

ANÁLISE TEMÁTICA DA PRODUÇÃO ACADÊMICA DA REDE FEDERAL COM MODELAGEM DE TÓPICOS

Anna Caroline Ribeiro Mota[1]
Gustavo Henrique Dornelas de Deus[2]
Felipe Alves Vila Verde Gomes[3]
Geovana Araujo Perazzo[4]

RESUMO: A Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) reúne um amplo volume de produção acadêmica distribuída em diferentes plataformas institucionais, dificultando análises integradas e identificação de padrões temáticos. Este trabalho apresenta o desenvolvimento de um sistema computacional voltado à coleta, processamento e análise de aproximadamente 680 mil registros acadêmicos, incluindo Trabalhos de Conclusão de Curso, artigos científicos e projetos de pesquisa e extensão provenientes de 33 instituições federais. A metodologia integra técnicas de web scraping assíncrono, processos de ETL, modelagem dimensional em Data Mart e Processamento de Linguagem Natural (PLN), com aplicação do algoritmo Latent Dirichlet Allocation (LDA) para identificação automática de tópicos. Durante o processo de transformação, foram removidos registros inconsistentes, duplicados, documentos em idiomas incompatíveis e produções sem conteúdo textual relevante, além da exclusão automatizada de registros não pertencentes à Rede Federal. Após o processo de limpeza e padronização, a base consolidada passou a contar com aproximadamente 91 mil TCCs, 255 mil artigos científicos e 243 mil projetos válidos para análise. Os resultados evidenciam predominância de temas relacionados a desenvolvimento de software, engenharia, sustentabilidade, inteligência artificial e análise de dados, além de forte alinhamento entre ensino, pesquisa e extensão.

Palavras-chave: análise temática, PLN, LDA, dados educacionais, dashboard

ABSTRACT: The Federal Network of Professional, Scientific and Technological Education (RFEPCT) gathers a large volume of academic production distributed across institutional platforms, making integrated analysis and thematic pattern identification difficult. This study presents the development of a computational system designed to collect, process and analyze approximately 680 thousand academic records, including undergraduate theses, scientific articles and research and extension projects from 33 federal institutions. The methodology

[1] Graduanda em Tecnologias em Sistemas para Internet pelo Instituto Federal de Brasília (IFB). e-mail: anna.mota@estudante.ifb.edu.br.
[2] Mestre em Sistemas de Informação e Gestão do Conhecimento. Professor do Instituto Federal de Brasília. e-mail: gustavo.deus@ifb.edu.br
[3] Graduando em Tecnologias em Sistemas para Internet pelo Instituto Federal de Brasília (IFB). e-mail: felipe63360@estudante.ifb.edu.br.
[4] Graduanda em Tecnologias em Sistemas para Internet pelo Instituto Federal de Brasília (IFB). e-mail: geovanaperazzo63@gmail.com.

integrates asynchronous web scraping, ETL processes, dimensional modeling in a Data Mart and Natural Language Processing (NLP) techniques, with the application of the Latent Dirichlet Allocation (LDA) algorithm for automatic topic identification. During the transformation stage, inconsistent and duplicated records, documents in incompatible languages and productions without relevant textual content were removed, in addition to the exclusion of records not associated with the Federal Network. After the cleaning and standardization process, the analytical base consisted of approximately 91 thousand undergraduate theses, 255 thousand scientific articles and 243 thousand academic projects. The results reveal thematic predominance related to software development, engineering, sustainability, artificial intelligence and data analysis, as well as strong alignment between teaching, research and extension activities.

Keywords: topic modeling; NLP; LDA; educational data; dashboard

Introdução

A produção acadêmica da Rede Federal de Educação Profissional, Científica e Tecnológica constitui um importante ativo informacional para análise de tendências científicas, tecnológicas e educacionais. Entretanto, essa produção encontra-se distribuída em diferentes plataformas institucionais, sem padronização estrutural ou integração analítica, dificultando o uso estratégico dos dados por pesquisadores e gestores.

Nesse contexto, justifica-se a utilização de técnicas computacionais capazes de consolidar, processar e interpretar grandes volumes de dados acadêmicos heterogêneos, permitindo identificar padrões temáticos e apoiar processos de tomada de decisão baseados em dados.

