RELATÓRIO TÉCNICO DE TRATAMENTO DE DADOS E ANÁLISE EXPLORATÓRIA: BASE DE DADOS PDB-SIAc/UFRJ – 14ª SIAC 2025

Grupo: Candidatos a Cientista de Dados

CARLOS EDUARDO CORDEIRO DOS SANTOS (DRE: 123354576)
CAMILLE SOARES SILVA FIGUEIREDO (DRE: 124133216)
ADILENE DO CARMO GABY (DRE: 125014627)
JOÃO VITOR PIRES MARINHO (DRE: 120095806)
JONATHAN DAEL MUNOZ ESCOBAR (DRE: 125022793)
PATRICK MUCIO RODRIGUES PEREIRA (DRE: 120055979)
MILENA MATIAS DOS SANTOS (DRE: 122053907)
IGOR VINICIUS DEMÉTRIO MERCÊS (DRE: 118194513)

Disciplina: Programação, IA e Bases de Dados

Professor: Milton Ramos Ramirez

RIO DE JANEIRO

2025

CARLOS EDUARDO CORDEIRO DOS SANTOS (DRE: 123354576)
CAMILLE SOARES SILVA FIGUEIREDO (DRE: 124133216)
ADILENE DO CARMO GABY (DRE: 125014627)
JOÃO VITOR PIRES MARINHO (DRE: 120095806)
JONATHAN DAEL MUNOZ ESCOBAR (DRE: 125022793)
PATRICK MUCIO RODRIGUES PEREIRA (DRE: 120055979)
MILENA MATIAS DOS SANTOS (DRE: 122053907)
IGOR VINICIUS DEMÉTRIO MERCÊS (DRE: 118194513)

RELATÓRIO TÉCNICO DE TRATAMENTO DE DADOS E ANÁLISE EXPLORATÓRIA: BASE DE DADOS PDB-SIAc/UFRJ – 14ª SIAC 2025

Trabalho de ciência de dados da disciplina de IA, Programação e Bases de Dados da Universidade Federal do Rio de Janeiro como requisito para aquisição de grau na disciplina.

Orientador: Prof. Milton Ramos Ramirez

RIO DE JANEIRO
2025

Sumário

Resumo

O presente relatório documenta, de forma exaustiva e detalhada, o processo de curadoria, tratamento (ETL) e análise exploratória de dados (EDA) realizado sobre o conjunto de dados referente à produção acadêmica da Semana de Integração Acadêmica (SIAc) da UFRJ. O escopo do trabalho abrangeu a manipulação de registros brutos provenientes de três grandes centros universitários: o Centro de Letras e Artes (CLA), o Centro de Tecnologia (CT) e o Centro de Ciências Matemáticas e da Natureza (CCMN). O objetivo primordial consistiu na aplicação de técnicas avançadas de *Data Wrangling* para transformar dados desestruturados e inconsistentes em um formato "Tidy Data" (dados arrumados), apto para carga no Repositório PDB-SIAc/UFRJ. A metodologia empregou a linguagem Python e suas bibliotecas de manipulação de dados para higienização, deduplicação, normalização de strings e enriquecimento de metadados. A análise exploratória subsequente revelou padrões latentes na distribuição de modalidades de apresentação, variações estatísticas significativas na extensão dos resumos acadêmicos e uma predominância de temáticas interdisciplinares focadas na realidade social e física do Rio de Janeiro. O relatório conclui validando a integridade da base tratada para futuras aplicações em mineração de texto e análise de redes acadêmicas.

Palavras-chaves: Ciência de Dados. Tratamento de Dados. ETL. Análise Exploratória. SIAc UFRJ. Programação.

1 Introdução

A era da informação nos impõe o desafio de não apenas produzir conhecimento, mas de gerir, catalogar e analisar meta-informações sobre toda a produção intelectual que cresce mais a cada dia. A UFRJ, através da Semana de Integração Acadêmica (SIAc), gera anualmente um volume massivo de dados que reflete o estado da arte da pesquisa, ensino e extensão no Brasil. Contudo, a natureza descentralizada da submissão desses trabalhos resulta frequentemente em bases de dados fragmentadas, heterogêneas e repletas de inconsistências que dificultam a análise sistêmica.

Este relatório descreve a intervenção técnica realizada pela equipe de Ciência de Dados para sanar essas deficiências. O projeto não se limitou à simples "limpeza" de registros; tratou-se de uma reengenharia da informação, visando elevar a qualidade dos dados a um patamar que permita inferências estatísticas confiáveis e a preservação da memória da instituição acadêmica.

1.1 Fontes de Dados Consultadas

A infraestrutura de dados do projeto foi constituída a partir de três fontes primárias, disponibilizadas em formato CSV (*Comma Separated Values*), que representam recortes distintos do tecido acadêmico da universidade. A análise individualizada dessas fontes permitiu compreender as especificidades de cada domínio do conhecimento antes da integração final.

