da aplicação de processos de ETL e técnicas de BI. Ambos os projetos reconhecem a importância da visualização de dados para subsidiar a tomada de decisões institucionais e utilizam ferramentas modernas de análise de dados. Porém, apesar de desenvolver um *dashboard* interativo, foca exclusivamente em dados administrativos do IFB (permanência, êxito, matrículas), com a ausência da análise de conteúdo acadêmico dos TCCs. Além disso, o trabalho enfrentou limitações relacionadas

à qualidade dos dados brutos, que impediram análises mais aprofundadas.

A plataforma "IFB em Números (2025)" converge com o sistema proposto no aspecto da transparência pública e da disponibilização estruturada de dados educacionais, ao demonstrar o potencial institucional e o reconhecimento oficial desse tipo de iniciativa no contexto dos Institutos Federais. Apesar de se configurar como uma referência em transparência, a plataforma não contempla a análise específica das temáticas dos trabalhos de conclusão de curso, além de apresentar a limitação de abranger exclusivamente dados do IFB, com ênfase em indicadores administrativos e acadêmicos institucionais de natureza geral.

O sistema em desenvolvimento apresenta contribuições específicas que o distinguem dos trabalhos correlatos. Primeiramente, a abrangência da coleta de dados se mostra significativamente maior, ao incluir as 42 instituições da Rede Federal. Essa amplitude possibilita análises comparativas entre diferentes instituições e a identificação de tendências nacionais nas temáticas de pesquisa. Além disso, a adoção da metodologia de *web scraping* representa um avanço tecnológico em relação aos métodos de coleta manual utilizados por Parnaíba (2020), por garantir maior escalabilidade, redução de erros humanos e atualizações periódicas dos dados. Essa abordagem também supera as limitações relacionadas aos dados brutos apontadas por Campos (2021), ao viabilizar a extração direta e estruturada das informações necessárias.

O foco específico na análise de conteúdo acadêmico dos TCCs, por meio do processamento de títulos, resumos e palavras-chave, preenche uma lacuna não contemplada pela plataforma "IFB em Números (2025)" e oferece uma perspectiva complementar sobre a produção acadêmica da Rede Federal. Adicionalmente, a estruturação do armazenamento dos dados e a aplicação de algoritmos de categorização temática automática posicionam o sistema como uma solução tecnologicamente avançada, com potencial de contribuição significativa para a análise e compreensão das áreas de conhecimento desenvolvidas pelos estudantes.

O sistema proposto integra os aspectos positivos observados nos trabalhos correlatos, como a relevância acadêmica da análise temática desenvolvida por Parnaíba (2020), a aplicação de tecnologias de visualização interativa no trabalho de Campos (2021) e o compromisso com a transparência institucional presente na plataforma "IFB em Números (2025)". Ao mesmo tempo, busca superar as limitações dessas iniciativas por meio de uma maior abrangência institucional, da automação na coleta de dados e do foco direcionado à análise de conteúdo acadêmico dos TCCs. Essa combinação resulta em uma contribuição original e relevante para a compreensão das tendências de pesquisa na Rede Federal de Educação Profissional, Científica e Tecnológica.

3 METODOLOGIA

A metodologia detalha a abordagem e os procedimentos planejados para o desenvolvimento da pesquisa e da solução proposta. A escolha e o delineamento metodológico são cruciais, uma vez que a metodologia pode ser compreendida como o estudo dos métodos e caminhos utilizados pelas ciências para alcançar o conhecimento. Ela representa uma preocupação instrumental da ciência, ao indicar os meios para captar a realidade e orientar a construção do saber. Assim, a escolha das metodologias e técnicas reflete a natureza do problema investigado e a necessidade de uma análise abrangente e aprofundada dos dados (Martins; Theóphilo, 2016).

3.1 Tipo e descrição geral da pesquisa

Quanto à finalidade, esta pesquisa se caracteriza como aplicada. Este tipo de estudo busca gerar conhecimento prático e uma ferramenta concreta que possa ser diretamente utilizada para solucionar um problema ou aprimorar uma prática existente, especificamente no desenvolvimento de *dashboards* interativos aplicados aos temas de Trabalhos de Conclusão de Curso. Conforme Gil (2022), a pesquisa aplicada tem como principal objetivo a resolução de problemas específicos e a aplicação de conhecimentos para o desenvolvimento de produtos e processos ou aprimoramento de práticas.

Em relação aos objetivos, a pesquisa classifica-se como descritiva e exploratória. Sua natureza descritiva reside na busca por caracterizar e quantificar os temas dos Trabalhos de Conclusão de Curso da Rede Federal, de modo a apresentar um panorama detalhado da distribuição e frequência desses assuntos ao longo do tempo e entre as diferentes instituições. Um estudo descritivo, segundo Richardson (2017), busca descrever sistematicamente uma situação, problema, fenômeno ou programa para revelar sua estrutura e comportamento. Isso é crucial para entender o cenário atual da produção de TCCs nas intituições que compõem a Rede Federal.

A pesquisa é, ao mesmo tempo, exploratória, dada sua natureza investigativa em um vasto e heterogêneo conjunto de dados acadêmicos, nomeadamente os metadados dos TCCs da Rede Federal, cuja análise em tal escala é ainda pouco consolidada. Conforme Richardson (2017), a pesquisa exploratória busca proporcionar maior familiaridade com o problema em estudo, pois torna-o mais claro e permite a formulação de novas hipóteses. Portanto,essa vertente exploratória é essencial para identificar e organizar as principais categorias temáticas dos TCCs, para que sirva como um ponto de partida para estudos mais detalhados e abrangentes.

Quanto à abordagem, adota-se uma perspectiva predominantemente quantitativa. Isso se justifica pela necessidade de coletar e analisar um grande volume de dados numé-

ricos e textuais (metadados de TCCs) que serão processados e convertidos para análise estatística e visualização. A abordagem quantitativa, como aponta Richardson (2017), permite a mensuração, a comparação de dados e a identificação de relações estatísticas. Complementarmente, o estudo incorpora elementos qualitativos, particularmente nas etapas de categorização temática e interpretação semântica dos conteúdos dos TCCs, uma vez que, segundo Richardson (2017), a pesquisa qualitativa possibilita uma compreensão mais profunda dos significados e contextos subjacentes aos dados, de modo a enriquecer a análise através da interpretação contextualizada dos fenômenos observados.

Quanto aos procedimentos técnicos, a pesquisa caracteriza-se como documental e bibliográfica. É documental pois utiliza TCCs e seus metadados como documentos primários extraídos de repositórios digitais, e bibliográfica pela revisão de literatura sobre dados abertos, mineração de texto e *Business Intelligence*. Conforme (Gil, 2022), a pesquisa documental utiliza materiais sem tratamento analítico prévio, enquanto a bibliográfica fundamenta-se em materiais já elaborados. O estudo também possui caráter técnico, com foco no desenvolvimento de sistema de *web scraping*, a estruturação de banco de dados e a criação de um *dashboard* interativo para visualização dos resultados.

3.2 Caracterização do objeto de estudo

O objeto central de estudo desta pesquisa são os Trabalhos de Conclusão de Curso defendidos nas diversas instituições que compõem a Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil. Esta rede abrange os Institutos Federais, a Universidade Tecnológica Federal do Paraná, os Centros Federais de Educação Tecnológica do Rio de Janeiro e de Minas Gerais, e o Colégio Pedro II. A escolha dos TCCs como objeto de estudo se justifica por sua representatividade como indicadores formais das áreas de interesse acadêmico, das competências desenvolvidas pelos estudantes e das tendências de pesquisa e desenvolvimento que emergem em cada instituição e na rede como um todo.

A análise terá foco nos metadados dos TCCs, como título, autor, ano de defesa, resumo e palavras-chave. Essas informações são cruciais por condensarem o conteúdo principal de cada trabalho e por estarem mais consistentemente disponíveis nos diferentes repositórios digitais. A coleta desses metadados busca abranger a totalidade das instituições da Rede Federal que disponibilizam tais informações, o que permite uma visão ampla e comparativa da produção acadêmica nacional. Este levantamento tem como objetivo criar um panorama inédito das linhas de pesquisa priorizadas e desenvolvidas no ensino profissional e tecnológico do país.

3.3 Instrumento de pesquisa

O principal instrumento para a coleta de dados neste estudo será um sistema de *web scraping*. Este sistema será desenvolvido especificamente para extrair, de forma

automatizada, os metadados dos TCCs disponíveis nos sistemas de cada instituição da Rede Federal vinculado ao Portal Integra (Educação, 2025). O *web scraping*, como técnica de extração programática de informações de *websites*, é justificado pela necessidade de coletar um grande volume de dados que estão dispersos em múltiplas plataformas com diferentes estruturas (Mitchell, 2018). Essa abordagem permite uma coleta eficiente e abrangente, fundamental para a análise de temas em escala nacional.

Na Tabela 1, apresentam-se os repositórios identificados por meio do portal Integra, os quais serviram como fonte primária para a extração dos metadados associados aos Trabalhos de Conclusão de Curso das instituições da Rede Federal. Diante da necessidade de mapeamento e verificação da disponibilidade desses repositórios, foi realizado, em 5 de setembro de 2025, um levantamento que abrangeu todas as instituições que, até essa data, mantinham seus acervos digitais integrados ao portal. Esses sistemas constituem, portanto, a base de dados fundamental utilizada na análise temática dos TCCs, o que viabiliza uma compreensão consolidada da produção acadêmica no contexto da Rede Federal de Educação Profissional, Científica e Tecnológica.

A Tabela 1 apresenta todas as instituições que compõem a Rede Federal, acompanhadas de seu respectivo estado, nome completo, sigla e URL.

Tabela 1 – **Instituições da Rede Federal e Disponibilidade de Dados no Portal Integra**

Estado	Instituição	Sigla	URL para a Biblioteca/TCCs
AC	Instituto Federal do Acre	IFAC	<https://integra.ifac.edu.br/>
AL	Instituto Federal de Alagoas	IFAL	<https://integra.ifal.edu.br/>
AP	Instituto Federal do Amapá	IFAP	<https://integra.ifap.edu.br/>
AM	Instituto Federal do Amazonas	IFAM	<https://integra.ifam.edu.br/>
BA	Instituto Federal da Bahia	IFBA	<https://integra.ifba.edu.br/>
DF	Instituto Federal de Brasília	IFB	<https://integra.ifb.edu.br/>
CE	Instituto Federal do Ceará	IFCE	<https://integra.ifce.edu.br/>
ES	Instituto Federal do Espírito Santo	IFES	<https://integra.ifes.edu.br/>
GO	Instituto Federal de Goiás	IFG	<https://integra.ifg.edu.br/>
GO	Instituto Federal Goiano	IFGOIANO	<https://integra.ifgoiano.edu.br/>
MA	Instituto Federal do Maranhão	IFMA	<https://integra.ifma.edu.br/>
MG	Instituto Federal de Minas Gerais	IFMG	<https://integra.ifmg.edu.br/>
MG	Instituto Federal do Norte de MG	IFNMG	<https://integra.ifnmg.edu.br/>
MG	Instituto Federal do Sudeste de MG	IFSUDESTEMG	<https://integra.ifsudestemg.edu.br/>
MG	Instituto Federal do Sul de MG	IFSULDEMINAS	<https://integra.ifsuldeminas.edu.br/>
MG	Instituto Federal do Triângulo Mineiro	IFTM	<https://integra.iftm.edu.br/>
MT	Instituto Federal de Mato Grosso	IFMT	<https://integra.ifmt.edu.br/>
MS	Instituto Federal de Mato Grosso do Sul	IFMS	<https://integra.ifms.edu.br/>
PA	Instituto Federal do Pará	IFPA	<https://integra.ifpa.edu.br/>
PB	Instituto Federal da Paraíba	IFPB	<https://integra.ifpb.edu.br/>
PE	Instituto Federal de Pernambuco	IFPE	<https://integra.ifpe.edu.br/>
PE	Instituto Federal do Sertão Pernambucano	IFSertãoPE	<https://integra.ifsertao-pe.edu.br/>
PI	Instituto Federal do Piauí	IFPI	<https://integra.ifpi.edu.br/>
PR	Instituto Federal do Paraná	IFPR	<https://integra.ifpr.edu.br/>
RJ	Instituto Federal do Rio de Janeiro	IFRJ	<https://integra.ifrj.edu.br/>
RJ	Instituto Federal Fluminense	IFFLUMINENSE	<http://integra.iff.edu.br/>
RN	Instituto Federal do Rio Grande do Norte	IFRN	<https://integra.ifrn.edu.br/>
RO	Instituto Federal de Rondônia	IFRO	<https://integra.ifro.edu.br/>
RR	Instituto Federal de Roraima	IFRR	<https://integra.ifrr.edu.br/>
RS	Instituto Federal do Rio Grande do Sul	IFRS	<https://integra.ifrs.edu.br/>
RS	Instituto Federal Farroupilha	IFFarroupilha	<https://integra.iffarroupilha.edu.br/>
RS	Instituto Federal Sul-rio-grandense	IFSul	<https://integra.ifsul.edu.br/>
SC	Instituto Federal de Santa Catarina	IFSC	<https://integra.ifsc.edu.br/>
SC	Instituto Federal Catarinense	IFC	<https://integra.ifc.edu.br/>
SP	Instituto Federal de São Paulo	IFSP	<https://integra.ifsp.edu.br/>
SE	Instituto Federal de Sergipe	IFS	<https://integra.ifs.edu.br/>
TO	Instituto Federal do Tocantins	IFTO	<https://integra.ifto.edu.br/>
RJ	Centro Federal de Educação Tecnológica Celso Suckow da Fonseca	CEFET-RJ	<https://integra.cefet-rj.br/>
MG	Centro Federal de Educação Tecnológica de Minas Gerais	CEFET-MG	<https://integra.cefetmg.br/>

Fonte: Elaborado pela autora.

As instituições apresentadas na Tabela 2 não estavam acessíveis nos sistemas digitais integrados ao Portal Integra na data de 5 de setembro de 2025, momento em que foi realizado o levantamento das URLs destinadas à coleta de dados. Em razão dessa indisponibilidade, tais instituições não foram incluídas nas etapas posteriores do processo de coleta.

Tabela 2 – **Instituições da Rede Federal com Dados Ausentes no Portal Integra**

Estado	Instituição	Sigla	Situação dos Dados
BA	Instituto Federal Baiano	IFBAI	Ausente
RJ	Colégio Pedro II	–	Ausente
PR	Universidade Tecnológica Federal do Paraná	UTFPR	Ausente

Fonte: Elaborado pela autora.

3.4 Coleta e análise de dados

O desenvolvimento do sistema de coleta de dados foi realizado em *Python* (Python, 2025), com a utilização de bibliotecas especializadas como *asyncio* (Asyncio, 2025) e

aiohttp (Aiohttp, 2025), as quais permitem a execução de requisições assíncronas e concorrentes aos *endpoints* de API JSON disponibilizados pelos repositórios digitais institucionais (Mitchell, 2018). Dessa forma, o sistema foi concebido de maneira flexível e robusta, de modo a se adaptar às variações nas estruturas de dados e às diferentes arquiteturas presentes entre os repositórios das instituições. Ademais, foram incorporados mecanismos específicos para enfrentar desafios de acesso, tais como limites de requisição, tratamento de exceções HTTP e inconsistências na disponibilidade dos dados. Com isso, busca-se maximizar a cobertura, a confiabilidade e a qualidade das informações coletadas, com o intuito de garantir uma base sólida para a etapa subsequente de análise.

Para assegurar a eficiência na implementação da coleta automatizada, será adotada uma estratégia de mapeamento sistemático dos dados disponíveis nas APIs dos repositórios das 39 instituições da Rede Federal de Educação Profissional, Científica e Tecnológica presentes no portal Integra. Desse modo, a coleta de dados levou em conta as particularidades de cada servidor, de forma a extrair metadados relevantes, tais como nome do autor, orientador, título, resumo, palavras-chave, área de conhecimento, curso e ano de publicação. O processo foi estruturado em duas etapas complementares, em que, na primeira, ocorreu a coleta dos metadados básicos dos docentes por meio do *endpoint* de listagem de professores e, na segunda, foi realizada a extração detalhada dos trabalhos de conclusão de curso vinculados a cada docente. Assim, os dados obtidos possibilitaram a categorização temática, a identificação de tendências e a comparação entre instituições, o que favorece interpretações estratégicas sobre a produção científica no âmbito da Rede Federal.

Os dados extraídos foram inicialmente armazenados em um banco de dados relacional *SQLite* (SQLite, 2025) de *staging*, o qual atuou como camada intermediária responsável pela persistência imediata das informações coletadas em seu formato original. Posteriormente, será executado um processo de ETL (*Extract, Transform, Load*) que, por sua vez, transformou esses dados brutos em um *Data Mart* analítico, também estruturado em *SQLite* (SQLite, 2025), em virtude de sua robustez, escalabilidade e suporte a operações analíticas avançadas. Além disso, a modelagem adotará o padrão *Star Schema*, composto por uma tabela fato central, dimensões bem definidas e tabelas ponte para representar relacionamentos muitos-para-muitos, o que possibilitará consultas otimizadas e análises detalhadas. Desse modo, conforme afirmam Elmasri e Navathe (2011), o modelo relacional oferece uma base conceitual sólida para a organização e o gerenciamento eficiente dos dados, de modo a assegurar a consistência, a integridade e a confiabilidade às informações processadas.