A utilização de técnicas de mineração textual e Processamento de Linguagem Natural (PLN) torna-se relevante para identificação automatizada de padrões temáticos em grandes volumes de documentos acadêmicos. Além disso, abordagens baseadas em ciência de dados e *Business Intelligence* permitem estruturar informações heterogêneas e transformá-las em conhecimento útil para apoio à tomada de decisão.

Diante desse cenário, este trabalho tem como objetivo desenvolver um sistema computacional capaz de coletar, estruturar e analisar automaticamente a produção acadêmica da Rede Federal, permitindo identificar tendências temáticas e apoiar processos de análise institucional baseados em dados.

Como diferencial, o trabalho propõe uma arquitetura analítica integrada composta por *web scraping* assíncrono, pipeline ETL, modelagem dimensional em *Data Mart*, validação automatizada de integridade dos dados e modelagem de tópicos baseada em LDA aplicada em larga escala sobre bases acadêmicas heterogêneas.

Referencial Teórico

A análise de grandes volumes de dados textuais demanda técnicas capazes de transformar dados em conhecimento útil para apoio à decisão (Davenport, 1998). Nesse contexto, o Processamento de Linguagem Natural (PLN) permite estruturar e analisar documentos textuais em larga escala (Jurafsky; Martin, 2021). Entre as técnicas mais utilizadas destaca-se o algoritmo *Latent Dirichlet Allocation* (LDA), empregado para identificação automática de tópicos em coleções documentais extensas (Blei; Ng; Jordan, 2003).

Além disso, abordagens baseadas em ETL e modelagem dimensional contribuem para consolidação, organização e exploração analítica de dados heterogêneos. Segundo Kimball e Ross (2013), a modelagem dimensional favorece ambientes analíticos orientados à tomada de decisão e aplicações de *Business Intelligence*.

Tratamento e Validação dos Dados

Antes da análise temática, os dados passaram por um processo estruturado de ETL responsável pela extração, transformação e validação das informações acadêmicas coletadas. Inicialmente, foram extraídos aproximadamente 228.295 registros de Trabalhos de Conclusão de Curso, 296.318 artigos científicos e 295.162 projetos acadêmicos provenientes de diferentes bases institucionais da Rede Federal.

Durante a etapa de transformação, foram aplicadas rotinas automatizadas de limpeza e validação dos dados. Foram removidos registros inconsistentes, duplicados, documentos em idiomas incompatíveis e produções sem conteúdo textual significativo após o pré-processamento. Além disso, foram realizadas validações institucionais para identificação de registros pertencentes à Rede Federal, resultando na exclusão de mais de 123 mil registros incompatíveis.

Adicionalmente, registros com inconsistências relacionadas ao mapeamento de campus e cursos foram descartados para garantir integridade referencial no modelo dimensional utilizado no *Data Mart*.

Na etapa de pré-processamento textual, foram removidos mais de 77 mil registros inconsistentes ou duplicados considerando os diferentes tipos documentais analisados. Após o processo de limpeza e padronização, a base consolidada passou a contar com aproximadamente 91 mil TCCs, 255 mil artigos científicos e 243 mil projetos acadêmicos válidos para análise temática.

Posteriormente, aplicou-se modelagem de tópicos utilizando o algoritmo LDA. Também foi realizada otimização automática da quantidade de tópicos para cada conjunto documental, resultando em 11 tópicos para TCCs e 5 tópicos para artigos científicos e projetos acadêmicos.

Por fim, os dados processados foram disponibilizados em dashboards interativos para exploração analítica e visualização temática.

Resultados

Os resultados evidenciam forte predominância de áreas tecnológicas, engenharias e ciências aplicadas nos TCCs da Rede Federal, especialmente relacionadas ao desenvolvimento de software, redes, energia, engenharia civil e produção agronômica. Observa-se também presença significativa de temas voltados à educação, saúde, sustentabilidade e ciências naturais, indicando perfil interdisciplinar e alinhamento entre formação profissional, inovação tecnológica e demandas sociais contemporâneas, como apresentado na Tabela 1.