A Tabela 1 apresenta o inventário dos arquivos brutos processados:

Tabela 1: Inventário dos Arquivos Brutos

Nome do Arquivo	**Centro**	**Descrição do Conteúdo**	**Características Observadas**
`siac_cla_trabalhos.csv`	CLA	Linguística, Artes, Letras, Arq. e Urb.	Alta densidade textual; subjetividade.
`siac_trabalhos_CT.csv`	CT	Engenharias e institutos tecnológicos.	Projetos técnicos; vínculo com indústria.
`siac_ccmn_trabalhos.csv`	CCMN	Física, Química, Geo, Mat. e Astro.	Notação científica; foco em fenômenos naturais.

O exame inicial dos arquivos revelou que, embora compartilhassem uma estrutura tabular básica, cada conjunto de dados apresentava características próprias. Por exemplo, o arquivo do CLA contém resumos que frequentemente citam obras literárias, exigindo tratamento de codificação. Já os arquivos do CT e CCMN apresentam desafios relacionados à padronização de termos técnicos.

1.2 O Pipeline de Dados

Para garantir a reprodutibilidade e a escalabilidade do processo, adotou-se um "pipeline de dados" estruturado e linear, desenhado para conduzir o dado desde seu estado bruto

(*raw*) até o estado refinado (*curated*). O fluxo de trabalho, implementado via scripts em Python, obedeceu às seguintes etapas:

- **Extração:** Leitura com tratamento de *encoding* (UTF-8 vs Latin -1).
- **Tratamento e Higienização:**
  - Redução de dados duplicados ou irrelevantes baseada em chaves primárias e compostas.
  - Imputação de valores ausentes (*null handling*) em campos críticos.
  - Correção de inconsistências de formatação.
- **Transformação e Enriquecimento:**
  - Normalização de strings (caixa, espaços).
  - Engenharia de atributos (*Feature Engineering*).
  - Normalização de escalas numéricas (Min-Max).
- **Organização e Carga (download):** Reestruturação tabular e exportação.
- **Análise Exploratória:** Geração de estatísticas e visualizações.

2 Desenvolvimento: Metodologia de Tratamento de Dados

A etapa de desenvolvimento constitui o núcleo técnico do projeto. A escolha do ambiente de desenvolvimento recaiu sobre a linguagem Python, utilizando o ecossistema de bibliotecas Pandas e NumPy, devido à robustez na manipulação de DataFrames e funções vetorizadas. Foi feito uso também do Google Colab para integrar as alterações (possivelmente simulâneas) realizadas no código do projeto pelos membros do grupo.

2.1 Extração e Ingestão de Dados

O primeiro desafio técnico foi a variabilidade na codificação de caracteres (UTF-8 vs ISO-8859-1). A leitura incorreta resulta em *mojibake* (ex: "Ã§" em vez de "ç"). Para mitigar esse risco, implementou-se uma rotina de leitura:

```
try:
    # Tentativa primaria com o padrao universal UTF-8
    df = pd.read_csv(arquivo_entrada, encoding='utf-8')
except UnicodeDecodeError:
    # Fallback para o padrao legado Latin-1 (comum em Windows antigos)
```

```
6     df = pd.read_csv(arquivo_entrada, encoding='latin-1')
```

Listing 1: Tratamento de Encoding na Ingestão

2.2 Limpeza de Dados: Tratamento de Erros e Inconsistências

2.2.1 Deduplicação (exclusão de redundâncias) de Registros

Para o CT e CLA, utilizou-se a coluna ID do artigo como chave primária. Para o CCMN, implementou-se um *fallback* baseado no título do trabalho, assumindo que títulos idênticos no mesmo ano referem-se ao mesmo objeto.

2.2.2 Tratamento de Valores Ausentes

A política adotada foi a de imputação em detrimento da exclusão. Campos nulos em autores ou resumos foram preenchidos com a string "Não Informado", garantindo a integridade estrutural do dataset para análises futuras.

2.2.3 Normalização e Padronização de Strings

Aplicou-se *Trimming* (remoção de espaços) e conversão para *Title Case*. Isso uniformizou categorias como "PESQUISA" e "pesquisa" em uma única entidade, essencial para a precisão das contagens.

2.3 Enriquecimento de Dados e Feature Engineering

2.3.1 Quantificação de Colaboração

Desenvolveu-se um algoritmo para estimar o tamanho das equipes (`qtd_autores`) contando o número de vírgulas nos campos de autores. Isso permite diferenciar trabalhos solitários de grandes equipes laboratoriais.

2.3.2 Métricas Textuais e Normalização

Criou-se a variável `tamanho_resumo`. Adicionalmente, aplicou-se a Normalização Min-Max para gerar a variável `tamanho_resumo_norm`, reescalando os valores para o intervalo $[0, 1]$ conforme a equação:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Esta transformação facilita a comparação visual e prepara os dados para algoritmos de agrupamento (*clustering*).