A etapa seguinte constitui no processamento e preparação dos dados para análise, por meio de técnicas de Processamento de Linguagem Natural (PLN). Nessa fase, foram aplicados procedimentos de normalização dos campos, identificação e correção de inconsistências com funções de validação e limpeza, além da categorização temática

baseada em palavras-chave, títulos e resumos dos TCCs. O pré-processamento textual incluiu tokenização, remoção de *stopwords* em português, exclusão de pontuação e caracteres não alfabéticos para garantir maior precisão semântica. Em seguida, os textos foram vetorizados pela técnica *Bag-of-Words* (BoW) com o *CountVectorizer* e submetidos à modelagem de tópicos por meio do algoritmo *Latent Dirichlet Allocation* (LDA), de modo a permitir a identificação automática dos principais temas da produção acadêmica. Dessa forma, a limpeza, transformação e o enriquecimento semântico dos dados tornam-se etapas indispensáveis para assegurar a coerência e a confiabilidade das análises. Conforme salientam Goldschmidt *et al.* (2015), a qualidade das descobertas depende diretamente da qualidade dos dados, o que reforça a necessidade de um tratamento criterioso antes da aplicação das técnicas de mineração.

Complementarmente à coleta e ao tratamento dos dados, o produto final da pesquisa consiste em um *dashboard* interativo, desenvolvido para viabilizar a visualização e a exploração dos metadados dos TCCs, após sua limpeza, categorização e agrupamento. Para o desenvolvimento deste dashboard, foi utilizado o *framework Streamlit* (Streamlit, 2025) em *Python* (Python, 2025), integrado com a biblioteca *Plotly* para geração de visualizações interativas, reconhecidos por sua capacidade de transformar dados em insights visuais e permitir análises exploratórias em tempo real. Os dados processados foram armazenados no formato *Apache Parquet* (Apache, 2025), escolhido por sua alta performance de leitura em aplicações analíticas. Com os dados devidamente estruturados, a investigação se volta para sua análise e interpretação, etapa em que reside o verdadeiro valor informacional, não na simples existência dos dados, mas em sua capacidade de responder de forma significativa às questões propostas pela pesquisa, por meio da aplicação criteriosa do raciocínio lógico (Marconi; Lakatos, 2021).

4 PROPOSTA

Este capítulo trata do desenvolvimento do projeto, com a apresentação os processos, métodos e etapas aplicados para a obtenção do resultado esperado com o artefato proposto. A elaboração da solução é conduzida de forma estruturada e documentada, ao considerar os elementos essenciais para seu planejamento e execução. A proposta metodológica está composta pelos seguintes itens: levantamento dos requisitos dos usuários, definição dos requisitos funcionais e não funcionais, elaboração do diagrama de casos de uso e apresentação das telas prototipadas da interface do sistema.

4.1 Requisitos

A especificação de requisitos constitui etapa fundamental no desenvolvimento de sistemas de informação, pois define as funcionalidades e características que o sistema deve possuir para atender às necessidades dos usuários (Sommerville, 2018). Segundo Pressman e Maxim (2021), os requisitos representam uma condição ou capacidade necessária para que um usuário resolva um problema ou alcance um objetivo, ou uma condição ou capacidade que deve ser atendida ou possuída por um sistema.

4.1.1 Requisitos de usuário

Os requisitos de usuário definidos foram:

RU01: O usuário deve poder visualizar a distribuição dos temas abordados nos TCCs da Rede Federal de forma clara e organizada, com opções de filtragem por instituição (IFs, CEFETs, UTFPR, Colégio Pedro II) , ano de defesa, cursos e temática.

RU02: O usuário deve poder pesquisar por TCCs específicos ou por temas relacionados por meio de palavras presentes nos títulos ou resumos dos trabalhos.

RU03: O usuário deve poder identificar e explorar tendências temáticas e *clusters* de temas, além de observar a evolução dos assuntos ao longo do tempo e as relações entre eles, para identificar nichos de pesquisa ou lacunas.

RU04: O usuário deve poder acessar os detalhes de um TCC específico, como título completo, autor, orientador, ano, instituição, curso, resumo e temática categorizada, ao interagir com a visualização.

RU05: O usuário deve poder exportar as informações completas dos TCCs, tais como título, autor, ano, instituição, resumo e palavras-chave.

RU06: O *dashboard* deve poder auxiliar os usuários (alunos e pesquisadores) na definição de temas de TCC, oferecer *insights* sobre áreas de interesse e lacunas na pesquisa acadêmica da Rede Federal.

RU07: O *dashboard* deve poder apoiar usuários (coordenadores de curso e professores orientadores) na análise da produção acadêmica, ao possibilitar a identificação de áreas de concentração e detectar lacunas em nível de curso ou instituição.

RU08: O *dashboard* deve poder subsidiar usuários (gestores institucionais e pesquisadores da educação) na tomada de decisões estratégicas, ao oferecer uma visão consolidada da produção científica da Rede Federal para planejamento e monitoramento de tendências macro.

4.1.2 Requisitos funcionais

Na Tabela 3, encontram-se os requisitos funcionais levantados no projeto, com sua respectiva correspondência aos requisitos de usuários atendidos:

Tabela 3 – **Requisitos Funcionais**

ID	Descrição do Requisito Funcional	RUs Atendidos
RF01	O sistema deve armazenar e gerenciar os metadados coletados em banco de dados estruturado e acessível, para garantir a integridade, disponibilidade e organização para análise posterior.	RU01, RU02, RU03, RU04, RU05
RF02	O sistema deve realizar a limpeza e o pré-processamento dos dados textuais, normalizar campos e padronizar os metadados dos TCCs.	RU01, RU02, RU03, RU05
RF03	O sistema deve categorizar e agrupar automaticamente os temas dos TCCs, por meio de algoritmos de clusterização, para facilitar a análise por área temática.	RU01, RU03, RU06
RF04	O sistema deve disponibilizar um *dashboard web* interativo para visualização dos dados temáticos dos TCCs, com recursos gráficos intuitivos, como gráficos de barras, linha, séries temporais ou mapas de calor, que evidenciem a distribuição, evolução e relações temáticas.	RU01, RU03, RU06, RU07, RU08
RF05	O *dashboard* deve permitir filtragem dos dados por múltiplas dimensões, como instituição, ano de defesa, cursos e áreas temáticas agrupadas.	RU01, RU06, RU07, RU08
RF06	O sistema deve oferecer funcionalidade de busca por termos específicos nos títulos e resumos dos trabalhos.	RU02, RU03
RF07	O *dashboard* deve permitir o acesso aos detalhes completos de um TCC, o que inclui título, autor, orientador, ano, instituição, curso, resumo e temática associada.	RU04, RU05
RF08	O sistema deve permitir a exportação dos dados completos dos TCCs disponíveis na análise, como título, autor, ano, instituição, resumo e temática, em formatos CSV.	RU05

Fonte: Elaborado pela autora.

4.1.3 Requisitos não funcionais

Na Tabela 4, apresenta os requisitos não funcionais, que especificam as qualidades do sistema e as restrições operacionais, de modo a complementar as funcionalidades detalhadas na Tabela 3.

Tabela 4 – **Requisitos Não Funcionais**

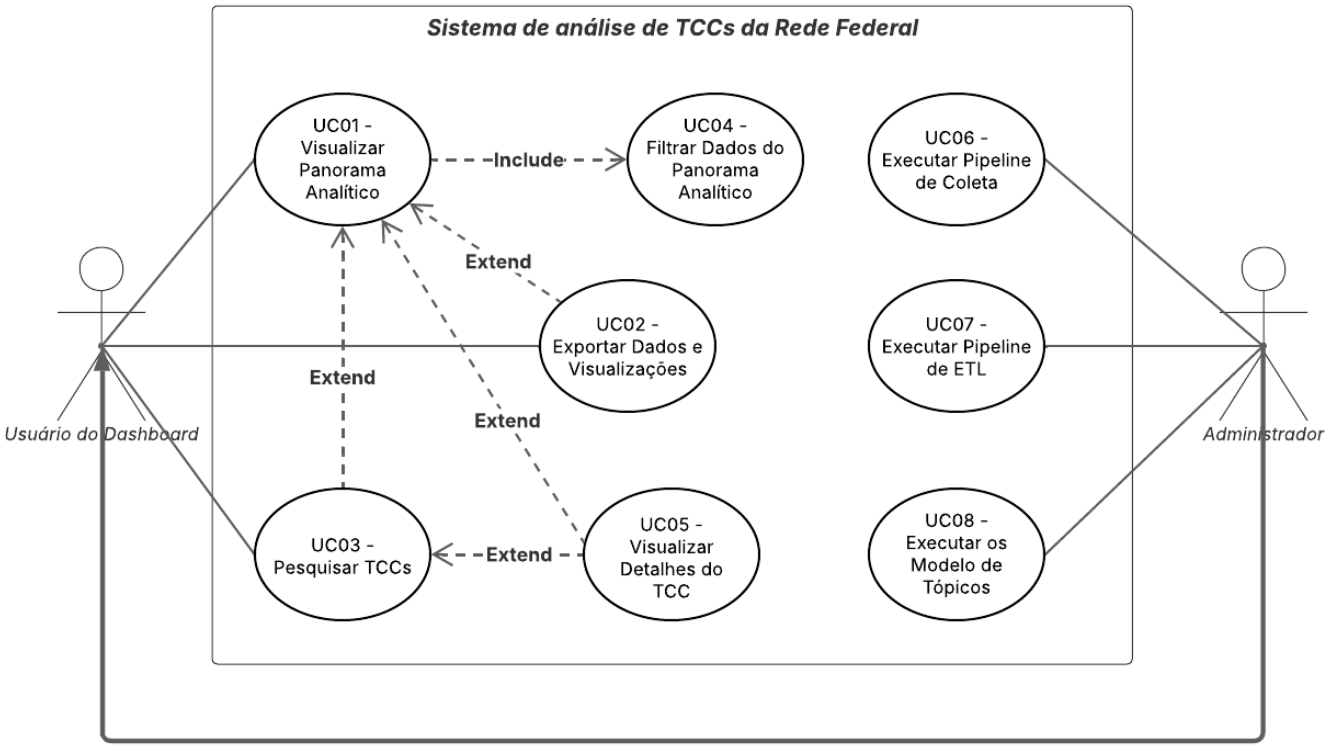
ID	Descrição do Requisito Não Funcional	RUs Atendidos
RNF01	O *dashboard* deve apresentar os dados e visualizações com um tempo de resposta máximo de 3 segundos para a maioria das interações (filtragens, pesquisas).	RU01, RU02, RU03, RU04, RU05, RU06, RU07
RNF02	O sistema deve garantir a disponibilidade do serviço em 99% do tempo, com a exclusão de janelas de manutenção planejadas.	RU01, RU05, RU06, RU07
RNF03	O *dashboard* deve ser usável e intuitivo, com uma interface clara e fácil de navegar para todos os perfis de usuário.	RU01, RU02, RU03, RU04, RU05, RU06, RU07
RNF04	O sistema deve ser escalável para suportar o crescimento do volume de dados (novos TCCs) e um número crescente de usuários simultâneos sem degradação significativa de desempenho.	RU01, RU07
RNF05	O sistema deve garantir a segurança dos dados coletados, com proteção contra acesso não autorizado, perda ou corrupção.	Implícito em todos os RUs (confiança nos dados)
RNF06	O sistema deve ser manutenível, com código bem documentado e arquitetura modular que permita futuras atualizações e correções de forma eficiente.	Implícito em todos os RUs (longevidade e confiabilidade do sistema)
RNF07	O *dashboard* deve ser compatível com os principais navegadores *web* (Chrome, Firefox, Edge, Safari) e responisvos, ou seja, com a interface adaptável a diferentes tamanhos de tela (*desktop*, *tablet*, *mobile*).	RU01, RU02, RU03, RU04, RU05, RU06, RU07
RNF08	O *dashboard* deve apresentar os dados com precisão, para que reflita fielmente as informações extraídas e processadas dos repositórios.	RU01, RU03, RU05, RU06, RU07

Fonte: Elaborado pela autora.

4.1.4 Diagrama de casos de uso

O diagrama de casos de uso é uma ferramenta fundamental na modelagem de sistemas orientados a objetos, pois permite representar, de forma clara e intuitiva, as funcionalidades esperadas do sistema sob a perspectiva dos usuários (atores). De acordo com Sommerville (2018), os casos de uso são descrições de interações típicas entre o sistema e seus usuários, e são úteis tanto para a validação de requisitos quanto para o planejamento de testes. A Figura 4 apresenta o diagrama de casos de uso do sistema desenvolvido neste trabalho a partir da utilização do Lucidchart (2025).

Figura 4 – **Diagrama de Casos de Uso do Sistema**



Fonte: Elaborado pela autora.

4.1.5 Protótipo de tela

A prototipação desempenha um papel fundamental no desenvolvimento de software, especialmente por facilitar a compreensão do sistema e o alinhamento entre a equipe técnica e os usuários logo nas etapas iniciais do projeto. De acordo com Pressman e Maxim (2021), o protótipo atua como uma representação funcional inicial do sistema, que possibilita aos usuários e desenvolvedores a visualização e a validação das funcionalidades antes da construção final. Essa abordagem contribui para a identificação precoce de falhas, a redução de retrabalho e a definição clara das expectativas dos *stakeholders*. A figura 5 apresenta o protótipo do sistema desenvolvido no Projeto de Trabalho de Conclusão de Curso (PTCC) com a ferramenta de prototipagem Figma (2025).

Figura 5 – **Protótipo de Tela**

Fonte: Elaborado pela autora.

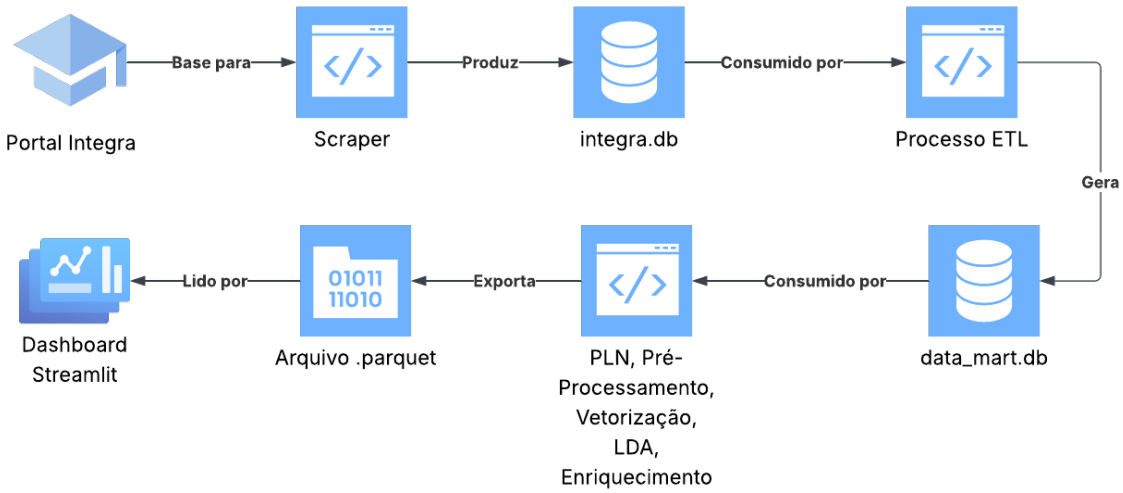
5 ARQUITETURA DO SISTEMA

Este capítulo apresenta os processos de engenharia e desenvolvimento de software empregados na construção do artefato proposto, com ênfase na implementação prática do sistema de análise temática de TCCs. A abordagem metodológica adotada para a condução da pesquisa, bem como os requisitos funcionais e o design conceitual do sistema, foram previamente definidos e detalhados em capítulos anteriores. Dessa forma, esta seção dedica-se à descrição da execução técnica do projeto, com o intuito de contemplar a arquitetura do sistema, as tecnologias empregadas, as etapas que compõem o *pipeline* de dados, as decisões de implementação e os desafios técnicos enfrentados ao longo do desenvolvimento.

5.1 Arquitetura Geral do Sistema

A Figura 6 apresenta a arquitetura geral do sistema, estruturada como um *pipeline* de processamento de dados que integra coleta, transformação, análise semântica e visualização interativa. O diagrama de componentes evidencia o papel de cada módulo e a forma como os artefatos gerados são encadeados até a construção do *dashboard* analítico.

Figura 6 – **Diagrama do Sistema**



Fonte: Elaborado pela autora.

A solução foi implementada como um *pipeline* de processamento de dados *end-to-end*, modularizado em etapas distintas e independentes para garantir manutenibilidade, escalabilidade e facilitar a depuração de problemas em cada fase. A linguagem de programação *Python* (Python, 2025), na versão 3.11, foi adotada como base tecnológica principal devido à sua vasta disponibilidade de bibliotecas especializadas em ciência de dados,

processamento de linguagem natural e desenvolvimento de aplicações web, além de sua sintaxe expressiva que facilita a prototipagem rápida e a manutenção do código.