Tabela 1 — Tópicos identificados nos Trabalhos de Conclusão de Curso (TCCs)

Tópico	**Nome semântico**	**Termos representativos**
T0	Educação, Matemática e Aprendizagem	Matemática, Fundamental, Aprendizagem, Pandemia, Cultural
T1	Produção Animal, Água e Saúde	Água, Qualidade, Zootecnia, Medicina, Produção, Veterinário
T2	Engenharia, Energia e Redes	Design, Rede, Eletrico, Engenharia, Energia, Industrial
T3	Desenvolvimento de Software e Web	Desenvolvimento, Análise, Internet, Estágio, Web, Software
T4	Gestão, Administração e Arquitetura	Gestão, Administração, Publicar, Arquitetura, Publico, Empresa
T5	Agronomia e Produção Rural	Agrônomo, Soja, Agronômico, Cultura, Curso, Agronomia

T6	Física, Química e Educação Científica	Física, Química, Químico, Aprendizagem, Escola, Biológico
T7	Meio Ambiente, Turismo e Sustentabilidade	Ambiental, Gestão, Turismo, Resíduo, Estado, Solar
T8	Pedagogia, Saúde e Gestão Educacional	Pedagogia, Revisão, Administração, Infantil, Gestão, Saúde
T9	Engenharia Civil e Produção Industrial	Engenharia, Civil, Mecânico, Construção, Análise, Produção
T10	Alimentos, Química e Caracterização	Alimento, Químico, Química, Letra, Físico, Caracterização

Fonte: Dados da pesquisa (2026).

Nos artigos científicos, observou-se predominância de estudos voltados à investigação aplicada, propriedades físico-químicas, produção e caracterização de materiais, além de pesquisas relacionadas à gestão ambiental, desenvolvimento urbano e sustentabilidade. Também se destacam temas associados à educação, aprendizagem, práticas escolares, saúde e questões sociais, evidenciando forte interdisciplinaridade na produção científica da Rede Federal.

Os projetos acadêmicos evidenciam forte direcionamento para ações de extensão, desenvolvimento regional, sustentabilidade e práticas educacionais interdisciplinares. Destacam-se iniciativas relacionadas à aprendizagem, energia, gestão de resíduos, desenvolvimento tecnológico e projetos voltados à integração social e comunitária, conforme apresentado na Tabela 2.

Tabela 2 — Tópicos identificados nos projetos acadêmicos

Tópico	**Nome semântico**	**Termos representativos**
T1	Produção Rural e Sustentabilidade	Produção, Água, Solo, Planta, Qualidade, Região
T2	Desenvolvimento, Energia e Tecnologia	Desenvolvimento, Energia, Utilizar, Resíduo, Dado, Aplicação
T3	Educação, Aprendizagem e Extensão	Curso, Aprendizagem, Matemática, Estudante, Extensão, Língua

T4	Programas de Extensão e Tecnologia	Extensão, Curso, Extensão, Programa, Edital, Tecnologia
T5	Inclusão Social, Cultura e Comunidade	Social, Extensão, Comunidade, Cultural, Mulher, Arte

Fonte: Dados da pesquisa (2026).

De modo geral, os resultados demonstram convergência entre ensino, pesquisa e extensão, com predominância de temas relacionados à inovação tecnológica, desenvolvimento regional, inclusão, sustentabilidade e transformação digital.

Conclusão

Os resultados demonstram o potencial das técnicas de Processamento de Linguagem Natural e modelagem de tópicos para análise de grandes volumes de produção acadêmica da Rede Federal. O sistema desenvolvido possibilitou a integração e interpretação temática de bases heterogêneas, contribuindo para identificação de tendências e apoio à gestão baseada em dados.

Dessa forma, o trabalho evidencia aplicações concretas de ciência de dados e mineração textual no contexto da educação profissional e tecnológica. Como perspectivas futuras, destacam-se a ampliação da base analisada, o refinamento dos modelos de análise temática e a incorporação de novas fontes de dados acadêmicos.

Referências bibliográficas

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, 2003.

DAVENPORT, Thomas H. *Ecologia da informação*. São Paulo: Futura, 1998.

FEW, Stephen**. Information dashboard design**. 2. ed. Burlingame: Analytics Press, 2013.

JURAFSKY, Daniel; MARTIN, James H. **Speech and language processing**. 3. ed. [S.l.]: Draft, 2021.

KIMBALL, Ralph; ROSS, Margy**. The Data Warehouse Toolkit**: The Definitive Guide to Dimensional Modeling. 3. ed. Indianapolis: Wiley, 2013.