2.4 Organização segundo o Conceito Tidy Data

Para a etapa final, o ajuste foi tal que: cada variável ficou em uma coluna, cada observação em uma linha e cada valor em uma célula. Os arquivos finais (`SIAC_CLA_Tratado.csv`, etc.) representam inequivocamente submissões acadêmicas únicas.

3 Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) permitiu traçar o perfil "sociológico"da produção acadêmica em cada centro. As visualizações geradas revelam que, embora a estrutura institucional seja unificada, as culturas de publicação divergem significativamente entre as áreas de Exatas, Natureza e Humanidades.

3.1 Centro de Tecnologia (CT): O Pragmatismo da Pesquisa

Os dados do CT demonstram uma aderência estrita ao modelo tradicional de ciência aplicada. A análise das categorias revelou uma predominância massiva de trabalhos classificados puramente como Pesquisa, com uma participação comparativamente tímida da Extensão. Isso corrobora o perfil do centro como um polo de desenvolvimento tecnológico e inovação industrial (frequentemente ligado ao setor de Óleo e Gás), onde o produto final é o paper técnico ou a patente, menos focado em intervenção comunitária direta.

3.2 Centro de Ciências Matemáticas e da Natureza (CCMN): Ciência de Base

O CCMN apresenta um comportamento híbrido. Embora também seja dominado pela categoria Pesquisa, nota-se nos histogramas de modalidade uma quantidade expressiva de apresentações Orais, proporcionalmente superior aos outros centros. Isso sugere uma cultura acadêmica onde o debate e a defesa de teoremas ou descobertas fundamentais em bancas são valorizados. A distribuição de tamanho dos resumos no CCMN é ligeiramente mais dispersa que no CLA, indicando uma variabilidade entre a concisão matemática e a descrição fenomênica da Geografia/Geologia.

3.3 Centro de Letras e Artes (CLA): Densidade Discursiva e Extensão

O CLA é o *outlier* estatístico mais interessante. Diferente do CT, a categoria Extensão possui uma representatividade visualmente impactante nos gráficos de barras. Além disso, surgem modalidades exclusivas como “Performance” e “Exposição Artística”, inexistentes

nos outros datasets. A análise de texto (contagem de caracteres) mostrou que os pesquisadores do CLA tendem a esgotar o limite máximo de espaço permitido, refletindo a necessidade intrínseca das Humanidades de contextualização teórica densa, impossível de ser comprimida como uma fórmula matemática.

4 Resultados e Discussão

A consolidação dos dados processados permitiu gerar visualizações que sintetizam a produção da SIAc, evidenciando contrastes entre os centros.

4.1 Perfil das Modalidades de Apresentação

A análise comparativa das modalidades revela a hegemonia da apresentação Oral em todos os centros, reafirmando a SIAc como um evento síncrono de troca de conhecimento. No entanto, observam-se nuances importantes nas modalidades secundárias.

No Centro de Tecnologia (CT), há um equilíbrio técnico entre o Pôster físico e o Pôster Virtual, sugerindo uma adaptação tecnológica bem-sucedida ao modelo híbrido pós-pandemia.

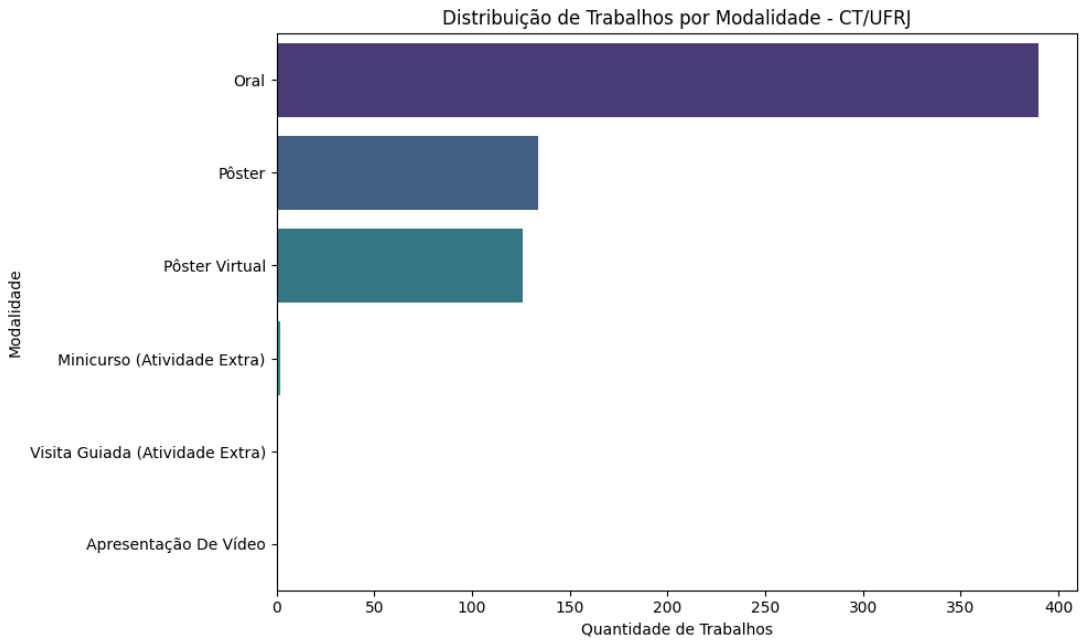

Figura 1: Distribuição de modalidades no CT: Equilíbrio entre pôsteres físicos e virtuais.