A arquitetura lógica do sistema segue o padrão de *pipeline* de dados em múltiplas camadas, no qual cada módulo possui responsabilidades bem definidas e comunica-se com os demais por meio de interfaces de dados padronizadas. Nesse sentido, essa abordagem arquitetural favorece a modularidade e a manutenção do sistema, de modo que cada componente seja desenvolvido, testado e otimizado de forma independente. Dessa forma, futuras expansões ou modificações tornam-se mais simples e seguras de implementar. Assim, o sistema é estruturado a partir dos seguintes módulos principais:

- **Módulo de Coleta de Dados por *Web Scraping*:** Componente assíncrono responsável pela extração automatizada de dados brutos dos repositórios digitais institucionais. Este módulo implementa lógica de navegação web, *parsing* de HTML, tratamento de erros de rede e mecanismos de retentativas para garantir robustez na coleta de dados de múltiplas fontes heterogêneas.
- **Módulo de Persistência Bruta na Camada de *Staging*:** Camada intermediária de armazenamento que utiliza um banco de dados relacional *SQLite* (SQLite, 2025), *integra.db*, para persistência temporária dos dados coletados em seu formato bruto, sem transformações. Esta camada de *staging* desacopla temporalmente a fase de coleta da fase de transformação, de modo que o *scraper* execute de forma independente e que o processo ETL possa ser executado quantas vezes forem necessárias sem necessidade de nova coleta. Adicionalmente, esta camada funciona como checkpoint de recuperação em caso de falhas no pipeline.
- **Módulo de Transformação de Dados com Processo de *Extract, Transform, Load*:** Possui *script* especializado responsável pela aplicação sistemática de regras de negócio, limpeza de dados, validação de qualidade, normalização de valores e modelagem dimensional dos dados. Este módulo consome os dados brutos do *staging* e produz um *Data Mart* analítico estruturado, nomeado como *datamart.db*, otimizado para consultas e análises, que seguem o modelo dimensional *Star Schema*.
- **Módulo de Análise e Enriquecimento com Técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina:** Etapa de processamento analítico avançado onde os dados textuais estruturados são submetidos a técnicas de Processamento de Linguagem Natural (PLN) e *Machine Learning*. Este módulo realiza pré-processamento textual, vetorização de documentos, modelagem de tópicos através do algoritmo LDA e enriquecimento dos dados com informações derivadas para que sejam preparados para consumo na camada de visualização.
- **Módulo de Visualização e Interação com Dashboard Interativo:** Interface gráfica de usuário final, desenvolvida com o *framework Streamlit* (Streamlit, 2025), que consome os dados processados e enriquecidos para apresentação através de visualizações interativas, filtros dinâmicos e funcionalidades analíticas avançadas. Este módulo

implementa a lógica de apresentação, interação com o usuário e renderização de gráficos e tabelas.

Os módulos comunicam-se por meio de dados persistidos, o que garante a integridade e a continuidade das informações. Nessa estrutura, o *scraper* grava os dados no arquivo *integra.db*, que é posteriormente lido pelo módulo de ETL, responsável por realizar as transformações e armazená-las em *datamart.db*. Em seguida, o módulo de Processamento de Linguagem Natural (PLN) acessa esse repositório e gera um documento no formato *parquet* (Apache, 2025). Por fim, o *dashboard* consome esse arquivo para exibição e análise interativa dos resultados. Dessa forma, a adoção de uma arquitetura baseada em artefatos persistidos contribui para a solidez do sistema e garante a rastreabilidade completa de todas as etapas do fluxo de dados.

5.2 Processamento ETL e Modelagem Dimensional

A transformação dos dados brutos coletados em um *Data Mart* analítico estruturado representa uma etapa crítica do *pipeline* de processamento, pois assegura a qualidade, a consistência e a adequação das informações às demandas analíticas do sistema. Essa fase foi desenvolvida por meio de um *script* dedicado, o qual materializa os princípios do processo de *Extract, Transform and Load* (ETL) e articula um conjunto integrado de operações de extração, transformação e carga de dados.

Do ponto de vista tecnológico, a implementação fundamenta-se em duas bibliotecas centrais do ecossistema *Python* (Python, 2025) voltadas à ciência de dados. A primeira delas, *Pandas* (Pandas, 2025), é responsável pela manipulação eficiente de dados tabulares em memória, por meio da utilização de estruturas do tipo *DataFrame*. A segunda, *SQLAlchemy* (SQLAlchemy, 2025), fornece uma camada de abstração que facilita a interação com diferentes sistemas de gerenciamento de bancos de dados relacionais. A integração dessas bibliotecas possibilita a execução de transformações de dados com elevado desempenho, ao mesmo tempo em que assegura a portabilidade entre distintos ambientes e mecanismos de armazenamento.

5.2.1 Fase de Extração

A etapa de coleta de dados constitui o ponto de entrada do *pipeline* e foi implementada por meio de um *web scraper* customizado desenvolvido especificamente para extrair informações dos repositórios digitais das instituições da Rede Federal disponíveis no portal Integra. A partir disso, o uso de um mesmo portal possibilitou abstrair a complexidade de lidar diretamente com distintos repositórios institucionais, cada um com seu próprio layout HTML e organização de informações, ao mesmo tempo em que demandou o desenvolvimento de uma lógica de *parsing* flexível e robusta para tratar eventuais variações estruturais.

5.2.1.1 Arquitetura Assíncrona de Coleta

A fim de otimizar o desempenho da coleta de dados provenientes das instituições com repositórios ativos no portal Integra, adotou-se uma arquitetura híbrida, que combina abordagens síncrona e assíncrona, fundamentada no paradigma de programação concorrente. As bibliotecas *asyncio* (Asyncio, 2025) e *aiohttp* (Aiohttp, 2025) foram utilizadas como base para a implementação, pois oferecem suporte nativo a operações assíncronas e à manipulação eficiente de requisições HTTP.

O processo foi projetado de modo que a coleta das instituições ocorra de forma sequencial, uma vez que cada repositório apresenta estrutura e pontos de acesso distintos. No entanto, dentro de cada instituição, a coleta dos dados dos professores e, subsequentemente, dos Trabalhos de Conclusão de Curso (TCCs) orientados é realizada de maneira assíncrona. Essa estratégia permite que múltiplas páginas sejam processadas simultaneamente, de modo a aproveitar o tempo ocioso entre as operações de entrada e saída (I/O), o que resulta em ganho expressivo de eficiência.

A execução assíncrona viabiliza o envio e o processamento concorrente de diversas requisições, sem bloquear o fluxo principal do programa. Em contraste com uma abordagem estritamente síncrona, na qual cada requisição precisa aguardar a resposta completa do servidor antes de iniciar a próxima, o modelo assíncrono permite a sobreposição de operações, de modo que outras tarefas continuem sendo executadas enquanto as respostas são recebidas. Essa característica é especialmente vantajosa em cenários *I/O bound*, como no caso do *web scraping*, em que a maior parte do tempo de execução é consumida na espera por respostas de rede.

Dessa forma, os ganhos de desempenho são substanciais. Em uma abordagem sequencial pura, o tempo total de execução seria proporcional a $N \times T$, em que N representa o número de instituições e T o tempo médio de coleta por instituição. Ao introduzir concorrência com grau C, a coleta assíncrona de múltiplos professores e TCCs reduz o tempo total aproximado para $(N/C) \times T$, com a consideração eventuais *overheads* de coordenação e sincronização.

5.2.1.2 Etapas do Processo de Coleta

O processo de *scraping* foi estruturado em duas fases hierárquicas e complementares:

- **Coleta de professores:** nesta fase inicial, o *scraper* acessa a página de listagem de docentes de cada instituição e extrai informações essenciais, como nome completo, identificador único no sistema e URL da página individual do professor. Essa etapa constrói um índice abrangente de orientadores potenciais, que serve como base para a coleta subsequente. A implementação utiliza técnicas de *HTML parsing* por meio da biblioteca *BeautifulSoup* (BeautifulSoup, 2025), de modo que sejam identificados

elementos estruturais específicos com o uso de seletores CSS e expressões *XPath*.

- **Coleta de detalhes dos TCCs:** a partir da lista de professores obtida na fase anterior, o sistema percorre as páginas individuais de cada docente para extrair os metadados completos de todos os Trabalhos de Conclusão de Curso orientados. Entre os metadados coletados estão o título do trabalho, o resumo, o ano de defesa, os autores, a instituição, o campus, o curso e o respectivo endereço eletrônico. Esta fase é computacionalmente mais intensiva, uma vez que pode envolver o processamento de centenas ou milhares de páginas individuais.

5.2.1.3 Estratégias de Robustez e Tratamento de Erros

Ao que se refere a variabilidade estrutural dos repositórios e a natureza imprevisível das operações de rede, o *scraper* incorpora múltiplas estratégias de robustez, entre as quais destacam-se:

- **Tratamento de exceções:** utilização de blocos *try-except* abrangentes para capturar e tratar erros específicos, como *timeouts*, falhas HTTP (404 e 500), erros de *parsing* e exceções inesperadas;
- **Mecanismos de *retry*:** repetição automática de requisições mal-sucedidas, com política de *backoff* exponencial;
- ***Logging* detalhado:** registro estruturado de eventos e erros para o rastreamento e a auditoria do processo de coleta;
- ***Rate limiting*:** aplicação de atrasos controlados entre requisições, com o intuito de prevenir sobrecarga e bloqueios por detecção de atividade automatizada.

5.2.1.3.1 Persistência na Camada de *Staging*

Os dados brutos extraídos são armazenados em um banco de dados *SQLite* (SQLite, 2025) (*integra.db*), gerenciado pela classe *DatabaseManager*. O Modelo Entidade-Relacionamento e o dicionário de dados que descrevem essa base estão apresentados, respectivamente, nos Apêndices A e B. A estrutura da tabela de *staging* foi projetada para acomodar dados em formato semiestruturado, com o intuito de garantir flexibilidade e rastreabilidade. Dessa forma, essa camada revelou-se essencial para garantir a atomicidade das operações, o desacoplamento temporal entre as etapas de coleta e transformação, além de assegurar a rastreabilidade integral dos dados brutos.

5.2.1.3.2 Extração para o ETL

A fase de extração inicia o processo ETL a partir da recuperação dos dados brutos previamente persistidos na *staging area*. O *script* estabelece conexão com o banco de dados *integra.db* por meio do *engine* do *SQLAlchemy* (SQLAlchemy, 2025) e executa uma consulta SQL completa sobre a tabela *tccs*, a fim de materializar o conjunto de dados

resultante em um *DataFrame Pandas* (Pandas, 2025) carregado integralmente na memória RAM.

Essa abordagem de extração total, também conhecida como *full extraction*, embora demandante em termos de recursos computacionais, mostrou-se adequada ao escopo do projeto, ao considerar o volume moderado de dados típico de repositórios institucionais acadêmicos. Dessa forma, a estratégia elimina a complexidade de implementar mecanismos de extração incremental, como *change data capture*, de modo que a lógica de processamento seja simplicada e garanta que cada execução do ETL reflita o estado completo e atualizado dos dados coletados.

O *DataFrame* resultante preserva integralmente a estrutura original da tabela de *staging*, de modo que todos os campos coletados sejam abrangidos. Entre esses campos, incluem-se os identificadores dos professores, os metadados institucionais, como sigla, instituição, unidade federativa e campus, além dos metadados temporais referentes ao ano, das informações de curso, dos dados de autoria que distinguem orientando e orientador, bem como dos campos textuais fundamentais à análise, tais como título, resumo e palavras-chave.

5.2.2 Fase de Transformação

A fase de transformação representa o núcleo do processo ETL, visto que aplica sistematicamente um conjunto abrangente de regras de negócio, procedimentos de limpeza, validações de qualidade e operações de modelagem de dados. Por conseguinte, essa etapa foi estruturada em cinco subfases distintas, cada uma responsável por aspectos específicos da preparação dos dados.

5.2.2.1 Validação de Integridade Institucional

A primeira subfase do processo foi responsável pela implementação de mecanismos de validação destinados a assegurar que apenas os trabalhos efetivamente vinculados às instituições que compõem o escopo deste estudo fossem processados. Com isso, foi utilizada uma função específica para realizar verificações heurísticas sobre o campo referente às instituições com o intuito de identificar expressões-chave como “instituto federal" ou as siglas específicas das instituições coletadas, a exemplo de “IFB” e “IFSP”.

Essa etapa de verificação revelou-se essencial diante das inconsistências observadas nos repositórios institucionais, em que trabalhos oriundos de outras organizações educacionais apareciam associados a perfis de docentes da Rede Federal. Para mitigar esse problema, os registros que não atendem aos critérios estabelecidos são descartados e devidamente registrados em um arquivo, o qual armazena informações detalhadas em um arquivo de *log* estruturado. Essa estratégia garante a rastreabilidade das exclusões para possíveis auditorias posteriores e o aprimoramento contínuo das regras de validação.

5.2.2.2 Limpeza e Padronização Textual

A segunda subfase concentrou-se na normalização e padronização dos campos textuais, etapa essencial para assegurar a consistência dos dados e possibilitar análises fundamentadas em técnicas de PLN. Nessa fase, foram implementadas duas funções principais, responsáveis por estruturar o processo de limpeza e uniformização textual.

A primeira função executa um encadeamento de transformações sobre cadeias de caracteres, o qual inclui a remoção de acentuação por meio da normalização *Unicode* na forma NFD, a conversão de todas as letras para minúsculas, a correção ortográfica automática e a eliminação de espaços duplicados. Essa normalização é aplicada tanto a campos institucionais quanto descritivos, o que assegura maior coerência e padronização textual em todo o conjunto de dados.

Por sua vez, a segunda função implementa uma lógica de capitalização inteligente voltada à correta formatação de nomes próprios, de modo a preservar conectores e preposições recorrentes na língua portuguesa, como “de”, “da”, “dos” e “e”. Essa função é empregada principalmente nos campos correspondentes aos nomes de discentes, orientadores e denominações de campus, o que contribui para a padronização e legibilidade das informações apresentadas no *dashboard* analítico final.

5.2.2.3 Derivação e Enriquecimento de Atributos

A terceira subfase tem como objetivo derivar novos atributos a partir das informações já existentes, de modo a enriquecer o modelo analítico com dados complementares não estruturados na origem. Nessa etapa, são aplicadas regras semânticas e heurísticas textuais voltadas à identificação e segmentação de entidades presentes nos campos brutos, com destaque para a distinção entre os autores dos trabalhos e seus respectivos orientadores.

O processo de extração baseia-se na análise de padrões recorrentes no texto, como a presença de expressões indicativas de vínculo docente, que permitem isolar o nome do orientador em relação aos demais colaboradores do trabalho. Essa separação aprimora significativamente a qualidade dos metadados e assegura maior precisão na representação das relações acadêmicas.

A derivação dessas informações complementares é essencial para a construção de uma base de dados analítica consistente, pois possibilita o estabelecimento de vínculos adequados entre trabalhos, discentes e docentes. Como resultado, o modelo passa a oferecer suporte a análises avançadas sobre a produtividade de orientação, o perfil de colaboração entre professores e a distribuição das orientações por curso, campus e instituição.

5.2.2.4 Construção de Tabelas Dimensionais

A quarta subfase representa a etapa de transição entre a estrutura desnormalizada proveniente da *staging area* e um modelo dimensional consolidado, preparado para consultas analíticas de alta eficiência. Nessa fase, os dados são reorganizados conforme os princípios do *Star Schema*, de modo a reduzir redundâncias, otimizar o desempenho de leitura e facilitar a interpretação das relações entre entidades. O Modelo Entidade-Relacionamento que fundamenta essa reorganização encontra-se descrito no Apêndice A, enquanto o dicionário de dados correspondente é apresentado no Apêndice B.

Com base nesse modelo, foram gerados *DataFrames* específicos para cada dimensão analítica, de modo a assegurar granularidade adequada e coerência referencial em toda a base:

- **dim_campus:** formada a partir da combinação entre campus, sigla institucional e unidade federativa, com a criação de chaves substitutas (*surrogate keys*) para garantir unicidade e independência lógica em relação aos identificadores originais;
- **dim_curso:** composta pelos valores únicos de cursos, submetidos a processos de normalização textual e consolidação de grafias equivalentes, a fim de eliminar duplicidades e inconsistências semânticas;
- **dim_pessoa:** integra em uma única estrutura as entidades correspondentes a alunos e orientadores, diferencia-as por categoria funcional e estabelece vínculo direto com o campus de origem, o que possibilita análises cruzadas entre o perfil docente e a distribuição de orientações.

Cada dimensão foi projetada com uma chave primária do tipo *surrogate key* e um conjunto de atributos descritivos que refletem as principais características das entidades modeladas. Essa abordagem confere ao sistema maior consistência semântica, além de simplificar futuras expansões do modelo e a integração com a tabela fato responsável por consolidar as métricas analíticas.

5.2.2.5 Geração de Estruturas de Mapeamento

A quinta subfase é responsável pela geração de estruturas auxiliares de mapeamento que asseguram a coerência entre os dados textuais e as tabelas dimensionais previamente construídas. Nessa etapa, são criadas tabelas de correspondência entre valores textuais originais e as chaves primárias de suas respectivas dimensões.

