ISSN 2395-8618

Procesamiento de lenguaje natural para la organización temática de programas de estudio de ingeniería: comparativa de modelado de tópicos y agrupamiento

César Iván Abraján Barraza*, Andrés Saúl de la Serna Tuya, Judith Karina López Nario, Juan Martín Sekisaka Millán

***Resumen*—Las instituciones de educación superior gestionan grandes repositorios de programas de estudio, pero carecen de herramientas computacionales para analizar sistemáticamente sus contenidos temáticos. Este trabajo presenta un flujo de Procesamiento de Lenguaje Natural (PLN) para la organización temática automática de programas de estudio de ingeniería en español. Se procesaron 51 programas de la carrera de Ingeniería en Gestión Empresarial del sistema de Institutos Tecnológicos de México, aplicando técnicas de modelado de tópicos (LDA y NMF), agrupamiento (K-Means y jerárquico aglomerativo con Ward) y análisis de similitud coseno sobre representaciones TF-IDF. Se evaluaron métricas intrínsecas de coherencia (C_v, C_{npmi}) y de calidad de agrupamiento (Silhouette, Davies-Bouldin, Calinski-Harabasz). NMF con K = 8 tópicos obtuvo la mayor coherencia (C_v=0.63), generando agrupaciones temáticas interpretables que cubren áreas como matemáticas, estadística, economía, mercadotecnia, marco jurídico, desarrollo humano, producción y metodología de investigación. La reducción dimensional mediante SVD con 10 componentes mejoró el Silhouette Score de 0.04 a 0.42 en agrupamiento, demostrando que es indispensable para corpus pequeños con alta dimensionalidad. Análisis adicionales de variabilidad por semilla aleatoria (CV<5% en métricas centrales), estabilidad estructural por Adjusted Rand Index (ARI=0.92 entre semillas) y comparación con representaciones contextuales (paraphrase-multilingual-MiniLM-L12-v2) confirman la robustez del enfoque. Los resultados sugieren que las técnicas de PLN pueden ofrecer una organización temática coherente de documentos curriculares, con potencial para apoyar procesos de gestión académica.**

***Palabras clave*—Procesamiento de lenguaje natural, modelado de tópicos, agrupamiento, programas de estudio, NMF, análisis curricular automático.**

Natural Language Processing for the Thematic Organization of Engineering Curricula: A Comparison of Topic Modeling and Clustering

***Abstract*—Higher education institutions manage large repositories of study programs, but lack computational tools to systematically analyze their thematic content. This work presents a Natural Language Processing (NLP) workflow for the automatic thematic organization of engineering study programs in Spanish. Fifty-one programs from the Business Management Engineering degree program of the Mexican Institutes of Technology system were processed, applying topic modeling techniques (LDA and NMF), clustering (K-Means and Ward's hierarchical agglomerative), and cosine similarity analysis on TF-IDF representations. Intrinsic coherence metrics (C_v, C_{npmi}) and clustering quality metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz) were evaluated. NMF with K = 8 topics yielded the highest coherence (C_v=0.63), generating interpretable thematic groupings that cover areas such as mathematics, statistics, economics, marketing, legal framework, human development, production, and research methodology. Dimensionality reduction using SVD with 10 components improved the Silhouette Score from 0.04 to 0.42 in clustering, demonstrating its indispensability for small corpora with high dimensionality. Additional analyses of variability by random seed (CV<5% in central metrics), structural stability by Adjusted Rand Index (ARI=0.92 between seeds), and comparison with contextual representations (paraphrase-multilingual-MiniLM-L12-v2) confirm the robustness of the approach. The results suggest that NLP techniques can offer a coherent thematic organization of curricular documents, with the potential to support academic management processes.**

Manuscript received on 02/02/2026, accepted for publication on 13/04/2026. Corresponding author is César Iván Abraján Barraza (cesar.ab@culiacan.tecnm.mx).

César Iván Abraján Barraza and Judith Karina López Nario are with the Instituto Tecnológico de Culiacán, Tecnológico Nacional de México, Culiacán, Sinaloa, México.

Andrés Saúl de la Serna Tuya is with the Universidad Popular Autónoma del Estado de Puebla, Puebla, México.

Juan Martín Sekisaka Millán is with the Instituto Tecnológico de Culiacán, Tecnológico Nacional de México, Culiacán, Sinaloa, México.