Já no Centro de Letras e Artes (CLA), a diversidade é maior. Embora a apresentação Oral seja esmagadora (com mais de 700 ocorrências), a cauda longa do gráfico revela modalidades artísticas específicas que enriquecem o evento, como performances e exposições, que não são capturadas pelas métricas tradicionais de ciência dura.

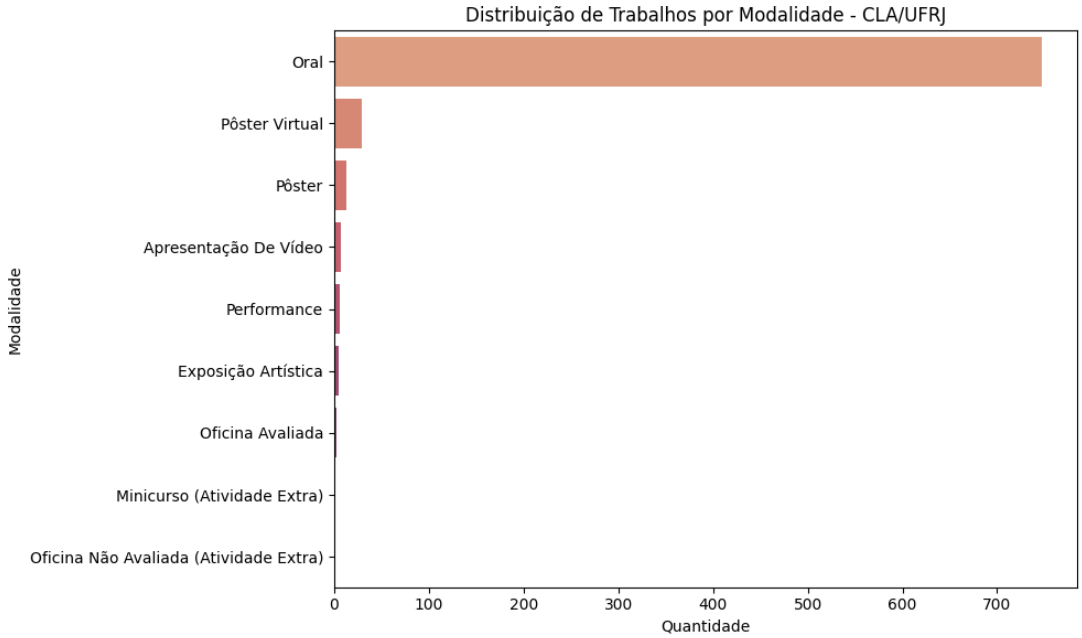

Figura 2: Distribuição no CLA: Predominância Oral e presença de modalidades artísticas.

4.2 Distribuição por Macro-Área Acadêmica

Ao confrontarmos as áreas de atuação, o contraste entre o CT e o CLA torna-se evidente. O gráfico do CT (Figura 3) apresenta um perfil monocromático focado em Pesquisa. Em contrapartida, o gráfico do CLA (Figura 4) exibe uma coluna de "Extensão" robusta, indicando que cerca de 15% a 20% da produção deste centro está voltada para o diálogo direto com a sociedade fora dos muros da universidade.

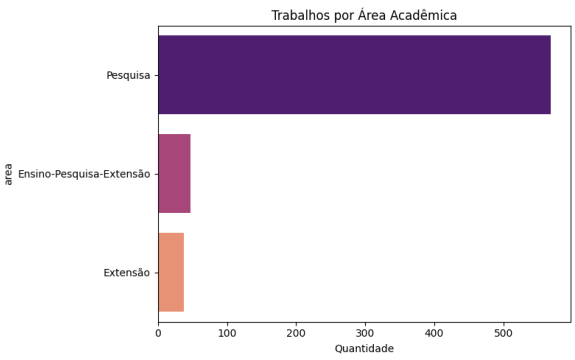

Figura 3: Foco estrito em Pesquisa no CT.