Esses mapeamentos são aplicados ao *DataFrame* principal de Trabalhos de Conclusão de Curso, de modo a promover a substituição dos valores textuais por identificadores numéricos correspondentes às chaves estrangeiras. Essa transformação reforça a integridade referencial entre as entidades e otimiza tanto o armazenamento quanto o desempenho das consultas analíticas, uma vez que reduz a redundância e padroniza as relações entre os conjuntos de dados.

5.2.3 Fase de Carga

A fase de carga representa a etapa de materialização física do modelo dimensional no banco de dados de destino *datamart.db*, o que permite a consolidação de todas as transformações realizadas nas etapas anteriores. Cada *DataFrame* dimensional, bem como a tabela fato, é persistido por meio do método *to_sql* da biblioteca *Pandas* (Pandas, 2025), configurado com o parâmetro *if_exists='replace'*, o que assegura a idempotência do processo e elimina a possibilidade de duplicidade de registros.

O modelo *Star Schema* estrutura-se em componentes relacionais bem definidos:

- **Tabela Fato *fato_tcc*** contém os fatos mensuráveis associados a cada Trabalho de Conclusão de Curso e mantém as referências às dimensões correspondentes por meio de chaves estrangeiras;
- **Tabelas Dimensionais** *dim_instituicao*, *dim_campus*, *dim_curso* e *dim_pessoa* armazenam os atributos descritivos que fornecem o contexto analítico necessário às consultas;
- **Tabelas Ponte** *ponte_tcc_aluno* e *ponte_tcc_orientador* resolvem as relações muitos para muitos entre TCCs e pessoas, de modo a garantir consistência e flexibilidade analítica.

Essa modelagem dimensional, embasada nos conceitos de Kimball discutidos na fundamentação teórica, aprimora de forma significativa a eficiência das consultas analíticas executadas pelo *dashboard*. A desnormalização controlada característica do modelo *Star Schema*, associada à adoção de tabelas ponte, possibilita consultas agregadas de alta performance com junções simplificadas. Essa combinação assegura a responsividade da interface e a escalabilidade do sistema mesmo diante do crescimento contínuo do volume de dados.

5.3 Análise Textual e Modelagem de Tópicos

A etapa de análise textual e modelagem de tópicos representa o componente de inteligência artificial do sistema, responsável por extrair automaticamente estruturas temáticas latentes a partir dos conteúdos textuais dos trabalhos de conclusão de curso. Esta fase implementa técnicas avançadas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina não supervisionado, o que enriquece os dados estruturados do *Data Mart* com classificações temáticas que viabilizam análises de alto nível sobre tendências e padrões de pesquisa.

5.3.1 Preparação da Visão Analítica

O *pipeline* de PLN inicia com a preparação de uma visão analítica consolidada dos dados, processo executado por uma função que fica responsável pela disponibilização dos dados no *Data Mart*. Esta função estabelece conexão com o banco de dados *datamart.db* e

executa uma consulta SQL complexa que desnormaliza o modelo *Star Schema* através de operações de junção (JOIN) entre a tabela fato e suas dimensões associadas.

A desnormalização recupera todos os atributos descritivos necessários para análise, como instituição, campus, curso, ano e orientadores, de modo a associá-los aos campos textuais centrais título, resumo e palavras-chave de cada TCC. Uma particularidade importante desta consulta é o tratamento dos relacionamentos muitos-para-muitos através da função de agregação, que consolida múltiplos autores de um mesmo trabalho em um campo textual concatenado para preservar a informação de autoria múltipla em formato adequado para processamento posterior.

O resultado desta operação é um *DataFrame Pandas (Pandas, 2025)* que contém uma visão plana e completa de cada TCC, otimizada para as operações subsequentes de análise textual. Esta abordagem de desnormalização, embora introduza certa redundância temporária em memória, simplifica significativamente a lógica de processamento e elimina a necessidade de múltiplas consultas ao banco de dados durante a fase analítica.

5.3.2 Pipeline de Pré-processamento Textual

O pré-processamento textual constitui etapa fundamental para garantir a qualidade e eficácia dos algoritmos de PLN, ao remover ruídos, padronizar representações e reduzir a dimensionalidade do espaço linguístico. O *pipeline* implementado segue práticas consolidadas da área e aplica uma sequência de transformações sobre os textos.

5.3.2.1 Consolidação do Corpus Textual

A primeira operação do pré-processamento consiste na criação de um campo unificado *texto_completo*, resultante da concatenação dos campos *titulo* e *resumo* de cada TCC. Esta decisão de design fundamenta-se no princípio de que tanto o título quanto o resumo contêm informações temáticas complementares e relevantes, em que o o título tipicamente expressa de forma sintética o foco principal da pesquisa, enquanto o resumo desenvolve de maneira mais detalhada os objetivos, metodologia e contribuições do trabalho.

A concatenação destes campos amplia o contexto textual disponível para a modelagem de tópicos, o que enriquece a representação semântica de cada documento e aumenta a robustez do modelo frente a resumos excessivamente concisos ou títulos ambíguos. Embora o campo *palavras_chave* também contenha informação temática relevante, sua inclusão foi criteriosamente avaliada ao considerar seu formato estruturado e densidade semântica diferenciada, que será objeto de análises específicas em outros módulos do sistema.

5.3.2.2 Transformações Linguísticas

O texto consolidado é submetido a um *pipeline* de transformações linguísticas aplicadas de forma sequencial, com o objetivo de prepará-lo para a vetorização:

- **Normalização de Caixa:** A primeira transformação aplica o método *.lower()* e converte todos os caracteres para minúsculas. Esta normalização é essencial para garantir que palavras como “Educação”, “EDUCAÇÃO” e “educação” sejam tratadas como *tokens* idênticos.
- **Remoção de Elementos não-alfabéticos:** Uma expressão regular é aplicada para remover pontuação, numerais e caracteres especiais, ainda com os espaços mantidos para delimitar *tokens*.
- **Tokenização:** O texto limpo é segmentado em *tokens* individuais através do método *.split()*, que utiliza espaços em branco como delimitadores.
- **Filtragem de *Stopwords*:** Utiliza-se a lista de *stopwords* do NLTK (NLTK, 2025) para o idioma português (*nltk.corpus.stopwords.words('portuguese')*) com o intuito de remover termos de alta frequência e baixo valor semântico.
- **Filtragem por Comprimento:** *Tokens* com menos de três caracteres são removidos para eliminar fragmentos e siglas desnecessárias.

Sendo assim, o texto pré-processado resultante é uma sequência limpa de *tokens* lexicalmente significativos, otimizada para servir como entrada aos algoritmos de vetorização e modelagem de tópicos subsequentes.

5.3.3 Vetorização Textual com o Modelo Bag-of-Words

A vetorização textual converte cada documento em uma representação numérica de dimensionalidade fixa, o que permite que algoritmos de aprendizado de máquina processem e identifiquem padrões linguísticos de forma estruturada. Nesta etapa, adota-se o modelo *Bag-of-Words* (BoW), implementado por meio da classe *CountVectorizer* da biblioteca *Scikit-learn* (Scikit-learn, 2025), cuja função consiste em quantificar a frequência de ocorrência dos termos em todo o corpus e construir a matriz termo-documento correspondente.

A configuração do *CountVectorizer* foi ajustada de modo a equilibrar abrangência vocabular e relevância estatística dos termos, conforme os seguintes hiperparâmetros:

- **max_df = 0.9** ignora termos excessivamente frequentes, presentes em mais de 90% dos documentos, por não contribuírem para a distinção temática;
- **min_df = 20** considera apenas palavras que ocorrem em, no mínimo, vinte documentos distintos, de modo a evitar ruídos e termos esparsos;
- **max_features = 2000** restringe o vocabulário às duas mil palavras mais representativas do corpus, o que assegura balanceamento entre desempenho e expressividade;
- **ngram_range = (1, 2)** inclui unigramas e bigramas, o que possibilita capturar expressões compostas e termos técnicos recorrentes, como “inteligência artificial” e

"aprendizado supervisionado".

Como resultado dessa etapa, obtém-se uma matriz termo-documento de alta dimensionalidade, estruturada de forma otimizada para armazenar apenas as ocorrências relevantes dos termos. Nessa matriz, de dimensão $N \times 2000$, N representa o número total de Trabalhos de Conclusão de Curso processados. Essa representação vetorial, por sua vez, constitui a base quantitativa da modelagem de tópicos, pois viabiliza a análise semântica e a identificação de padrões latentes no conjunto textual.

5.3.4 Modelagem de Tópicos via LDA

A descoberta automática das estruturas temáticas é conduzida pelo algoritmo *Latent Dirichlet Allocation* (LDA), o qual pressupõe que cada documento constitui uma combinação de múltiplos tópicos, enquanto cada tópico se caracteriza por uma distribuição probabilística sobre um conjunto de palavras. A implementação utiliza a classe *LatentDirichletAllocation* da biblioteca *Scikit-learn*(Scikit-learn, 2025), configurada com parâmetros ajustados para garantir equilíbrio entre desempenho e interpretabilidade. Os principais parâmetros empregados são:

- **n_components = 10** define o número de tópicos latentes para assegurar a granularidade adequada das categorias identificadas;
- **random_state = 42** estabelece um ponto fixo de inicialização que garante a reprodutibilidade dos resultados;
- **n_jobs = -1** habilita a execução paralela em todos os núcleos disponíveis, o que promove maior eficiência computacional durante a inferência.

O processo de modelagem resulta em duas matrizes fundamentais. A primeira expressa a relação entre tópicos e palavras, com dimensão $K \times 2000$, e a segunda descreve a distribuição de tópicos em cada documento, com dimensão $N \times 10$. Ambas constituem a base para as análises temáticas subsequentes e fornecem o alicerce para a etapa de interpretação dos tópicos.

5.3.5 Atribuição e Nomenclatura de Tópicos

Para cada Trabalho de Conclusão de Curso, o tópico principal é identificado a partir do maior valor de probabilidade na distribuição documento-tópico, obtido por meio do método *argmax axis=1*. É utilizada uma função que analisa os termos de maior peso em cada tópico e atribui rótulos interpretáveis, como "Educação e Tecnologias Digitais" ou "Redes e Segurança da Informação". Apesar de envolver certo grau de subjetividade, esse processo é essencial para tornar os resultados compreensíveis e proporcionar informações relevantes aos usuários do sistema.

5.3.6 Persistência do Dataset Enriquecido

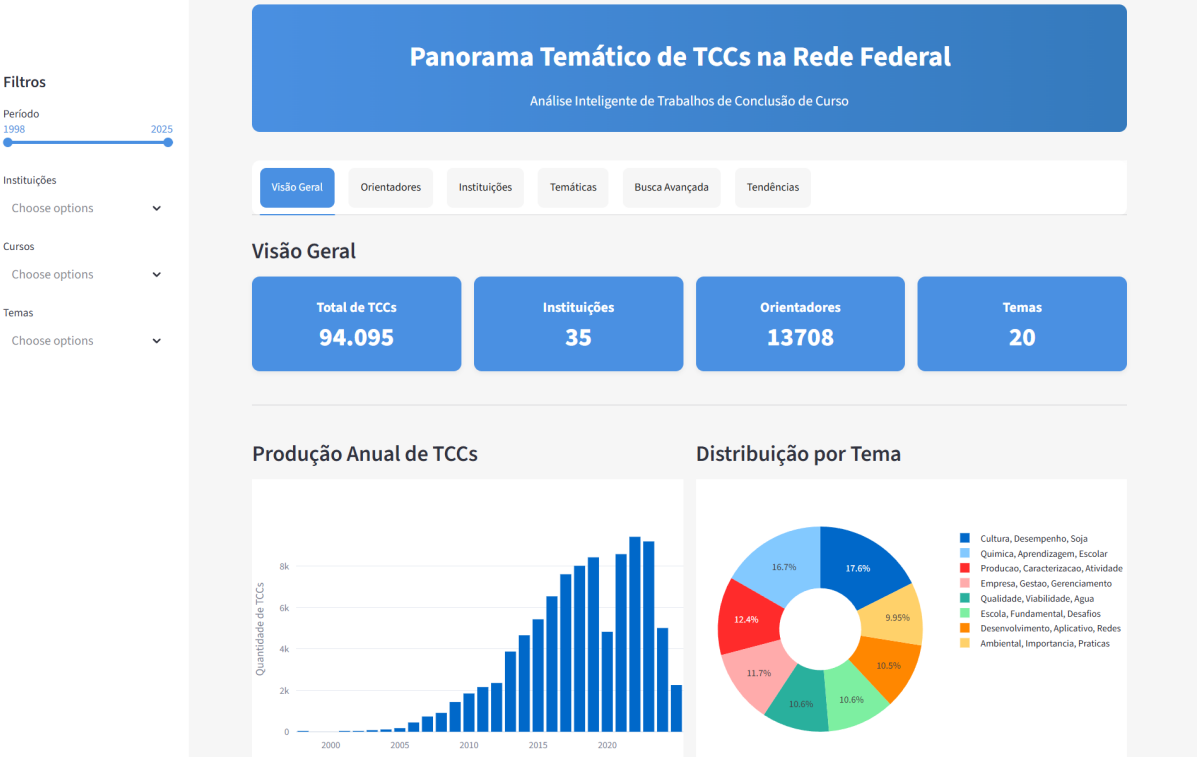
O *DataFrame* final, enriquecido com o campo *nome_topico*, é persistido no formato *Apache Parquet*, cuja escolha fundamenta-se em suas propriedades técnicas superiores, como compressão eficiente, leitura colunar seletiva e compatibilidade nativa com ferramentas analíticas. Além disso, a materialização do *dataset* enriquecido estabelece um ponto de *checkpoint* no *pipeline* de processamento, de modo que o *dashboard* consome exclusivamente esse artefato e permanece desacoplado das etapas de coleta, ETL e modelagem. Esse desacoplamento, por sua vez, possibilita otimizações independentes e simplifica significativamente a manutenção e evolução do sistema.

5.4 Implementação do *Dashboard* Interativo

O *dashboard* interativo constitui a camada de apresentação do sistema, o que garante a materialização do artefato final de visualização. Essa interface web foi desenvolvida integralmente em *Python* (Python, 2025), por meio do *framework Streamlit* (Streamlit, 2025), o que assegura alto nível de integração com as estruturas analíticas implementadas nas etapas anteriores. Além disso, o sistema oferece aos usuários finais acesso intuitivo e responsivo às análises temáticas dos trabalhos de conclusão de curso, o que possibilita a exploração dos dados por meio de visualizações interativas, filtros dinâmicos e funcionalidades analíticas avançadas.

A Figura 7 apresenta a tela inicial da aplicação, que organiza os principais indicadores e fornece o ponto de entrada para as diferentes perspectivas analíticas disponíveis no painel.

Figura 7 – **Tela inicial**



Fonte: Elaborado pela autora.

5.4.1 Estrutura da Aplicação

A aplicação adota uma arquitetura modular e reativa que define de forma nítida as responsabilidades entre os componentes, favorece a manutenção e amplia a testabilidade e a capacidade de evolução do sistema. Essa estrutura organiza o fluxo de execução da interface, administra o estado global e coordena a exibição dos módulos analíticos que compõem o ambiente interativo.

A concepção arquitetural apoia-se em três princípios fundamentais. O primeiro é a separação de responsabilidades, segundo o qual cada módulo independente concentra a lógica de uma dimensão analítica específica, de modo a garantir coesão e clareza funcional. O segundo princípio baseia-se na reatividade, pois alterações em filtros, seleções ou parâmetros de entrada propagam efeitos automáticos em toda a aplicação, com atualização imediata das visualizações impactadas. O terceiro princípio refere-se à eficiência computacional, obtida por meio de estratégias de *cache* que evitam o reprocessamento de dados e reduzem a carga de operações intensivas.

A adoção do *Streamlit* (Streamlit, 2025) como *framework* fundamenta-se em sua filosofia de construção simplificada de interfaces analíticas em *Python* (Python, 2025), a qual elimina a necessidade de controle explícito de estado e de manipulação direta de eventos. Seu modelo de execução sequencial, reavaliado a cada interação do usuário, mantém o desempenho elevado graças a mecanismos internos de *cache* e memorização

que asseguram fluidez, responsividade e estabilidade no uso do sistema.

5.4.2 Interface de Usuário e Controles Interativos

A interface de usuário foi concebida com base em princípios de design responsivo e usabilidade orientada à análise exploratória de dados, com o intutito de garantir clareza visual, coerência estrutural e adaptabilidade a diferentes tamanhos de tela. A configuração inicial da página, definida por meio do comando *st.set_page_config*, especifica o layout em formato *wide*, o que amplia o espaço horizontal destinado às visualizações e favorece a disposição simultânea de múltiplos painéis analíticos. Essa escolha arquitetural reduz a necessidade de rolagem vertical e melhora a leitura comparativa entre gráficos e indicadores, especialmente em análises temporais ou em seções que apresentam métricas correlacionadas. Além disso, o layout amplo contribui para uma experiência interativa mais fluida, poi permite que os usuários explorem as informações de forma intuitiva e contextual e mantenha a consistência visual e a harmonia estética do *dashboard*.

5.4.2.1 Personalização Visual

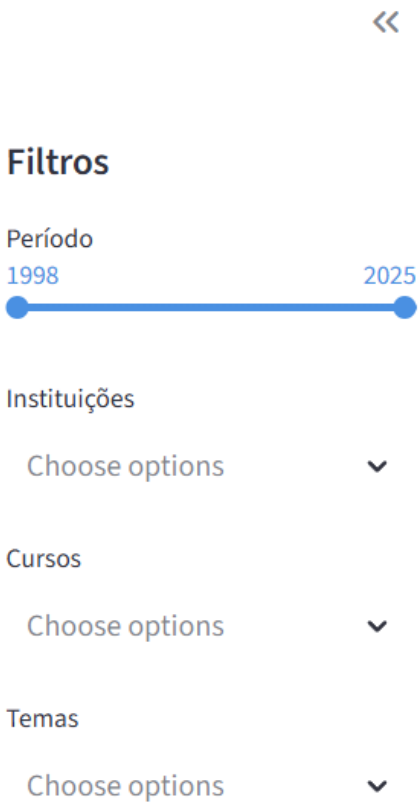
O módulo responsável pela estilização concentra as definições personalizadas em CSS, incorporadas à aplicação por meio de comandos do *Streamlit* (Streamlit, 2025) que permitem a injeção controlada de código HTML. Essa estratégia amplia as possibilidades de personalização além dos recursos nativos da biblioteca, o que assegura alinhamento estético com a identidade visual institucional, legibilidade aprimorada para textos técnicos e controle preciso sobre espaçamentos e interações dos elementos na interface. Além disso, a centralização das regras de estilo em um componente dedicado contribui para a manutenção do código, facilita o versionamento e preserva a separação entre a camada lógica e a de apresentação.

5.4.2.2 Sistema de Filtros Globais

A barra lateral, implementada por meio do componente *st.sidebar*, reúne os controles de filtragem global que influenciam de forma transversal as diversas perspectivas analíticas do painel. Essa decisão de design segue boas práticas consolidadas em interfaces voltadas à exploração de dados, nas quais os filtros de escopo permanecem permanentemente visíveis, de modo a facilitar a compreensão imediata dos parâmetros aplicados e reforçar a transparência das análises. Além disso, a disposição lateral contribui para a organização espacial da interface, o que evita sobrecarga visual e mantém as áreas de visualização centralizadas no conteúdo analítico.

Abaixo, na Figura 8 observa-se a disposição dos filtros presentes na barra lateral

Figura 8 – **Filtros Barra Lateral**



Fonte: Elaborado pela autora.

Os principais mecanismos de filtragem são descritos a seguir:

- **Filtro Temporal:** desenvolvido com o componente deslizante *st.slider*, permite ao usuário selecionar de forma interativa o intervalo de anos de interesse. Essa abordagem se mostra particularmente adequada à natureza contínua da dimensão temporal, pois a análise de períodos ou tendências é mais recorrente e informativa do que a observação de anos isolados. A configuração do controle foi otimizada para garantir precisão na seleção, com limites dinâmicos definidos conforme os dados disponíveis.
- **Filtros Categóricos:** implementados com o componente *st.multiselect*, possibilitam a seleção simultânea de múltiplas instituições, cursos e temas. A interface incorpora busca incremental, funcionalidade essencial diante da alta cardinalidade de certas dimensões, como a variedade de cursos ofertados. Além disso, a ausência de seleção implica a inclusão de todas as categorias, o que simplifica o acesso a visões agregadas e amplia a flexibilidade da exploração analítica.

A aplicação dos filtros é tratada por uma função dedicada, que realiza operações booleanas sobre o *DataFrame* principal e retorna o subconjunto *df_filtrado*. Essa estrutura modular assegura a consistência e a reprodutibilidade das análises, uma vez que todas as visualizações consomem o mesmo conjunto filtrado. Desse modo, evita-se a redundância de

código e preserva-se a coerência entre as diferentes representações visuais apresentadas no painel.

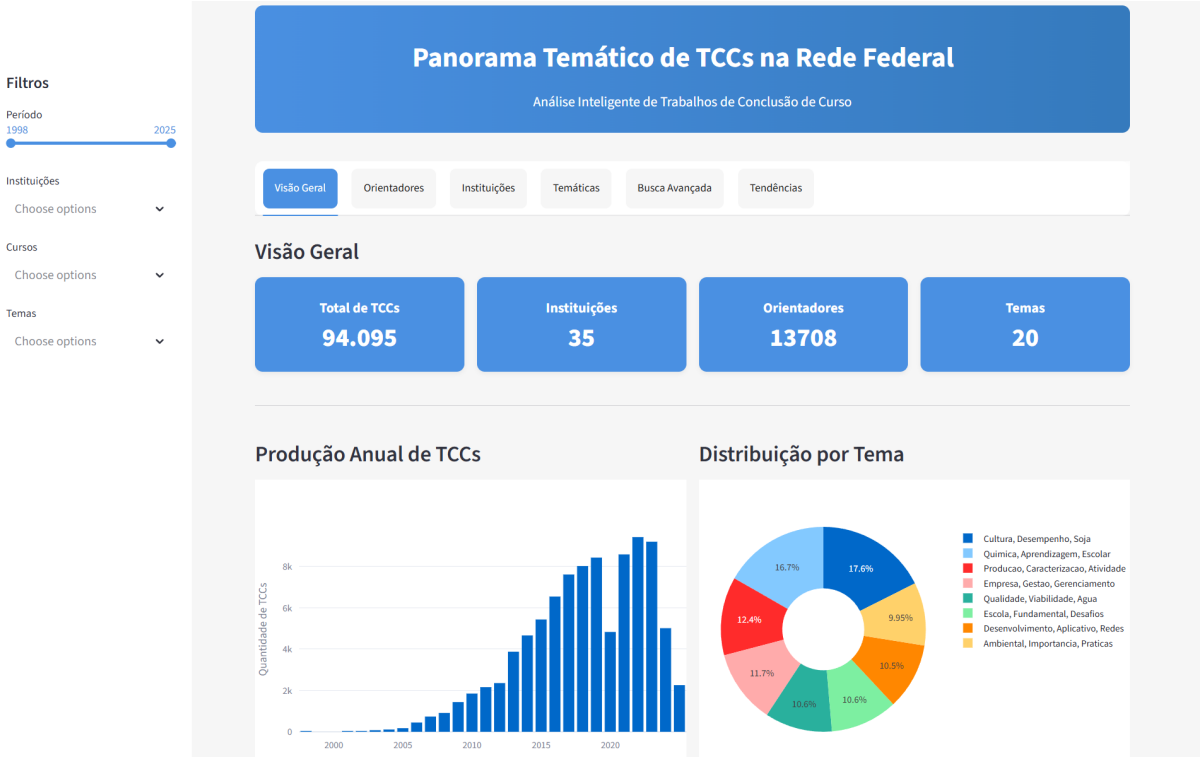
5.4.3 Sistema de Navegação e Módulos Analíticos

A navegação principal da aplicação é estruturada por meio de abas interativas, implementadas com o componente *st.tabs*, que segmentam o *dashboard* em múltiplas perspectivas analíticas complementares. Essa organização permite ao usuário transitar de forma instantânea entre diferentes dimensões de análise sem perda de contexto, uma vez que todas as abas compartilham o mesmo estado global de filtros e parâmetros. Além de favorecer a fluidez da experiência de uso, essa estrutura evita a fragmentação do fluxo cognitivo e reforça a coerência entre os módulos analíticos.

Cada aba corresponde a um módulo independente, responsável por uma camada específica de interpretação dos dados, conforme descrito a seguir:

- **Visão Geral:** apresenta indicadores-chave e métricas agregadas, oferece uma síntese inicial do conjunto de trabalhos e um panorama quantitativo do corpus analisado. Essa aba serve como ponto de partida para a exploração e destaca padrões gerais e tendências iniciais. Para ilustrar a organização inicial do painel e os principais indicadores apresentados ao usuário, a Figura 9 a seguir exibe a tela correspondente à aba "Visão Geral".

Figura 9 – **Visão Geral**



Fonte: Elaborado pela autora.

- **Orientadores:** concentra as análises relacionadas aos docentes, de forma a abranger a produtividade, o volume de orientações e a distribuição por instituição. Os gráficos permitem identificar a concentração de orientações e potenciais relações entre áreas temáticas e atuação docente. A Figura 10 seguinte apresenta a aba "Orientadores"que fica responsável pelos indicadores associado aos docentes.

Figura 10 – **Orientadores**

Fonte: Elaborado pela autora.

- **Instituições:** examina a distribuição dos Trabalhos de Conclusão de Curso (TCCs) por instituição, de modo a possibilitar comparações temporais e temáticas. Essa visão pode ser visualizada na aba de "Insituições"apresentada na Figura 11.

Figura 11 – **Instituições**

Fonte: Elaborado pela autora.

- **Temáticas:** apresenta os resultados da modelagem de tópicos com LDA, detalha a evolução e a relevância relativa dos temas ao longo dos anos. Essa visualização evidencia dinâmicas de interesse científico e emergências de novas áreas de pesquisa. A Figura 12 a seguir demonstra a aba "Temáticas".

Figura 12 – **Temáticas**

Fonte: Elaborado pela autora.

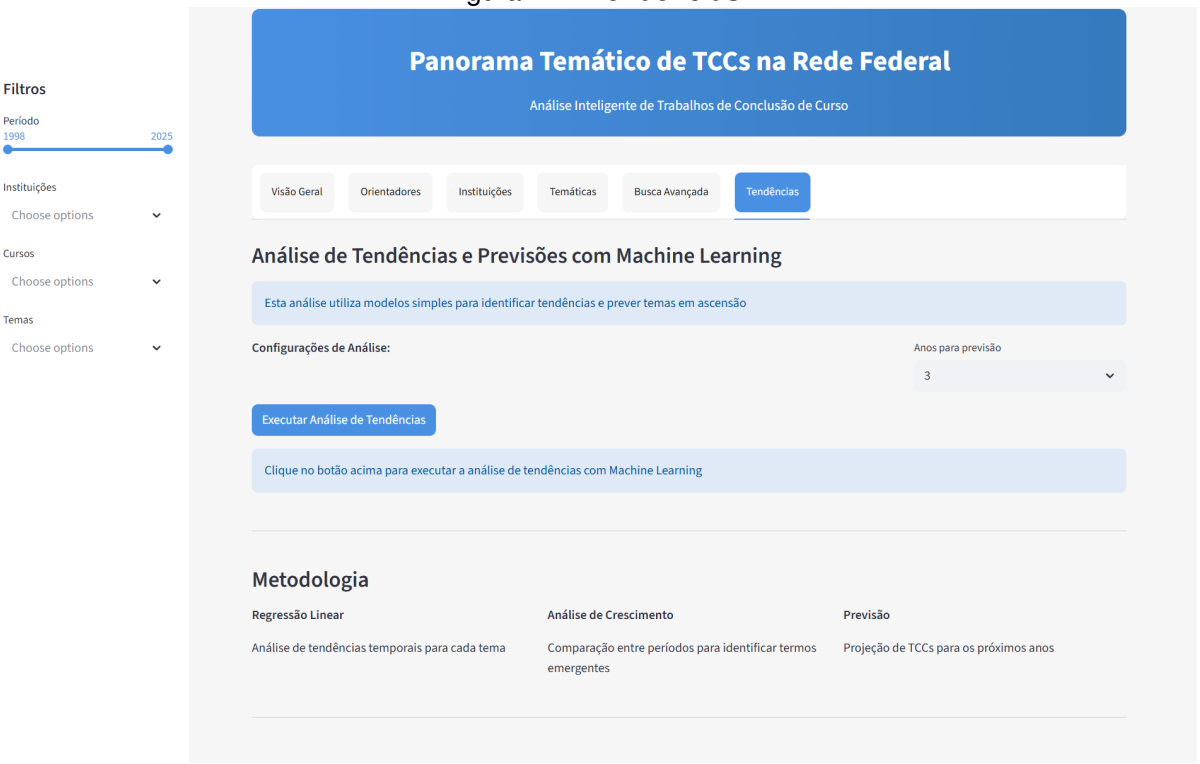
- **Busca Avançada:** oferece recursos de busca semântica, que permitem localizar trabalhos semelhantes por contexto e significado,e a busca por trabalhos similares, o que amplia as possibilidades de descoberta de informações relevantes. A Figura 13 apresenta a aba "Busca Avançada".

Figura 13 – **Busca Avançada**

Fonte: Elaborado pela autora.

- **Tendências:** dedica-se às análises preditivas ao permitir que seja estimada a trajetória futura de determinadas temáticas com base em séries históricas e modelos de regressão temporal. Com foco na análise temporal e preditiva, a Figura 14 abaixo mostra a aba "Tendências".

Figura 14 – **Tendências**



Fonte: Elaborado pela autora.

As visualizações de dados são produzidas com a biblioteca *Plotly* (Plotly, 2025), por meio das APIs *plotly.express* e *plotly.graph_objects*, que proporcionam alto nível de interatividade e expressividade visual. Os gráficos resultantes permitem operações como zoom dinâmico, exibição de informações contextuais via *hover tooltips*, seleção de intervalos e exportação direta em diferentes formatos. Essa escolha tecnológica reforça a proposta de uma análise exploratória interativa, acessível e visualmente consistente, alinhada aos princípios de transparência e usabilidade adotados em todo o sistema.

5.4.4 Funcionalidades Analíticas Avançadas

Além das visualizações descritivas, o *dashboard* integra funcionalidades analíticas de maior complexidade, que aplicam métodos de aprendizado de máquina e técnicas de estatística computacional voltadas à extração de padrões, correlações e tendências subjacentes nos dados. Essas capacidades ampliam o papel da ferramenta, que deixa de se restringir à exploração visual para atuar também como um instrumento de apoio à descoberta de conhecimento, o que oferece subsídios quantitativos à tomada de decisão e ao entendimento da dinâmica temática da produção acadêmica.

5.4.4.1 Busca por Similaridade Semântica

Essa funcionalidade realiza recomendações automáticas de trabalhos semelhantes e fundamenta-se na análise da similaridade semântica entre resumos acadêmicos. O processo envolve a vetorização dos textos com a utilização do método TF-IDF (*Term Frequency-Inverse Document Frequency*) e o cálculo da similaridade por cosseno entre os vetores resultantes, permite identificar os documentos mais próximos semanticamente ao trabalho selecionado. Além disso, essa abordagem possibilita a descoberta de pesquisas conexas, o mapeamento de áreas temáticas correlatas e a identificação de possíveis lacunas no conhecimento, o que oferece suporte à ampliação de perspectivas acadêmicas e à construção de redes de colaboração.

5.4.4.2 Análise de Tendências e Predição

O módulo de tendências emprega técnicas de análise estatística e modelagem preditiva para identificar áreas de pesquisa em crescimento, declínio ou estabilidade. A metodologia baseia-se na agregação de séries temporais da produção acadêmica por tema e ano, seguida da aplicação de regressão linear simples, cujo coeficiente angular representa a taxa de variação da produção ao longo do tempo. De forma complementar, são realizadas análises diferenciais de vocabulário, que comparam períodos históricos e recentes a fim de identificar termos emergentes e destacar conceitos, tecnologias ou metodologias em ascensão. Esses resultados fornecem subsídios para a tomada de decisão estratégica, a definição de prioridades de pesquisa e a antecipação de novos campos de interesse acadêmico.

Ao integrar essas funcionalidades, o *dashboard* transcende sua função descritiva inicial e passa a ser visto como uma plataforma de inteligência analítica robusta, capaz de fornecer recomendações automatizadas, identificar padrões de evolução temática e orientar de forma estratégica o planejamento acadêmico.

5.5 Ferramentas e Tecnologias Utilizadas

A implementação do sistema de análise temática de TCCs foi conduzida com base em um conjunto integrado de ferramentas e tecnologias que sustentaram todas as etapas do processo, desde a coleta de dados até a visualização interativa dos resultados. As principais tecnologias empregadas são descritas a seguir.

- **Coleta de Dados:** *asyncio* (Asyncio, 2025), biblioteca do *Python* (Python, 2025) voltada para programação assíncrona, que permite a execução concorrente de tarefas de I/O; e *aiohttp* (Aiohttp, 2025), cliente HTTP assíncrono utilizado para realizar múltiplas requisições simultâneas de forma eficiente.
- **Armazenamento de Dados:** SQLite (SQLite, 2025), um sistema de banco de dados relacional leve e embutido; SQLAlchemy (SQLAlchemy, 2025), ferramenta ORM que

facilita o mapeamento entre objetos *Python* (Python, 2025) e tabelas relacionais; e *Apache Parquet* (Apache, 2025), formato de armazenamento colunar otimizado para leitura e processamento de grandes volumes de dados.