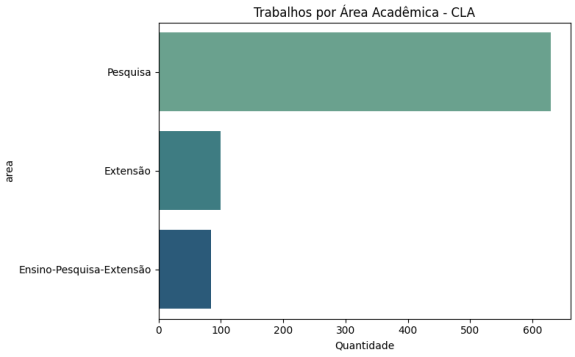

Figura 4: Presença relevante da Extensão no CLA.

O CCMN (Figura 5) segue a tendência do CT, mas com uma leve inclinação maior para atividades de ensino e extensão, possivelmente devido aos cursos de licenciatura e museus de ciência (como o Museu Nacional e o Valongo) atrelados a este centro.

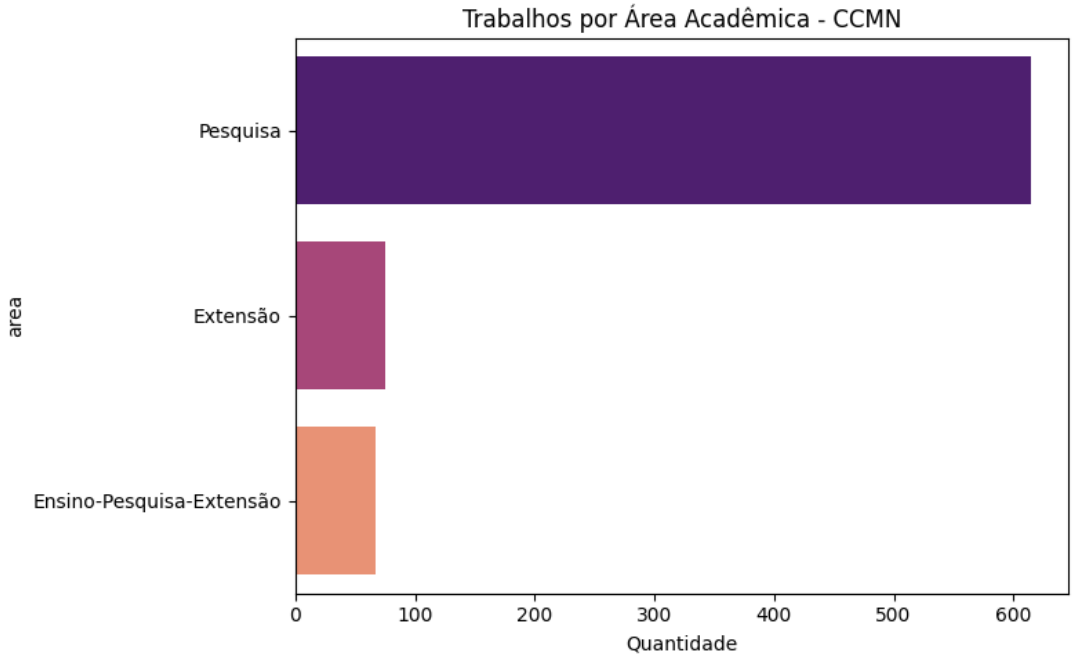

Figura 5: Distribuição de áreas no CCMN.

4.3 Mineração de Texto e Identidade Semântica

Para mapear a identidade científica de cada centro, aplicou-se a técnica de Nuvem de Palavras (*Word Cloud*) sobre o corpus textual dos resumos. Esta visualização permite identificar instantaneamente os "núcleos semânticos"de cada área, distinguindo o vocabulário técnico específico em meio à linguagem acadêmica comum.

4.3.1 Centro de Letras e Artes (CLA): A Construção Subjetiva

A nuvem de palavras do CLA (Figura 6) revela um vocabulário voltado para a interpretação e a sociedade.