- **Transformação e Análise de Dados:** Pandas (Pandas, 2025), biblioteca amplamente empregada para manipulação e análise de dados estruturados; NLTK (NLTK, 2025), utilizada para tarefas de processamento de linguagem natural, como tokenização e limpeza textual; e *Scikit-learn* (Scikit-learn, 2025), biblioteca de aprendizado de máquina responsável por algoritmos de modelagem, classificação e agrupamento.
- **Visualização de Dados e Interface Interativa:** *Streamlit* (Streamlit, 2025), *framework* para criação de interfaces web voltadas à exploração de dados; e *Plotly* (Plotly, 2025), biblioteca usada para construir visualizações gráficas interativas.

Em conjunto, essas ferramentas possibilitaram a integração eficiente entre as etapas de coleta, transformação, análise e visualização, de modo a assegurar reprodutibilidade, escalabilidade e consistência ao sistema desenvolvido.

6 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados obtidos após a implementação integral do sistema de análise temática dos TCCs da Rede Federal de Educação Profissional, Científica e Tecnológica. Inicialmente, descrevem-se o conjunto de dados coletado e sua caracterização quantitativa. Em seguida, realiza-se a validação funcional e de desempenho do sistema desenvolvido. Posteriormente, discutem-se os resultados analíticos da modelagem de tópicos obtidos por meio do algoritmo LDA. Por fim, apresentam-se as descobertas temáticas e as tendências identificadas, que confirmam o atendimento aos requisitos estabelecidos na proposta.

6.1 Caracterização do Conjunto de Dados Coletado

A execução completa do *pipeline* de coleta resultou em um conjunto substantivo de Trabalhos de Conclusão de Curso da Rede Federal. O sistema coletou um total de 233.310 TCCs na data de 31 de outubro de 2025, os quais foram distribuídos ao longo do período de 1977 a 2025, provenientes de 37 instituições da Rede Federal, de modo a contemplar 24 estados brasileiros mais o Distrito Federal. Estes trabalhos foram orientados por 78.621 docentes, com uma média de 2,96 orientações por docente, distintos e estão associados a 6.315 cursos de graduação únicos identificados durante o processo de coleta conforme apresentado na Tabela 5.

Tabela 5 – **Quantidade de Professores e TCCs por Instituição da Rede Federal**

Instituição	Professores	TCCs	Média TCCs por Professor
CEFET-MG	1687	8528	5,05
CEFET-RJ	1253	6492	5,18
IFAC	771	1326	1,72
IFAL	1780	3788	2,13
IFAM	2082	3032	1,46
IFAP	681	1510	2,22
IFB	1162	4391	3,78
IFBA	5963	12281	2,06
IFC	3897	11769	3,02
IFCE	4407	15573	3,53
IFES	2863	10823	3,78
IFFARROUPILHA	1548	4914	3,17
IFFLUMINENSE	1575	3449	2,19
IFG	2098	8604	4,10
IFGOIANO	1596	7131	4,47
IFMA	2799	8275	2,96
IFMG	1958	8179	4,18
IFMT	1883	5647	2,99
IFNMG	991	0	0
IFPA	2431	6089	2,50
IFPB	2369	6930	2,92
IFPE	2003	1114	0,56
IFPI	2392	6979	2,92
IFPR	2539	8945	3,52
IFRJ	2128	6199	2,91
IFRN	2644	5823	2,20
IFRO	2723	4280	1,57
IFRR	619	1048	1,69
IFRS	2420	9028	3,73
IFSC	2740	8191	2,99
IFSP	5576	19640	3,52
IFSUDESTEMG	558	3094	5,54
IFSUL	1791	4159	2,32
IFSULDEMINAS	1155	5377	4,65
IFSertaoPE	1007	1857	1,84
IFTM	1259	5046	4,01
IFTO	1273	3799	2,98
TOTAL	**78621**	**233310**	**2,96**

Fonte: Elaborado pela autora.

A análise da distribuição institucional dos TCCs coletados evidenciou uma expressiva assimetria no volume de produção entre as instituições que compõem a Rede Federal. As três instituições com maior número de trabalhos, IFSP (8,42%), IFCE (6,67%) e IFBA (5,26%), concentram aproximadamente 20,35% do total dos dados, o que faz com que sejam institutos de elevada produtividade acadêmica. Em contrapartida, nove instituições, entre as quais se destacam IFNMG (0%), IFRR (0,45%), IFPE (0,48%), IFAC (0,57%), IFAP (0,65%), IFSertãoPE (0,80%), IFAM (1,30%), IFSUDESTEMG (1,33%), IFFLUMINENSE (1,48%), contribuíram individualmente com menos de 1,5% dos trabalhos. Essa discrepância pode ser explicada tanto por diferenças estruturais, como o porte institucional e a quantidade de cursos ofertados, quanto por fatores relacionados às políticas de registro, disponibilização e publicização digital da produção acadêmica. No caso específico do IFNMG, a ausência de TCCs decorre da indisponibilidade de dados sobre trabalhos acadêmicos, visto que estava acessíveis apenas informações referentes aos docentes, o que explica o total zerado na coleta.

A Tabela 6 apresenta o volume de trabalhos coletados ao longo dos anos, o que permite a visualização de forma detalhada da evolução histórica na volumetria das orientações.

Tabela 6 – **Quantidade de TCCs por Ano**

Ano	TCCs
1977	1
1983	3
1984	1
1987	5
1989	4
1990	11
1991	18
1992	27
1993	21
1994	35
1995	90
1996	116
1997	160
1998	268
1999	440
2000	648
2001	1004
2002	1609
2003	2614
2004	4155
2005	6520
2006	7779
2007	8949
2008	9591
2009	10255
2010	10000
2011	9503
2012	9533
2013	11440
2014	12157
2015	12973
2016	13247
2017	13997
2018	14068
2019	13875
2020	8530
2021	13420
2022	13345
2023	12639
2024	7160
2025	3099

Fonte: Elaborado pela autora.

Do ponto de vista temporal, observou-se um crescimento consistente no volume de orientações disponibilizadas digitalmente ao longo do período analisado, com uma taxa média anual de crescimento de 20,8%. Contudo, é importante destacar que essa taxa é influenciada pelos valores iniciais extremamente baixos, o que intensifica artificialmente o percentual de crescimento quando considerado todo o período histórico.

Os anos de 2017, 2018 e 2019 concentram 17,98% da produção total, o que indica que a digitalização e disponibilização sistemática dos trabalhos é um fenômeno relativamente recente na Rede Federal. Este padrão temporal é consistente com a expansão dos repositórios digitais institucionais e a adoção de políticas de acesso aberto ao conhecimento

científico produzido nas instituições públicas brasileiras.

Quanto à qualidade e completude dos dados coletados, o processo de validação implementado no módulo ETL demonstrou robustez adequada. Do total de 233.310 registros inicialmente extraídos pelo *scraper*, 94.095 (40,33%) passaram pelos critérios de validação estabelecidos, os quais foram efetivamente incorporados a camada analítica. Dentre os principais motivos de rejeição de registros estão instituição não integrantes da Rede Federal e trabalhos de outro grau acadêmico que não o nível superior.

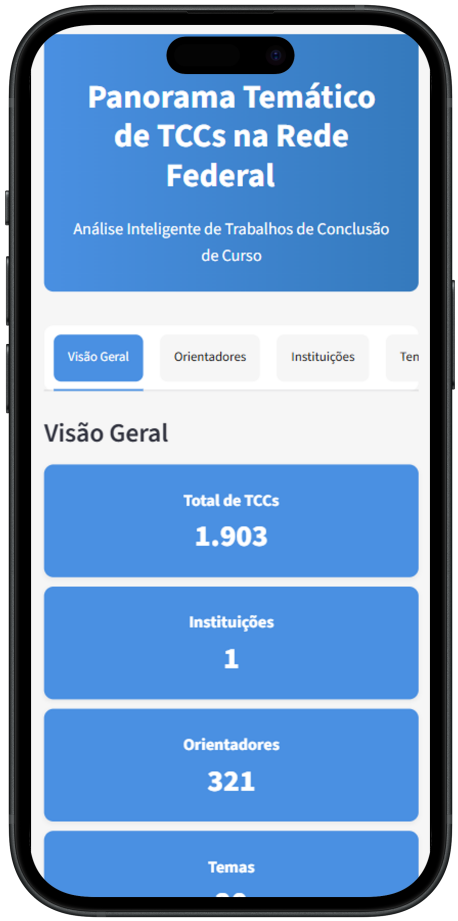
6.2 Validação Funcional e Performance do Sistema

A validação funcional do sistema implementado foi conduzida em múltiplas dimensões com o intuito de medir a performance, robustez e eficiência de cada módulo do *pipeline*. O módulo de coleta, foi desenvolvido baseado em arquitetura assíncrona com *asyncio* (Asyncio, 2025) e *aiohttp* (Aiohttp, 2025). O tempo total de execução para a coleta completa dos 233.310 TCCs foi de vinte e cinco minutos, o que resultou em um *throughput* médio de 9.332,4 trabalhos por minuto com sucesso das requisões HTTP, sem falhas registradas associadas ao processo de extração dos dados. Isso se deve ao mecanismo de tratamento de exceções implementado permitiu que o *scraper* continuasse a operar mesmo diante destas falhas pontuais, sem comprometer a coleta geral.

O módulo ETL, responsável pela transformação e estruturação dos dados no modelo *Star Schema*, processou 233.310 registros com uma taxa de sucesso de 100%. O tempo total de execução do pipeline ETL completo, desde a extração do banco de *staging* até a carga final no *Data Mart*, foi de 47,5 segundos. A modularização do processo em funções especializadas facilitou significativamente a depuração e o refinamento iterativo das regras de transformação. O modelo dimensional resultante 94.095 registros na tabela fato referente aos TCCs, apoiada pelas tabelas dimensionais de instituição, curso, pessoa e campus.

A performance do *dashboard* interativo desenvolvido em *Streamlit* (Streamlit, 2025) atendeu aos requisitos de responsividade estabelecidos, como apresentado na Figura 15.

Figura 15 – **Visão Geral com Apresentação Mobile**



Fonte: Elaborado pela autora.

O tempo médio de carregamento inicial da aplicação, o que inclui a leitura do arquivo *parquet* (Apache, 2025) com os dados finais foi de 3,83 segundos. Ademais, testes de interatividade demonstraram que a aplicação de filtros multidimensionais (ano, instituição, curso, tema) e a atualização dos gráficos correspondentes ocorrem em tempo médio de 1 segundo, o que garante uma experiência de usuário fluida. A decisão de adotar o formato *Apache Parquet* (Apache, 2025), mostrou-se acertada, visto que permite que o *Streamlit* (Streamlit, 2025) carregue os dados em tempo de execução sem a necessidade de firmar uma conexão com um banco e sem o risco de indisponibilidade.

6.3 Resultados da Modelagem de Tópicos

A aplicação do algoritmo *Latent Dirichlet Allocation* (LDA) sobre o conjunto de dados previamente pré-processado resultou na identificação de vinte tópicos temáticos distintos. O vocabulário final utilizado pelo modelo, definido a partir de critérios de filtragem com *max_df* igual a 0,9, *min_df* igual a 20, *max_features* limitado a 2.000 e *ngram_range* entre 1 e 2, compreende 2.000 termos únicos, o que abrange tanto unigramas quanto bigramas. Esse

vocabulário estabelece um equilíbrio entre cobertura semântica e tratabilidade computacional, uma vez que exclui termos excessivamente comuns e termos demasiadamente raros, ambos de baixo poder discriminativo para a modelagem temática.

A Tabela 7 apresenta a lista completa dos vinte tópicos identificados pelo modelo LDA, cada um sintetizado pelos três termos de maior relevância em sua composição.

Tabela 7 – **Temáticas geradas pelo modelo LDA**

Tópico	**Temática**
0	*Cultura, Desempenho, Soja*
1	*Meio, Ambiente, Revisao*
2	*Escola, Fundamental, Desafios*
3	*Saude, Vida, Regiao*
4	*Fisica, Rede, Infantil*
5	*Percepcao, Professores, Matematica*
6	*Ambiental, Importancia, Praticas*
7	*Construcao, Residuos, Solidos*
8	*Aplicacao, Supervisionado, Relatorio*
9	*Tratamento, Agua, Livros*
10	*Turismo, Comportamento, Aplicaçoes*
11	*Tecnologia, Mercado, Custo*
12	*Empresa, Gestao, Gerenciamento*
13	*Software, Digital, Desenvolvimento*
14	*Qualidade, Viabilidade, Agua*
15	*Quimica, Aprendizagem, Escolar*
16	*Energia, Eletrica, Geração*
17	*Desenvolvimento, Aplicativo, Redes*
18	*Produção, Caracterizacao, Atividade*
19	*Revisao, Dados, Literatura*

Fonte: Elaborado pela autora.

A análise qualitativa dos tópicos identificados revelou agrupamentos semânticos coerentes e interpretáveis, em que cada um representa uma área temática distinta da produção acadêmica da Rede Federal. O tópico com maior incidência na coleção foi o Tópico 0 (*"Cultura, Desempenho e Soja"*), que reúne 8.567 Trabalhos de Conclusão de Curso (TCCs), o que corresponde a 9,10% do conjunto total. Entre os títulos associados a esse agrupamento observam-se produções relacionadas a práticas culturais, desempenho produtivo e investigações aplicadas ao setor agrícola. A análise dos resumos indica que esse tópico concentra pesquisas voltadas a metodologias de produção, manejo, análises de desempenho e estudos comparativos em ambientes educacionais e produtivos.

A distribuição dos TCCs entre os vinte tópicos identificados revelou-se heterogênea, de tal modo que reflete a concentração temática característica da produção acadêmica da Rede Federal. Os três tópicos mais recorrentes, Tópico 0 (*"Cultura, Desempenho e Soja"*), Tópico 15 (*"Quimica, Aprendizagem e Escolar"*) e Tópico 18 (*"Producao, Caracterizacao e Atividade"*), somam 22.732 TCCs, que representam 24,15% da coleção total, o que evidencia áreas consolidadas e de elevada produtividade científica. Em contraposição, os tópicos menos frequentes, Tópico 19 (*"Revisao, Dados, Literatura"*), Tópico 9 (*"Tratamento, Agua, Livros"*) e Tópico 16 (*"Energia, Eletrica, Geracao"*), correspondem a 9.277 TCCs, 12,4% do total, o que sugere áreas temáticas emergentes ou de caráter mais especializado. Essa distribuição desigual, entretanto, é inerente a corpora acadêmicos reais e não compromete

a validade ou a consistência da modelagem proposta.

A avaliação da qualidade da modelagem de tópicos foi conduzida prioritariamente por meio de análise qualitativa da coerência semântica e da interpretabilidade dos resultados, uma vez que métricas quantitativas tradicionais, como *perplexidade* e *topic coherence*, não capturam adequadamente a relevância prática dos agrupamentos temáticos para os usuários finais. A inspeção manual dos termos mais representativos de cada tópico, aliada à análise de amostras de TCCs classificados em cada categoria, confirmou que os tópicos identificados correspondem a domínios temáticos reconhecíveis e semanticamente coesos. A sobreposição temática entre tópicos mostrou-se limitada, ao serem observados casos pontuais de ambiguidade entre Tópico 15 (*"Quimica, Aprendizagem, Escolar"*), o Tópico 5 (*"Percepcao, Professores, Matematica"*) e o Tópico 2 (*"Escola, Fundamental, Desafios"*), em que trabalhos situados na interseção entre escola e o ensino de ciências e ensino de exatas apresentaram probabilidades similares de associação a ambos os tópicos. Essa ambiguidade moderada é considerada aceitável, o que reflete a natureza multidisciplinar de parte da produção acadêmica analisada.

6.4 Análises Temáticas e Descobertas

A análise exploratória dos dados, enriquecida com as classificações temáticas proporcionadas pelo modelo LDA, revelou padrões significativos na distribuição das áreas de pesquisa na Rede Federal. De modo geral, a distribuição temática indica que o Tópico 0 (*“Cultura, Desempenho, Soja”*) constitui a área de maior concentração de trabalhos, com 8.567 TCCs, o que corresponde a 9,10% do conjunto total de trabalhos analisados.

Em contraste, temas como o Tópico 16 (*“Energia, Eletrica, Geração”*), com 2.760 TCCs (2,93%), e o Tópico 9 (*“Tratamento, Agua, Livros”*), com 3.135 TCCs (3,33%), representam as menores parcelas do conjunto de dados. Embora esses tópicos apresentem frequências reduzidas, podem ser interpretados como áreas temáticas mais especializadas ou emergentes, a depender da evolução temporal observada. Nesse contexto, a análise longitudinal, realizada no módulo de *dashboard* de "Tendências", permitiu identificar padrões de crescimento para cada um dos tópicos ao considerar um período de cinco anos para a análise, em que apresetaram percentual de mudança de 274% para o Tópico 16 e de 302,9% para o Tópico 9.

Os resultados da análise de tendências, realizada por meio de regressão linear simples, evidenciaram três tópicos com crescimento estatisticamente significativo e sustentado. Entre eles, o Tópico 15 (*“Quimica, Aprendizagem, Escolar”*) apresentou o maior coeficiente angular, igual a 32,81, o que indica um aumento médio de aproximadamente 32,81 TCCs por ano. Na sequência, destacaram-se o Tópico 0 (*“Cultura, Desempenho, Soja”*), com coeficiente angular de 31,40, e o Tópico 2 (*“Escola, Fundamental, Desafios”*), com 21,43. Dessa forma, projeções baseadas nesses modelos de regressão indicam que, mantidas as tendências atuais, esses três tópicos deverão representar cerca de 24,41% do conjunto

total até o ano de 2030.

Por outro lado, nenhum tópico apresentou tendência de declínio estatisticamente significativa, o que sugere resiliência e manutenção do interesse nas áreas temáticas já estabelecidas. Além disso, os demais tópicos demonstraram comportamento estável, com variações interanuais não sistemáticas, de modo a indicar a existência de áreas consolidadas caracterizadas por uma produção relativamente constante ao longo do período analisado.

A análise institucional evidenciou especializações temáticas expressivas entre as diferentes unidades da Rede Federal. O IFGoiano apresentou forte concentração no Tópico 4 (*"Cultura, Desempenho, Soja"*), com média de 176,2 TCCs por tema distribuídos nas vinte temáticas. De modo semelhante, o IFSULDEMINAS destacou-se pelo mesmo tópico, com média de 170,9 TCCs por tema associados a todos os vintes temáticas. Essa especialização pode estar relacionada à vocação regional das instituições, às características dos Arranjos Produtivos Locais (APLs) em suas áreas de abrangência ou a políticas institucionais orientadas para o fortalecimento de determinadas linhas de pesquisa.

A análise da produção por orientador identificou 3.135 docentes classificados como os mais produtivos do conjunto de dados, ou seja, aqueles que registraram dez ou mais orientações, com variação entre dez e cento e doze trabalhos. O perfil temático desses orientadores revelou tanto padrões de especialização quanto de diversidade temática. O docente Tiago Henrique Faccio Segato, vinculado ao Instituto Federal de Brasília, por exemplo, concentrou todas as suas orientações no Tópico 4 (*"Desenvolvimento, Aplicativo, Redes"*), de modo a demonstrar expertise altamente especializada. Esses resultados auxiliam estudantes em processo de escolha de orientador ao permitir a identificação de docentes com expertise consolidada em áreas específicas, de modo a contribuir para uma alocação mais estratégica das orientações no contexto institucional.

A Tabela 8 apresenta os cinco docentes mais produtivos, juntamente com suas respectivas quantidades de orientações e temáticas predominantes.

Tabela 8 – **Top 5 docentes com maior número de orientações e suas respectivas temáticas predominantes**

Docente	Número de TCCs	Temática Predominante
Auzuir Ripardo DE Alexandria	112	Tópico 17 (*Desenvolvimento, Aplicativos, Redes*)
Brenno Vitorino Costa	100	Tópico 10 (*Turismo, Comportamento, Aplicacoes*)
*R*onaldo Ribeiro Corrêa	98	*T*ópico 14 (*Qualidade, Viabilidade, Agua*)
Lilian Vilela Andrade Pinto	98	Tópico 0 (*Cultura, Desempenho, Soja*)
José Sérgio De Araújo	92	Tópico 0 (*Cultura, Desempenho, Soja*)

Fonte: Elaborado pela autora.

6.5 Validação do Atendimento aos Requisitos

A validação sistemática do sistema implementado em relação aos requisitos demonstra o cumprimento integral dos objetivos estabelecidos. O Requisito Funcional RF01, que especificava a capacidade de coletar automaticamente dados de múltiplos repositórios institucionais, foi plenamente atendido pela implementação do *scraper* assíncrono, evidenci-

ado pela coleta bem-sucedida de 233.310 TCCs de 37 instituições distintas. O Requisito Funcional RF02, que demandava a identificação automatizada de temas através de técnicas de PLN, foi satisfeito pela implementação do pipeline de modelagem de tópicos via LDA, que identificou 20 tópicos temáticos interpretáveis e semanticamente coesos.

O Requisito Funcional RF03, relativo à implementação de um *dashboard* interativo para visualização dos dados, foi atendido pelo desenvolvimento da aplicação *Streamlit*, que oferece interface gráfica intuitiva e responsiva. O Requisito Funcional RF04, que especificava a capacidade de buscar TCCs similares a um trabalho de referência, foi implementado através do módulo chamado "Busca Avançada", que utiliza vetorização TF-IDF e similaridade cosseno para identificar os trabalhos semanticamente mais próximos. O Requisito Funcional RF05, que demandava filtragem multidimensional dos dados, foi plenamente satisfeito pela implementação de filtros interativos na barra lateral do *dashboard* de modo a permitir seleção simultânea por ano, instituição, curso e tema.

Do ponto de vista dos Requisitos de Usuário, o RU01, que estabelecia a necessidade de visualizar um panorama geral das temáticas mais frequentes, foi atendido pelo módulo "Visão Geral"do *dashboard*, que apresenta gráficos de distribuição temática e estatísticas descritivas dos dados. O Requisito de Usuário RU02, relativo à exploração da produção acadêmica por orientador, foi satisfeito pelo módulo "Orientadores", que oferece visualizações de produtividade, perfil temático e detalhamento das orientações de cada docente. O Requisito de Usuário RU03, que demandava a identificação de tendências temporais nas áreas temáticas, foi plenamente atendido pelo módulo "Tendências", que implementa análise de séries temporais, regressão linear para projeção de tendências futuras e identificação de termos emergentes.

6.6 Síntese dos Resultados

A execução completa do pipeline proposto resultou em um sistema funcional e robusto para análise temática de TCCs da Rede Federal de Educação Profissional, Científica e Tecnológica. Nesse contexto, o conjunto de 233.310 trabalhos coletados, distribuídos ao longo de 48 anos e provenientes de 37 instituições, constitui uma base de dados substantiva e representativa da produção acadêmica da Rede Federal. Além disso, a modelagem de tópicos via LDA demonstrou efetividade na identificação de agrupamentos temáticos coerentes e interpretáveis, uma vez que permitiu a categorização automática dos dados textuais em vinte áreas temáticas distintas. Por fim, as análises temporais revelaram tendências de crescimento em áreas emergentes, declínio em áreas tradicionais e estabilidade em áreas consolidadas, o que possibilitou a geração de observações valiosas acerca da evolução das prioridades de pesquisa na Rede Federal.

Ademais, o *dashboard* interativo desenvolvido materializa, portanto, o objetivo central da pesquisa, uma vez que proporciona aos gestores institucionais, coordenadores de curso e pesquisadores uma ferramenta prática e acessível para a exploração do panorama

temático da produção acadêmica. Além disso, os resultados obtidos demonstram que técnicas de ciência de dados, processamento de linguagem natural e visualização interativa podem ser integradas de forma sinérgica, de modo a extrair conhecimento significativo de grandes volumes de produção acadêmica textual e, assim, apoiar processos de tomada de decisão baseados em evidências.

7 CONSIDERAÇÕES FINAIS

7.1 Síntese dos Resultados Obtidos

Este trabalho apresentou uma solução para a análise temática dos Trabalhos de Conclusão de Curso em toda a Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil, o que permitiu preencher a lacuna na compreensão consolidada da produção acadêmica dessas instituições ao possibilitar o apoio a decisões de gestão acadêmica, a redução de assimetrias regionais. Além disso, a análise demonstrou que a dispersão e a heterogeneidade dos dados em múltiplos repositórios digitais institucionais representavam um obstáculo relevante para gestores, pesquisadores e coordenadores, uma vez que dificultavam a compreensão das tendências de pesquisa, a identificação de áreas emergentes e a fundamentação de decisões estratégicas baseadas em evidências sobre os direcionamentos científicos da Rede Federal.

O objetivo geral do estudo, voltado à análise sistemática dos temas dos TCCs e ao desenvolvimento de uma ferramenta de visualização interativa para exploração desses dados, foi plenamente alcançado por meio da consecução dos objetivos específicos estabelecidos. Cabe destacar que esse total inclui trabalhos orientados em outras instituições e também tipos de produção acadêmica distintos do TCC. Além disso, verificou-se que três instituições não constavam na listagem inicial (Colégio Pedro II, Instituto Federal Baiano e Universidade Tecnológica Federal do Paraná), uma instituição (Instituto Federal de Mato Grosso do Sul) estava com o Portal Integra indisponível, e outra (Instituto Federal de Sergipe) não disponibiliza dados de professores, e, consequentemente, não apresenta trabalhos orientados, motivo pelo qual não foi possível realizar a coleta para essas unidades.

Em seguida, a implementação de um *pipeline* robusto de persistência e transformação de dados, estruturado segundo os princípios de ETL e da modelagem dimensional *Star Schema*, assegurou a qualidade, consistência e rastreamento das informações processadas. Após as etapas de validação e filtragem, mantiveram-se listados 94.095 trabalhos provenientes das 37 instituições efetivamente coletadas, o que confere solidez à base consolidada e viabiliza, assim, as etapas posteriores de análise e visualização interativa.

O processamento e categorização temática dos TCCs, fundamentados em técnicas consolidadas de Processamento de Linguagem Natural (PLN) e mineração de texto, culminaram na aplicação bem-sucedida do algoritmo *Latent Dirichlet Allocation* (LDA) para identificação de 20 tópicos temáticos distintos e interpretáveis. Sendo assim, a análise qualitativa desses tópicos revelou agrupamentos semânticos coerentes que capturam áreas consolidadas, emergentes e especializadas da produção acadêmica da Rede Federal, de modo a validar a efetividade da abordagem metodológica adotada. As análises quantitativas subsequentes permitiram identificar padrões de distribuição temática, especializações

institucionais, perfis de orientadores e, especialmente, tendências temporais de crescimento e declínio de áreas temáticas, de modo a cumprir com o terceiro objetivo específico estabelecido.

O dashboard interativo, desenvolvido em *Streamlit* (Streamlit, 2025), concretiza a contribuição prática central desta pesquisa, ao oferecer uma interface intuitiva, responsiva e multidimensional para a exploração dos dados processados. Além disso, a ferramenta atende integralmente aos requisitos funcionais e de usuário especificados, pois disponibiliza funcionalidades de filtragem multidimensional, visualização de tendências temporais, busca por similaridade semântica, análise da produtividade de orientadores e identificação de termos emergentes. Por fim, a validação funcional do sistema comprovou desempenho adequado para uso em ambiente real, com tempos de resposta compatíveis com os padrões de interatividade exigidos em aplicações analíticas contemporâneas.

A metodologia empregada mostrou-se apropriada para o desenvolvimento do artefato proposto, uma vez que proporcionou um marco metodológico consistente para a concepção, implementação e avaliação da solução. Além disso, a combinação de técnicas de web *scraping* assíncrono, *Business Intelligence* (BI), modelagem dimensional, Processamento de Linguagem Natural (PLN) e visualização interativa configura uma abordagem tecnicamente sólida e replicável, capaz de ser aplicada a problemas análogos de consolidação e análise de produção acadêmica dispersa. Por conseguinte, a arquitetura modular do sistema, caracterizada pela separação clara de responsabilidades entre os módulos de coleta, transformação, análise e visualização, favorece a manutenibilidade, escalabilidade e evolução futura da solução, visto que é assegurada sua sustentabilidade técnica e adaptabilidade a novos contextos institucionais.

Os resultados obtidos demonstram de forma contundente que a integração de técnicas de ciência de dados, aprendizado de máquina e visualização da informação constitui um mecanismo eficaz para extrair conhecimento significativo de grandes volumes de dados textuais não estruturados, uma vez que possibilita sua transformação em informações acionáveis voltados ao apoio à tomada de decisão no âmbito da gestão educacional. Ademais, a identificação de tendências de crescimento em áreas emergentes, aliada à detecção de termos emergentes relacionados a tecnologias contemporâneas, fornece aos gestores institucionais auxílio para a revisão curricular, o planejamento de infraestrutura laboratorial, a contratação de docentes especializados e o fomento a linhas de pesquisa estratégicas. Desse modo, os achados desta pesquisa reforçam o potencial transformador do uso articulado de técnicas analíticas e computacionais na gestão do conhecimento acadêmico.

7.2 Trabalhos futuros

Para trabalhos futuros, diversas direções de aprofundamento e expansão podem ser exploradas. Dentre elas, a incorporação de dados adicionais, como a análise das temáticas e

descrições de projetos de ensino, pesquisa, extensão e inovação permite identificar aspectos semelhantes aos presentes nos TCCs, bem como estabelecer correlações entre eles. Adicionalmente, a integração com dados de produção científica docente, como publicações em periódicos e participação em eventos acadêmicos, possibilitaria investigar correlações entre a atividade de orientação e a produtividade científica dos professores com o intuito de identificar padrões de desenvolvimento de linhas de pesquisa consolidadas.

Do ponto de vista metodológico, a exploração de técnicas mais avançadas de aprendizado de máquina configura uma oportunidade promissora para o aperfeiçoamento da abordagem proposta. Nesse sentido, a aplicação de modelos de *embeddings* contextuais, como o BERT (*Bidirectional Encoder Representations from Transformers*) ou outros modelos baseados em *transformers* pré-treinados para a língua portuguesa, poderia ampliar a capacidade de captura das nuances semânticas e das relações contextuais entre conceitos, de modo a contribuir para o refinamento da categorização temática. Além disso, a implementação de algoritmos de detecção de comunidades e de análise de redes sobre grafos de co-ocorrência de termos ou de colaboração entre orientadores e cursos tenderia a revelar estruturas latentes de relacionamento, o que possilitaria uma compreensão mais profunda das interconexões entre áreas temáticas e agentes institucionais.

A validação empírica do *dashboard* com usuários reais, gestores institucionais, coordenadores de curso, docentes e alunos, constitui uma etapa essencial para refinar a usabilidade da interface e garantir que os insights gerados sejam efetivamente compreensíveis, relevantes e acionáveis para os diferentes perfis de *stakeholders*. Estudos de caso qualitativos poderiam investigar como as informações disponibilizadas pela ferramenta influenciam processos decisórios concretos, como a definição de eixos temáticos prioritários em planejamentos estratégicos institucionais ou a escolha de temas de TCC por alunos.

A expansão geográfica da coleta de dados com a adição de outras redes de instituições públicas de ensino superior além da Rede Federal, permitiria análises comparativas entre diferentes contextos institucionais, o que pode revelar especificidades da produção acadêmica de institutos federais em contraste com universidades federais ou estaduais. Também a incorporação da dimensão temporal de forma ainda mais granular, com análises de sazonalidade intra-anual ou detecção de eventos que influenciaram deslocamentos temáticos abruptos, como a pandemia de COVID-19, enriqueceria a compreensão dos fatores contextuais que moldam as agendas de pesquisa.

Por fim, a evolução do sistema para uma arquitetura de atualização incremental e contínua, de modo a substituir o paradigma atual de processamento em lote por um pipeline de dados em tempo real, transformaria a ferramenta em um observatório permanente da produção acadêmica da Rede Federal. Esta evolução demandaria a implementação de mecanismos de detecção de mudanças nos repositórios fonte, orquestração automatizada de pipelines de dados e estratégias de retreinamento periódico dos modelos de tópicos para capturar a emergência de novas áreas temáticas ao longo do tempo.

7.3 Conclusão

Este estudo configura-se como uma contribuição científica e tecnológica relevante para o campo da gestão da informação educacional e da ciência de dados aplicada ao contexto acadêmico. Ao demonstrar a viabilidade de consolidar, processar e analisar automaticamente grandes volumes de produção acadêmica textual dispersa, por meio da extração de padrões e tendências interpretáveis, a pesquisa estabelece um precedente metodológico sólido e replicável para outras redes institucionais e diferentes domínios de conhecimento.

Além disso, o artefato desenvolvido transcende o papel de uma ferramenta isolada e constitui-se em um instrumento efetivo para a promoção de uma cultura de tomada de decisão orientada por dados no âmbito da educação profissional e tecnológica brasileira. Por conseguinte, os resultados alcançados contribuem diretamente para o fortalecimento da qualidade, relevância e alinhamento estratégico da pesquisa conduzida nos Institutos Federais de Educação, Ciência e Tecnologia, o que reforça o papel dessas instituições na produção e difusão do conhecimento científico nacional.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. **Mining Text Data**. [S.l.]: Springer, 2012.

AIOHTTP. **aiohttp — Asynchronous HTTP Client/Server**. 2025. <https://docs.aiohttp.org/>. Acesso em: 09 nov. 2025.

ALLAN, J. *et al.* Topic detection and tracking pilot study final report. In: **Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop**. [S.l.: s.n.], 1998. p. 194–218.

APACHE. **Apache Parquet — Columnar Storage File Format**. 2025. <https://parquet.apache.org/>. Acesso em: 09 nov. 2025.

ASYNCIO. **asyncio — Asynchronous I/O**. 2025. <https://docs.python.org/3/library/asyncio.html>. Acesso em: 09 nov. 2025.

ATEFEH, F.; KHREICH, W. A survey of techniques for event detection in twitter. **Computational Intelligence**, v. 31, n. 1, p. 132–164, 2015.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval: The Concepts and Technology Behind Search**. 2nd. ed. [S.l.]: Addison-Wesley, 2011.

BEAUTIFULSOUP. **Beautiful Soup 4.12.0 documentation**. 2025. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 09 nov. 2025.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77–84, 2012. Disponível em: <https://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003. Disponível em: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis: Forecasting and Control**. 5th. ed. [S.l.]: Wiley, 2015.