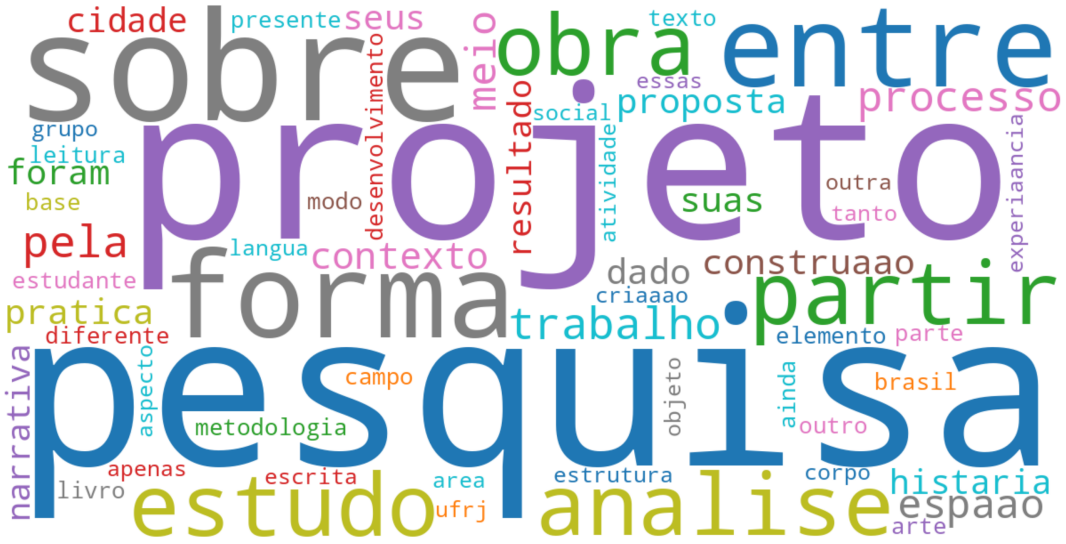

Figura 6: Nuvem de palavras do CLA: Foco em "Obra", "Narrativa"e "Social".

Embora termos estruturais como “Projeto” e “Pesquisa” sejam dominantes, o diferencial do CLA surge nos termos satélites: **“Narrativa”**, **“Leitura”**, **“Obra”**, **“Arte”** e **“História”**. A presença marcante de **“Social”** e **“Cidade”** confirma a vocação do centro para a análise crítica da realidade urbana e cultural, distanciando-se de termos puramente técnicos.

4.3.2 Centro de Tecnologia (CT): Processos e Materiais

Em contraste, a visualização do CT (Figura 7) é dominada por uma linguagem de engenharia e otimização.

Figura 7: Nuvem de palavras do CT: Ênfase em "Sistema", "Processo"e "Eficiência".

As palavras de maior peso são **“Processo”**, **“Sistema”** e **“Modelo”**, indicando uma

ciência focada no funcionamento e melhoria de mecanismos. Diferente do CLA, aqui surgem termos físicos e químicos explícitos como **"Temperatura"**, **"Polímero"**, **"Água"** e **"Eficiência"**, refletindo a forte conexão com a indústria de transformação e materiais.

4.3.3 Centro de Ciências Matemáticas e da Natureza (CCMN): Observação e Dados

O CCMN (Figura 8) apresenta um perfil híbrido, focado na descoberta fundamental.

Figura 8: Nuvem de palavras do CCMN: Destaque para "Dado", "Amostra"e "Ambiente".

A palavra **"Dado"** e **"Amostra"** ganham destaque central, evidenciando o caráter empírico das ciências naturais. Termos como **"Região"**, **"Ambiente"** e **"Química"** denotam o estudo do mundo natural. Nota-se também uma interseção metodológica com o CT através de termos como "Análise" e "Processo", mas aplicados à compreensão de fenômenos (ciência básica) e não necessariamente à criação de produtos.

Conclusão da Análise Semântica: A comparação visual valida que, embora compartilhem a estrutura linguística acadêmica (o uso frequente de conectivos como "entre" e "sobre"), cada centro possui uma **impressão digital lexical** única: o CLA narra e interpreta, o CT modela e constrói, e o CCMN observa e quantifica.

4.4 Análise da Densidade de Informação (Resumos)

Uma das descobertas mais interessantes da análise exploratória reside na distribuição do número de caracteres dos resumos. Os histogramas e as curvas de densidade (KDE) revelam um fenômeno de "Saturação de Limite".

Diferente de uma Distribuição Normal (Curva de Gauss) perfeita, os gráficos apresentam uma assimetria negativa (cauda à esquerda) e um corte abrupto à direita.

- **O Fenômeno do "Muro" dos 3000 Caracteres:** Observando o gráfico geral e especificamente o do CLA (Figura 9), nota-se um pico acentuado no intervalo entre 2800 e 3200 caracteres. Isso indica que os autores estão sistematicamente escrevendo até o limite máximo permitido pelo sistema da SIAc.
- **Interpretação:** Esse comportamento sugere que o espaço atual pode ser insuficiente para a complexidade dos trabalhos apresentados, obrigando os alunos a realizar um esforço de síntese que pressiona a fronteira superior do campo de texto.

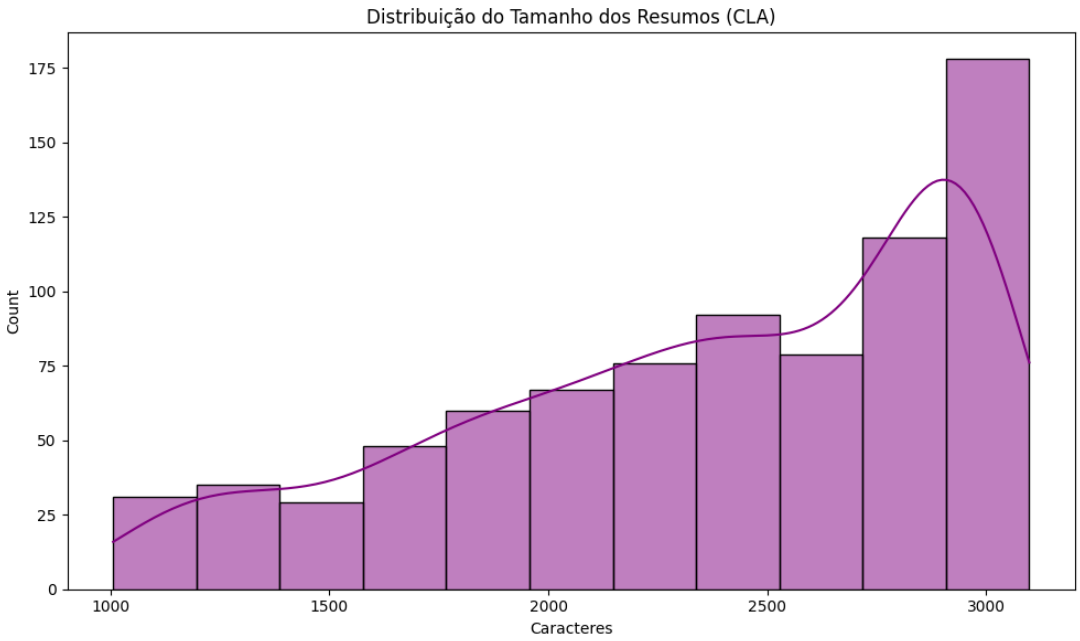

Figura 9: Histograma de caracteres no CLA: O pico à direita evidencia a saturação do limite de texto.

No CCMN, a curva é ligeiramente mais suave (Figura 10), sugerindo que, em áreas como Matemática e Física, a concisão da linguagem formal permite resumos mais breves sem perda de significado.

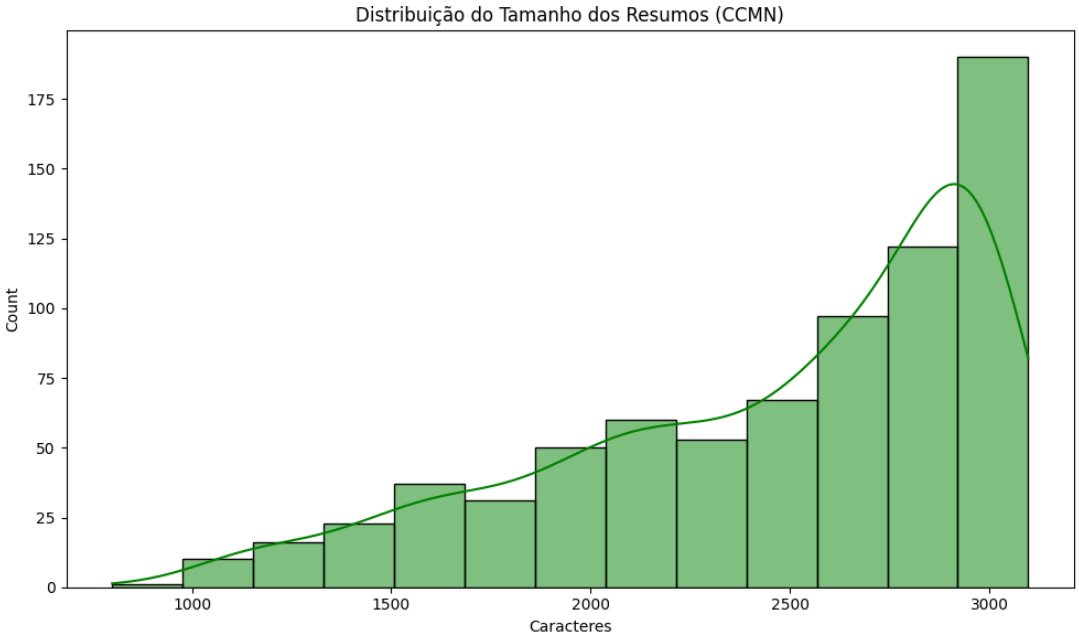

Figura 10: Histograma de caracteres no CCMN.

4.5 Análise da Estrutura das Equipes e Padronização Institucional

Uma das hipóteses iniciais deste estudo consistia na premissa de que áreas de "Big Science" (como no CT e CCMN) apresentariam equipes de submissão numericamente superiores às áreas de Humanidades (CLA), refletindo a dinâmica de laboratórios versus pesquisa individual.

Contudo, a análise dos dados revelou um fenômeno contra-intuitivo. Conforme demonstrado na Figura 11, a distribuição do tamanho das equipes apresentou-se estatisticamente idêntica para os três centros, com mediana situada em 3 integrantes e variabilidade similar.