Brasil. **Lei nº 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica, cria os Institutos Federais de Educação, Ciência e Tecnologia, e dá outras providências**. 2008. Diário Oficial da União, Brasília, DF, 30 dez. 2008. Acesso em: 12 jun. 2025. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm>.

Brasil. **Decreto nº 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo federal**. 2016. Diário Oficial da União, Brasília, DF, 12 maio 2016. Seção 1, p. 21. Acessado em: 20/06/2025. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm>.

BROUCKE, S. V.; BAESENS, B. **Practical Web Scraping for Data Science: Best Practices and Examples with Python**. 1. ed. New York: Apress, 2018.

CAIRO, A. **The Truthful Art: Data, Charts, and Maps for Communication**. Berkeley: New Riders, 2016.

CAMPOS, T. M. da S. Trabalho de Conclusão de Curso (Graduação em Tecnologia em Sistemas para Internet), **Aplicação das metodologias de Business Intelligence para análise dos dados abertos governamentais do Instituto Federal de Brasília**. 2021. Acessado em: 08/05/2025. Disponível em: <https://bdtcbra.omeka.net/items/show/530>.

CHATFIELD, C. **The Analysis of Time Series: An Introduction**. 6th. ed. [S.l.]: Chapman & Hall/CRC, 2003.

CRESWELL, J. W. **Investigação qualitativa e projeto de pesquisa : escolhendo entre cinco abordagens**. [S.l.]: Penso Editora Ltda, 2014. v. 3.

CROFT, W. B.; METZLER, D.; STROHMAN, T. **Search Engines: Information Retrieval in Practice**. 2nd. ed. [S.l.]: Pearson Education, 2015.

EDUCAçãO, M. da. **Rede Integra MEC**. 2025. <https://redeintegra.mec.gov.br/>. Acesso em: 13 nov. 2025.

ELMASRI, R.; NAVATHE, S. B. **Sistema de Banco de Dados**. [S.l.]: Pearson, 2011. v. 6.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data**. [S.l.]: Cambridge University Press, 2007.

FEW, S. **Information Dashboard Design: The Effective Visual Communication of Data**. Sebastopol: O'Reilly Media, 2006.

FIGMA. **Figma**. 2025. <https://www.figma.com/>.

FUNG, G. P. C. *et al.* Parameter free bursty events detection in text streams. In: **Proceedings of the International Conference on Very Large Data Bases (VLDB)**. [S.l.: s.n.], 2005. p. 181–192.

GIL, A. C. **Como elaborar projetos de pesquisa**. [S.l.]: Editora Atlas Ltda, 2022. v. 7.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. [S.l.]: Elsevier Editora Ltda, 2015. v. 2.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. 2nd. ed. [S.l.]: OTexts, 2018.

IFB. **Página institucional do Instituto Federal de Brasília**. 2025. Disponível em: <https://www.ifb.edu.br/reitori/24013-ifb-em-numeros-tem-nova-plataforma>. Acessado em: 12/06/2025.

IFB em Números. **IFB em Números - Pesquisa**. 2025. Disponível em: <https://ifbemnumeros.ifb.edu.br/>. Acessado em: 05/07/2025.

INMON, W. **Building the Data Warehouse**. [S.l.]: John Wiley Sons, Inc, 2002. v. 3.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. [S.l.]: Stanford University, 2024. v. 3.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. [S.l.]: Wiley Publishing, Inc, 2004.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit The Definitive Guide to Dimensional Modeling**. [S.l.]: John Wiley Sons, Inc, 2013. v. 3.

KLEINBERG, J. Bursty and hierarchical structure in streams. **Data Mining and Knowledge Discovery**, Springer, v. 7, n. 4, p. 373–397, 2003.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [S.l.]: Springer, 2013.

LUCIDCHART. **Lucidchart**. 2025. <https://www.lucidchart.com/pages/pt>.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval International Student Edition**. [S.l.]: Cambridge University Press, 2008.

MARCONI, M. de A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. [S.l.]: Editora Atlas, 2021. v. 9.

MARTINS, G. de A.; THEóPHILO, C. R. **Metodologia da investigação científica para ciências sociais aplicadas**. [S.l.]: Editora Atlas, 2016. v. 3.

MEC. **Rede Federal de Educação Profissional, Científica e Tecnológica**. 2024. Portal do Ministério da Educação. Acesso em: 12 jun. 2025. Disponível em: <https://www.gov.br/mec/pt-br/assuntos/ept/rede-federal>.

MITCHELL, R. **Web Scraping with Python**. [S.l.]: O'Reilly Media, Inc, 2018. v. 2.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 5th. ed. [S.l.]: Wiley, 2012.

MURRAY, S. **Interactive Data Visualization for the Web**. 2nd. ed. Sebastopol: O'Reilly Media, 2017.

NLTK. **NLTK: Natural Language Toolkit**. 2025. <https://www.nltk.org/>. Acesso em: 09 nov. 2025.

Open Knowledge Foundation. **Open Knowledge Foundation**. 2025. Site institucional. Acessado em: 20/06/2025. Disponível em: <https://opendefinition.org/>.

PACHECO, E. **Institutos Federais Uma Revolução na Educação Profissional e Tecnológica**. [S.l.]: Editora Moderna Ltda, 2011.

PANDAS. **Pandas: Python Data Analysis Library**. 2025. <https://pandas.pydata.org/>. Acesso em: 09 nov. 2025.

PARNAíBA, R. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis), **Uma análise das áreas temáticas do trabalho de conclusão de curso (TCC) em Ciências Contábeis Campus I da UFPB no quadriênio 2016.1-2019.1**. 2020. Acessado em: 13/06/2025. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/17375>.

PLAYWRIGHT. **Playwright — End-to-End Testing for Web Applications**. 2025. <https://playwright.dev/>. Acesso em: 01 nov. 2025.

PLOTLY. **Plotly — Interactive Graphing Library**. 2025. <https://plotly.com/python/>. Acesso em: 09 nov. 2025.

PRESSMAN, R. S.; MAXIM, B. R. **Engenharia de Software: Uma Abordagem Profissional**. [S.l.]: AMGH Editora LTDA, 2021. v. 9.

PUPPETEER. **Puppeteer — Headless Chrome Node API**. 2025. <https://pptr.dev/>. Acesso em: 01 nov. 2025.

PYTHON. **Python Programming Language**. 2025. <https://www.python.org/>. Acesso em: 09 nov. 2025.

RICHARDSON, R. J. **Pesquisa social : métodos e técnicas**. [S.l.]: Editora Atlas Ltda, 2017. v. 4.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 513–523, 1988. Disponível em: <https://ecommons.cornell.edu/server/api/core/bitstreams/fc18789c-6a03-48e6-8226-7dba0ce94e32/content>.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, 1975. Disponível em: <https://openlib.org/home/krichel/courses/lis618/readings/salton75.pdf>.

SCIKIT-LEARN. **Scikit-learn: Machine Learning in Python**. 2025. <https://scikit-learn.org/>. Acesso em: 09 nov. 2025.

SELENIUM. **Selenium — Browser Automation Tool**. 2025. <https://www.selenium.dev/>. Acesso em: 01 nov. 2025.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business Intelligence e análise de dados para gestão do negócio**. 4. ed. [S.l.]: Bookman, 2019.

SOMMERVILLE, I. **Engenharia De Software**. [S.l.]: Pearson, 2018. v. 10.

SQLALCHEMY. **SQLAlchemy — SQL Toolkit and ORM**. 2025. <https://www.sqlalchemy.org/>. Acesso em: 09 nov. 2025.

SQLITE. **SQLite: SQL Database Engine**. 2025. <https://www.sqlite.org/>. Acesso em: 09 nov. 2025.

STREAMLIT. **Streamlit — The fastest way to build data apps**. 2025. <https://streamlit.io/>. Acesso em: 09 nov. 2025.

TUFTE, E. R. **The Visual Display of Quantitative Information**. 2nd. ed. Cheshire: Graphics Press, 2001.

APÊNDICE A – Modelo Entidade-Relacionamento (MER)

Nesta seção, apresentam-se os diagramas MER utilizados para modelar o banco de dados do projeto, tanto o banco de *staging* (*integra.db*) quanto o banco destinado à modelagem dimensional (*datamart.db*).

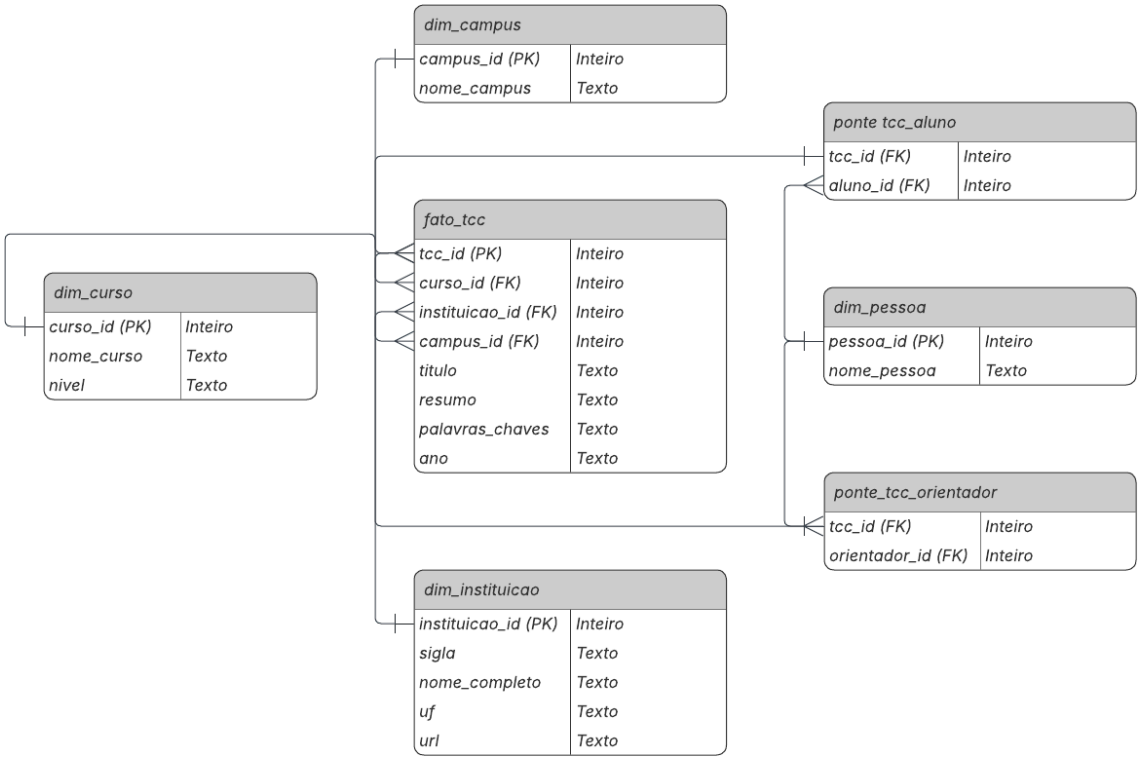
A Figura 16 ilustra a estrutura principal das entidades do banco *integra.db*, correspondente aos dados de *staging*.

Figura 16 – **Modelo Entidade-Relacionamento para *Staging* (*integra.db*)**

Fonte: Elaborado pela autora.

Por conseguinte, a Figura 17 apresenta a estrutura principal das entidades do *datamart.db* e seus respectivos relacionamentos.

Figura 17 – **Modelo Entidade-Relacionamento para Modelagem Star Schema (datamart.db)**



Fonte: Elaborado pela autora.

APÊNDICE B – Dicionário de Dados

O dicionário de dados apresenta a descrição detalhada dos campos das principais tabelas do banco de dados.

Dicionário de Dados Logo Após Coletados

1. Tabela: `professores`

Campo	Tipo	Descrição
id	Inteiro	Identificador único do professor, gerado automaticamente (*PRIMARY KEY AUTOINCREMENT*).
sigla	Texto	Sigla da instituição à qual o professor está vinculado.
nome	Texto	Nome completo do professor.
campus	Texto	Nome do campus onde o professor atua.
cargo	Texto	Cargo do professor (ex.: Professor EBTT, Coordenador).
slug	Texto	Identificador textual padronizado do professor.
url_final	Texto	URL final do perfil do professor.
UNIQUE(slug, sigla)	Restrição	Garante unicidade entre slug e sigla.

Fonte: Elaborado pela autora.

2. Índices da tabela `professores`

Índice	Descrição
idx_professores_slug	Indexa o campo `slug`.
idx_professores_sigla	Indexa o campo `sigla`.

Fonte: Elaborado pela autora.

3. Tabela: *tccs*

Campo	Tipo	Descrição
id	Inteiro	Identificador único do TCC.
slug_professor	Texto	Slug do professor autor/orientador.
nome_professor	Texto	Nome do professor autor/orientador.
sigla	Texto	Sigla da instituição.
instituicao	Texto	Nome da instituição.
UF	Texto	Unidade da Federação.
campus	Texto	Campus do TCC.
ano	Texto	Ano de publicação/defesa.
curso	Texto	Curso ao qual o TCC pertence.
autores	Texto	Nome(s) do(s) autor(es).
titulo	Texto	Título do TCC.
resumo	Texto	Resumo do trabalho.
palavras_chaves	Texto	Palavras-chave do TCC.
UNIQUE(slug_professor, titulo)	Texto	Impede duplicidade de TCCs por professor.

Fonte: Elaborado pela autora.

4. Índices da tabela *tccs*

Índice	Descrição
idx_tccs_slugprof	Indexa `slug_professor`.
idx_tccs_sigla	Indexa `sigla`.

Fonte: Elaborado pela autora.

Dicionário de Dados Referente aos Dados Tratados

Tabela: *dim_campus*

Campo	Tipo	Descrição
campus_id	Inteiro	Identificador único do campus (chave primária).
nome_campus	Texto	Nome completo do campus da instituição.

Fonte: Elaborado pela autora.

Tabela: *dim_pessoa*

Campo	Tipo	Descrição
pessoa_id	Inteiro	Identificador único da pessoa (autor ou orientador).
nome_pessoa	Texto	Nome da pessoa.

Fonte: Elaborado pela autora.

Tabela: *dim_instituicao*

Campo	Tipo	Descrição
instituicao_id	Inteiro	Identificador único da instituição.
sigla	Texto	Sigla da instituição (ex.: IFCE, IFMG, CEFET-MG).
nome_completo	Texto	Nome completo da instituição.
uf	Texto	Unidade da Federação (ex.: CE, MG, RJ).
url	Texto	URL oficial da instituição.

Fonte: Elaborado pela autora.

Tabela: *dim_curso*

Campo	Tipo	Descrição
curso_id	Inteiro	Identificador único do curso.
nome_curso	Texto	Nome completo do curso.
nivel	Texto	Nível do curso (ex.: Técnico, Superior, Especialização).

Fonte: Elaborado pela autora.

Tabela: *ponte_tcc_aluno*

Campo	Tipo	Descrição
tcc_id	Inteiro	Referência ao TCC na tabela fato.
aluno_id	Inteiro	Referência ao aluno na tabela `dim_pessoa`.

Fonte: Elaborado pela autora.

Tabela: *ponte_tcc_orientador*

Campo	Tipo	Descrição
tcc_id	Inteiro	Referência ao TCC na tabela fato.
orientador_id	Inteiro	Referência ao orientador na tabela `dim_pessoa`.

Fonte: Elaborado pela autora.

Tabela: *fato_tcc*

Campo	**Tipo**	**Descrição**
tcc_id	Inteiro	Identificador único do TCC (chave primária).
titulo	Texto	Título completo do TCC.
resumo	Texto	Resumo textual do TCC.
palavras_chaves	Texto	Palavras-chave associadas ao trabalho.
ano	Texto	Ano de defesa/publicação do TCC.
curso_id	Inteiro	Chave estrangeira para `dim_curso`.
instituicao_id	Inteiro	Chave estrangeira para `dim_instituicao`.
campus_id	Inteiro	Chave estrangeira para `dim_campus`.

Fonte: Elaborado pela autora.

APÊNDICE C – Reconhecimento do uso de tecnologias e ferramentas de Inteligência Artificial (IA) generativa, softwares e outras ferramentas de apoio.

Reconheço que, no desenvolvimento deste Trabalho de Conclusão de Curso, foram utilizadas tecnologias de Inteligência Artificial (IA) generativa exclusivamente como ferramentas de apoio, sem geração automática de conteúdo acadêmico, ideias originais ou resultados técnicos.

Foram empregadas soluções de IA para revisão linguística, refinamento de clareza textual, troca de ideias para definição de caminhos metodológicos, sugestão de autores e materiais relevantes para a revisão bibliográfica e tradução do resumo. As ferramentas utilizadas incluem, por exemplo, ChatGPT e Gemini.

Reconheço, também, a utilização do Overleaf para organização e escrita do conteúdo em LaTeX, a ferramenta Figma para desenvolvimento do protótipo e o Lucidchart para a criação dos diagramas.

Ressalto que nenhuma dessas tecnologias foi utilizada para gerar conteúdo acadêmico pronto, mas apenas como instrumentos auxiliares de revisão, organização e suporte ao processo de escrita.