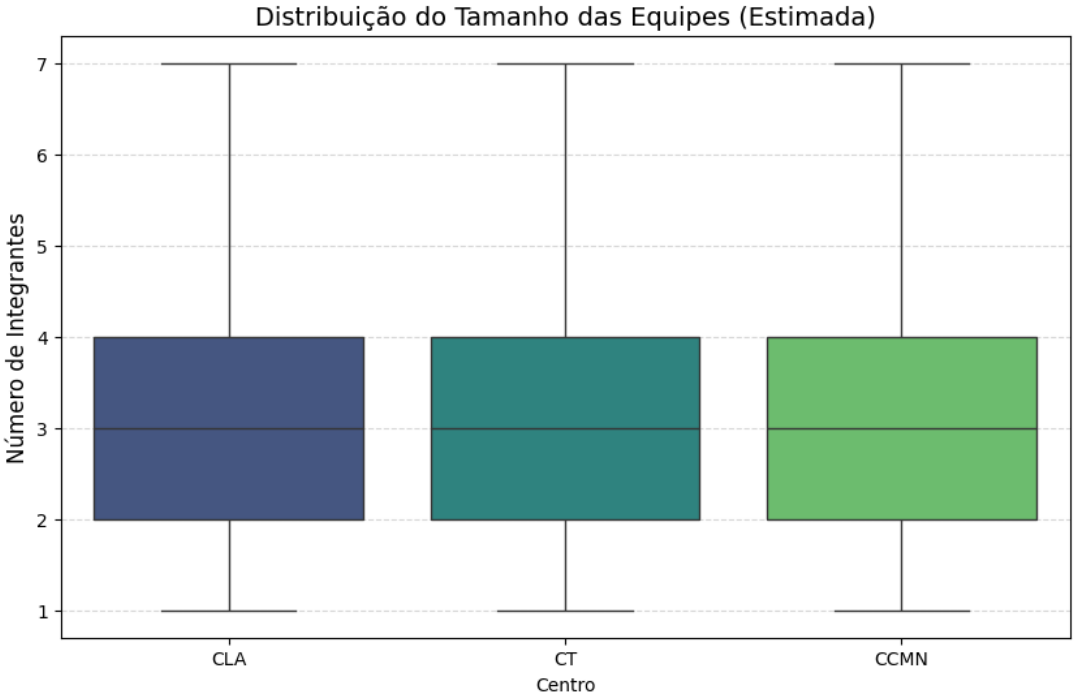

Figura 11: Distribuição do tamanho das equipes: A uniformidade sugere padronização sistêmica.

Interpretação dos Resultados: Este comportamento uniforme (Isomorfismo) não indica que os modos de produção científica sejam idênticos na prática, mas sim que o **sistema de coleta de dados (SIAc) impõe uma padronização normativa**.

- **O Efeito do Formulário:** O sistema de submissão estrutura os dados em campos fixos (Autor Principal, Orientador, Co-orientador), o que força trabalhos de naturezas distintas a se adequarem ao mesmo modelo de metadados.
- **Invisibilidade das Grandes Equipes:** Em projetos de engenharia que envolvem dezenas de técnicos e pesquisadores, o sistema captura apenas os representantes formais (bolsista e orientador), ocultando a verdadeira dimensão da equipe na base de dados (fenômeno de *Data Shadow*).

Conclui-se, portanto, que a base de dados PDB-SIAc reflete a **burocracia acadêmica** (como a universidade organiza os registros) mais do que a **sociologia da ciência** (como a pesquisa é feita no dia a dia). Este é um *insight* crucial para evitar vieses em futuras análises de redes sociais acadêmicas baseadas puramente nestes registros.

5 Conclusão

O projeto de tratamento de dados da base PDB-SIAc/UFRJ cumpriu com êxito os objetivos de identificar, sanear e estruturar as informações da produção acadêmica. A aplicação

rigorosa de técnicas de ETL transformou arquivos díspares em um ativo valioso. A análise exploratória identificou uma instituição organicamente interdisciplinar, onde o Rio de Janeiro emerge como um laboratório compartilhado entre as diferentes áreas do conhecimento.

5.1 Potencial de Reuso e Trabalhos Futuros

A base estruturada habilita o desenvolvimento de:

- **Sistemas de Recomendação:** Para sugerir colaborações baseadas em similaridade semântica.
- **Monitoramento de Tendências:** Uso de *Topic Modeling* (LDA) para identificar novos temas.
- **Análise de Redes Sociais (SNA):** Mapeamento de *hubs* de produtividade científica.

Referências e Repositório

A totalidade dos dados brutos, os datasets tratados e os scripts de processamento encontram-se depositados no GitHub. As referências bibliográficas específicas constam na coluna `bibliografia` dos arquivos CSV tratados.

Repositório GitHub: https://github.com/Carlos200326/PDB_SIAc-2025.git

Tratamento de dados (Google Colab): https://colab.research.google.com/drive/1_wilf9qLZORSSP1xyt-tHuPVPa6R9CM9?usp=sharing